Neighborhood Co-regularized Multi-view Spectral Clustering of Microbiome Data

Evgeni Tsivtsivadze[1]*, Hanneke Borgdorff[2]*, Janneke van de Wijgert[3], Frank Schuren[1], Rita Verhelst[4], and Tom Heskes[5]

[1] MSB Group, The Netherlands Organization for Applied Scientific Research,
Zeist, The Netherlands
{e.tsivtsivadze,f.schuren}@tno.nl

[2] Academic Medical Center of the University of Amsterdam and
Amsterdam Institute for Global Health and Development,
Amsterdam, The Netherlands
h.borgdorff@aighd.org

[3] Institute of Infection and Global Health, University of Liverpool,
Liverpool, UK
j.vandewijgert@liverpool.ac.uk

[4] International Centre of Reproductive Health,
Faculty of Medicine and Health Sciences, Ghent University,
Ghent, Belgium
Rita.Verhelst@UGent.be

[5] Institute for Computing and Information Sciences, Radboud University,
Nijmegen, The Netherlands
t.heskes@science.ru.nl

Abstract. In many unsupervised learning problems data can be available in different representations, often referred to as views. By leveraging information from multiple views we can obtain clustering that is more robust and accurate compared to the one obtained via the individual views. We propose a novel algorithm that is based on neighborhood co-regularization of the clustering hypotheses and that searches for the solution which is consistent across different views. In our empirical evaluation on publicly available datasets, the proposed method outperforms several state-of-the-art clustering algorithms. Furthermore, application of our method to recently collected biomedical data leads to new insights, critical for future research on determinants of the cervicovaginal microbiome and the cervicovaginal microbiome as a risk factor for the transmission of HIV. These insights could have an influence on the interpretation of clinical presentation of women with bacterial vaginosis and treatment decisions.

1 Introduction

The multi-view paradigm [1–3] is particularly suitable for learning on datasets having more than a single data representation. A classic example is a web document classification task [1], where documents are represented via two different

* These authors contributed equally to this work.

views - one that is based on the links and another one based on the text document. Complex, structured data with multiple representations are frequently encountered in the biomedical domain, making multi-view methods a natural application choice. Although in many circumstances the individual data representation can be sufficient for training a model, a combination of the multiple views can lead to more robust and accurate predictions compared to the ones obtained via the individual views.

The multi-view paradigm has been successfully applied to various learning problems such as semi-supervised classification e.g. [4], regression e.g. [5], preference learning e.g. [6] and clustering e.g. [3, 7]. Our work concerns an unsupervised multi-view learning algorithm that builds upon the spectral clustering method [8, 9]. The proposed algorithm is conceptually different from the existing methods as it uses novel neighborhood co-regularization technique to adapt cluster assignments for different views. Unlike in previous studies that consider aggregation of clusters based on the individual data representations e.g. [10–13], our method promotes consistent cluster assignments across multiple views and penalizes solutions that differ significantly. The closest in spirit to our method is the recent work described in [7], where for the first time a co-regularization framework was successfully applied to a clustering task. However, our co-regularization approach is fundamentally different and is geared towards solutions that capture local/neighborhood-based relations in the dataset. Furthermore, the optimization problem in our work differs from [7] and leads to a simpler closed form solution with fewer terms involved.

We apply the proposed method to a recently collected biomedical dataset from a study aimed at investigating cervicovaginal microbiome compositions. As the determination of microbial community compositions is becoming increasingly complex due to new molecular laboratory methods, unsupervised learning techniques have become an essential part of microbiome studies, e.g.[14, 15]. We demonstrate that, unlike in previous studies, the proposed neighborhood co-regularized multi-view spectral clustering algorithm (NCMSC) identifies distinct clusters within the group of women with a "healthy" cervicovaginal microbiome and within the group of women with bacterial vaginosis (BV). Our observation will aid the analysis of the determinants of the cervicovaginal microbiome and the cervicovaginal microbiome as risk factor for other adverse outcomes, such as transmission of sexually transmitted infections (STIs) and HIV.

2 Neighborhood co-regularized multi-view spectral clustering

Consider we are given a dataset containing multiple representations. Let $X^{(v)} = \{x_i^{(v)}\}_{i=1}^n$. Note that here superscript v denotes the representation for a single view. Let $A^{(v)}$ denote an adjacency matrix of the graph constructed using the data representation in a view v. We can write the normalized Laplacian matrix as $L^{(v)} = D^{(v)-1/2}A^{(v)}D^{(v)-1/2}$, where $D^{(v)}$ is the corresponding degree matrix.

Following [8] the standard special clustering problem (or single view spectral clustering [7]) solves the optimization problem

$$\min_{Q^{(v)} \in \mathcal{R}^{n \times c}} tr\left(Q^{(v)T} L^{(v)} Q^{(v)}\right), \quad \text{s.t.} \quad Q^{(v)T} Q^{(v)} = I \tag{1}$$

where $Q^{(v)} \in \mathcal{R}^{n \times c}$ denotes the cluster assignment matrix and c is number of predefined clusters. In spectral clustering the final cluster membership is obtained by applying the k-means algorithm on the rows of the matrix $Q^{(v)}$. The algorithm we propose extends the standard spectral clustering framework by using neighborhood co-regularization techniques that naturally allow to leverage information from multiple views. Let us denote the cluster assignment matrix $Q^{(v)} = (\mathbf{q}_1^{(v)}, \ldots, \mathbf{q}_c^{(v)})^T$. Slightly overloading our notation, we denote the confidence that a data point $\mathbf{x}_j$ belongs to the cluster c as $[\mathbf{q}_c]_j$. For simplicity, in the derivations below we omit the cluster index.

We define the k neighbors of data point $\mathbf{x}_i^{(v)}$ as $N(\mathbf{x}_i^{(v)}) = \{\mathbf{x}_{i_1}^{(v)}, \ldots, \mathbf{x}_{i_k}^{(v)}\}$ or $X_i^{(v)} = (\mathbf{x}_{i_1}^{(v)}, \ldots, \mathbf{x}_{i_k}^{(v)})^T \in \mathcal{R}^{k \times d}$ where d is the dimensionality of the data point in a view v. Also, the corresponding cluster assignments can be written as $\mathbf{q}_i^{(v)} = (q_{i_1}^{(v)}, \ldots, q_{i_k}^{(v)})^T \in \mathcal{R}^k$. Below we generalize local linear regularization [16, 17] to a multi-view setting. In our setting, for each data point $\mathbf{x}_i^{(v)}$, projections $W_i = (\mathbf{w}_i^{(1)}, \ldots, \mathbf{w}_i^{(M)})^T \in \mathcal{R}^{M \times d}$ are leaned via $\min_{\mathbf{w}_i^{(v)}} J(W_i)$, where

$$\begin{aligned} J(W_i) = &\sum_{v=1}^{M} \sum_{\mathbf{x}_j^{(v)} \in N(\mathbf{x}_i^{(v)})} \|[\mathbf{q}^{(v)}]_j - \mathbf{x}_j^{(v)T} \mathbf{w}_i^{(v)}\|_F^2 + \lambda \sum_{v=1}^{M} \mathbf{w}_i^{(v)T} \mathbf{w}_i^{(v)} \\ &+ \nu \sum_{\substack{v,u=1 \\ v \neq u}}^{M} \sum_{\mathbf{x}_j^{(v)} \in N(\mathbf{x}_i^{(v)})} \|\mathbf{x}_j^{(v)T} \mathbf{w}_i^{(v)} - \mathbf{x}_j^{(u)T} \mathbf{w}_i^{(u)}\|_F^2. \end{aligned} \tag{2}$$

The first term in equation (2) stands for a multi-view version of the problem where we aim to find a weight vector that corresponds as closely as possible to the optimal clustering solution, the second term is the $L2$ regularization on the weight vector $\mathbf{w}_i^{(v)}$, and the third term is a co-regularization that promotes agreement among different views on the obtained clustering. Once the local predictors for all views have been constructed (see Appendix) we can compute the sum of the prediction errors for all clusters

$$\begin{aligned} J_c &= \sum_{v=1}^{M} \sum_{l=1}^{c} \|H^{(v)} \mathbf{q}_l^{(v)} - \mathbf{q}_l^{(v)}\|^2 \\ &= \sum_{v=1}^{M} \sum_{l=1}^{c} [\mathbf{q}_l^{(v)T}((H^{(v)} - I)^T (H^{(v)} - I)) \mathbf{q}_l^{(v)}] \\ &= tr[\mathbf{Q}^T((\mathbf{H} - \mathbf{I})^T (\mathbf{H} - \mathbf{I}))\mathbf{Q}], \end{aligned}$$

where $\mathbf{Q}$ is a $(Mn \times c)$ matrix containing the cluster assignments for all views and $\mathbf{H}$ is a $(Mn \times Mn)$ matrix containing predictions of the linear classifiers estimated via minimization of (2). Thus, the optimization problem we solve to determine cluster assignment matrices for all views is

$$\min_{\mathbf{Q} \in \mathcal{R}^{Mn \times c}} tr[\mathbf{Q}^T((\mathbf{H}-\mathbf{I})^T(\mathbf{H}-\mathbf{I}))\mathbf{Q}] \quad \text{s.t.} \quad \mathbf{Q}^T\mathbf{Q} = \mathbf{I} \tag{3}$$

The above problem is closely related to the standard spectral clustering and the solutions for all views are given by top-c eigenvectors of the matrix $\mathbf{L} = (\mathbf{H}-\mathbf{I})^T(\mathbf{H}-\mathbf{I})$. Similarly to [7] we can use any of the obtained cluster assignment matrices $Q^{(v)}$ in the final k-means step of the clustering algorithm. In our experiments, we observe no major dependence of the clustering performance on the choice of a particular $Q^{(v)}$.

2.1 Related work

There have been a number of multi-view clustering algorithms proposed that build on the idea of leveraging information from different graphs/data representations. For example, in [10] the authors obtain a graph cut which on average is the most suitable for multiple graphs and provide a random walk formulation of the clustering problem. A clustering algorithm that constructs a graph based on nodes from two views and solves a standard spectral problem, is proposed in [11]. Other approaches for multi-view clustering fuse the information from multiple graphs based on matrix factorization [12] or consensus learning techniques [13].

The central idea behind all these methods is to construct clustering for the individual views and to reconcile them in the final solution. Our neighborhood co-regularized multi-view clustering algorithm is conceptually different from these methods as the cluster assignments for the individual views are adapted based on neighborhood co-regularization implemented via the third term in the equation (2). Informally, our method promotes consistent cluster assignments across multiple views and penalizes solutions that differ significantly.

Recently an unsupervised learning algorithm using a co-regularization approach has been proposed in [7]. The co-regularization we propose here is notably different and is geared towards clustering that captures local/neighborhood-based relations in the dataset. This in turn leads to a simpler closed form solution with fewer terms involved. Furthermore, our algorithm can be straightforwardly formulated as a kernel-based method, which can make it suitable for learning non-linear dependencies when estimating cluster assignments.

We also note that the co-regularization framework has recently attracted considerable attention in the machine learning community and proved to work well on a wide range of learning problems e.g. [18, 5, 6]. Moreover, theoretical investigations demonstrate that the co-regularization approach reduces the Rademacher complexity by an amount that depends on the “distance” between the views [19, 20]. We expect that a similar type of analysis can be applied to our algorithm and we aim to investigate this in the near future.

3 Experiments on benchmark datasets

To empirically validate the performance of the NCMSC algorithm, we compare the results of five algorithms on four benchmark datasets. Following [17] we select publicly available datasets, namely Newsgroup, UMIST, WebACE, and USPS.

To evaluate the performance of the algorithms, we compare the obtained clusters with the true class labels in each of the datasets. For this purpose we use two performance measures - clustering accuracy (ACC) and normalized mutual information (NMI) [13]. The clustering accuracy performance measure estimates the relationship between computed clusters and the class labels. Informally, it measures the extent to which data points contained in the clusters correspond to the class label and sums up matches between all class-cluster pairs. The normalized mutual information criterion reports mutual information between the obtained clustering and the true clustering, normalized by the cluster entropies. NMI ranges between 0 and 1 with a higher value indicating a closer match to the true clustering.

In our experiments we compare the performance of the proposed NCMSC algorithm with four other clustering methods, namely k-means (K-Means), hierarchical clustering (HC), spectral clustering (SC), and co-regularized multi-view spectral clustering (CMSC) [7]. For HC we use Euclidean distance. For NCMSC, CMSC and SC, the weights on data graph edges are computed by Gaussian functions. Similarly to [17] the variance is determined by local scaling. All regularization parameters in CMSC and our approach are determined by searching the grid $\{0.1,1,10\}$, and the neighborhood size is set by searching the grid $\{20, 40, 80\}$. In the experiments we consider two views for the NCMSC and CMSC algorithms, which are created by partitioning of the feature vector into two parts (random partitioning has been successfully used in previous studies e.g. [5, 6]).

Table 1: Clustering accuracy results

Algorithm	HC	K-means	SC	CMSC	NCMSC
UMIST	0.4127	0.4372	0.6432	0.6824	**0.6914**
USPS	0.7354	0.7399	0.9312	0.9561	**0.9732**
Newsgroup	0.3223	0.3234	0.5211	0.5734	**0.5811**
WebAce	0.3123	0.3131	0.4553	0.5625	**0.5893**

It can be observed from the Table 1 and Table 2 that the NCMSC algorithm performs better compared to other clustering methods. We suggest that our algorithm outperforms CMSC due to the employed co-regularization procedure, which uses neighborhood-based cluster assignment models and, therefore, captures additional "local" relations in the data. Also, multi-view algorithms in general tend to perform better than their single view counterparts and k-means or hierarchical clustering tend to perform poorer than SC-based approaches.

Table 2: Normalized mutual information results

Algorithm	HC	K-means	SC	CMSC	NCMSC
UMIST	0.6441	0.6481	0.7623	0.8121	**0.8236**
USPS	0.8231	0.8521	0.9732	0.9832	**0.9917**
Newsgroup	0.2114	0.2212	0.4962	0.5213	**0.5283**
WebAce	0.1323	0.1431	0.3842	0.5125	**0.5298**

4 Experiments on a cervicovaginal microbiome dataset

We also apply the proposed NCMSC algorithm to a new biomedical dataset containing results from experiments on a single channel phylogenetic microarray designed to characterize cervicovaginal microbiota [21]. The dataset is from an ongoing study aimed at identifying groups of women with similar cervicovaginal microbial compositions, the determinants of these compositions and their possible association with adverse reproductive health outcomes.

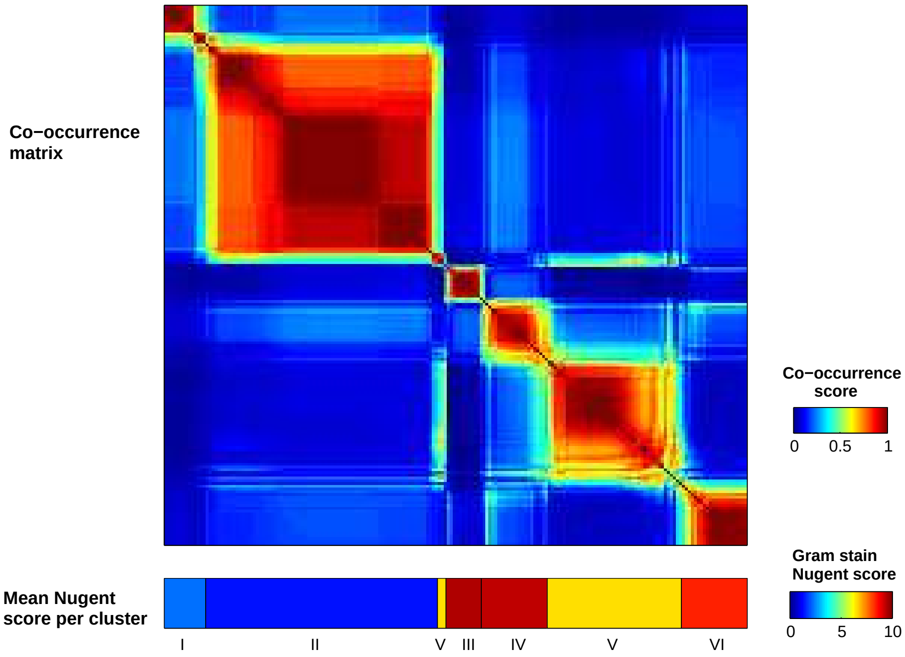
4.1 Dataset preparation

Data from microarray experiments are available for 196 cervical samples from women who participated in an observational prospective cohort study aimed at estimating the HIV-1 incidence in Kigali, Rwanda [22]. BV was diagnosed using Gram stain and microscopy using Nugent scoring [23], which is considered the golden standard. BV is a disruption of the cervicovaginal microbiome characterized by a reduction of lactobacilli and an increase of mostly anaerobic bacteria that leads to an increased risk of preterm birth and sexually transmitted infections [24–26]. A Nugent score of 0-3 was considered as BV negative, 4-6 as intermediate microbiota and 7-10 as BV positive.

We have followed standard microarray preprocessing strategies as described, for example, in [27]. For each spot, signal over background ratios (S/B) are calculated. If the signal is not confidently above background [27], the S/B ratio is set at 1. Samples for which the positive controls (general bacterial probes) show a low S/B ratio, are discarded from the analysis. Lowess smoothing was performed for plate normalization.

The NCMSC algorithm is applied to the preprocessed dataset. The parameters and views are obtained as described in Section 3. We run NCMSC with predefined number of clusters ranging from 3 to 10. Next, we compute the co-occurrence matrix [13] (see Figure 1) that reveals the true number of clusters in the data. For example, former research on the vaginal microbiome of asymptomatic American women described five clusters [15].

Fig. 1: Co-occurrence matrix based on the clusters obtained by the NCMSC algorithm. Patients that often co-occur in different solutions receive a high co-occurrence score.



I-VI: Number of the cervicovaginal microbiome community cluster. Average Nugent scoring per cluster based on gram stained microscopy denoted in colored bar below the co-occurrence matrix. A Nugent score of 0-3 is considered negative for bacterial vaginosis (BV), 4-6 as an intermediate state and 7-10 as BV-positive.

4.2 Results

Based on the co-occurrence matrix six separate clusters are identified (see Figure 1). As would be expected, lactobacilli are present in most samples. The first cluster is dominated by *Lactobacillus crispatus* (all 3 probes targeted at *L. crispatus/ L. kefiranofaciens* give a positive signal in 8/11 samples), the second and largest by *Lactobacillus iners* (75-78/83 samples have a positive signal for the three probes targeted at *L. iners*) and the third to sixth cluster by bacteria which are known to be associated with bacterial vaginosis, such as *Gardnerella vaginalis*, *Atopobium vaginae*, *Mobiluncus mulieris*, *Prevotella* spp., *Dialister* spp., *Sneathia* spp. and *Megasphaera* spp. Furthermore, comparing the obtained clustering to the diagnosis of BV by gram stain Nugent scoring shows that this diverse group is indeed associated with BV, while cluster I and II are associated with the absence of BV. Figure 1 shows the average Nugent score per cluster. Interestingly, the BV-associated group also divides into several clusters.

Our analysis confirms the results of two recent studies, that is [14] in which 93% of women without BV had a vaginal microbial community dominated by either *Lactobacillus crispatus* or *Lactobacillus iners* and [15] where clusters with low Nugent scores were dominated by respectively *Lactobacillus crispatus*, *Lactobacillus iners*, *Lactobacillus gasseri* and *Lactobacillus jensenii*. Unlike in these studies, our approach clearly distinguishes between clusters within the lactobacilli-dominated group and within the BV-associated group.

Contrary to the benchmark datasets, there is no standardized way of comparing clustering methods on this biomedical dataset. Although most of the methods were able to separate the BV-negative from the BV-positive group, only NCMSC is able to clearly identify and differentiate between six cluster groups. Other methods do not separate the *Lactobacillus crispatus* from the *Lactobacillus iners* cluster, even though scientific evidence increasingly shows that these should be separate clusters [14, 15].

5 Conclusion

As the determination of microbial community compositions is becoming increasingly complex due to new molecular laboratory techniques, unsupervised learning methods have become an essential part of microbiome studies. In this paper we propose a novel algorithm for the analysis of complex microbiome data. Our work extends the spectral clustering method [8, 9] to a multi-view setting. We propose neighborhood co-regularization approach to promote consistent cluster assignments across multiple views and to penalize the solutions that differ significantly. Our approach is fundamentally different from existing multi-view methods and is geared towards solutions that capture local/neighborhood-based relations in the dataset.

We have evaluated the performance of the proposed algorithm on several publicly available datasets and applied our method to a recently collected microarray dataset. On all these datasets the NCMSC algorithm outperformed other clustering methods.

When applied to the microbiome data, besides confirming previous studies, the NCMSC algorithm identifies clusters within the lactobacilli-dominated group and within the BV-associated group. This observation will help to identify determinants of the cervicovaginal microbiome and cervicovaginal microbiome compositions associated with other adverse outcomes, such as transmission of STIs and HIV. BV is a difficult to treat condition and the separation of BV-positive women into clusters with different microbial compositions can potentially be important for clinical presentation and treatment decisions.

Acknowledgment

We acknowledge support from the Netherlands Organization for Scientific Research (grant number 639.023.604). Funding for the cervicovaginal microbiome study was received from the European and Developing Countries Clinical Trials Partnership (EDCTP), European Commission 7th Framework CHAARM project and the Aids Fonds Netherlands (grant number 201102). The views expressed in this paper are those of the authors and do not necessarily represent the official position of EDCTP, the EU, or the Aids Fonds Netherlands.

6 Appendix

Given the matrix formulation of our optimization problem, we can find the following closed form for the solution. Taking the partial derivative of $J(W_i)$ with respect to $\mathbf{w}_i^{(v)}$ we get

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}_i^{(v)}} J(W_i) &= -2X_i^{(v)}(\mathbf{q}_i^{(v)} - X_i^{(v)T}\mathbf{w}_i^{(v)}) + 2\lambda \mathbf{w}_i^{(v)} \\ &\quad -4\nu \sum_{u,v=1,u\neq v}^{M} X_i^{(v)}(X_i^{(v)T}\mathbf{w}_i^{(v)} - X_i^{(u)T}\mathbf{w}_i^{(u)}). \end{aligned}$$

By defining $G^{\nu} = 2\nu(M-1)X_i^{(v)}X_i^{(v)T}$, $G^{\lambda} = \lambda X_i^{(v)T}$ and $G = X_i^{(v)}X_i^{(v)T}$, we can rewrite the above term as

$$\begin{aligned} \frac{\partial}{\partial \mathbf{w}_i^{(v)}} J(W_i) &= 2(G + G^{\nu} + G^{\lambda})\mathbf{w}_i^{(v)} - 2X_i^{(v)T}\mathbf{q}_i^{(v)} \\ &\quad -4\nu \sum_{u,v=1,u\neq v}^{M} X_i^{(v)}X_i^{(u)T}\mathbf{q}_i^{(u)}. \end{aligned}$$

At the optimum we have $\frac{\partial}{\partial \mathbf{w}_i^{(v)}} J(W_i) = 0$ for all views, thus we get the exact solution by solving

$$\begin{pmatrix} G_1 & -2\nu X_i^{(1)}X_i^{(2)T} & \cdots \\ -2\nu X_i^{(2)}X_i^{(1)T} & G_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \mathbf{q}_i^{(1)} \\ \mathbf{q}_i^{(2)} \\ \vdots \end{pmatrix} = \begin{pmatrix} X_i^{(1)T}\mathbf{q}_i^{(1)} \\ X_i^{(2)T}\mathbf{q}_i^{(2)} \\ \vdots \end{pmatrix}$$

with respect to $\mathbf{w}_i^{(1)}, \ldots, \mathbf{w}_i^{(M)}$. Note that the left-hand side matrix is positive definite and therefore invertible.

References

1. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory, New York, NY, USA, ACM (1998) 92–100
2. Sindhwani, V., Niyogi, P., Belkin, M.: A co-regularization approach to semi-supervised learning with multiple views. In: Proceedings of ICML Workshop on Learning with Multiple Views. (2005)
3. Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th Annual International Conference on Machine Learning, ACM (2009) 129–136
4. Krishnapuram, B., Williams, D., Xue, Y., Hartemink, A.J., Carin, L., Figueiredo, M.A.T.: On semi-supervised classification. In: Advances in Neural Information Processing Systems 17. (2004)
5. Brefeld, U., Gärtner, T., Scheffer, T., Wrobel, S.: Efficient co-regularised least squares regression. In: Proceedings of the International Conference on Machine learning, New York, NY, USA, ACM (2006) 137–144
6. Tsivtsivadze, E., Pahikkala, T., Boberg, J., Salakoski, T., Heskes, T.: Co-regularized least-squares for label ranking. In Hüllermeier, E., Fürnkranz, J., eds.: Preference Learning. (2010) 107–123
7. Kumar, A., Rai, P., III, H.D.: Co-regularized multi-view spectral clustering. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., Weinberger, K., eds.: Advances in Neural Information Processing Systems 24. (2011) 1413–1421
8. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems 14. (2001) 849–856
9. Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4) (2007) 395–416
10. Zhou, D., Burges, C.J.C.: Spectral clustering and transductive learning with multiple views. In: Proceedings of the 24th international conference on Machine learning. (2007) 1159–1166
11. de Sa, V.R.: Spectral clustering with two views. In: Workshop on Learning with Multiple Views, International Conference on Machine Learning. (2005)
12. Tang, W., Lu, Z., Dhillon, I.S.: Clustering with multiple graphs. In: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining. (2009) 1016–1021
13. Strehl, A., Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research **3** (March 2003) 583–617
14. Srinivasan, S., Hoffman, N.G., Morgan, M.T., Matsen, F.A., Fiedler, T.L., Hall, R.W., Ross, F.J., McCoy, C.O., Bumgarner, R., Marrazzo, J.M., Fredricks, D.N.: Bacterial communities in women with bacterial vaginosis: high resolution phylogenetic analyses reveal relationships of microbiota to clinical criteria. PLoS ONE **7**(6) (2012) e37818
15. Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., Brotman, R.M., Davis, C.C., Ault, K., Peralta, L., Forney, L.J.: Vaginal microbiome of reproductive-age women. PNAS **108 Suppl 1** (Mar 2011) 4680–4687
16. Wu, M., Schölkopf, B.: A local learning approach for clustering. In Schölkopf, B., Platt, J., Hoffman, T., eds.: Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA (2007) 1529–1536

17. Wang, F., Zhang, C., Li, T.: Clustering with local and global regularization. In: Proceedings of the 22nd national conference on Artificial intelligence, AAAI Press (2007) 657–662
18. Sindhwani, V., Niyogi, P.: A co-regularized approach to semi-supervised learning with multiple views. In: Proceedings of the ICML Workshop on Learning with Multiple Views. (2005)
19. Rosenberg, D., Bartlett, P.L.: The Rademacher complexity of co-regularized kernel classes. In Meila, M., Shen, X., eds.: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics. (2007) 396–403
20. Sindhwani, V., Rosenberg, D.: An RKHS for multi-view learning and manifold co-regularization. In McCallum, A., Roweis, S., eds.: Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008), Helsinki, Finland, Omnipress (2008) 976–983
21. Dols, J.A., Smit, P.W., Kort, R., Reid, G., Schuren, F.H., Tempelman, H., Bontekoe, T.R., Korporaal, H., Boon, M.E.: Microarray-based identification of clinically relevant vaginal bacteria in relation to bacterial vaginosis. American Journal Obstetrics and Gynecology **204**(4) (Apr 2011) 1–7
22. Braunstein, S.L., Ingabire, C.M., Kestelyn, E., Uwizera, A.U., Mwamarangwe, L., Ntirushwa, J., Nash, D., Veldhuijzen, N.J., Nel, A., Vyankandondera, J., van de Wijgert, J.H.: High human immunodeficiency virus incidence in a cohort of Rwandan female sex workers. Sexually Transmitted Diseases **38**(5) (May 2011) 385–394
23. Nugent, R.P., Krohn, M.A., Hillier, S.L.: Reliability of diagnosing bacterial vaginosis is improved by a standardized method of gram stain interpretation. Journal of Clinical Microbiology **29**(2) (Feb 1991) 297–301
24. Hauth, J.C., Macpherson, C., Carey, J.C., Klebanoff, M.A., Hillier, S.L., Ernest, J.M., Leveno, K.J., Wapner, R., Varner, M., Trout, W., Moawad, A., Sibai, B.: Early pregnancy threshold vaginal pH and Gram stain scores predictive of subsequent preterm birth in asymptomatic women. American Journal Obstetrics and Gynecology **188**(3) (Mar 2003) 831–835
25. Cohen, C.R., Lingappa, J.R., Baeten, J.M., Ngayo, M.O., Spiegel, C.A., Hong, T., Donnell, D., Celum, C., Kapiga, S., Delany, S., Bukusi, E.A.: Bacterial vaginosis associated with increased risk of female-to-male HIV-1 transmission: a prospective cohort analysis among African couples. PLoS Medicine **9**(6) (Jun 2012)
26. Wiesenfeld, H.C., Hillier, S.L., Krohn, M.A., Landers, D.V., Sweet, R.L.: Bacterial vaginosis is a strong predictor of Neisseria gonorrhoeae and Chlamydia trachomatis infection. Clinical Infectious Diseases **36**(5) (Mar 2003) 663–668
27. Quackenbush, J.: Microarray data normalization and transformation. Nature Genetics **32 Suppl** (Dec 2002) 496–501