An Experimental Evaluation of Lifted Gene Sets

Ondřej Kuželka and Filip Železný

Czech Technical University, Prague, Czech Republic
{kuzelon2,zelezny}@fel.cvut.cz,

Abstract. We introduce so-called *lifted gene sets* which are relational descriptions of sets of genes. We show that lifted gene sets are, to some extent, able to capture common structures in correlation matrices. In experiments with real-life microarray data, we show that, in fact, lifted gene sets can provide slightly better estimates of correlation than *ground gene sets*.

1 Introduction

In this paper, we study the question whether small sets of correlated genes can be better characterized by direct conventional methods or by indirect methods using first-order-logic descriptions. We introduce so-called lifted gene sets and show that they are able to predict correlations of genes slightly better than conventional methods. The motivation for this study is not directly the estimation of correlations of genes but rather the possibility to find relational descriptions of sets of highly correlated genes because such sets are important in machine learning applications such as set-level predictive classification methods or group-lasso-based regression and classification methods [3].

2 Lifted Gene Sets

Let us start with an intuitive explanation of lifted and ground gene sets. A ground gene set is simply a set of genes, e.g. $\{IL1A, IL2, IL3\}$ is a ground gene set. Lifted gene sets are little more complicated - as we shall see, a single lifted gene set may correspond to multiple ground gene sets. An example of a lifted gene set (expressed in natural language) is *all pairs of genes sharing a transcription factor*. Thus, this lifted gene set can correspond to multiple ground gene sets where each of these ground gene sets corresponds to a pair of genes with this transcription-factor-relation. Our hypothesis is that certain lifted gene sets correspond to sets of ground gene sets with (on average) similar correlation matrices. For example, it is reasonable to expect that expressions of most pairs of genes with a transcription factor in common will be positively correlated. Similarly, we may expect positive correlation between the expression of a gene and the expression of its transcription factor. We may also expect positive correlation between a gene A and a gene B where the protein coded by the gene A activates

a protein C (by binding to it) which is a transcription factor of B etc. In the rest of this section we will be concerned with formalization of this idea.

In order to be able to predict correlations of gene sets based on their relational descriptions we need to work with training examples which have both *structure* and *real parameters*. One example may e.g. describe a measurement of the expression of several genes; here the structure would describe functional relations between the genes and the parameters would describe their measured expressions. Note that we allow different structures in different examples. For example, a training set thus may consist of measurements pertaining to different gene sets, each giving rise to a different structure of mutual relations between the genes.

To describe the training examples as well as the lifted gene sets, we use a conventional first-order logic language $\mathcal{L}$ whose alphabet contains two distinguished sets of constants $\{r_1, r_2, \ldots r_n\}$ and $\{g_1, g_2, \ldots g_n\}$ and two distinguished sets of variables $\{R_1, R_2, \ldots R_m\}$ and $\{G_1, G_2, \ldots G_m\}$. The constants g_i are meant to be used for gene names and the constants r_i are meant for the expression levels of these genes. Any substitution in our framework must map variables (other than) R_i only to terms (other than) r_j and variables (other than) G_i only to terms (other than) g_j. The structure of an example is described by a (Herbrand) interpretation H, in which the constants r_i represent uninstantiated real parameters and g_i represent the genes. The parameter values are then determined by a real vector $\boldsymbol{\theta}$. Thus each example is a pair $(H, \boldsymbol{\theta})$. A *lifted gene set* is simply a logic formula which has some free distinguished variables. Intuitively, a lifted gene set extracts some of the genes g_i and their respective expression levels r_i from the examples. For example, the intentionally simplistic lifted gene set

$$\exists G_1, G_2 \quad \mathrm{expr}(G_1, R_1) \wedge \mathrm{expr}(G_2, R_2) \wedge \mathrm{regulates}(G_1, G_2) \tag{1}$$

contains just two distinguished gene variables G_1, G_2 and two distinguished variables R_1, R_2 corresponding to gene-expression levels of G_1 and G_2.

Let us now first introduce some more notation so that we could clarify what we mean by *extracting genes and their expressions from examples*. If v is a real vector (an ordered list of genes, respectively) then v_i denotes the i-th element of v. If $I \subseteq [1; n]$ then $v_I = (v_{i_1}, v_{i_2}, \ldots v_{i_{|I|}})$ where $i_j \in I$. For the largest k such that $\{R_1/r_{i_1}, R_2/r_{i_2}, \ldots, R_k/r_{i_k}\} \subseteq \vartheta$ we denote $I_R(\vartheta) = (i_1, i_2, \ldots i_k)$ and analogically for the largest k such that $\{G_1/g_{i_1}, G_2/g_{i_2}, \ldots, G_k/g_{i_k}\} \subseteq \vartheta$ we denote $I_G(\vartheta) = (i_1, i_2, \ldots i_k)$. Given an example $e = (H, \boldsymbol{\theta})$ and a lifted gene set φ, the *sample set* of φ and e is the multi-set $\mathcal{S}(\varphi, e) = \{\boldsymbol{\theta}_{I_R(\vartheta)} | H \models \varphi\vartheta\}$ where ϑ are r-substitutions grounding all free variables[1] in φ, and $H \models \varphi\vartheta$ denotes that $\varphi\vartheta$ is true under H. Similarly, we define the *ground gene sets of a lifted gene set φ and example e* as $\mathcal{G}(\varphi, e) = \{\boldsymbol{\theta}_{I_G(\vartheta)} | H \models \varphi\vartheta\}$. For example, the the set of

[1] Note that an interpretation H does not assign domain elements to variables in $\mathcal{L}$. The truth value of a *closed* formula (i.e., one where all variables are quantified) under H does not depend on variable assignment. For a general formula though, it does depend on the assignment to its free (unquantified) variables.

ground gene sets corresponding to the example lifted gene set 1 is the set of all pairs of genes which are in the relation of *regulation*.

Our aim in this paper will be to discover lifted gene sets which correspond to highly correlated ground gene sets. Therefore we need to be able to compute *correlations* of genes within the lifted gene sets. For this we employ methods from our recently introduced framework called *Gaussian logic* [5]. Here we only briefly mention the way a covariance matrix is computed for a lifted gene set from which correlations may be easily obtained. The theoretical justification of the procedure can be found in [5].

Given a non-empty sample set $\mathcal{S}(\varphi, e)$, we define the $\boldsymbol{\Sigma}$-*matrix* as

$$\boldsymbol{\Sigma}(\varphi, e) = \frac{1}{|\mathcal{S}(\varphi, e)|} \sum_{\boldsymbol{\theta} \in \mathcal{S}(\varphi, e)} (\boldsymbol{\theta} - \boldsymbol{\mu}(\varphi, e)) (\boldsymbol{\theta} - \boldsymbol{\mu}(\varphi, e))^T \tag{2}$$

Using the above, we define the estimate $\widehat{\boldsymbol{\Sigma}}_\varphi$ over the entire training set $\{e_1, e_2, \ldots e_m\}$

$$\widehat{\boldsymbol{\Sigma}}_\varphi = \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{\Sigma}(\varphi, e_i) + \boldsymbol{\mu}(\varphi, e_i) \boldsymbol{\mu}(\varphi, e_i)^T \right) - \widehat{\boldsymbol{\mu}}_\varphi \widehat{\boldsymbol{\mu}}_\varphi^T \tag{3}$$

where

$$\boldsymbol{\mu}(\varphi, e) = \frac{1}{|\mathcal{S}(\varphi, e)|} \sum_{\boldsymbol{\theta} \in \mathcal{S}(\varphi, e)} \boldsymbol{\theta} \text{ and } \widehat{\boldsymbol{\mu}}_\varphi = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\mu}(\varphi, e_i). \tag{4}$$

We can extract the *lifted-gene-set correlations* from this matrix $\widehat{\boldsymbol{\Sigma}}_\varphi$.

Example 1. Let us have the following two examples $e_1 = (H_1, \boldsymbol{\theta}_1)$ and $e_2 = (H_2, \boldsymbol{\theta}_2)$ where

$$H_1 = g(g_1, r_1), g(g_2, r_2), g(g_3, r_3), regulates(g_1, g_2), regulates(g_2, g_3), \boldsymbol{\theta}_1 = [1, 1, 3]^T$$

$$H_2 = g(g_1, r_1), g(g_2, r_2), regulates(g_1, g_2), \boldsymbol{\theta}_2 = [2, 2]^T$$

and a lifted gene set $\varphi_1 = g(G_1, R_1), regulates(G_1, G_2), g(G_2, R_2)$. Then $\mathcal{S}(\varphi_1, e_1) = \{(1, 1), (1, 1)\}$ and $\mathcal{S}(\varphi_1, e_2) = \{(1, 0)\}$. From this we have $\boldsymbol{\mu}_{G_1} = \frac{1}{2}(\boldsymbol{\mu}(\varphi_1, e_1) + \boldsymbol{\mu}(\varphi_1, e_2)) = \frac{1}{2}([1, 1]^T + [1, 0]^T) = [1, 0.5]^T$. Furthermore, we have

$$\boldsymbol{\Sigma}(\varphi_1, e_1) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \ \boldsymbol{\Sigma}(\varphi_1, e_2) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

From this we may finally get

$$\boldsymbol{\Sigma}_{\boldsymbol{\varphi_1}} = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.5 \end{bmatrix}.$$

We could thus conclude from this limited set of three examples that the average correlation of pairs of genes where one gene regulates the other one is zero.

3 Experiments

In this section we will experimentally assess the next three questions which should give us clues about the usefulness of lifted gene sets.

- Q1: *Are lifted gene sets better predictors of correlation than ground gene sets?*
- Q2: *Are complex lifted gene sets constructed by relational machine learning techniques better predictors of correlation than gene sets corresponding to biological pathways?*
- Q3: *Are lifted-gene-set correlations stable across gene-expression datasets?*

We will answer each of these questions experimentally in the following three subsections. In all the experiments we used algorithms from [5] for construction of lifted gene sets. In each lifted gene set φ we added $\neq$ constraints for all pairs of variables except the R_i variables but for brevity we will not display them. So when we will write $\varphi = g(G_1, R_1), g(G_2, R_2)$ we will mean $\varphi = g(G_1, R_1), g(G_2, R_2), G_1 \neq G_2$ etc. In all the experiments we used a set of six gene-expression datasets obtained from the *Gene-Expression Omnibus* database [1] and relational descriptions of biological pathways from the *KEGG* database [4]. We worked with a subset of genes contained in a set of 50 KEGG pathways.

In the following sections, we will call the correlations estimated using lifted gene sets *lifted correlations* and correlations estimated conventionally *ground correlations* for brevity.

3.1 Comparison of Lifted Gene Sets and Ground Gene Sets

Let us start with the first question: *Are lifted gene sets better predictors of correlation among genes than ground gene sets?* First of all, it is not likely that lifted correlations would be better estimates of correlation for very homogeneous datasets than ground correlations (i.e. when training on a subset and then testing on another subset of a homogeneous dataset coming from the same tissue from the same individual). It is more likely that lifted gene sets will be useful in less homogeneous situations. For example, they might be useful when *transferring knowledge* from one dataset to another (which can be useful for construction of priors or for construction of shrinkage targets or more generally whenever there are too few examples in the target data we are interested in). Therefore in our experiments, we will be interested in this latter case. We will estimate correlations on one more or less homogeneous dataset and then we will use them to estimate correlations in a different dataset.

We performed the following experiment. We constructed a set of lifted gene sets with just 2 gene-variables (to make the interpretation of the results as simple as possible) and estimated the *lifted correlations*, i.e. the correlations computed according to Formula 3, for each gene-expression dataset separately. We also estimated the correlations using a conventional shrinkage-based method [6] (separately for each pathway). After that we used these values for prediction

of correlations in datasets not used for training. We compared the errors incurred when using lifted gene sets and when using the conventional estimates (ground gene sets).

Let us now explain in detail what we mean by the *errors* and how we computed them. For two datasets *Train* and *Test* we computed the errors of the lifted gene sets as follows. For each constructed lifted gene set φ, we computed its covariance matrix $\boldsymbol{\Sigma}_\varphi$ using only the dataset *Train* and from this matrix we extracted the correlations. Then we found the corresponding two-element ground gene sets (recall that we have deliberately limited the number of genes to two for the sakes of easier interpretability) and for each of these ground gene sets we computed the correlations using the dataset *Test* (this is considered the golden standard in our experiments since, of course, we do not know the *true* correlations) and using this we computed the root mean squared errors for all the lifted gene sets. This gave us a somehow interpretable number for each lifted gene set and a dataset - we present these results in the form of *heat maps* in Fig. 1. However, we would also like to have a single number which would represent the overall quality of the lifted and ground gene sets. We used a weighted average of the root mean squared errors of the individual lifted gene sets where the weight of a lifted gene set was the number of the respective ground gene sets. In what follows, we will also present more detailed views on this error - we will present the errors as bar-graphs where the height of each bar will correspond to the contribution of one lifted gene set to the overall error.

The results are shown in Fig. 1. On average, the error incurred when estimating the correlations using lifted gene sets was lower in four out of six datasets as can be seen from Fig. 2. This was a promising result. However, it was perhaps still not yet conclusive enough. So we reasoned as follows. Some biological pathways might need to be controlled more tightly than others, so maybe, if we estimated the correlations for the Gaussian features from individual pathways and then

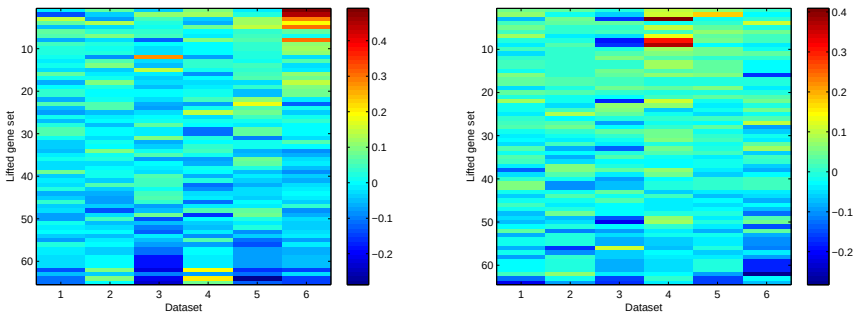

Fig. 1. Difference between the error of ground gene sets and lifted gene sets across 6 gene-expression datasets (the more positive the number the better for lifted gene sets). **Left:** lifted correlations estimated on the whole set of pathways. **Right:** Lifted correlations estimated on individual pathways.

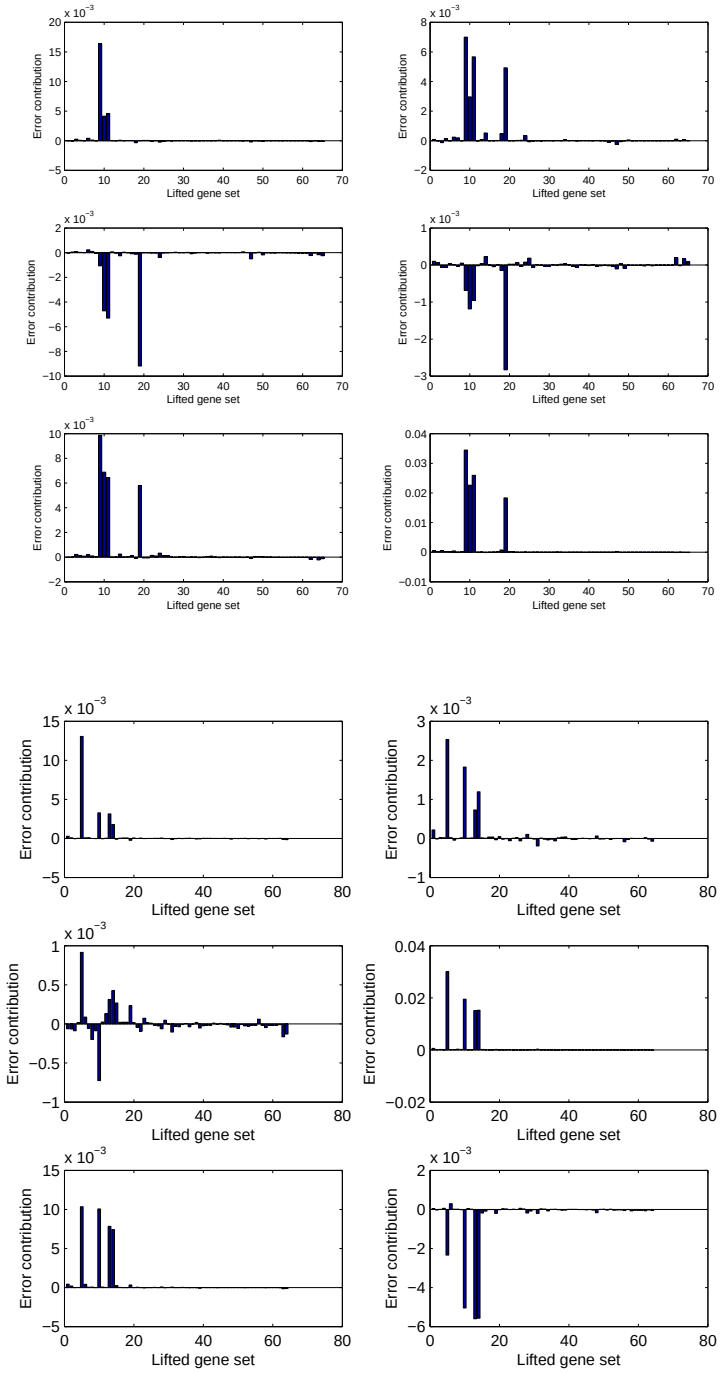

Fig. 2. Contributions to differences between the errors of ground gene sets and lifted gene sets across 6 gene-expression datasets (the more positive the number the better for lifted gene sets). **Top:** lifted correlations estimated on the whole set of pathways. **Bottom:** Lifted correlations estimated on individual pathways.

used them as estimates for the same pathways (but of course from different datasets), we might be able to capture this phenomenon and thus improve the predictive ability. When we actually did this, we indeed obtained slightly better results (cf. Fig. 2 where the summary errors are represented through individual lifted gene sets' contributions) - the lifted correlations turned out to be better estimates of correlation than ground gene sets on five out of the six datasets.

3.2 Comparison of Lifted Gene Sets and Pathway-based Gene Sets

The results from the previous subsection indicate that lifted gene sets perform better than ground gene sets. A simple type of a lifted gene set is a lifted gene set which assumes all pairs of genes which are contained in some pathway P. This type of gene sets is, in fact, quite popular in machine learning applications (pathway-based gene sets are often used in set-level predictive classification methods, e.g. [2]). It is therefore an interesting question whether our more complex lifted gene sets based on relations from KEGG are able to estimate correlations between genes more reliably than the traditional pathway-based gene sets. In order to answer this question we performed the same procedure as described in the previous subsection but instead of ground gene sets we used the pathway-based gene sets. The results are shown in Fig. 4 and 3. In this case the lifted gene sets do not perform very well, as can be seen from Fig. 4, for many combinations *lifted gene set - dataset*. This means that the information about the pathway in which a given pair of genes is contained is useful for estimation of correlations. We should not be surprised by this because in the previous subsection we have witnessed something similar when accuracy of lifted correlations improved when we estimated them on individual pathways rather than on the whole sets of pathways. Therefore we performed an additional experiment in which we tested the estimates obtained by averaging the predictions of lifted gene sets and pathway-based gene sets (*combined gene sets*). Indeed, the accuracy improved quite significantly as can be seen from Fig. 3 where the combined gene sets outperformed the pathway-based gene sets on four out of the six datasets. The results could be further improved by estimating the lifted gene sets again only on the individual pathways.

3.3 Stability of Lifted-Gene-Set Correlations across Datasets

Finally, we also performed a set of experiments in order to determine stability of lifted-gene-set correlations across various gene-expression datasets. In order to do so we again estimated the correlations of the individual lifted gene sets separately for each dataset and plotted them in Fig. 5. We can notice that on average the correlation of the lifted gene sets does not change very much across datasets with some notable exceptions which may or may not be interesting from the biological point of view.

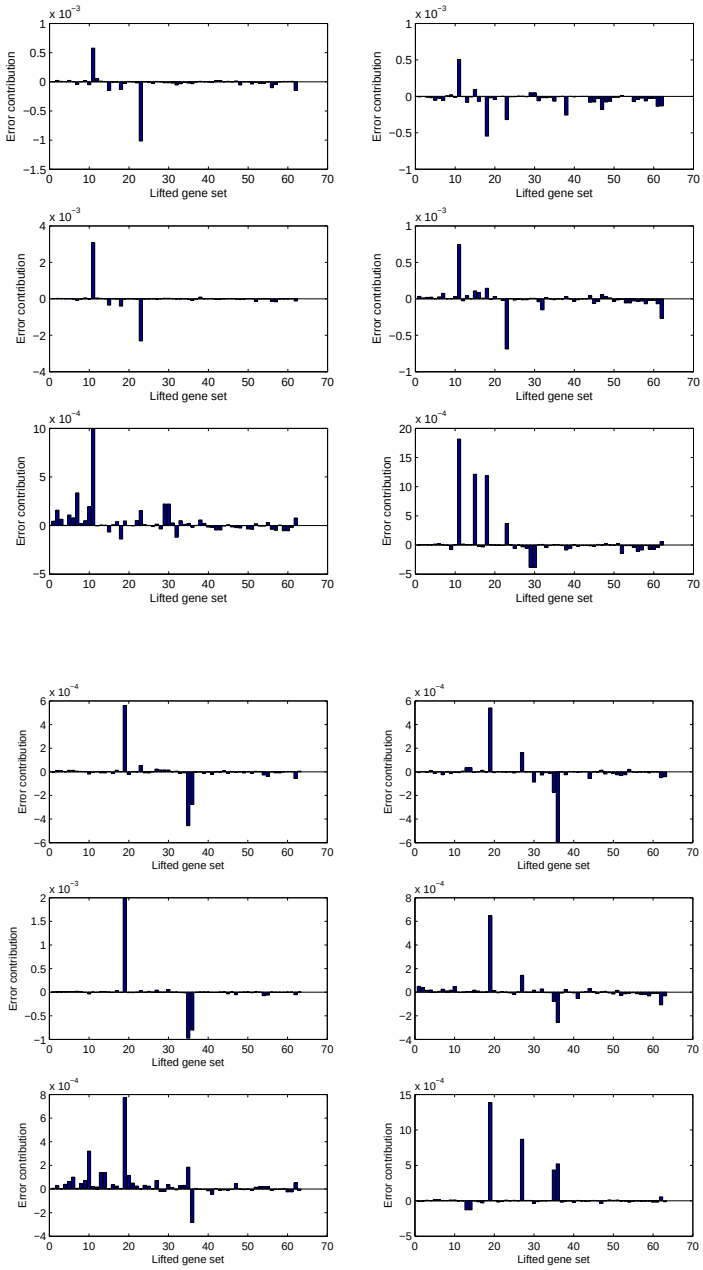

Fig. 3. Top: Contributions to differences between the errors of pathway-based gene sets and lifted gene sets (the more positive the number the better for lifted gene sets). **Bottom:** Contributions to differences between the errors of pathway-based gene sets and lifted gene sets combined with the pathway-based gene sets.

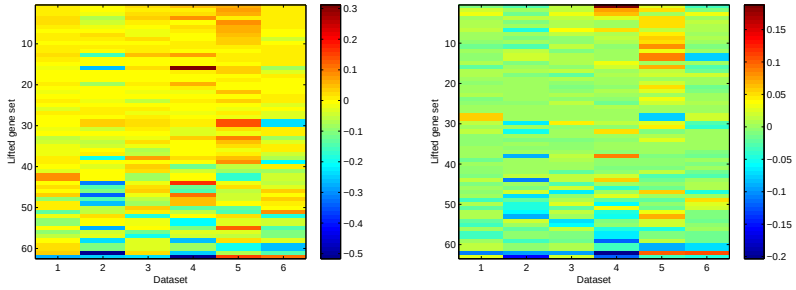

Fig. 4. Differences between the error of pathway-based gene sets and lifted gene sets across 6 gene-expression datasets (left panel) and average difference between the error of pathway-based gene sets and combined gene sets - the more positive the number the better for the combined gene sets.

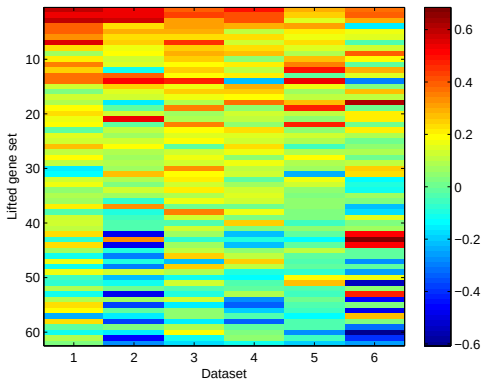

Fig. 5. Average correlations of lifted gene sets across 6 gene-expression datasets.

4 Discussion and Conclusions

The experiments that we have performed in this paper give some clues as for the potential of lifted gene sets. However, they also show that simple gene sets like pathway-based gene sets can perform similarly well as the more complex lifted gene sets. They also show that combinations of the simple and the more complex gene sets can sometimes improve ability to predict correlation between genes.

Acknowledgement: This work was supported by the Czech Grant Agency through project 201/09/1665 *Bridging the Gap between Systems Biology and Machine Learning* and project 103/11/2170 *Transferring ILP techniques to SRL.*

References

1. R. Edgar, M. Domrachev, and A. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 1, 2002.
2. M. Holec, F. Železný, J. Kléma, and J. Tolar. Integrating multiple-platform expression data through gene set features. In *Proceedings of the 5th International Symposium on Bioinformatics Research and Applications*, ISBRA '09, pages 5–17. Springer-Verlag, 2009.
3. L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 433–440. ACM, 2009.
4. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The kegg resource for deciphering the genome. *Nucleic Acids Research*, 1, 2004.
5. O. Kuželka, A. Szabóová, M. Holec, and F. Železný. Gaussian logic for predictive classification. In *To appear in Proceedings of the European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases*, ECML PKDD, 2011.
6. J. Schäffer and K.Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 2005.