Kernel conditional embeddings for associating omic data types

Ferran Reverter[⋆], Esteban Vegas and Josep M. Oller

Department of Genetics, Microbiology and Statistics. University of Barcelona
Diagonal, 643, 08028 Barcelona, Spain
{freverter,evegas,joller}@ub.edu

Abstract. Computational methods are needed to combine diverse type of genome-wide data in a meaningful manner. Based on the kernel embedding of conditional probability distributions, a new measure for inferring the degree of association between two multivariate data sources is introduced. We analyze the performance of the proposed measure to integrate mRNA expression, DNA methylation and miRNA expression data.

1 Introduction

Modern genomic and clinical studies are in a strong need of integrative machine learning models for better use of big volumes of heterogeneous information in the deep understanding of biological systems and the development of predictive models. For example, in current biomedical research, it is not uncommon to have access to a large amount of data from a single patient, such as clinical records (e.g. age, gender, medical histories, pathologies and therapeutics), high-throughput omics data (e.g. genomics, transcriptomics, proteomics and metabolomics measurements) and so on. How data from multiple sources are incorporated in a learning system is a key step for successful analysis.

Some of the most powerful methods for integrating heterogeneous data types are kernel-based methods [1]. Kernel-based data integration approaches can be described using two basic steps. Firstly, the right kernel is chosen for each data set. Secondly, the kernels from the different data sources are combined to give a complete representation of the available data for a given statistical task.

In this paper we propose a new measure (to the best of our knowledge) for inferring the degree of association between two multivariate data sources based on the embedding of conditional probability distributions in the framework of kernel methods.

2 Kernel conditional embeddings

The Reproducing Kernel Hilbert Space (RKHS) methods provide a general and rigorous foundation to learn predictive models, where models are determined by specifying a kernel function, a loss function and a penalty function [2]. Representer theorem [2]

[⋆] Corresponding author

shows that solutions of a large class of optimization problems in RKHS can be expressed as kernel expansions over the sample points. A question that arises in a natural manner in the context of inference refers to the representation of a probability distribution P in a RKHS. With this goal Smola et al. [3], Fukumizu et al. [4] among others, have introduced the RKHS versions of the fundamental multivariate statistics, the mean vector and the covariance matrix. These RKHS-counterparts of the mean vector and the covariance matrix are called mean element and covariance operator, respectively.

Let $\mathcal{H}$ be an RKHS on the separable metric space $\mathcal{X}$, with continuous feature mapping $\boldsymbol{\varphi}(\boldsymbol{x}) \in \mathcal{H}$ for each $\boldsymbol{x} \in \mathcal{X}$. The inner product between feature mappings is given by the kernel function $k(\boldsymbol{x}, \boldsymbol{z}) := \langle \boldsymbol{\varphi}(\boldsymbol{x}), \boldsymbol{\varphi}(\boldsymbol{z}) \rangle$. Let P be a probability distribution on $\mathcal{X}$. We can represent $P(X)$ for an element in the RKHS associated with a kernel k:

$$\mu_X := E_X[\boldsymbol{\varphi}(X)] = \int_{\mathcal{X}} \boldsymbol{\varphi}(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}.$$

It has been shown that if $E_X[k(\boldsymbol{x}, \boldsymbol{x})] < \infty$, μ_X is guaranteed to be an element of RKHS. The embedding μ_X of $P(X)$ enjoys two attractive properties. First, if the kernel is characteristic, the mapping from $P(X)$ to μ_X is injective, which means that different distributions are mapped to different points in a RKHS. An example of characteristic kernel is the gaussian kernel. Second, the expectation of any function $f \in \mathcal{H}$ can be evaluated as a scalar product in $\mathcal{H}$

$$\langle \mu_X, f \rangle_{\mathcal{H}_k} := E_X[f(X)], \qquad \forall f \in \mathcal{H}.$$

Let $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_m\}$ be a sample i.i.d from P, an empirical estimator $\hat{\mu}_X$ is defined through

$$\hat{\mu}_X := \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\varphi}(\boldsymbol{x}_i). \tag{1}$$

Let (X, Y) be a random variable taking values on $\mathcal{X} \times \mathcal{Y}$ and $(\mathcal{H}, k)$ and $(\mathcal{G}, l)$ be RKHSs with measurable kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Let $\boldsymbol{\varphi}(\boldsymbol{x}) = k(\cdot, \boldsymbol{x})$ and $\boldsymbol{\phi}(\boldsymbol{y}) = l(\cdot, \boldsymbol{y})$ denote the feature maps. According to the definition of the kernel embedding of a probability distribution $P(X)$, for the kernel embedding of a conditional distribution $P(Y|X)$ we have

$$\mu_{Y|\boldsymbol{x}} := E_{Y|\boldsymbol{x}}\Big(\boldsymbol{\phi}(Y)\Big) = \int_{\mathcal{Y}} \boldsymbol{\phi}(\boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y}.$$

Given a data set $\mathcal{S} = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_m, \boldsymbol{y}_m)\}$ drawn i.i.d from $P(X, Y)$, and where $\Phi := (\boldsymbol{\phi}(\boldsymbol{y}_1), ..., \boldsymbol{\phi}(\boldsymbol{y}_m))$ and $\Upsilon := (\boldsymbol{\varphi}(\boldsymbol{x}_1), ..., \boldsymbol{\varphi}(\boldsymbol{x}_m))$ are implicitly formed feature matrix, and $K = \Upsilon^\intercal \Upsilon$ is the kernel matrix for samples from variable X, Song et al. [5] estimate the conditional embedding as

$$\hat{\mu}_{Y|\boldsymbol{x}} = \sum_{i=1}^{m} \beta_i(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{y}_i) = \sum_{i=1}^{m} \beta_i(\boldsymbol{x}) l(\cdot, \boldsymbol{y}_i) = \Phi B(\boldsymbol{x}) \tag{2}$$

where

$$B(\boldsymbol{x}) = (\beta_1(\boldsymbol{x}), ..., \beta_m(\boldsymbol{x}))^\intercal = (K + \lambda I)^{-1} K_{:\boldsymbol{x}} \tag{3}$$

and $K_{:\boldsymbol{x}} = (k(\boldsymbol{x}, \boldsymbol{x}_1), ..., k(\boldsymbol{x}, \boldsymbol{x}_m))^\intercal$. The empirical estimator of the conditional embedding is similar to the estimator of the ordinary embedding from equation (1). The difference is that, instead of applying uniform weights $\frac{1}{m}$, the former applies non-uniform weights, $\beta_i(\boldsymbol{x})$, on observations which are, in turn, determined by the value $\boldsymbol{x}$ of the conditioning variable. These non-uniform weights reflect the effects of conditioning on the embeddings.

2.1 Measuring the discrepancy between conditional embeddings

Conditional embeddings allows us to quantify the differential effect on the response vector Y, when the values of the conditioning vector X varies. For instance, the conditioning values on which the vector X is fixed, may correspond to the mean vector of X measured in different experimental conditions.

We propose the quantity $||\mu_{Y|\mathbf{x}_1} - \mu_{Y|\mathbf{x}_2}||^2_{\mathcal{G}}$ for measuring the differential effect on Y when conditioning X to $\boldsymbol{x}_1$ or when conditioning X to $\boldsymbol{x}_2$. From (2) we can estimate this quantity by using the statistic:

$$\begin{aligned} T &= ||\hat{\mu}_{Y|\boldsymbol{x}_1} - \hat{\mu}_{Y|\boldsymbol{x}_2}||^2_{\mathcal{G}} \\ &= \langle \hat{\mu}_{Y|\boldsymbol{x}_1} - \hat{\mu}_{Y|\boldsymbol{x}_2}, \hat{\mu}_{Y|\boldsymbol{x}_1} - \hat{\mu}_{Y|\boldsymbol{x}_2} \rangle_{\mathcal{G}} \\ &= \langle \hat{\mu}_{Y|\boldsymbol{x}_1}, \hat{\mu}_{Y|\boldsymbol{x}_1} \rangle_{\mathcal{G}} + \langle \hat{\mu}_{Y|\boldsymbol{x}_2}, \hat{\mu}_{Y|\boldsymbol{x}_2} \rangle_{\mathcal{G}} - 2\langle \hat{\mu}_{Y|\boldsymbol{x}_1}, \hat{\mu}_{Y|\boldsymbol{x}_2} \rangle_{\mathcal{G}} \\ &= \sum_{i,j=1}^{m} \big(\beta_i(\boldsymbol{x}_1)\beta_j(\boldsymbol{x}_1) + \beta_i(\boldsymbol{x}_2)\beta_j(\boldsymbol{x}_2) - 2\beta_i(\boldsymbol{x}_1)\beta_j(\mathbf{x}_2) \big) l(\mathbf{y}_i, \mathbf{y}_j). \end{aligned} \tag{4}$$

To assess the significance, we generate a null distribution by taking permutation of the rows of Y but keeping the rows of X. Thus, after B permutations we have B datasets $S_1 = (X, Y_1), ..., S_B = (X, Y_B)$, where each Y_i results from a random permutation of the rows of Y. Thus we get $T_1, ..., T_B$ and we can estimate a p-value by computing the number of times that T_i , $i = 1, ..., B$, are greater than T.

3 A case study: glioblastoma multiforme cancer

We used data glioblastoma multiforme cancer type available from TCGA [6] (The Cancer Genome Atlas, 2008), preprocessed and provided by Wang et al. (2014) in [7]. We downloaded data sets containing mRNA expression (12,042 genes), miRNA expression (534 miRNAs) and DNA methylation (1,305 genes) from 215 patiens.

We aim to determine the degree of association between methylation and mRNA expression. To this goal, we measure the effect on mRNA (Y) when conditioning on different conditions of DNA methylation (X). In particular, DNA methylation conditions are fixed by the centers of the clusters discovered by using spectral clustering of DNA methylation data. According with [7], we set the number of clusters to be three. Patients were grouped in three clusters with 18, 140 and 57 patients each one. Using (3) we computed $B(\boldsymbol{x}_i)$, where $\boldsymbol{x}_i$, $i = 1, 2, 3$, denotes the mean vectors (centroids) of the clusters. Then, from (4) we computed $T_{ij} = ||\hat{\mu}_{Y|\boldsymbol{x}_i} - \hat{\mu}_{Y|\boldsymbol{x}_j}||^2_{\mathcal{G}}$ where indexes i and j denote on which pair of vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ the conditional embeddings were compared.

In addition, we estimate $||\hat{\mu}_{Y|\boldsymbol{x}_i} - \hat{\mu}_{Y|\bar{\boldsymbol{x}}}||^2_{\mathcal{G}}$, $i = 1, 2, 3$, where $\bar{\boldsymbol{x}}$ denotes the overall mean vector of DNA methylation data.

We used gaussian kernel for both X and Y, kernel parameters were adjusted using the `sigest` function in Kernlab package [8]. In Figure 1 it is shown the heatmap of the kernel matrix corresponding to the methylation data. We observe that the kernel matrix also reveals the same patterns of similarities found by spectral clustering. In fact, when samples are ordered according the clusters found by spectral clustering, identified in the heatmap by the upper color bar, we observe that the similarity values in the kernel matrix shows three homogenous groups that coincide with clusters. A small group of samples, left bottom corner, we call this group as group 1. The largest group, in the central part of the heatmap, we call this group as group 2, and a group of samples, upper right corner, that we identify group as group 3. Figure 2 shows vectors $B(\boldsymbol{x}_i)$, $i = 1, 2, 3$ and $B(\bar{\boldsymbol{x}})$ in the last column, that define the weights of the conditional embeddings (3). Samples, grouped according cluster they belong, are in rows. For each sample, row-normalized weights are displayed. Observe that the normalized weights change consistently across conditions (cluster centroids). That is, samples with highest weights belong to the same cluster on which we are conditioning. To asses the statistical significance of the empirical values T_{ij} we applied a permutation based test, using 5000 permutation samples. We observe (Table 1) that are significant pairwise comparisons that involve group 3. On the other hand, comparisons with respect the conditional embedding on the overall mean are only significant in clusters 2 and 3. Table 1 also includes a summary of the null distribution of the test.

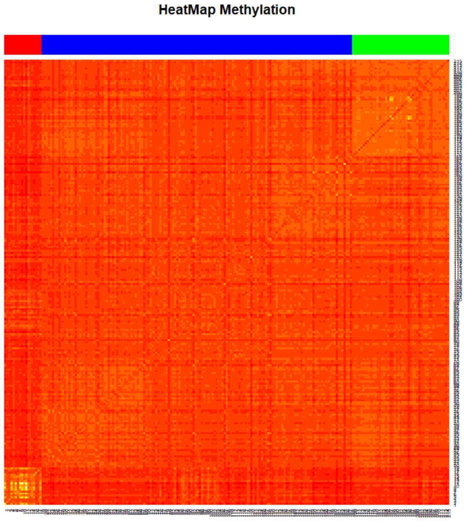

Fig. 1. Heatmap of the kernel matrix from DNA methylation data. Clusters found by spectral clustering are also supported by the kernel matrix.

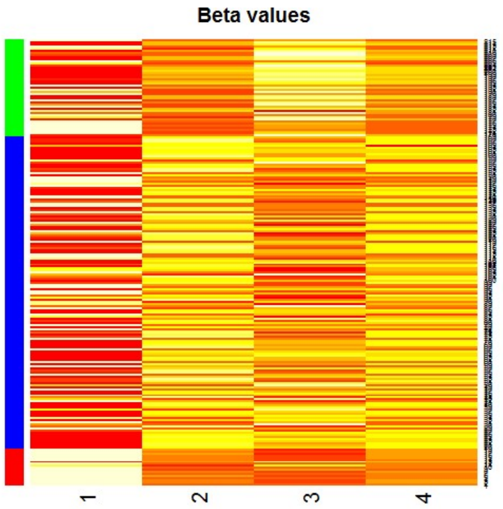

Fig. 2. Gene expression and DNA methylation analysis. Weights that determine the conditional embedding.

comparison	T_{ij}	raw p-value	min	q1	med	q3	max
1.vs.2	0.0332	0.19561	0.0128	0.0218	0.0258	0.0313	0.0699
1.vs.3	0.0301	0.01397	0.0081	0.0148	0.0170	0.0200	0.0471
2.vs.3	0.0155	0.00200	0.0035	0.0061	0.0072	0.0084	0.0146
1.vs.$\bar{\boldsymbol{x}}$	0.0318	0.20958	0.0143	0.0215	0.0256	0.0301	0.0595
2.vs.$\bar{\boldsymbol{x}}$	0.0005	0.00399	0.0001	0.0002	0.0002	0.0002	0.0006
3.vs.$\bar{\boldsymbol{x}}$	0.0131	0.00200	0.0042	0.0064	0.0076	0.0088	0.0128

Table 1. Gene expression and DNA methylation analysis. Summary of the permutation test.

In addition, we study the association between gene expression (Y) and miRNA (X). In analogy with the previous analysis, the miRNA conditions were determined by the centroids of the clusters from the spectral clustering of the miRNA dataset. In accordance with [7], we set the number of clusters to be three. Clusters have 70, 84 and 61 patients each one. Next, from (4) we computed $T_{ij} = ||\hat{\mu}_{Y|\mathbf{x}_i} - \hat{\mu}_{Y|\mathbf{x}_j}||^2_{\mathcal{G}}$ where indexes i and j denote on which pair of vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ the conditional embeddings were compared. Figure 3 shows vectors that define the weights of the conditional embeddings (3). Samples in rows are grouped according cluster they belong. For each sample, row-normalized weights are displayed. Normalized weights change almost consistently across conditions (cluster centroids). We applied a permutation based test, using 5000 permutation samples, to evaluate the significance of the empirical values T_{ij}. We observe (Table 2) that are significant only the comparison between groups 1 and 2. Any other comparison is not significant neither comparisons between conditional embeddings and mean embedding.

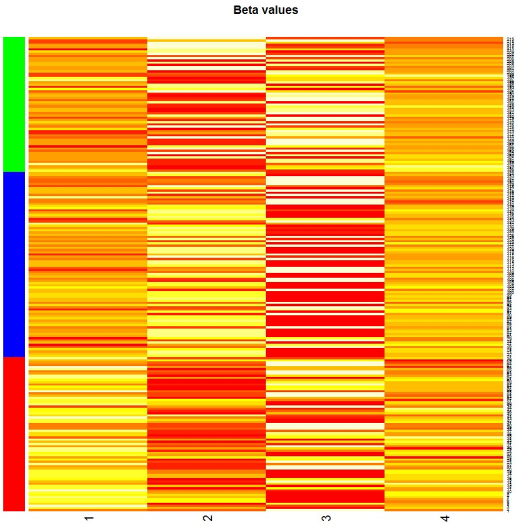

Fig. 3. Gene expression and miRNA analysis. Weights that determine the conditional embedding.

4 Conclusions

We propose a measure to integrate data in the framework of kernel methods. This methodology is based on the kernel embedding of conditional probability distributions. Our measure allows us to infer the degree of association between two types of multivariate measurements by measuring the effect on the mean element associated with the response vector when it is conditioned on different values of the explanatory vector, representing different experimental or clinical conditions.

comparison	T_{ij}	raw p-value	min	q1	med	q3	max
1vs2	0.0027	0.00200	0.0007	0.0013	0.0015	0.0017	0.0025
1vs3	0.0013	0.55090	0.0009	0.0012	0.0013	0.0016	0.0030
2vs3	0.0042	0.79840	0.0029	0.0043	0.0050	0.0059	0.0102
1.vs.$\bar{x}$	0.0001	0.84431	0.0000	0.0001	0.0001	0.0001	0.0001
2.vs.$\bar{x}$	0.0016	0.90818	0.0010	0.0018	0.0021	0.0024	0.0034
3.vs.$\bar{x}$	0.0010	0.49102	0.0006	0.0008	0.0009	0.0011	0.0021

Table 2. Gene expression and miRNA analysis. Summary of the permutation test.

References

1. Gonen M., Alpaydin E. Multiple Kernel Learning Algorithms. Journal of Machine Learning Research. 12, 2211-2268. (2011)
2. Schoelkopf, B., Smola, A. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press. (2001)
3. Smola A., Gretton A., Song L., Schoelkopf B. A Hilbert Space Embedding for Distributions. In: Hutter M., Servedio R.A., Takimoto E. (eds) Algorithmic Learning Theory. ALT 2007. Lecture Notes in Computer Science, vol 4754. Springer, Berlin, Heidelberg. (2007)
4. Fukumizu, Kenji; Bach, Francis R.; Jordan, Michael I. Kernel dimension reduction in regression. Ann. Statist. 37 , no. 4, 1871–1905. doi:10.1214/08-AOS637. https://projecteuclid.org/euclid.aos/1245332835. (2009)
5. Song, L., Fukumizu, K., Gretton. A.: Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. Signal Processing Magazine, IEEE, 30(4). (2013)
6. The Cancer Genome Atlas Network. The Cancer Genome Atlas. http:// cancergenome.nih.gov/. (2006)
7. Wang, Bo., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., Goldenberg, A.:. Similarity network fusion for aggregating data types on a genomic scale. Nature Methods. 11, 333. http://dx.doi.org/10.1038/nmeth.2810. (2014)
8. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A. kernlab – An S4 Package for Kernel Methods in R. Journal of Statistical Software. 11,9,1–20. (2004)