Canonical Correlation Methods for Exploring Microbe-Environment Interactions in Deep Subsurface

Viivi Uurtio[1,2], Malin Bomberg[3], Kristian Nybo[1,2], Merja Itävaara[3], and Juho Rousu[1,2(✉)]

[1] Helsinki Institute for Information Technology HIIT Department of Computer Science, Aalto University, P.O.Box 15400, FI-00076 Aalto, Finland
{viivi.uurtio,kristian.nybo,juho.rousu}@aalto.fi
[2] Helsinki Institute for Information Technology HIIT Department of Computer Science, Aalto University, Konemiehentie 2, 02150 Espoo, Finland
[3] VTT Technical Research Centre of Finland, Espoo, Finland
{malin.bomberg,merja.itavaara}@vtt.fi

Abstract. In this study, we apply non-linear kernelized canonical correlation analysis (KCCA) as well as primal-dual sparse canonical correlation analysis (SCCA) to the discovery of correlations between sulphate reducing bacterial taxa and their geochemical environment in the deep biosphere. For visualization of canonical patterns, we demonstrate the applicability of the correlation plot technique on kernelized data. Finally, we provide an extension to the visual analysis by clustergrams. The presented framework and visualization tools enabled extraction of latent canonical correlation patterns between the salinity of the groundwater and the bacterial taxonomic orders *Desulfobacterales*, *Desulfovibrionales* and *Clostridiales*.

Keywords: Canonical correlation · Kernel methods · Sparsity · Deep biosphere

1 Introduction

Multivariate analysis methods are becoming increasingly popular in uncovering the complex network of microbe-environment interactions. Various settings have been studied concerning the human microbiome [14], soil microbes related to agricultural practice [15] and microbiota in sediments associated with europhication [16]. Canonical correlation analysis (CCA) [14–16] and combinations of univariate and multivariate regression [14] including principal component analysis (PCA) [15] have been among the popular methods. Despite their popularity, these methods are limited by the assumption of linear dependencies among the variables and the fact that the resulting models are often overly complicated for human interpretation.

N. Japkowicz and S. Matwin (Eds.): DS 2015, LNAI 9356, pp. 299–307, 2015.
DOI: 10.1007/978-3-319-24282-8_25

In this paper, we examine sulphate reducing bacteria (SRB)-environment data arising from the deep biosphere research. Our data originates from deep bedrock drill holes of the Fennoscandian shield. There, SRB are observed up to several kilometer's deep [7]. SRB affect their anoxic living habitats for example by producing corrosive hydrogen sulfides. In deep geological storage of nuclear waste they may impact the long-term safety of the spent nuclear fuel storage canisters and other metallic radioactive waste [11]. In order to efficiently abate or estimate the effects of SRB the factors driving the SRB communities residing deep in the bedrock environment the physicochemical parameters driving the SRB communities must be identified. Better understanding of the deep biosphere has potential ramifications to application fields such as climate research [13] and biotechnology [8].

In microbe-environment interaction studies, the sample size is generally relatively modest and the number of variables is large. Thus, we choose to analyse microbe-environment interactions by kernel CCA (KCCA) [4] and sparse CCA (SCCA) [5], recent extensions of CCA, designed to tackle high-dimensional data through regularization, to extract non-linear dependencies and to find sparse solutions facilitating interpretation. In order to ensure statistical validity of the results, we apply in addition cross-validation to optimize model hyperparameters and randomization through permutation tests to determine the statistical significance of the discovered patterns.

Visualization of the results of multivariate analysis is challenging due to the typical high dimensionality. Frequent visualization approaches of projection-based methods, such as PCA or CCA, include score plots [15] biplots [14,16], and, more recently, correlation plots [2,3,10], all designed to project the data on a two-dimensional scatter plot, where similarity of variables or data points can be visually deduced. In this study, we first show that the correlation plot technology naturally extends to kernelized CCA variants, and go on to introduce a new clustergram visualization to represent the results of CCA-based methods, including kernelized ones. A clustergram provides an alternative dimension to the analysis of the results since it does not suffer from the problem of visual clutter that occurs in correlation plots when multiple variables have similar correlation coefficients with the projections.

2 Canonical Correlation Analysis Methods

We first present the Canonical Correlation Analysis (CCA) methods used in this paper. Let the data matrices X_a and X_b, of sizes $n \times p$ and $n \times q$, denote the views a and b respectively. The row vectors $\mathbf{x}_a^k \in \mathbb{R}^p$ and $\mathbf{x}_b^k \in \mathbb{R}^q$ for $k = 1, 2, \ldots, n$ denote the sets of empirical observations, or samples, of X_a and X_b respectively and the column vectors $\mathbf{a}_i \in \mathbb{R}^n$ for $i = 1, 2, \ldots, p$ and $\mathbf{b}_j \in \mathbb{R}^n$ for $j = 1, 2, \ldots, q$ denote centered variable vectors of the n samples respectively.

In canonical correlation analysis [6], two projection directions $\mathbf{w}_a \in \mathbb{R}^p$ and $\mathbf{w}_b \in \mathbb{R}^q$ that maximize the correlation

$$\rho = \max_{\mathbf{w}_a, \mathbf{w}_b} \frac{\mathbf{w}_a^T X_a^T X_b \mathbf{w}_b}{\|\mathbf{w}_a^T X_a\| \|\mathbf{w}_b^T X_b\|} \tag{1}$$

between the two datasets are sought for. Extending the CCA framework, recent years have put forward regularized, sparse and kernelized variants of CCA, widening the applicability of the method and to overcome limitations of CCA on high-dimensional problems [4,5].

Kernel Canonical Correlation Analysis (KCCA). [4] performs CCA by first mapping the original observations through a feature map $\phi_a : \mathbb{R}^p \mapsto \mathcal{H}_a$ to a Hilbert Space $\mathcal{H}_a$. The similarity of the objects is captured by a symmetric positive semi-definite kernel function, corresponding to the inner product in $\mathcal{H}_a$

$$K_a(\mathbf{x}_a^i, \mathbf{x}_a^j) = \langle \boldsymbol{\phi}_a(\mathbf{x}_a^i), \boldsymbol{\phi}_a(\mathbf{x}_a^j) \rangle_{\mathcal{H}_a}.$$

Using kernels K_a and K_b to map the objects in view a and b, respectively, one can express the KCCA objective by [4]

$$\rho = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^T K_a K_b \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^T K_a^2 \boldsymbol{\alpha} \cdot \boldsymbol{\beta}^T K_b^2 \boldsymbol{\beta}}}, \tag{2}$$

where $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ denote the dual variables that assign weights to the training examples. However, this optimisation problem results in a trivial correlation coefficient of value 1 when either K_a or K_b is invertible. Following [4], we solve this problem through partial Gram-Schmidt orthogonalisation (PGSO) to reduce the dimensionality of the kernels and by penalising the norms of the weight vectors by a convex combination of constraints based on Partial Least Squares to enforce non-trivial learning of the projection directions:

$$\rho = \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \frac{\boldsymbol{\alpha}^T \tilde{K}_a \tilde{K}_b \boldsymbol{\beta}}{\sqrt{(\boldsymbol{\alpha}^T \tilde{K}_a^2 \boldsymbol{\alpha} + \kappa \boldsymbol{\alpha}^T \tilde{K}_a \boldsymbol{\alpha}) \cdot (\boldsymbol{\beta}^T \tilde{K}_b^2 \boldsymbol{\beta} + \kappa \boldsymbol{\beta}^T \tilde{K}_b \boldsymbol{\beta})}}$$

Above, the kernel matrices are substituted by product of lower-triangular matrices R_a (resp. R_b) arising from PGSO approximation: $\tilde{K}_a = R_a R_a^T \cong K_a$ and $\tilde{K}_b = R_b R_b^T \cong K_b$, respectively.

Primal-Dual Sparse Canonical Correlation Analysis (SCCA). [5] seeks to maximise the correlation among subsets of features by discarding the features that do not contribute to the correlation sufficiently in comparison to others. In primal-dual SCCA, one of the views is given in the primal representation (using features) while the other is given in dual (using kernels). We denote the non-kernelized view by X_a and the kernelized view by K_b. The primal weights for X_a are denoted by $\mathbf{w}_a$ and the dual weights for K_b by $\boldsymbol{\beta}$.

$$\rho = \max_{\mathbf{w}_a, \boldsymbol{\beta}} \frac{\mathbf{w}_a^T X_a^T K_b \boldsymbol{\beta}}{\sqrt{\mathbf{w}_a^T X^2 \mathbf{w}_a \cdot \boldsymbol{\beta}^T K_b^2 \boldsymbol{\beta}}}$$

The correlation is maximised between the vectors $X_a\mathbf{w}_a$ and $K_b\boldsymbol{\beta}$. This is equivalent to minimising the 2-norm between the vectors subject to $||K_b\boldsymbol{\beta}||^2 = 1$. Since this would not result in a convex optimisation problem, the constraint is replaced by $||\boldsymbol{\beta}||_\infty = 1$. In order to force a non-trivial solution, the dual weight of one selected example that has an index k is fixed to $\beta_k = 1$ and the a constraint of 1-norm is put on the remaining entries of dual weight vector, denoted by $\tilde{\boldsymbol{\beta}} = (\beta_\ell)_{\ell \neq k}$. The 1-norm of $\mathbf{w}_a$ is also constrained to favour a sparse solution in view a. The final optimisation problem is then

$$\min_{\mathbf{w}_a, \boldsymbol{\beta}} ||X_a\mathbf{w}_a - K_b\boldsymbol{\beta}||^2 + \mu||\mathbf{w}_a||_1 + \gamma||\tilde{\boldsymbol{\beta}}||_1 \tag{3}$$

subject to $||\boldsymbol{\beta}||_\infty = 1$ where μ and γ are regularisation parameters controlling the trade-off between the function objective and the level of sparsity. The scalar μ represents the level of sparsity that controls how many of the features in X_a are discarded. The parameter γ is determined directly from the data in K_b.

3 Experiments

3.1 Data

Data consists of 43 deep bedrock groundwater samples obtained at different time points from three different sites around Finland: 15 and 11 samples from Outokumpu, Finland in 2007 and 2009, respectively, 13 samples from Olkiluoto in years 2009–2013, one from Onkalo, Finland, and three samples from Palmottu, Finland. Bacterial species were identified by dissimilatory sulphate reduction dsrB marker gene targeting which is used to identify specifically sulphate reducing microbial species. Denaturing Gradient Gel Electrophoresis (DGGE) was used to separate taxonomically variable genes which were used to construct operational taxonomic units (OTUs). The 58 DGGE bands corresponded to the binary bacterial variables that were paired with 15 geochemical variables.

3.2 Training Settings

In the SCCA algorithm [5] the constraint $||\boldsymbol{\beta}||_\infty = 1$ is fulfilled by selecting a seed example $\boldsymbol{x}_b^k$, and fixing its dual variable at $\beta_k = 1$. Following [12], we used spectral clustering to compute three medoids from the second view b to be used as candidate seeds. We compared two settings of kernel combinations. First, we performed linear analysis by applying linear kernel $K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T\mathbf{x}'$ to both views in KCCA and the kernelized view of SCCA. These settings will be referred to as L-KCCA and L-SCCA respectively. In the second setting, we used Gaussian kernel $K(\mathbf{x}_b, \mathbf{x}_b') = exp(-\frac{||\mathbf{x}_b - \mathbf{x}_b'||^2}{2\sigma^2})$ on the second (geochemical data) view b and let the first view (microbial communities) view to remain linear for both SCCA and KCCA. The motivation for this setup was to keep the data representation comparable for the two methods, whilst allowing us to impose primal sparsity on the microbial community view. These settings will be referred to as G-KCCA and G-SCCA respectively.

Table 1. Results obtained at the optimal parameter values that yielded a maximal predictive canonical correlation coefficient.

	Projections		
	1	2	3
L-SCCA	0.863	0.835	0.826
G-SCCA	0.910	0.908	0.751
L-KCCA	0.838	0.802	0.798
G-KCCA	0.965	0.962	0.948

3.3 Parameter Estimation and Statistical Significance Testing

We optimized hyperparameter μ that controls sparsity of view a in L-SCCA and G-SCCA, as well as the width of the Gaussian kernel σ in G-SCCA by 3-fold cross-validation. The same kernel width σ was also applied to G-KCCA. The model selection criterion was predictive canonical correlation, that is, the canonical correlation of test fold, using the canonical weights computed from the training fold.

The canonical correlation coefficients given the optimized hyperparameters of the first three leading projection directions are shown in Table 1. We observe that correlation coefficient of the leading projection ($k = 1$) is the greatest, as expected. The use of Gaussian kernels improve the correlations for both SCCA and KCCA. The statistical significance of the canonical correlation coefficients, given the optimized hyperparameters, were estimated using permutation tests [12]. A background data distribution consistent with the null hypothesis, H_0 : "There is no correlation between the two views of the data", was generated by permuting the rows of one view 500 times and computing the correlation coefficients for such randomized data. According to the permutation tests, the leading three canonical correlations obtained from the dataset were statistically significant in all settings at 99 % significance level. We note that the present setup corrects for multiple testing with respect to the optimal projection directions $(\mathbf{w}_a, \mathbf{w}_b)$. However, it omits multiple testing correction of the hyperparameters (μ, σ) due to large computational resource requirement of permutation tests.

3.4 Visualization of the Correlations

We visualized the canonical projections arising from SCCA and KCCA with correlation plots (c.f. [3,9,10]) based on Pearson's correlation coefficients of single variables and the projections. In particular, here we show that the technique is immediately applicable to kernelized projections, despite the fact that we do not have access to the projection weights of the variables.

Within view a, correlation coefficient between values of a single variable and the primal ($\mathbf{s}_a^k = X_a \mathbf{w}_a^k$) and dual ($\mathbf{s}_a^k = K_a \boldsymbol{\alpha}_k$) representation of the k'th

canonical projection scores are computed by

$$\rho(\mathbf{a}_i, \mathbf{s}_a^k) = \frac{\langle \mathbf{a}_i, X_a \mathbf{w}_a^k \rangle}{||\mathbf{a}_i|| ||X_a \mathbf{w}_a^k||} = \frac{\langle \mathbf{a}_i, K_a \boldsymbol{\alpha}_k \rangle}{||\mathbf{a}_i|| ||K_a \boldsymbol{\alpha}_k||} \tag{4}$$

Correlations of the two leading canonical projections in view a are used as coordinates for plotting the single variables $\mathbf{a}_i$ in view a by $(\rho(\mathbf{a}_i, \mathbf{s}_a^1)), \rho(\mathbf{a}_i, \mathbf{s}_a^2))$. The correlations and coordinates regarding view b are computed analogously.

A correlation plot showing the relations between the variables in the data, obtained by L-SCCA, is shown in Fig. 1. For example, a high correlation is observed between the DGGE band number 57, that represents the *Peptococcaceae* bacterial family, and Ca^{2+} measurements. On the other hand, there is a high negative correlation between the DGGE band number 68, that represents the *Desulfobacteraceae* family, and Ca^{2+}.

The similarities and dissimilarities between the samples can be analysed by score plots. In a score plot, the axes are the first two leading projections, $\mathbf{s}_a^k = X_a \mathbf{w}_a^k$ and $\mathbf{s}_a^k = K_a \boldsymbol{\alpha}_k$ for $k = 1, 2$ for SCCA and KCCA respectively [2]. In this case, the clusters of the samples can be interpreted by their positions in relation to the variables on the correlation plot.

A score plot on the results of L-SCCA on the data is shown in Fig. 2. In general, samples obtained from the same site at the same time point cluster together. The positions of the samples on the score plot can be explained by analysing which of the variables on the correlation plot are found in the same position. The samples obtained from very deep, OK-2-23, OK-2-21 and OK-2-19, are located in similar positions with respect to the projection axes as the

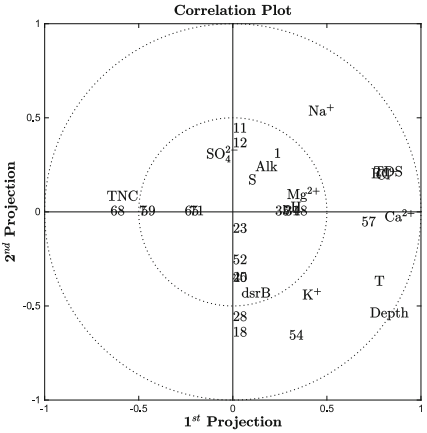

Fig. 1. Correlation plot showing L-SCCA results on dataset. The numbers represent the DGGE bands of the bacterial species and the geochemical measurements are given by their names.

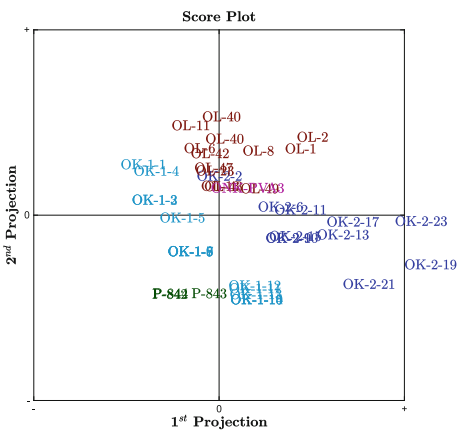

Fig. 2. Score plot showing L-SCCA results on dataset. The names of the different sampling sites are abbreviated by OK-1, OK-2, OL, P and ONK for Outokumpu 1, Outokumpu 2, Olkiluoto, Palmottu and the ONKALO respectively. The drill hole number is given after the name.

depth variable on the correlation plot. In addition, since salinity and temperature increase with depth, also the temperature and Ca^{2+} variables explain the separation of OK-2-23, OK-2-21 and OK-2-19 from the other samples.

Clustergram Visualization. In a correlation plot, visual clutter may occur since the coordinates are defined by the correlation coefficients of the variables with the two projections. In particular, this is a problem in KCCA, where no sparsity is enforced on the number of variables. Also, it is not easy to obtain an overall picture of correlations picked up by a set of projections by examining correlation plots. Here, we propose using clustergrams, frequently used in gene expression data analysis [1], in a novel way: to visualize the overall correlation of two sets of variables in a set of canonical projections. Clustergrams combine heatmaps and hierarchical clustering for visualization. To compute an entry $cg(i, j)$ in the clustergram heatmap for two variables $\mathbf{a}_i$ and $\mathbf{b}_j$ and k leading canonical projections, we compute

$$cg(i,j) = \frac{\langle \rho(i,\ell), \rho(j,\ell) \rangle}{||\rho(i,\ell)|| ||\rho(i,\ell)||}, \tag{5}$$

where $\rho(i, \ell)$ denotes the correlation of $\mathbf{a}_i$ with the ℓ'th canonical projection.

Clustergrams representing results of L-SCCA and G-SCCA are shown in Figs. 3 and 4. Both methods find similar correlation patterns but the Gaussian kernel induces more sparsity on the results than the linear kernel. When comparing the two clustergrams, the DGGE band 54 that represents the *Peptococcaceae* taxonomic family correlates positively with Ca^{2+}, depth and temperature (T). In addition, the DGGE bands 7, 59 and 68 representing the *Desulfobacteraceae* and *Desulfovibrionaceae* families, correlate negatively with depth, Cl^{-}, Ca^{2+}, total dissolved solids (TDS), electrical conductivity (EC) and temperature, in both clustergrams.

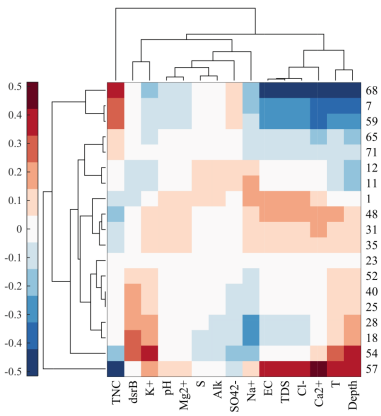

Fig. 3. Clustergram showing L-SCCA results.

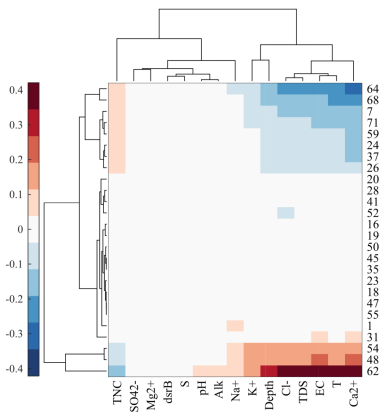

Fig. 4. Clustergram showing G-SCCA results.

4 Discussion

In this paper, we have studied primal-dual sparse (SCCA) and kernel canonical correlation (KCCA) analysis of the deep groundwater SRB communities and their geochemical environment. We have presented a data analysis framework including model selection, parameter tuning, statistical testing through permutation tests that allowed us to distill statistically significant patterns, despite a relatively modest-sized dataset. For visualization, we showed that correlation plots [3,9,10] are also applicable for kernelized setting in primal-dual SCCA and KCCA models. Finally, we introduced an alternative way to summarize the correlations in two or more projections through the use of clustergrams which provide an accessible overview to the correlations induced by the CCA projections. Indeed, the sparsifying effect of the Gaussian kernel in SCCA was first spotted by the authors from the clustergram.

Analyzing the models, higher canonical correlation coefficients were observed for the Gaussian kernel than the for linear kernel, which indicates that the data contains significant non-linear dependencies. We also observed that predictive canonical correlation coefficient assessed through cross-validation provided a good model selection criterion for SCCA.

The results in this paper help in characterizing the sulphate-reducing bacterial communities and their biochemical processes in their habitat. The discovered canonical patterns related the salinity of the groundwater, defined by the geochemical measurements Ca^{2+}, Cl^{-}, total dissolved solids and electrical conductivity, to the bacterial taxonomic orders *Desulfobacterales*, *Desulfovibrionales* and *Clostridiales*. Salinity seemed to be a unique characteristic of each of the drill hole sites based on the sample clusters of the score plots. In general, depth and temperature measurements co-occurred close together on the correlation plot which was expected, since temperature is known to increase with increasing depth below ground surface.

The software used to produce the results in this paper are available for download at https://github.com/aalto-ics-kepaco/DeepBiosphere.

Acknowledgements. Microbiology data used in this article has been generated in several earlier projects. We acknowledge financial support from Finnish Research Programme on Nuclear Waste Management KYT2010 (2006–2010) (GEOMOL project) and KYT2014 (2011–2014) Geobioinfo and GEOMICRO projects. Finnish Academy is acknowledged for funding Deep Life project (2009–2014). The work by Viivi Uurtio has been supported in part by Helsinki Doctoral Network in Information and Communication Technology HICT.

References

1. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. PNAS **95**(25), 14863–14868 (1998)

2. González, I., Déjean, S., Martin, P.G., Gonçalves, O., Besse, P., Baccini, A.: Highlighting relationships between heterogeneous biological data through graphical displays based on regularized canonical correlation analysis. J. Biol. Syst. **17**(02), 173–199 (2009)
3. González, I., Lê Cao, K.A., Davis, M.J., Déjean, S.: Visualising associations between paired omic data sets. BioData Min. **5**(1), 1–23 (2012)
4. Hardoon, D., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
5. Hardoon, D.R., Shawe-Taylor, J.: Sparse canonical correlation analysis. Mach. Learn. **83**(3), 331–353 (2011)
6. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3–4), 321–377 (1936)
7. Itävaara, M., Nyyssönen, M., Kapanen, A., Nousiainen, A., Ahonen, L., Kukkonen, I.: Characterization of bacterial diversity to a depth of 1500 m in the outokumpu deep borehole, fennoscandian shield. FEMS Microbiol. Ecol. **77**(2), 295–309 (2011)
8. Kalogerakis, N., Arff, J., Banat, I.M., et al.: The role of environmental biotechnology in exploring, exploiting, monitoring, preserving, protecting and decontaminating the marine environment. New Biotechnol. **32**(1), 157–167 (2015)
9. Lê Cao, K.A., Martin, P.G., Robert-Granié, C., Besse, P.: Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinformatics **10**(1), 34 (2009)
10. Mevik, B.H., Wehrens, R.: The pls package: principal component and partial least squares regression in r. J. Stat. Softw. **18**(2), 1–24 (2007)
11. Rajala, P., Carpén, L., Vepsäläinen, M., Raulio, M., Sohlberg, E., Bomberg, M.: Microbially induced corrosion of carbon steel in deep groundwater environment. Front. Microbiol. **6**, 647 (2015)
12. Rousu, J., Agranoff, D.D., Sodeinde, O., Shawe-Taylor, J., Fernandez-Reyes, D.: Biomarker discovery by sparse canonical correlation analysis of complex clinical phenotypes of tuberculosis and malaria. PLoS Comput. Biol. **9**(4), e1003018 (2013)
13. Waldron, P.J., Petsch, S.T., Martini, A.M., Nüsslein, K.: Salinity constraints on subsurface archaeal diversity and methanogenesis in sedimentary rock rich in organic matter. Appl. Environ. Microbiol. **73**(13), 4171–4179 (2007)
14. Wang, X., Eijkemans, M.J., Wallinga, J., Biesbroek, G., Trzciński, K., Sanders, E.A., Bogaert, D.: Multivariate approach for studying interactions between environmental variables and microbial communities. PloS One **7**(11), e50267 (2012)
15. Ye, R., Wright, A.L.: Multivariate analysis of chemical and microbial properties in histosols as influenced by land-use types. Soil and Tillage Res. **110**(1), 94–100 (2010)
16. Zeng, J., Yang, L., Li, J., Liang, Y., Xiao, L., Jiang, L., Zhao, D.: Vertical distribution of bacterial community structure in the sediments of two eutrophic lakes revealed by denaturing gradient gel electrophoresis (dgge) and multivariate analysis techniques. World J. Microbiol. Biotechnol. **25**(2), 225–233 (2009)