Nora K. Speicher[1] / Nico Pfeifer[1,2]

Towards Multiple Kernel Principal Component Analysis for Integrative Analysis of Tumor Samples

[1] Computational Biology and Applied Algorithmics, Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany, E-mail: nora@mpi-inf.mpg.de, pfeifer@informatik.uni-tuebingen.de

[2] Methods in Medical Informatics, Department of Computer Science, University of Tübingen, Tübingen, Germany, E-mail: pfeifer@informatik.uni-tuebingen.de

Abstract:
Personalized treatment of patients based on tissue-specific cancer subtypes has strongly increased the efficacy of the chosen therapies. Even though the amount of data measured for cancer patients has increased over the last years, most cancer subtypes are still diagnosed based on individual data sources (e.g. gene expression data). We propose an unsupervised data integration method based on kernel principal component analysis. Principal component analysis is one of the most widely used techniques in data analysis. Unfortunately, the straightforward multiple kernel extension of this method leads to the use of only one of the input matrices, which does not fit the goal of gaining information from all data sources. Therefore, we present a scoring function to determine the impact of each input matrix. The approach enables visualizing the integrated data and subsequent clustering for cancer subtype identification. Due to the nature of the method, no hyperparameters have to be set. We apply the methodology to five different cancer data sets and demonstrate its advantages in terms of results and usability.

Keywords: Cancer subtyping, Dimensionality reduction, Integrative kernel principal component analysis, Multiple kernel learning, Patient clustering

DOI: 10.1515/jib-2017-0019

Received: March 22, 2017; **Revised:** April 20, 2017; **Accepted:** April 27, 2017

1 Introduction

Despite improvements in its therapy, cancer remains a major health threat worldwide. One reason for this is the complexity of the disease, i.e. depending on their molecular bases, two tumors can behave very differently despite originating from the same tissue. The molecular basis of a tumor comprises alterations in a number of different characteristics of the cell, for instance gene expression [1], DNA methylation [2], and miRNA expression [3]. Although these characteristics influence each other, each of them adds new information concerning the biology of the tumor. Therefore, in recent years, the amount of data that is available for cancer patients has increased largely, not only in the number of samples and features, but also in the number of different platforms used. One challenge that goes along with this mass of multidimensional data (i.e. data describing different molecular levels of the tumor), is how to integrate and visualize it in a comprehensive and sensitive manner.

In the field of cancer research, a number of different methods have been developed that aim at using several data sources in combination in an unsupervised fashion. Consensus clustering was applied to six different data types (DNA copy number, DNA methylation, exome sequencing, mRNA expression, miRNA expression, and protein expression) for breast cancer samples from TCGA [4]. In this approach, each data type is clustered individually before generating a "cluster of clusters". This final cluster was mainly correlated to the mRNA expression clusters, which indicates that this approach is not able to capture structures that are weak but similar in different data sources. In the iCluster+ framework [5], these common associations are considered by the use of latent variables representing molecular driving factors of the cancer. Applying iCluster+ to four data types (exome sequencing, DNA copy number, promoter methylation, and mRNA expression) for colorectal carcinoma samples revealed clusters that have distinct signatures in different data types. Another data integration approach, similarity network fusion (SNF), was proposed by Wang et al. [6]. Here, each data source is represented as a sample (or patient) similarity network and these networks are combined iteratively using methods

Nora K. Speicher, Nico Pfeifer are the corresponding authors.

BY-NC-ND ©2017, Nora K. Speicher and Nico Pfeifer, published by De Gruyter.
This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

from message-passing theory. The authors applied SNF to five different cancer types with DNA methylation, miRNA expression and mRNA expression data available. In this study, they identified subtypes based on the combination of data types with distinct survival times. In contrast to these previous approaches, we employ multiple kernel learning, which provides a useful framework to optimize a weight for each input data type. In many applications, this optimization led to better results than approaches that give equal weights to the different data sources. Moreover, due to the flexibility of kernel functions the multiple kernel learning allows the integration of arbitrary data types and one does not need to know interactions or relationships between the different data sources used. In order to visualize biological data, multiple kernel learning has been used in combination with different dimensionality reduction schemes [7].

However, kernel principal component analysis (kPCA) [8], which is a widely used dimensionality reduction algorithm, is not easily extended using multiple kernel learning. The approach is based on the directions of maximum variance in the data and benefits from several advantages: kPCA does not suffer from the out-of-sample problem, one does not need to fix parameters that determine a neighborhood as in local dimensionality reduction techniques like *locality preserving projections*, but due to the use of the kernel function, the method provides enough flexibility to model different types of data. Although kPCA can be implemented in the graph embedding framework [9], due to an ill-posed eigenvalue problem, multiple kernel PCA cannot be solved using the extended framework presented in [10]. To be able to use this algorithm with multidimensional data, we introduce a scoring function for the optimization of the kernel weights before one applies kPCA to the ensemble kernel matrix. The results of subsequent k-means clustering of the projected data show that the presented method offers some advantages compared to naive kPCA approaches.

2 Methods

2.1 Multiple Kernel Learning

In general, multiple kernel learning describes the optimization of weights $\{\beta_1, ..., \beta_M\}$ for a fixed set of input kernel matrices $\{K_1, ..., K_M\}$ according to their importance [11]. The aim is to find an optimal ensemble kernel matrix $\boldsymbol{K}$, which is a weighted linear combination of the individual input kernel matrices, i.e.

$$\boldsymbol{K} = \sum_{m=1}^{M} \beta_m K_m; \;\; \beta_m \geq 0, m = 1, \ldots, M; \text{ and } \sum_{m=1}^{M} \beta_m = 1.$$

In this specific setting, each kernel matrix can be used to represent one data type. Due to the variety of available kernel functions, data with different characteristics can be included, for instance quantitative data from gene expression measurements, or sequences from genome sequencing approaches.

2.2 Kernel Principal Component Analysis

Principal component analysis (PCA) is a global dimensionality reduction approach, which uses the directions of maximum variance in the centered data. Having a data matrix X with data points x_i, the first principal component is found by optimizing

$$\arg\max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} ||\boldsymbol{\alpha}^T x_i||^2, ||\boldsymbol{\alpha}|| = 1.$$

The solution of this optimization problem is the eigenvector corresponding to the largest eigenvalue of the sample covariance matrix, subsequent principal components are calculated analogously. In the kernelized version, the data is implicitly projected into some (potentially higher dimensional) feature space using a mapping function $\phi{:}x_i \rightarrow \phi(x_i)$ with $k(x_i, x_j) = \langle\phi(x_i), \phi(x_j)\rangle$ where k is the positive semi-definite kernel function [8]. The directions of maximum variance are identified in the feature space, which is achieved by considering the largest eigenvalues of the kernel matrix and their respective eigenvectors.

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

2.3 Extending Kernel Principal Component Analysis

An extension of kPCA using multiple kernel learning would enable the user to project the data points into a low-dimensional subspace that is based on several data sources and, thereby, to visualize different characteristics of the data points in combination. However, the direct implementation, i.e.

$$\arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\| \left(\sum_{m=1}^{M} \boldsymbol{\beta}_m K_m \right) \boldsymbol{\alpha} \right\|^2,$$

$$\|\lambda_1 \boldsymbol{\alpha}^T \boldsymbol{\alpha}\| = 1;\ \boldsymbol{\beta}_m \geq 0,\ m = 1, \dots, M;\ \sum_{m=1}^{M} \beta_m = 1$$

does not allow for data integration. This becomes clear when looking at Thompson's inequality concerning the eigenvalues of sums of matrices [12]. Consider A and B being $n \times n$ Hermitian matrices and C=A+B, with their respective eigenvalues $\lambda(A)_i$, $\lambda(B)_i$, and $\lambda(C)_i$ sorted decreasingly. Then, for any $p \geq 1$,

$$\sum_{i=1}^{p} \lambda(C)_i \leq \sum_{i=1}^{p} \lambda(A)_i + \sum_{i=1}^{p} \lambda(B)_i$$

holds. If we extend this formula by including the kernel weight β_1 with C=$\beta_1 A$+$(1$–$\beta_1)B$ and $0 \leq \beta_1 \leq 1$, we obtain the following inequality

$$\sum_{i=1}^{p} \lambda(C)_i \leq \beta_1 \sum_{i=1}^{p} \lambda(A)_i + (1 - \beta_1) \sum_{i=1}^{p} \lambda(B)_i.$$

One can see, that the right hand side is maximized if the kernel matrix with the highest sum of the p largest eigenvalues has a weight of 1. In that setting, the right hand side is equal to the left hand side and, thus, this would also be the maximum of the left hand side. The extension to more than two kernel matrices can be made recursively. Therefore, optimizing Problem [1] leads to weight vectors $\boldsymbol{\beta}$ with β_i=1 and β_j=0 for all $j \neq i$, where i is the index of the matrix with the p largest eigenvalues.

Although this behavior maximizes the variance, it might not be the best choice for biological data, where we assume, that different data types can give complementing information and should therefore be considered jointly. Hence, in the following, we will introduce a scoring function, that combines the idea of kPCA with the assumption of different data supplementing each other.

2.4 Scoring Function

Dimensionality reduction methods are helpful in reducing noise while keeping the actual structure in the data. Proceeding in the spirit of kPCA, the aim of this approach is to find the ensemble kernel matrix that best preserves the global variance, but also integrates data from different sources such that they complement each other. So, for integrating M different kernel matrices $K_1, \ldots, K_M$ to an ensemble kernel matrix $\boldsymbol{K} = \sum_{m=1}^{M} \boldsymbol{\beta}_m K_m$ with $\sum_{m=1}^{M} \beta_m = 1$, we propose the following gain function:

$$g_i = \exp\left(\frac{\lambda(\boldsymbol{K})_i}{\max(\max\limits_m(\lambda(K_m)_i), 1)} - 1 \right)$$

for each dimension i, with $\lambda(K_m)_i$ being the i-th eigenvalue of K_m. Then the overall score for a projection into a p-dimensional space is calculated as $G = 1/p \sum_{i=1}^{p} g_i$, which is the average gain over the p dimensions considered. The main idea is that we define a baseline, i.e. $\max(\max\limits_m(\lambda(K_m)_i), 1)$, that represents the variance we can have by using only one matrix. Due to the use of the exponential function, gains of variance in comparison to this baseline have a strong positive impact on the score while losses of variance are penalized only slightly. Thereby, we can account for the fact that small losses of variance in one direction often do not change the global structure of the data, but allow for more variance in a subsequent direction. Additionally, we ensure that the baseline is not smaller than 1, which is the variance each direction would have in case of an equal distribution of the variance. This scoring function G is maximized to find the best kernel weights β.

Automatically generated rough PDF by ProofCheck from River Valley Technologies Ltd

3 Application

3.1 Materials

We applied the approach to five different cancer sets from TCGA [13], which were preprocessed by Wang et al. [6]. The data sets are breast invasive carcinoma (BIC; 105 samples), colon adenocarcinoma (COAD; 92 samples), glioblastoma multiforme (GBM; 213 samples), kidney renal clear cell carcinoma (KRCCC; 122 samples), and lung squamous cell carcinoma (LSCC; 106 samples). For all cancer types, gene expression, DNA methylation, and miRNA expression measurements, as well as survival data are available. The first three data types are used in the dimensionality reduction and clustering process, the latter is used to perform survival analysis in the evaluation of the identified clusters. Each of the three input data types is represented by a kernel matrix generated using the Gaussian radial basis kernel function $k(x, x')=\exp(-\gamma * ||x-x'||^2)$. The kernel width parameter γ was chosen according to the rule of thumb $\gamma = 1/2d^2$, with d being the number of features in the matrix [14]. The kernel matrices were centered in the feature space and normalized using spherical normalization [15]. Since these two steps affect each other, they were repeated iteratively until we obtained a normalized kernel matrix that is centered at the origin.

3.2 Results

For each cancer type, we generated results using a two-step procedure: First, we optimized the kernel weights according to the proposed scoring function and ran the dimensionality reduction approach in order to integrate the three data types and reduce the noise in the final projection. In similar approaches, the number of projection dimensions is usually determined either using the *elbow method* [16] or based on a chosen threshold for the remaining variance [17]. Here, we benefit from the scoring function, which indicates if we gain variance in comparison to using only one matrix. Since this function does not have a global maximum, we start by considering only the first eigenvalue (corresponding to a projection into one dimension) and increase the number of eigenvalues considered. Then, we use its first local maximum to determine the number of projection directions. Thereby, we avoid adding directions with no gain in combined variance.

In order to be able to evaluate the dimensionality reduction, we clustered the projected samples using k-means in the second step. The number of clusters was determined using the silhouette width [18] of all results from 2 to 15 clusters. For each cancer type, we evaluated the resulting clusterings by comparing the survival of the patients among the different groups using the log rank test of the Cox regression model [19]. Note that this test is based on a χ^2 distribution whose degrees of freedom is equal to the number of clusters, such that correction for multiple testing is not necessary. For comparison, we also used kPCA based on the average kernel (i.e. fixed kernel weights of $1/M$) and based on the kernel with the highest variance in the first p dimensions, which would be the trivial solution to multiple kernel PCA. The number of dimensions with the highest gain G and the p-values of the survival analysis for all three approaches can be seen in Table 1.

Table 1: Survival analysis of clustering results of kPCA used with an integrated kernel (gain function PCA), the kernel with the largest variance in the first p dimensions (max variance PCA) and average kernel PCA (average kPCA).

Cancer type	p (dimensions)	Gain function kPCA	Max variance kPCA	Average kPCA
BIC	3	**6.65E−3** (4)	0.59 (2)	**5.70E−4** (4)
COAD	2	**6.47E−3** (2)	**6.47E−3** (2)	3.28E−2 (3)
GBM	3	0.11 (5)	0.11 (5)	1.59E−2 (4)
KRCCC	3	1.37E−2 (14)	2.27E−2 (14)	0.17 (8)
LSCC	4	**7.52E−3** (3)	**7.52E−3** (3)	**9.22E−3** (3)

In brackets, the number of clusters determined by the silhouette value are given. Bold p-values refer to significant results with respect to the threshold $\alpha = 0.01$.

The number of dimensions p determined by the scoring function was for all cancer types rather small, at most four, which can be due to the fact that we used three input data types. However, as we can see in the survival analysis, this projection allowed the identification of biologically significant clusters within the cancer types.

Using the conservative significance threshold of $\alpha \leq 0.01$ we can see that our method was able to find significant clusters in three cancer types (all but GBM and KRCCC), while both other methods identified significant clusters only for two out of the five cancer types. In the GBM data, the gene expression kernel is very domi-

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

nant in terms of variance, therefore, it obtains a high weight. However, there is no clear group structure in this matrix, such that neither max variance kPCA nor gain function kPCA is able to find a clustering that correlates with the survival of the patients. For KRCCC, there is a very small group of patients with different survival behavior. The survival analysis of this clustering shows that trend for max variance kPCA and gain function kPCA (p-values < 0.05) but due to the small number of samples in each cluster, the result is not significant according to $\alpha \leq 0.01$. However, in this example, one can see that with the unweighted average of the kernel functions, the signal of interest is not captured (p-value $= 0.17$). In general, the results for the LSCC data are very stable; for all other cancer types, at least one of the naive approaches results in a clustering with no significant difference in survival times between the patient groups. This shows that using this gain function is beneficial since in cases, where only one of the naive approaches results in a significant clustering, it allows the flexibility to determine appropriate weights for the different kernel matrices.

4 Discussion

In this work, we presented a data integration method based on kernel principal component analysis. We showed that the direct extension of kPCA for several data sources does not allow for data integration. Thus, we proposed a scoring function to determine the best combination of the input data. On five cancer data sets, we showed that this procedure works in most cases better or as good as naive approaches in terms of survival analysis. Additionally, our scoring function can help to determine the number of projection dimensions. New samples from the same cancer type can be easily projected into the learnt subspace to observe similarities in the neighborhood.

Besides the survival data, the clusters could also vary in other clinical aspects (e.g. response to treatment). Investigating the response to treatment in combination with an analysis of the molecular foundation of the clusterings, for instance the identification of differentially methylated sites or differentially expressed genes, could reveal beneficial insights concerning the molecular mechanisms in tumor cells and consequently their treatment.

Conflict of interest statement: Authors state no conflict of interest. All authors have read the journal's Publication ethics and publication malpractice statement available at the journal's website and hereby confirm that they comply with all its parts applicable to the present scientific work.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd

References

[1] Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, et al. Gene expression profiles in normal and cancer cells. Science. 1997;276:1268–72.

[2] Esteller M. Epigenetics in cancer. N Engl J Med. 2008;358:1148–59.

[3] Garzon R, Marcucci G, Croce CM. Targeting microRNAs in cancer: rationale, strategies and challenges. Nat Rev Drug Discov. 2010;9:775–89.

[4] The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumors. Nature. 2012;490:61–70.

[5] Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci. 2013;110:4245–50.

[6] Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods. 2014;11:333–7.

[7] Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics. 2015;31:i268–75.

[8] Schölkopf B, Smola AJ, Müller K-R. Kernel principal component analysis. In: Schölkopf B, Burges CJC, Smola AJ, editor(s). Advances in kernel methods. Cambridge, MA, USA: MIT Press, 1999:327–52.

[9] Yan S, Xu D, Zhang B, Zhang HJ, Yang Q, Lin S. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell. 2007;29:40–51.

[10] Lin Y-YY, Liu T-LT, Fuh CC-S. Multiple kernel learning for dimensionality reduction. IEEE Trans Pattern Anal Mach Intell. 2011;33:1147–60.

[11] Gönen M, Alpaydin E. Multiple kernel learning algorithms. J Mach Learn Res. 2011;12:2211–68.

[12] Zhang F. Matrix theory: basic results and techniques, 2nd ed New York, NY, USA: Springer, 2011.

[13] The Cancer Genome Atlas. Website, Available from: http://cancergenome.nih.gov/.

[14] Gärtner T, Flach PA, Kowalczyk A, Smola AJ.. Multi-Instance Kernels. Proc 19th International Conf on Machine Learning, Morgan Kaufmann; 2002;179–86.

[15] Shawe-Taylor J, Cristianini N. Kernel methods for pattern analysis. New York, NY, USA: Cambridge University Press, 2004.

[16] Abdi H, Williams LJ. Principal component analysis. WIREs Comp Stat. 2010;2:433–59.

[17] Cangelosi R, Goriely A. Component retention in principal component analysis with application to cDNA microarray data. Biol Direct. BioMed Central. 2007;2:2.

[18] Rousseeuw PJ. Silhouettes-a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.

[19] Hosmer DW, Lemeshow S, May S. Applied survival analysis: regression modeling of time to event data, 2nd ed New York, NY, USA: John Wiley & Sons, Inc., 2008.

Automatically generated rough PDF by *ProofCheck* from River Valley Technologies Ltd