Integrating *Tara* Oceans data sets using a multiple kernel approach

Jérôme Mariette, Héléne Chiapello and Nathalie Villa-Vialaneix

18 mai 2016

1 Introduction

The *Tara* Oceans expedition [Sunagawa et al., 2015] facilitated the study of plankton communities by providing the scientists an ocean metagenomic data set combined with environmental measures. During the expedition, 243 seawater samples were collected from 68 locations representing all main oceanic regions from three depth layers : the surface (SRF), the deep chlorophyll maximum (DCM) layer and the mesopelagic (MES) zone.

In [Sunagawa et al., 2015], several approaches have been used to show a vertical stratification of the ocean microbial composition mainly driven by temperature. In the presented work, we propose to take advantage of kernel methods to integrate taxonomic and functional community composition with environmental factors of 139 ocean samples in a single exploratory analysis.

2 Method

2.1 Integrating multiple data sets

In the following, we consider M sets of input data, $(x_i^m)_{i=1,\dots,N,m=1,\dots,M}$ all measured on the same observations $(1,\dots,N)$ which take values in an arbitrary space $(\mathcal{X}^m)_m$. $\mathcal{X}^m$ is equipped with a kernel $K^m : \mathcal{X}^m \times \mathcal{X}^m \to \mathbb{R}$ that provides pairwise similarity between the observations, $K_{ij}^m := K^m(x_i, x_j)$. K^m is assumed symmetric ($K_{ij}^m = K_{ji}^m$) and positive ($\forall N \in \mathbb{N},\ \forall (\alpha_i)_{i=1,\dots,N} \subset \mathbb{R},\ \forall (x_i)_{i=1,\dots,N} \subset \mathcal{X}^m,\ \sum_{i,i'} \alpha_i \alpha_{i'} K_{ii'}^m \geq 0$).

The M kernels can be combined into a single one K^*, defined as a convex combination as following :

$$K_{ij}^* = \sum_{m=1}^{M} \alpha_m K_{ij}^m \tag{1}$$

where $\alpha_m \geq 0$ and $\sum_{m=1}^M \alpha_m = 1$. In this first version of the work, α_m are chosen such that all kernels have the same weight : $\alpha_m = \frac{1}{M}$.

Kernel methods allow to implement different versions of any algorithm which can be expressed in terms of dot products. This is the case of the kernel PCA, a PCA analysis performed in the feature space $(\mathcal{H}, \langle ., . \rangle)$ induced by the kernel. The kernel PCA was first introduced in [Schölkopf et al., 1998] and suppose that the kernel K^* is centered. Then, the following eigenvalue problem is solved : $K^* \boldsymbol{\alpha} = \lambda \boldsymbol{\alpha}$. The eigenvectors $(\boldsymbol{\alpha})_{k=1,\dots,n} \in \mathbb{R}^n$ and the corresponding eigenvalues $(\lambda_k)_{k=1,\dots,n}$ are obtained from the centered matrix.

2.2 Data sets and kernels

The present section is dedicated to describe the data sets and the kernels used. Three data sets provided by the *Tara* Oceans companion website [1] are used to evaluate the proposed method :

1. http ://ocean-microbiome.embl.de/companion.html

— the environmental dataset which contains 22 numeric features (*i.e.*, temperature, salinity, ...). Similarity between ocean samples are computed using the linear kernel, given by $K(x_i, x_j) =< x_i, x_j >$.

— the taxonomic community composition that contains more than 35,000 species. The operational taxonomic units (OTUs) provided, allow to compute both compositional and phylogenetic dissimilarities. The first one is obtained using the Bray-Curtis distance given by

$$d_{BC} = \frac{\sum_s |n_s^A - n_s^B|}{\sum_s n_s^A + n_s^B}, \tag{2}$$

where $(n_s^A)_{s=1,...,S}$ and $(n_s^B)_{s=1,...,S}$ respectively represent the count of species s in community A and B. The phylogenetic dissimilarities requires first to build the phylogenetic trees using fasttree [Morgan N. Price, 2010] with the OTUs sequences as inputs. The *weighted Unifrac* distance is then applied. This one is given by

$$d_{wUF} = \frac{\sum_e l_e |p_e - q_e|}{\sum_e p_e + q_e},$$

where for each branch e, l_e represent its length and p_e (respectively q_e) the fraction of community A (respectively community B) below branch e. Kernels can then be obtained from these two dissimilarities using

$$K_{ij} = -\frac{1}{2}\left(d_{ij} - \frac{1}{N}\sum_{k=1}^{N}(d_{ik} + d_{kj}) + \frac{1}{N^2}\sum_{k,\,k'=1}^{N} d_{kk'}\right),$$

as suggested in [Lee and Verleysen, 2007], where d can either be d_{BC} or d_{wUF}.

— the functional community composition which contains more than 63,000 KEGG orthologous groups. To obtain a functional composition kernel, the Bray-Curtis dissimilarity (Equation (2)) is used.

3 Results and discussion

The results obtained confirm those exposed in [Sunagawa et al., 2015]. Samples are not clearly grouped by their ocean regions but are separated by their depth layer of origin.

Figure 1 presents the kernel PCA results performed on the combined kernel K^*. The histogram (left) presents the entropy supported by the first 15 axes of the kernel PCA and shows that the first two axes are enough to provide relevant information on the data. Figure 1 (right) displays the projections of the ocean samples on the first two principal component. The first axis represents 18.08% of the total entropy and opposes samples from the mesopelagic (MES) zone and the epipelagic layers (SRF and DCM).

To conclude, integrating taxonomic and functional community composition with environmental factors allows to have a fast insight of the different data sets within a single analysis. Future work should investigate multiple kernel learning (MKL) in order to learn kernels weights, α_m, defined in Equation 1 as proposed in [Speicher and Pfeifer, 2015]. MKL would allow to have a better insight on the different data sets by sorting the kernels considering their weight and thus their significance. Also, the self-organizing map algorithm [Kohonen, 2001] and its extension to data described by a kernel will be investigated to classify the different ocean samples.

Références

[Kohonen, 2001] Kohonen, T. (2001). *Self-Organizing Maps, 3rd Edition*, volume 30. Springer, Berlin, Heidelberg, New York.

[Lee and Verleysen, 2007] Lee, J. and Verleysen, M. (2007). *Nonlinear Dimensionality Reduction*. Information Science and Statistics. Springer, New York ; London.

[Morgan N. Price, 2010] Morgan N. Price, Paramvir S. Dehal, A. P. A. (2010). Fasttree 2 ? approximately maximum-likelihood trees for large alignments. *PloS One*.

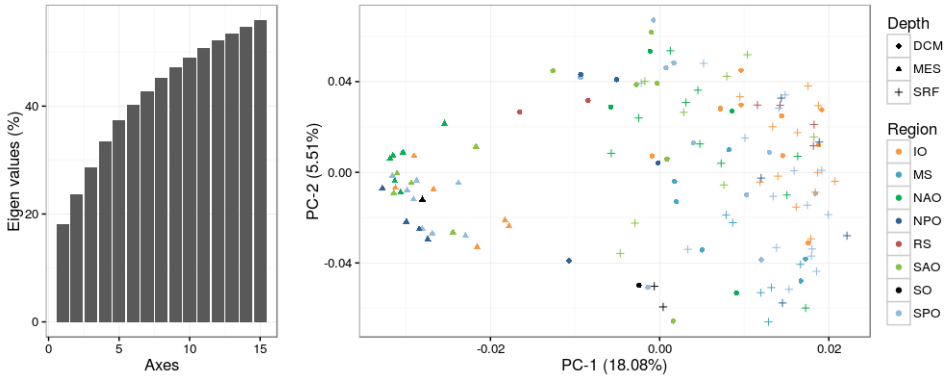

FIGURE 1 – Kernel PCA Entropy preserved by the 15 first axes on the left and projection of the observations on the first two principal components on the right. Colors represent the oceanic regions and shapes the depth layers.

[Schölkopf et al., 1998] Schölkopf, B., Smola, A., and Müller, K. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10 :1299–1319.

[Speicher and Pfeifer, 2015] Speicher, N. and Pfeifer, N. (2015). Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 12(31) :i268–i275.

[Sunagawa et al., 2015] Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D. R., Alberti, A., Cornejo-Castillo, F. M., Costea, P. I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J. M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B. T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., , Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M. B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S. G., and Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, 348(6237).