Artificial Intelligence in Medicine 70 (2016) 1–11

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

Traveling on discrete embeddings of gene expression

Pietro Lovato[a,*], Manuele Bicego[a], Maria Kesa[b], Nebojsa Jojic[c], Vittorio Murino[d], Alessandro Perina[c]

[a] Department of Computer Science, University of Verona, Strada le Grazie 15, 37134 Verona, Italy
[b] Tallinn University of Technology, Ehitajate tee 5, 19086 Tallinn, Estonia
[c] Microsoft Research, One Microsoft Way, 98052 Redmond, WA, USA
[d] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia (IIT), Via Morego 30, 16163 Genova, Italy

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form 20 May 2016
Accepted 21 May 2016

Keywords:
Data visualization
Bag of words clustering
Bag of words classification
Gene selection
Gene expression matrix

ABSTRACT

Objective: High-throughput technologies have generated an unprecedented amount of high-dimensional gene expression data. Algorithmic approaches could be extremely useful to distill information and derive compact interpretable representations of the statistical patterns present in the data. This paper proposes a mining approach to extract an informative representation of gene expression profiles based on a generative model called the Counting Grid (CG).
Method: Using the CG model, gene expression values are arranged on a discrete grid, learned in a way that "similar" co-expression patterns are arranged in close proximity, thus resulting in an intuitive visualization of the dataset. More than this, the model permits to identify the genes that distinguish between classes (e.g. different types of cancer). Finally, each sample can be characterized with a discriminative signature – extracted from the model – that can be effectively employed for classification.
Results: A thorough evaluation on several gene expression datasets demonstrate the suitability of the proposed approach from a twofold perspective: numerically, we reached state-of-the-art classification accuracies on 5 datasets out of 7, and similar results when the approach is tested in a gene selection setting (with a stability always above 0.87); clinically, by confirming that many of the genes highlighted by the model as significant play also a key role for cancer biology.
Conclusion: The proposed framework can be successfully exploited to meaningfully visualize the samples; detect medically relevant genes; properly classify samples.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Technologies such as gene expression microarrays and RNA-seq provide scientists with a way to measure the expression levels of thousands of genes simultaneously. Computational approaches are increasingly needed to manage this amount of data, and are effectively helping researchers to unravel the complexity of biological systems. Examples of computational problems related to the analysis of a gene expression matrix (a matrix containing the expression level of different genes under different experimental conditions) are classification of samples [1–4], clustering of genes or pathological subtypes [5,6], and selection of differentially expressed or discriminative genes [7].

Other than sophisticated methods of quantitative analysis, high-throughput experiments brought also the need for visualization, thoughtful validation, and, more generally, a deeper understanding of the phenomenon under investigation. For these reasons, interpretable models are required. In this context, generative models (in particular, topic models and latent process models [8]) have been shown to provide highly interpretable solutions, more than achieving high accuracy for classification tasks [9,10]. Within this literature, topic models have been either designed ad hoc for gene expression analysis [11,12], or exported from Natural Language Processing by postulating an analogy between textual documents and microarray samples [10,13]. In the latter case, the starting point is to see a gene expression profile (i.e. a sample) as a "bag of words" vector [14] – a numerical vector in which every entry counts how many times each "word" of a pre-defined dictionary occurs in the considered document. Similarly to text documents, a gene expression profile can be seen as a bag of words vector – genes now represent the words – since each entry measures the

* Corresponding author. Tel.: +39 0458027804; fax: +39 0458027068.
E-mail address: pietro.lovato@univr.it (P. Lovato).

http://dx.doi.org/10.1016/j.artmed.2016.05.002
0933-3657/© 2016 Elsevier B.V. All rights reserved.

intensity of expression of each gene (which indirectly reflects the amount of mRNA transcripts). This analogy also permits to exploit topic models in this context [10,13], which, by introducing the concept of "topic", allow to model co-occurrence (or co-expression) patterns within the data. Topics are latent distributions that assign high probability to co-occurring "words", and act as intermediate descriptors of samples (in the gene expression case, they can be associated to biological processes, as shown in [10,13]).

However, a common assumption of most topic models is that the topics act independently of each other. While this assumption is often needed to simplify computations and inference, it may be too simplistic in the gene expression scenario, where it is known that biological processes are tightly co-regulated and interdependent in a complex way. In this paper we make a step forward along this research line – pursuing the topic model philosophy, but coping with the afore-described limitation – presenting a novel strategy to extract an informative representation for a set of experimental samples through a recent generative model called Counting Grid (CG – [15]). The Counting Grid represents a probabilistic model for objects represented as "bag of words", that was recently introduced for text mining [15] and image processing [16]. The idea behind the model is that the topics are arranged on a discrete grid, learned in a way that "similar" topics are closely arranged. Similar biological samples, i.e. sharing topics and active genes, are mapped close on the grid, allowing for an intuitive visualization of the data set. More specifically, the CG seems to be very suitable in the gene expression scenario for the following reasons:

- The CG provides a powerful representation, which permits to capture evolution of patterns in experiments, and can be clearly visualized.
- The CG is well suited for data that exhibit smooth variation between samples. Expression values are biologically constrained to lie within certain bounds by purifying selection [17] and variation in only a few expression values can cause a pathology. This specific property of the data is captured well by the model.
- The CG permits a principled and founded way to extract the most relevant genes that are associated with a disease [18].
- Last, but not least, it is possible to achieve a better classification accuracy with respect to other topic model approaches, as well as to the recent state of the art.

In this paper, we comprehensively evaluate the CG model for mining and modeling gene expression data; we start from the preliminary findings which appeared in the literature [18,19], but we thoroughly evaluate the capabilities of the model with respect to the following novel aspects:

1. By visualizing different data sets, we show that samples belonging to different biological conditions (such as different types of cancer) cluster together on the grid, supporting this claim with a numerical validation (Section 4.1).
2. We systematically tested the accuracy of the CG model both in a gene selection and in a classification setting, experimenting on 7 different benchmark datasets, obtaining results comparable with the recent state-of-the-art.
3. We prove that the model is able to highlight genes that are involved in the pathology or in the phenomenon which motivated the experiment; moreover, the selected genes have a beneficial effect when used for classification, quantitatively comparable with other gene selection techniques.
4. We evaluate the sensitivity of the model to parameters such as grid and window size and the robustness of the model to overfitting.

2. Methods

2.1. The Counting Grid model

In machine learning research, a data point is often represented as a "bag of words": the representation is obtained by counting how many times each "word" (i.e. constituting feature) occurs in the object. This paradigm can represent in a vector space many types of objects, even ones that are non-vectorial in nature. However, one drawback is that in some domains and applications it destroys the possible structure of objects. A clear example can be found in the Natural Language Processing domain (where the bag of words has been originally introduced): by representing a document as a vector of word counts, the ordering of the words in such document is lost.

Recently, an analogy has been established between the Natural Language Processing and the gene expression contexts [10]: the idea is to directly interpret the gene expression matrix as a bag of words, where genes represent words and a sample $\mathbf{s}^t$ represents a document. The expression value can be seen as a count: the higher the expression, the higher the number of transcripts that will be translated into fully functional proteins. In the past, such bag of words representation of gene expressions has been successfully modeled with topic models [10]: these models, introduced in the text mining community, learn a small number of topics which correlate related genes particularly active in a subset of samples. However, there are no strong constraints in how topics are mixed, because they are assumed to be statistically independent. This is a strong drawback, overcame in the Counting Grid model by arranging these distributions representing topics on a discrete grid with topological constraints: intuitively, similar "topics" are located nearby on the grid, and have similar genes' distributions.

Formally, the Counting Grid $\pi_{\mathbf{i},z}$ is a D-dimensional discrete grid, of size $\mathbf{E} = (E_1, \ldots, E_D)$. Each position on the grid is indexed by $\mathbf{i} = (i_1, \ldots, i_D)$, where $i_d \in \{1, \ldots, E_d\}$. Each cell represents a tight distribution over genes (indexed by z), so $\sum_z \pi_{\mathbf{i},z} = 1$. A given sample $\mathbf{s}^t$, represented by expression values $\{g_z^t\}$ is assumed to follow a distribution found in a *window* of dimensions $\mathbf{W} = (W_1, \ldots, W_D)$ somewhere in the counting grid. The window is identified by the location $\mathbf{k}$ (upper-left corner of the window) and includes the grid region $W_{\mathbf{k}} = [\mathbf{k} \ldots \mathbf{k} + \mathbf{W}]$, that is the region starting from the location $\mathbf{k}$ (upper-left corner of the window) and extending in each direction d by W_d grid positions. For example, in Fig. 1 we show a bidimensional CG containing 10 × 10 cells ($\mathbf{E} = (10, 10)$), where the window has size $\mathbf{W} = (3, 3)$. Assuming that the sample $\mathbf{s}^t$ is generated from the window which starts in position $\mathbf{k} = (3, 8)$, the distribution of its genes is defined as the average of all the distributions from $\pi_{(3,8),z}$ to $\pi_{(5,10),z}$ (zoomed in the right part of Fig. 1). Mathematically, this average – given a gene indexed by z in sample $\mathbf{s}^t$ – is computed as:

$$h_{\mathbf{k},z} = \frac{1}{\prod_d W_d} \sum_{\mathbf{i} \in W_{\mathbf{k}}} \pi_{\mathbf{i},z} \tag{1}$$

A consequence of this can be seen in Fig. 2: if we consider two samples located nearby (s^1 and s^2 in the figure), we note that they share some cells on the grid, and for this reason their genes' distributions will be similar. In other words, spatial proximity implies similarity of expression values.

More formally, the position (upper-left corner) of the window $\mathbf{k}$ in the grid is a latent variable, given which the probability of the bag of words $\{g_z^t\}$ for sample $\mathbf{s}^t$ is

$$p(\{g_z^t\}|\mathbf{k}) = \prod_z (h_{\mathbf{k},z})^{g_z^t} = \left(\frac{1}{\prod_d W_d}\right) \prod_z \left(\sum_{\mathbf{i} \in W_{\mathbf{k}}} \pi_{\mathbf{i},z}\right)^{g_z^t} \tag{2}$$

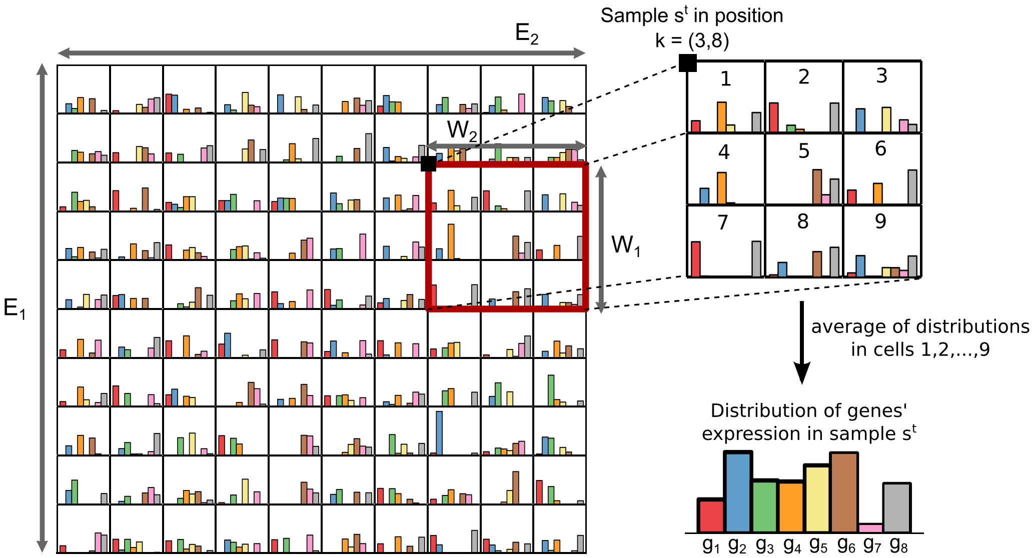

Fig. 1. On the left, an example of a Counting Grid: each cell in the grid correspond to a topic, i.e. a genes' distribution. On the right, we sketch the generative process: to generate a sample, we first select a fixed-size window (individuated by its upper-left corner **k**), and then we generate the expression of a gene by taking the average of its values in the window.

which can be thought of as the probability that the observed genes' expressions $\{g_z^t\}$ are generated by a multinomial distribution with parameter $h_{\mathbf{k},z}$.

Relaxing the terminology, we will refer to **E** and **W** respectively as the counting grid size and the window size, indicating with $W_{\mathbf{k}}$ the particular window placed at location **k**. We will refer to the ratio of the window volumes, $\kappa = \prod_d (E_d/W_d)$, as the capacity of the model in terms of an *equivalent number of topics*, as this represents how many non overlapping windows can be fit onto the grid.

To learn a Counting Grid, we need to maximize the log likelihood of the data:

$$\log P = \sum_{t=1}^{T} \log \left(\sum_{\mathbf{k}} \prod_{z} h_{\mathbf{k},z}^{g_z^t} \right) \tag{3}$$

The sum over the latent variables **k** makes it difficult to perform assignment to the latent variables while also estimating the model parameters. The problem is solved by employing an EM procedure [20], which iteratively learns the model by minimizing a bound $\mathcal{F}$ on the log likelihood $\log P$ by alternating the E and M-step. $\mathcal{F}$ is often referred to as the free energy of the model and, for the CG model, is equal to

$$\log P \geq \mathcal{F} = -\sum_{t}\sum_{\mathbf{k}} q_{\mathbf{k}}^t \cdot \log q_{\mathbf{k}}^t + \sum_{t}\sum_{\mathbf{k}} q_{\mathbf{k}}^t \sum_{z} g_z^t \cdot \log \sum_{\mathbf{i} \in W_{\mathbf{k}^t}} \pi_{\mathbf{i},z} \tag{4}$$

where $q_{\mathbf{k}}^t = P(\mathbf{k}|\mathbf{s}^t)$ is the distribution over the latent mapping onto the counting grid of the tth sample. The E-step estimates $q_{\mathbf{k}}^t$, a quantity representing the probability of a bag t being generated from a position **k** of the grid. The M-step re-estimates the counting grid π given the current q. As a final consideration, it is important to consider the counting grid as a torus, and perform all windowing operations accordingly. We sketched the pseudo-code for learning a CG in Algorithm 1; interested readers can also refer to [15].

Algorithm 1. EM-algorithm to learn a counting grid

```
Require: Gene expression matrix $\{\mathbf{s}^t\}$, $t = 1 \ldots T$
while EM not converged do
  % E-step
  for each sample $t = 1 \ldots T$ do
    Update $q_{\mathbf{k}}^t \propto \exp\left[\sum_z g_z^t \log h_{\mathbf{k},z}\right]$
  end for
  % M-step
  Update $\pi_{\mathbf{i},z} \propto \pi_{\mathbf{i},z}^{\text{old}} \cdot \sum_t g_z^t \sum_{\mathbf{k}|\mathbf{i} \in W_{\mathbf{k}}} \frac{q_{\mathbf{k}}^t}{h_{\mathbf{k},z}}$
  Compute $h_{\mathbf{k},z} = \frac{1}{\prod_d W_d} \sum_{\mathbf{i} \in W_{\mathbf{k}}} \pi_{\mathbf{i},z}$
  Compute the free energy $\mathcal{F}$ with Eq. (3)
  Check for convergence: $|\mathcal{F} - \mathcal{F}^{\text{old}}| < \epsilon$
end while
return $\pi_{\mathbf{i},z}$, $q_{\mathbf{k}}^t$
```

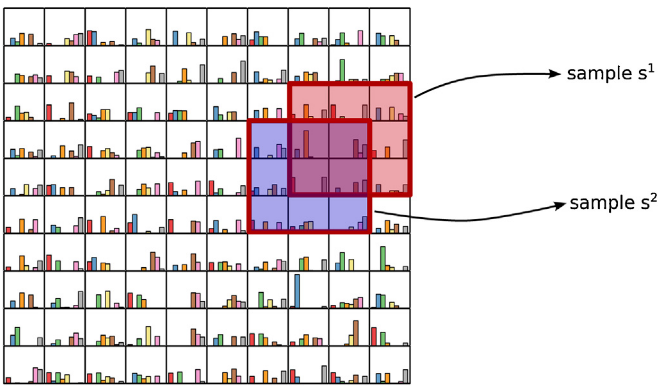

Fig. 2. Samples located nearby on the grid share some locations: in this way, the average of their gene expression will be similar.

2.2. *Computational efficiency*

Careful examination of the steps in Algorithm 1 reveals that by the efficient use of cumulative sums, both the E and M steps are linear in the size of the counting grid. Both steps require computing $\sum_{\mathbf{i} \in W_{\mathbf{k}}} \pi_{\mathbf{i}}$ (or a quantity related to π), which can be done by first computing – in linear time – the cumulative sums of π and then computing appropriate linear combinations. For example, in the 2D case we have $\mathbf{i} = (i, j)$, $\mathbf{k} = (k, l)$ and one can compute the cumulative sum $F_{m,n} = \sum_{(i,j) \leq (m,n)} \pi_{i,j}$ and then set $\sum_{(i,j) \in W_{(k,l)}} \pi_{i,j} = F_{k+W_1+1,l+W_2+1} - F_{k,l+W_2+1} - F_{k+W_1+1,l} + F_{k,l}$. An intuitive explanation of this is portrayed in Fig. 3. Finally, note that the E-step is linear

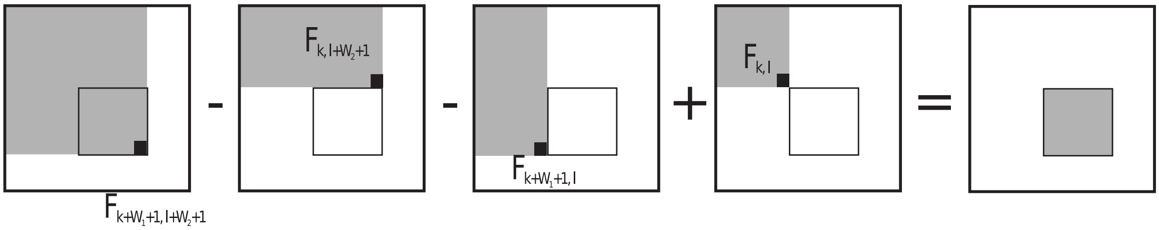

Fig. 3. Cumulative sums allow to efficiently compute the sum of the elements in an arbitrary window inside the grid.

in the number of genes, making the learning procedure efficient in cases of high-dimensional gene expression data.

2.3. Class embedding and biomarker identification

In the majority of the datasets of gene expression, samples are equipped with a label that reflects its category, such as the pathological subtype, or the different tissue of the organism under analysis. Let us call these labels $y^t = l$, $l = [1, \ldots, L]$, each y^t representing the class index of the sample t. Once a Counting Grid is learned and the location of each sample is probabilistically located on the grid (by looking at $q^t_{\mathbf{k}}$), it is possible to obtain a posterior probability of each class $p(l|\mathbf{i}) = \gamma_l(\mathbf{i})$ in each position $\mathbf{i}$. This is achieved using the posterior probabilities $q^t_{\mathbf{k}}$ already inferred in the EM:

$$\gamma_l(\mathbf{i}) = \frac{\sum_t \sum_{\mathbf{k}|\mathbf{i} \in W_{\mathbf{k}}} q^t_{\mathbf{k}} \cdot [y^t = l]}{\sum_t \sum_{\mathbf{k}|\mathbf{i} \in W_{\mathbf{k}}} q^t_{\mathbf{k}}} \tag{5}$$

where $[\cdot]$ is the indicator function, that indicates membership of an element in the class: the output is 1 if sample t belongs to class l, 0 otherwise. Intuitively, we are "averaging" the posterior (i.e. the probability of being generated from a specific location) of all points belonging to class l. This results in a map for every class l, of the same size of the CG, which clearly indicates which regions of the CG better "explain" the class labeled by l (the higher the posterior $p(l|\mathbf{i})$ in certain zone, the higher the probability of finding samples of class l in such part of the grid).

Starting from this $\gamma_l(\mathbf{i})$, it is now possible to derive information about which genes are more relevant for class l. The main idea is the following: genes that are mostly distinguishing of the class l are the ones which vary most along the boundary that separate the classes. Once a class is embedded in the grid, computing the gradient $\vec{v}$ of this embedding, $\vec{v} = \nabla\gamma_l(\mathbf{i})$, may provide information about where and how this class separates from the others. Genes that are mostly distinguishing of the class are the ones which vary most in the direction of the class separation. This information can be captured mathematically by the directional derivatives of the grid $\pi_{z,\mathbf{i}}$ in the direction $\vec{v}$ of the class gradient. Thus, we can derive a score for each gene (the higher the score, the more important the gene in discriminating the class from the others) by summing such quantity over all the locations $\mathbf{i}$, multiplying it by the module of v to reward more the variation in expression where we have a high variation between classes. In formulae, the score for each gene g_z is equal to:

$$F_z = \sum_{\mathbf{i}} \left| |\vec{v}| \cdot \frac{\vec{v}}{|\vec{v}|} \cdot \nabla\pi_{z,\mathbf{i}} \right| = \sum_{\mathbf{i}} \left| \vec{v} \cdot \nabla\pi_{z,\mathbf{i}} \right| \tag{6}$$

In the formula, we take the absolute value because we considered as equally relevant genes which under express in the transition to class l and genes which over express in the transition to class l.[1]

3. An illustrative example: mining yeast expression

To illustrate the main features of the proposed framework we present a simple example, where we studied a dataset by DeRisi et al. [21], measuring the gene expression of 6400 genes in *Saccharomyces cerevisiae* during the metabolic shift from fermentation to respiration. Expression values have been measured at 7 different time points. From our point of view, each time point is a bag $\mathbf{s}^t = \{g^t_z\}, z = 1, \ldots, 6400$. As done in other applications, we performed a filtering of the genes,[2] obtaining a final refined dataset of 310 gene expression values at 7 time points.

We learned the CG using these 7 samples, by setting the parameter κ to 4: specifically, we opted for a 12×12 grid with a 6×6 window for a clearer visualization. In the left part of Fig. 4 we provide a visualization of the mapping position on the learned CG of the 7 experiments – each red dot corresponds to the maximum of the q^t, i.e. to the most probable position of a given time point t. The highlighted path connects the temporal transition between the 7 time points, permitting a clear understanding of the dataset. By looking at this embedding, it seems that the more pronounced transition occurs between the 3rd and the 4th time points. Thus, we roughly divided the dataset in 2 classes: we can see the distribution of the "respiration" class (samples 4-5-6-7), i.e. the map $\gamma_l(\mathbf{i})$, in the right part of Fig. 4. From this map, we computed the gradient of $\gamma_l(\mathbf{i})$ (portrayed in Fig. 5), and identified the genes which vary the most across the direction of the gradient, as described in Section 2.3. For example, the gene highlighted in the zoomed portion of the grid (Fig. 5) is gad1, which seems to rapidly activate during the transition from fermentation to respiration. This is in line with previous findings reported in the literature [22,23].

Then, we performed a selection of relevant genes by using the framework described in Section 2.3, and reported the list of the top 10 genes selected in Table 1. To prove that these genes are indeed relevant from a biological point of view, we looked for terms in the Gene Ontology [24] which are highly over-represented among these 10 genes, with respect to all other terms pertaining the remaining 300 genes. Statistically significant ($p < 0.05$) terms are reported in Table 2, and they are interestingly related to synthesis of sugar and response to oxidative stress. The p-value is computed employing a chi-squared test with Benjamini multiple hypothesis correction (more details can be found in [25]).

This simple example permits to show the main features of the proposed framework: (i) the 7 experiments are projected in the grid

[1] With a slight abuse of terminology, by under expression we intend a gene whose expression tend to decrease in class l compared to the expression of the same gene in the other classes.

[2] Following http://www.mathworks.com/help/bioinfo/examples/gene-expression-profile-analysis.html (accessed 19 May 2016).

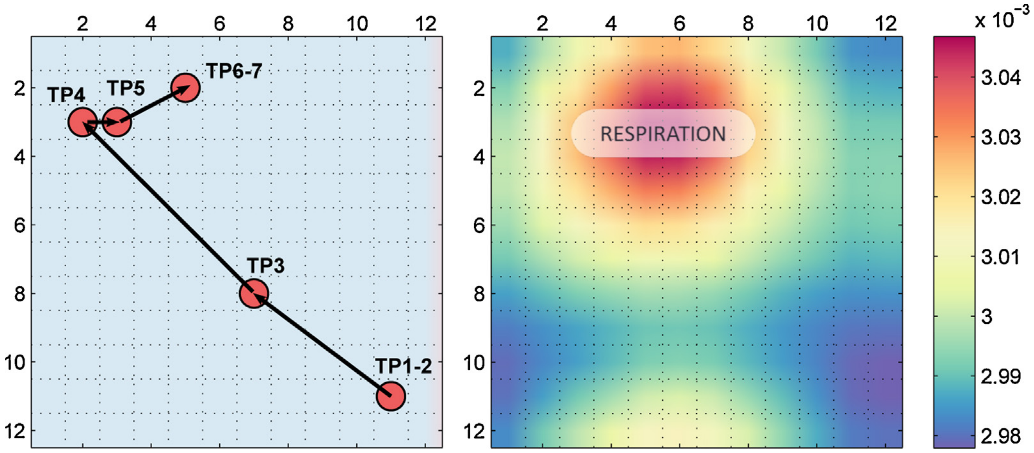

Fig. 4. Yeast dataset embedding. Each time point is placed in a location of the grid, highlighted in red (left part of the figure). There is a clear path connecting the dots: since the most pronounced transition occurs between the 3rd and the 4th time points, in the right part we show the class embedding $\gamma_l(\mathbf{i})$ of samples $l = \{4, 5, 6, 7\}$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

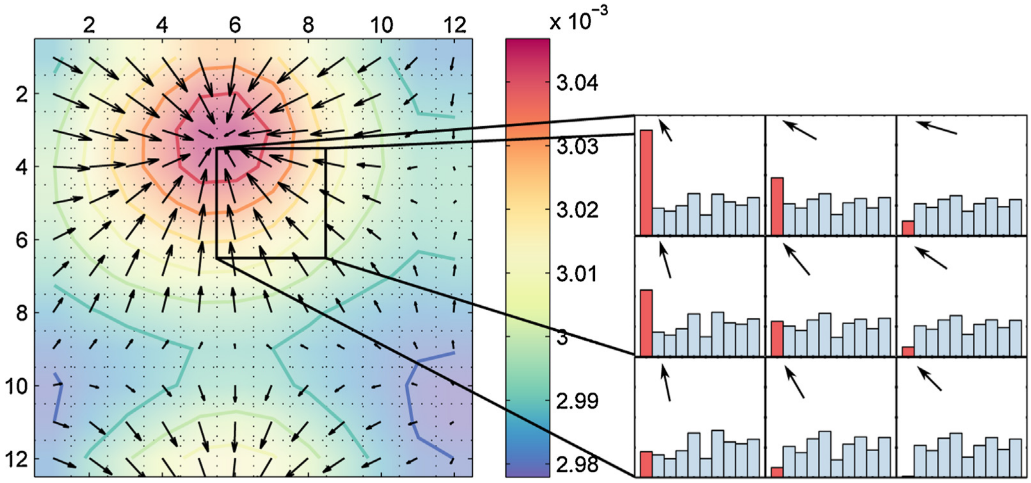

Fig. 5. Derivatives computed on the map $\gamma_l(\mathbf{i})$, where l represents the class of "respiration" samples. On the right, a zoom of an area of the Counting Grid where the gradient is high. The highlighted gene is the one which varies the most in the gradient direction.

Table 1
Top genes selected with the proposed approach.

Rank	Gene name	Description
1	gad1 (YMR250W)	Glutamate decarboxylase
2	hsp12 (YFL014W)	Heat shock protein
3	gsy1 (YFR015C)	Glycogen synthase
4	ygp1 (YNL160W)	Yeast glycoprotein
5	ctt1 (YGR088W)	Cytosolic catalase T
6	sam4 (YPL273W)	S-adenosylmethionine metabolism
7	gsy2 (YLR258W)	Glycogen synthase
8	sol4 (YGR248W)	6-Phosphogluconolactonase
9	hsp30 (YCR021C)	Heat shock protein
10	pgm2 (YMR105C)	Phosphoglucomutase

Table 2
Statistically significant GO terms over-represented in the 10 selected genes' set.

GO	Description	Genes	*p*-value
0005978	Glycogen biosynth. process	gsy1,pgm2,gsy2	0.0225
0006979	Response to oxidative stress	hsp12,ctt1,gad1	0.0235
0006950	Response to stress	hsp30,ygp1,hsp12,ctt1,gad1	0.0247

in a meaningful way, with a clear path which indicates the temporal evolution of the gene expressions; (ii) by looking at the gradient of the class embeddings we can highlight genes which are responsible of the transition of the gene expressions from "fermentation" to "respiration", this being qualitatively confirmed by the GO analysis.

4. Experimental evaluation

The merits of the proposed framework has been extensively tested to solve a wide range of tasks, from both a quantitative and a qualitative perspective. In the following, we first show that the model is able to properly embed the samples on separated parts of the grid, where different areas reflect different sample class/conditions – this shows that the framework well captures the differences in gene expressions related to different classes; then, we extract the most relevant genes with the approach of Section 2.3, validating them from a medical point of view; finally, we report classification accuracies obtained by using descriptors extracted from the model, reaching the state-of-the-art performances.

4.1. Embedding and clustering performances

To test the model on a clustering setting, we performed several experiments on three datasets widely employed in the literature.

Table 3
Summary of the datasets used. The columns Z, T, and L represents the total number of genes, total number of samples, and number of classes present in the dataset.

Dataset name	Z	T	L	Reference
Prostate1 cancer	9984	53	6	[26]
Lung cancer	12,600	203	5	[27]
Brain tumor	7129	90	5	[28]

The first one is the prostate cancer dataset by [26], containing the expression of 9984 genes in 53 different samples: 14 samples labeled for benign prostatic hyperplasia (BPH), three normal adjacent prostate (NAP), one normal adjacent tumor (NAT), 14 localized prostate cancer (PCA), one prostatitis (PRO), and 20 metastatic tumors (MET). The second is a lung cancer dataset [27], consisting of 203 gene expression profiles from normal and tumor samples, with the tumors labeled as squamous, COID, small cell, and adenocarcinoma (5 classes in total). Finally, the brain tumor dataset [28] contains the expression levels of 7129 genes measured in 90 different patients classified in 5 classes (normal, primitive neuroectodermal tumor – PNET, atypical teratoid/rhabdoid tumors – Rhab, and malignant gliomas). A quick summary of the datasets used can be found in Table 3.

As done in many microarray studies (e.g. [11]), we first reduced the dimensions of the original data sets by retaining the top 500 genes ranked by variance. Then, following the original recipe of [15], a single CG is learned using all samples (but ignoring their labels). Data samples are embedded into the CG space: we show some embeddings on a 15×15 grid (using a 3×3 window) in Fig. 6, to have an immediate insight into the datasets. To evaluate how well samples cluster on the grid, we resort to the external criterion of purity [29]. In few words, we leave out one sample and estimate $\gamma_l(k)$ on the remaining data by employing Eq. (3). Then, we assign a label to the test sample by computing

$$y^{\text{test}} = \text{argmax}_l \sum_{\mathbf{k}} q_{\mathbf{k}}^{\text{test}} \cdot \gamma_l(\mathbf{k}) \tag{7}$$

Table 4
Complexities used in the Experiment 4.1 (Fig. 5).

Dimensionality	W	Counting Grid sizes considered
5	[2 2 2 2 2]	[3 3 3 3 3], [4 4 4 4 4], [5 5 5 5 5] [6 6 6 6 6], [7 7 7 7 7], [8 8 8 8 8]
4	[2 2 2 2]	[3 3 3 3], [4 4 4 4], [5 5 5 5], [6 6 6 6] [7 7 7 7], [8 8 8 8], [9 9 9 9] [10 10 10 10], [12 12 12 12]
3	[3 3 3]	[4 4 4], [5 5 5], [6 6 6], [7 7 7], [8 8 8] [10 10 10], [12 12 12], [15 15 15] [20 20 20], [25 25 25], [30 30 30]
2	[5 5]	[7 7], [9 9], [10 10], [12 12], [15 15] [20 20], [25 25], [30 30], [40 40], [50 50] [60 60], [70 70], [80 80], [90 90] [100 100], [120 120]
1	[5]	[7], [10], [12], [15], [20], [25], [35] [45], [60], [80], [100], [140], [180], [220] [260], [300], [340]

The accuracy obtained with this nearest neighbor strategy is our purity score. We considered grids of dimensionality from 1 to 5, testing systematically up to 40 complexities per dimensionality (a more detailed list of parameters settings can be found in Table 4).

Results, shown in Fig. 7, confirms the capabilities of the proposed framework to embed the different classes of each dataset in different regions of the grid; moreover, except for the Brain tumor dataset, it seems that the grid size and the choice of the capacity do not affect much clustering abilities (with only 1-dimensional counting grids being slightly worse). Interestingly, performances do not drop even for very large complexities, suggesting that the model is robust with respect to overfitting.

4.2. Qualitative evaluation of highlighted genes

With the approach proposed in Section 2.3, we extracted the 10 most relevant genes that are involved in a particular tumor class

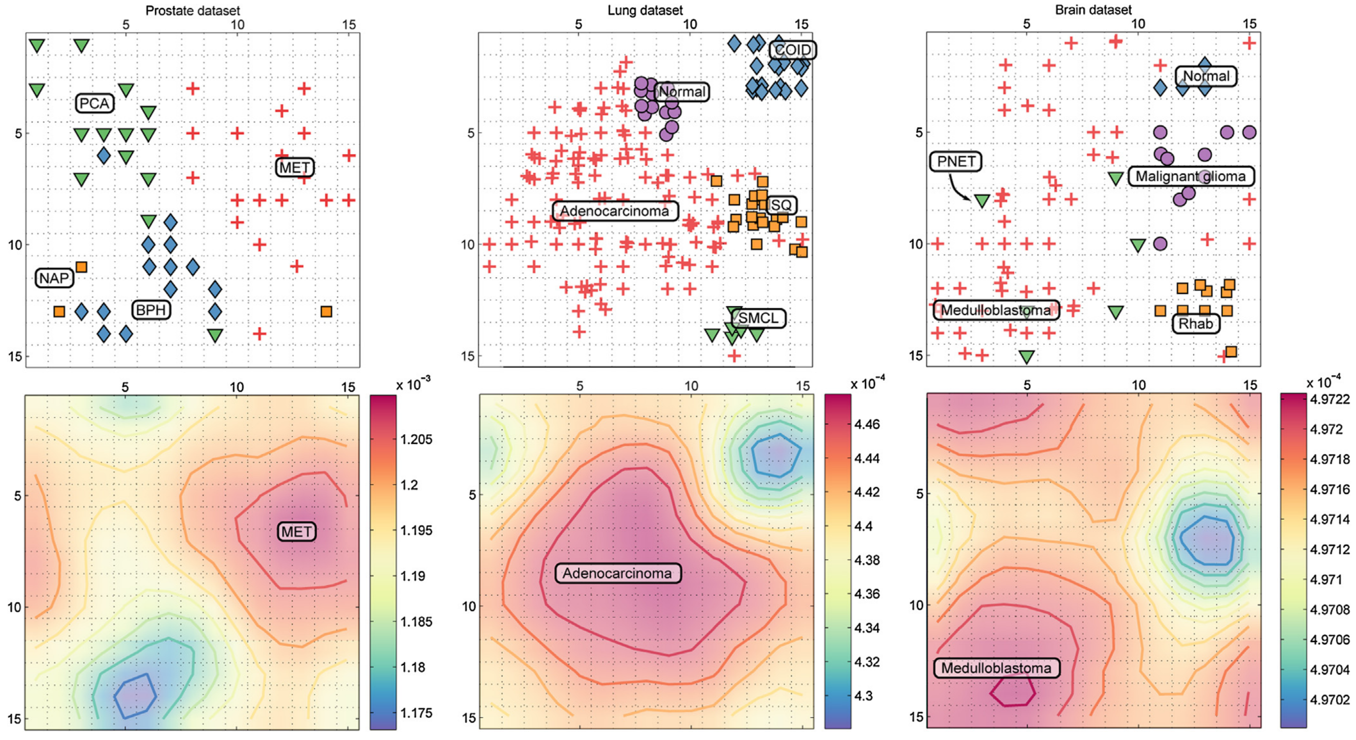

Fig. 6. Counting Grid embeddings for the three studied datasets.

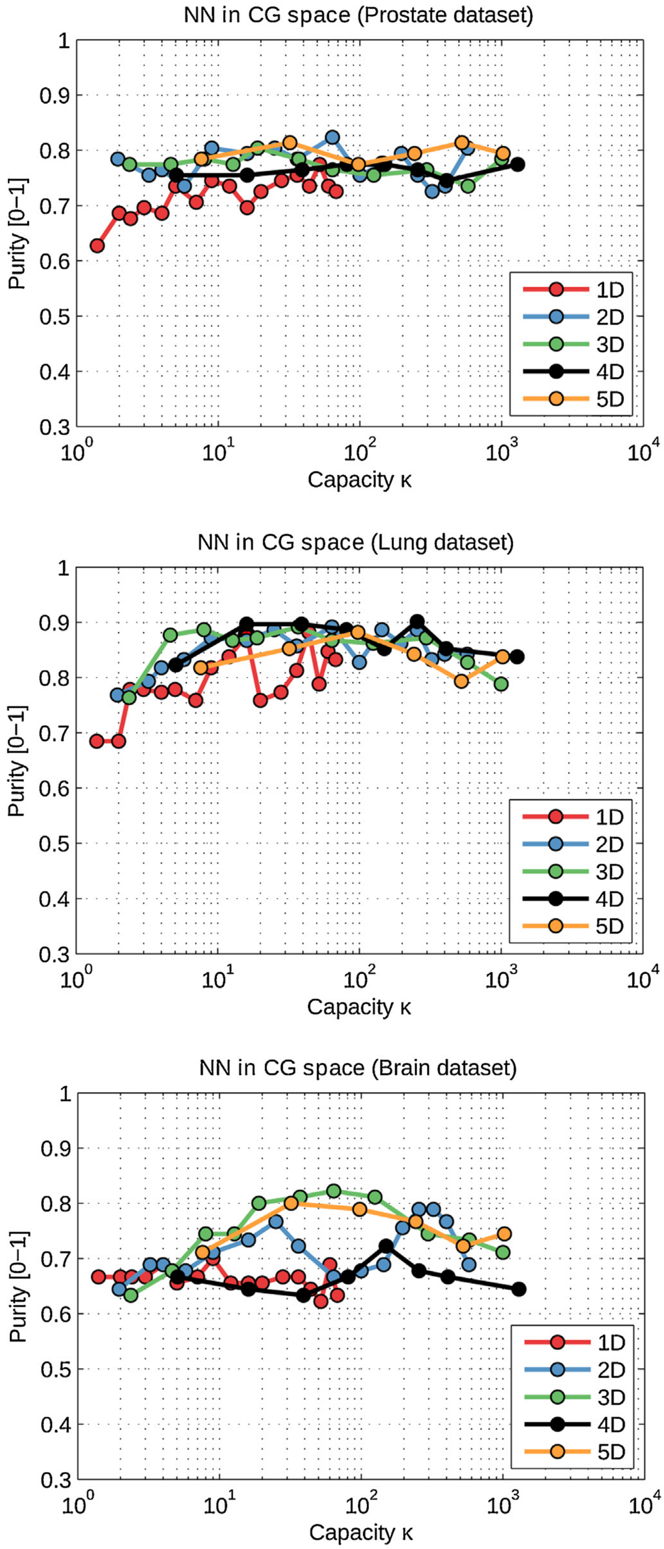

Fig. 7. Purity results.

Table 5
Top genes selected with the proposed approach on the Prostate1 dataset (stability index: 0.89).

Rank	Gene name	Description
1	CTGF	Connective tissue growth factor
2	EGR1	Early growth response 1
3	AMACR	Alpha-methylacyl-CoA racemase
4	ATF3	Activating transcription factor 3
5	LUM	Lumican
6	MMP7	Matrix metalloproteinase 7
7	SPRY4	Sprouty (Drosophila) homolog 4
8	FOSB	FBJ murine osteosarcoma viral oncogene homolog B
9	FGG	Fibrinogen, gamma polypeptide
10	DCT	Dopachrome tautomerase

Table 6
Top genes selected with the proposed approach on the Lung dataset (stability index: 0.91).

Rank	Gene name	Description
1	GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
2	MAPK3	Mitogen-activated protein kinase 3
3	IL13RA2	Interleukin 13 receptor
4	NCAM1	Neural cell adhesion molecule 1
5	TIE1	Tyrosine kinase
6	CYP2C19	Cytochrome P450
7	SLC20A1	Solute carrier family 20
8	YWHAE	Tyrosine 3-monooxygenase
9	ERF	Ets2 repressor factor
10	CXCR5	Chemokine (C-X-C motif) receptor 5

(metastasis for the prostate dataset, adenocarcinoma for the lung, and medulloblastoma for the brain). Ideally, the genes selected by our framework should not vary too much when varying the model capacity – thus confirming results shown in Section 4.1 – Fig. 7. To investigate this aspect we run the gene selection several times using CG of different complexities, and validate the "stability" of the selected through the index proposed by [30]: this index takes values in the range [−1, 1], and the higher its value, the larger the number of commonly selected genes during different training of the algorithm. More in detail, given two sets of genes $\mathbf{f}_1$ and $\mathbf{f}_2$, the stability index is defined as follows:

$$KI(\mathbf{f}_1, \mathbf{f}_2) = \frac{r - (s^2/N)}{s - (s^2/N)} \quad (8)$$

where s denotes the signature size, $r = |\mathbf{f}_1 \cap \mathbf{f}_2|$ and N is the total number of genes in the dataset.

For every dataset, such stability index was never below 0.8, as reported in Tables 5–7, confirming a preliminary investigation carried out in [18]. In the tables, we report the most frequently selected genes while varying the model complexity: on these genes we carried out a detailed investigation in order to assess their potential significance for cancer biology.

Prostate1 Cancer dataset: The top gene highlighted by the algorithm for prostate cancer is CTGF. This gene belongs to the CNN protein family which is involved in functions such as cell adhesion, proliferation, differentiation and apoptosis [31]. CNN family proteins have been identified as diagnostic and therapeutic agents for cancer [31]. Expression of CCN family proteins is altered in

Table 7
Top genes selected with the proposed approach on the Brain dataset (stability index: 0.81).

Rank	Gene name	Description
1	MAPK3	Mitogen-activated protein kinase 3
2	CXCR5	Chemokine (C-X-C motif) receptor 5
3	TIE1	Tyrosine kinase
4	CYP2C19	Cytochrome P450
5	DUSP1	Dual specificity phosphatase 1
6	HINT1	Histidine triad nucleotide binding protein 1
7	MAPK11	Mitogen-activated protein kinase 11
8	RABGGTA	Rab geranyltransferase
9	EIF2AK2	Eukaryotic translat. initiation factor 2-alpha kinase 2
10	IL13RA2	Interleukin 13 receptor, alpha 2

various cancers, including breast, colorectal, gallbladder, gastric, ovarian, pancreatic, and prostate cancers, gliomas, hepatocellular carcinoma, non-small cell lung and squamous cell carcinoma, lymphoblastic leukemia, melanoma, and cartilaginous tumors [32]. CTGF specifically has been shown to be involved in the invasiveness of cancer cells [32]. Moreover, experimental inhibition of CTGF showed that it is critical for tumor growth in pancreatic tumors [33]. It has been shown to correlate with patient survival in gliomas [34] and lung cancer [35]. Experimental inhibition of CTGF showed that it is critical for tumor growth in pancreatic tumors [33]. The existing literature thus indicates that this protein plays an important role in a variety of cancers and confirms that the top gene selected by the model is highly relevant. Similarly, it has been reported [36] that tumor angiogenesis and tumor growth are critically dependent on the activity of EGR1, the second gene selected. Gene ATF3 codes for a transcription factor, that affects cell death and cell cycle progression. There is some evidence [37] that this gene can suppress ras-mediated tumor genesis. Lumican levels in breast cancer are associated with disease progression and have been used to predict survival ([38] reported that low levels of lumican are related to tumor size), while FOSB has been found to drive ovarian cancer [39], and can be used as a prognostic indicator for epithelial ovarian cancer [40]. Finally, MMP7 is involved in cancer metastasis and has been proposed to be used as a target for drug intervention in cancer [41].

A similar analysis has been reported in [11] – where a topic model specifically designed for microarray data analysis is proposed. When analyzing the list of the 10 most relevant genes contained in such paper, only the CTGF and EGR1 are relevant from a medical point of view, being moreover ranked in lower position (in our case they have been ranked in the top positions). This suggests once more that CG may represent a powerful method to unravel the complexity contained in gene expression datasets.

Lung Cancer dataset: Also in the case of the Lung Cancer dataset, many of the most relevant genes extracted by the proposed framework show a medical relevance. For example, GAPDH expression was found to be strongly elevated in human lung cancer cells [42]; it is also correlated with breast cancer cell proliferation and aggressiveness [43]. IL13RA was found to be one of the genes that mediate the metastasis of breast cancer to lung [44]. NCAM has been researched as a target for immunotherapy for cancer as it is expressed in small cell lung cancer, neuroblastoma, rhabdomyosarkoma, brain tumors, multiple myelomas and acute myeloid leukemia. TIE1 is involved in angiogenesis, the creation of new blood vessels, which as in important process also in tumor progression [45]. The experimental deletion of these gene from mice inhibits tumor angiogenesis and growth [45]. Finally, YWHAE is correlated with survival in breast cancer: it was found to be enriched in metastatic tumor cell pseudopods [46], and is involved in the pathology of small cell lung cancer.

Brain Tumor dataset: The promising results obtained with Prostate and Colon datasets are confirmed by looking at the list of genes extracted for the Brain Tumor dataset. Actually, MAPK3 belongs to a family of proteins that regulate cell proliferation, differentiation and cell cycle progression. It was shown to be a prognostic biomarker in gastric cancer and implicated in the progression of hepatocellular carcinoma [47]. CXCR5 is a protein in CXC chemokine receptor family, which plays a role in the spread of cancer, including metastases [48]. TIE1 was implicated as a prognostic marker for gastric cancer [49] and showed over-expression in breast cancer. DUSP1 is a promoter for tumor angiogenesis, invasion and metastasis in non-small-cell lung cancer [50] and plays a prognostic role in breast cancer [51]. Finally, HINT1 is a tumor suppressor gene [52].

Table 8
Summary of the datasets used for the classification task. The columns Z, T, and L represents the total number of genes, total number of samples, and number of classes present in the dataset. The datasets are publicly available on the Kent Ridge Biomedical repository: http://datam.i2r.a-star.edu.sg/datasets/krbd/ (accessed: March 2016).

Dataset name	Z	T	L	Reference
Prostate2 cancer	10,509	102	2	[53]
Lung cancer	12,600	203	5	[27]
Brain tumor	7129	90	5	[28]
Colon Cancer	2000	62	2	[54]
DLBCL	6285	77	2	[55]
Leukemia	7129	72	2	[56]
Breast tumor	24,481	97	2	[57]

4.3. *Quantitative evaluation of highlighted genes*

We assessed numerically the performances of the gene selection approach presented in Section 2.3 by performing a classification experiment, where we classify samples using only the genes selected with the proposed approach. In order to perform an extensive investigation of the CG classification power, we enlarged the evaluation by considering seven benchmark datasets found in the literature, where other gene selection approaches have already been tested; the datasets are summarized in Table 8.

We compared our results with state-of-the-art methodologies for gene selection. We adopted the testing protocol of [58]: the data set was randomly split 2:3/1:3 (training/testing). Please note that the CG is learned on the whole dataset, without any pre-filtering of the genes': in this way, we also prove that the CG can efficiently deal with gene expression data without the need of such pre-processing.

More in detail, we employed the whole dataset to train a CG (of course labels are ignored in this phase), from which we computed the F_z score for each gene: after that, only the top-ranked genes have been extracted: in particular, we retain the top `[10 50 100 200]` genes. Then classification is performed using a linear Support Version Machine (SVM) with the parameter $C = 1$, using the area under the ROC curve (AUC) as an estimate for the classification performance. The test has been repeated 100 times, and the mean of the computed AUCs is shown in Tables 9 and 10, along with

Table 9
Classification results (AUC) for the datasets used. Underlined values indicate statistically significant improvements.

Gene sel. method	Gene signature size				
	10	50	100	150	200
Prostate2 dataset					
SVM-RFE [58]	89.8	91.3	92.1	92.1	92.2
ReliefF [58]	93.3	93.0	91.4	91.4	91.7
Mrmr [59]	**93.6**	**95.3**	**94.8**	94.9	95.2
Zero-norm [60]	52.4	93.0	92.8	94.6	95.0
Our method	78.2	88.3	92.5	**95.0**	**95.7**
Lung dataset					
SVM-RFE [58]	82.2	88.6	95.3	97.3	97.9
ReliefF [58]	91.0	95.3	96.4	97.0	97.4
Mrmr [59]	**92.8**	93.9	97.6	98.0	98.2
Zero-norm [60]	82.4	91.1	94.2	97.3	97.5
Our method	81.5	**96.0**	**98.6**	**98.5**	**98.7**
Brain dataset					
SVM-RFE [58]	86.2	92.6	93.0	93.4	94.0
ReliefF [58]	77.6	87.8	89.0	90.1	90.6
Mrmr [59]	86.1	90.5	91.3	91.7	91.9
Zero-norm [60]	83.3	83.0	85.8	**94.0**	94.7
Our method	**88.3**	**93.7**	**93.1**	**94.0**	**95.1**

Table 10
Classification results (AUC) for the datasets used. Underlined values indicate statistically significant improvements.

Method	Colon dataset					DLBCL dataset				
	Gene signature size					Gene signature size				
	10	50	100	150	200	10	50	100	150	200
SVM-RFE [58]	76.4	77.5	79.2	79.4	80.1	58.2	93.8	98.4	98.8	98.4
ReliefF [58]	78.8	80.1	78.5	77.5	76.1	93.3	95.9	96.6	96.1	96.3
Mrmr [59]	**85.1**	86.5	87.0	87.4	86.6	**95.9**	96.0	96.1	96.9	97.3
Zero-norm [60]	78.9	80.2	79.1	77.3	76.1	84.4	87.5	98.4	98.8	98.4
Our method	81.4	**89.5**	**89.6**	**89.3**	**89.0**	92.9	**97.2**	**98.5**	**99.0**	**99.0**

Method	Leukemia dataset					Breast tumor dataset				
	Gene signature size					Gene signature size				
	10	50	100	150	200	10	50	100	150	200
SVM-RFE [58]	45.4	81.5	91.6	95.5	98.5	57.7	52.0	50.3	49.1	54.4
ReliefF [58]	98.3	99.5	99.6	**99.8**	99.3	70.1	69.3	67.9	**71.3**	71.7
Mrmr [59]	**98.8**	99.5	99.7	99.7	99.0	**71.7**	66.0	69.0	68.1	68.8
Zero-norm [60]	54.2	81.9	91.4	96.8	98.8	45.1	49.2	49.5	51.2	51.0
Our method	69.2	**99.6**	**99.7**	99.6	**99.7**	62.1	**71.5**	**71.7**	69.9	**73.0**

comparative state-of-the-art results (see the references between brackets). As for the Counting Grid size, we varied its dimensions by selecting κ between 5 and 40, reporting in the table the mean of the obtained AUCs.

Given the 100 repetitions, we also performed a paired t-test, to assess if the difference between the CG and the best performing state-of-the-art method is statistically significant (with a confidence level $\alpha = 0.05$). We addressed such t-test in a paired setting because each of the 100 repetitions represent a different partition of the data, and the comparison should be carried out on the same partition. In the tables, such significant results have been underlined.

From the tables, it is evident that the proposed approach produces results comparable, and in many cases superior, with state-of-the-art techniques. Furthermore, when looking at the stability, we can observe that our approach is very competitive: the obtained indices are shown in Table 11. Since the proposed approach is aimed at explaining the data through a generative model, and labels are used later on, the stability index is very high: for every dataset, and all different signature sizes, it is always above 0.85.

Table 11
Stability of the proposed approach.

Dataset	Gene selection method	Gene signature size				
		10	50	100	150	200
Prostate2	Best (SVM-RFE [58])	0.68	0.65	0.68	0.69	0.69
	Our method	**0.90**	**0.94**	**0.96**	**0.96**	**0.96**
Lung	Best (Zero-norm [60])	0.94	0.92	0.92	0.93	0.93
	Our method	**0.95**	**0.95**	**0.95**	**0.98**	**0.99**
Brain	Best (SVM-RFE [58])	0.93	0.96	0.97	0.97	0.97
	Our method	**0.94**	**0.98**	**0.98**	**0.99**	**0.99**
Colon	Best (Mrmr [59])	0.78	0.75	0.70	0.69	0.67
	Our method	**0.94**	**0.92**	**0.92**	**0.91**	**0.91**
DLBCL	Best (Mrmr [59])	0.25	0.40	0.46	0.49	0.51
	Our method	**0.89**	**0.92**	**0.94**	**0.94**	**0.94**
Leukemia	Best (Mrmr [59])	0.39	0.50	0.55	0.56	0.57
	Our method	**0.87**	**0.87**	**0.90**	**0.89**	**0.88**
Breast	Best (Relieff [58])	0.31	0.37	0.37	0.37	0.38
	Our method	**0.91**	**0.92**	**0.93**	**0.92**	**0.93**

4.4. *Classification results*

As a last experiment, we employed the Counting Grid in a classification setting. We followed the standard hybrid generative-discriminative recipe [61]: the idea is to characterize every sample with a feature vector obtained from the learned CG, so that samples are projected in a highly informative space where standard discriminative classifiers such as the SVM can be used.

In our experiments, we employed two strategies, both based on the definition of a kernel to be used with a SVM classifier. In the former case [62], the kernel is defined on the basis of a geometric reasoning on the grid of the learned CG: in few words, the idea is to spread the posterior $q^t_{\mathbf{k}}$ evaluated on a single grid-point around its neighborhood. In this fashion, when two samples are compared, an implicit cross-count evaluation is introduced by avoiding a fully grid alignment constraint. The second kernel employed is the Fisher Kernel [63], whose derivation in the CG case has been proposed in [19]. In the original formulation, the authors first define the Fisher score for a gene $FS^t_{\mathbf{k},z}$

$$FS^t_{\mathbf{k},z} = g^t_z \cdot \sum_{\mathbf{i} \in W_{\mathbf{k}^t}} \frac{q^t_{\mathbf{i}}}{h_{\mathbf{i},z}} \qquad (9)$$

and the concatenation of the FS, computed for all genes z, comprises the Fisher score for a sample. Then, the standard linear kernel is computed between these Fisher score vectors.

These two classification strategies have been applied on the seven datasets. Accuracies have been computed using the dataset author's protocol: Leave-One-Out for the Prostate and Colon datasets, 5-fold cross-validation for the Lung dataset, 4-fold cross-validation for the Brain dataset, 10-fold cross-validation for the DLBCL, Breast and Leukemia datasets.

The best result obtained by varying the complexity of the grid is reported in Table 12. Moreover, in order to have a clear insight of the gain obtained by explicitly consider the relation between topics, as done in the CG case, we applied the same hybrid classification strategies to the classic probabilistic Latent Semantic Analysis (PLSA, [10]); finally, we compare our results to those obtained with the approach proposed in [11] – we took the results from [10], and with non topic-based approaches. In five datasets out of seven, the CG model (equipped with the Fisher kernel) was able to outperform other topic-based approaches, as well as approaches that do not

Table 12
Classification results.

	Prostate2	Lung	Brain	Colon	DLBCL	Leukemia	Breast
Geom PLSA	0.826	0.911	0.858	0.939	0.961	0.972	0.617
Geom CG	0.773	0.918	0.869	0.939	0.961	0.983	0.639
Fisher PLSA	0.921	0.938	0.862	0.917	0.845	0.972	0.634
Fisher CG	0.940	**0.959**	**0.900**	**0.951**	**0.974**	**0.986**	0.660
LPD	0.951	0.942	0.890	0.807	0.932	0.961	0.526
Non topic-based approach	**0.982** [64]	0.938 [65]	0.865 [66]	0.935 [67]	0.960 [68]	0.975 [69]	**0.667** [68]

derive from the topic models literature (see the references between brackets in the table).[3]

5. Conclusions

This paper proposed an approach based on the Counting Grid model for the analysis of gene expression samples. We have shown with different experiments that the proposed framework can be successfully exploited to (i) meaningfully visualize the samples; (ii) detect medically relevant genes, and (iii) properly classify samples, thus representing a valid alternative to classical gene expression analysis strategies. The proposed approach also finds a very promising application in analyzing expression data produced by rna-Seq: using this technology, gene expression is estimated from the number of reads mapped on each gene, and represents a proper count value. We are planning to develop a study on RNA-seq datasets as a future work.

References

[1] Lee JW, Lee JB, Park M, Song SH. An extensive comparison of recent classification tools applied to microarray data. Comput Stat Data Anal 2005;48(4):869–85.

[2] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 2005;21(5):631–43.

[3] Wu M-Y, Dai D-Q, Shi Y, Yan H, Zhang X-F. Biomarker identification and cancer classification based on microarray data using Laplace naive Bayes model with mean shrinkage. IEEE/ACM Trans Comput Biol Bioinform 2012;9(6):1649–62.

[4] Tong M, Liu K-H, Xu C, Ju W. An ensemble of SVM classifiers based on gene pairs. Comput Biol Med 2013;43(6):729–37.

[5] Kerr G, Ruskin H, Crane M, Doolan P. Techniques for clustering gene expression data. Comput Biol Med 2008;38(3):283–93.

[6] de Souto M, Costa I, de Araujo D, Ludermir T, Schliep A. Clustering cancer gene expression data: a comparative study. BMC Bioinformatics 2008;9(1):497.

[7] Lazar C, Taminau J, Meganck S, Steenhoff D, Coletta A, Molter C, et al. A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans Comput Biol Bioinform 2012;9(4):1106–19.

[8] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Mach Learning 2001;42(1–2):177–96.

[9] Fasoli M, Santo SD, Zenoni S, Tornielli GB, Farina L, Zamboni A, et al. The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program. Plant Cell Online 2012;24(9):3489–505.

[10] Bicego M, Lovato P, Perina A, Fasoli M, Delledonne M, Pezzotti M, et al. Investigating topic models' capabilities in expression microarray data classification. IEEE/ACM Trans Comput Biol Bioinform 2012;9(6):1831–6.

[11] Rogers S, Girolami M, Campbell C, Breitling R. The latent process decomposition of cDNA microarray data sets. IEEE/ACM Trans Comput Biol Bioinform 2005;2(2):143–56.

[12] Perina A, Lovato P, Murino V, Bicego M. Biologically-aware latent Dirichlet allocation (BALDA) for the classification of expression microarray. In: Proceedings of the international conference on pattern recognition in bioinformatics, LNCS. Springer; 2010. p. 230–41.

[13] Bicego M, Lovato P, Ferrarini A, Delledonne M. Biclustering of expression microarray data with topic models. In: Proceedings of the international conference on pattern recognition. IEEE; 2010. p. 2728–31.

[14] Joachims T. Text categorization with support vector machines: learning with many relevant features. In: Proceedings of the European conference on machine learning. 1998. p. 137–42.

[15] Jojic N, Perina A. Multidimensional counting grids: Inferring word order from disordered bags of words. In: Uncertainty in artificial intelligence; 2011.

[16] Perina A, Jojic N. Image analysis by counting on a grid. In: International conference on computer vision and pattern recognition. 2011. p. 1985–92.

[17] Jordan I, Marino-Ramirez L, Koonin E. Evolutionary significance of gene expression divergence. Gene 2005;345(1):119–26.

[18] Lovato P, Bicego M, Cristani M, Jojic N, Perina A. Feature selection using counting grids: application to microarray data. In: Structural, syntactic, and statistical pattern recognition, vol. 7626 of LNCS. 2012. p. 629–37.

[19] Perina A, Kesa M, Bicego M. Expression microarray data classification using counting grids and fisher kernel. In: Proceedings of the international conference on pattern recognition. 2014. p. 1770–5.

[20] Frey BJ, Jojic N. A comparison of algorithms for inference and learning in probabilistic graphical models. IEEE Trans Pattern Anal Mach Intell 2005;27:2005.

[21] DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278(5338):680–6.

[22] Rossignol T, Dulau L, Julien A, Blondin B. Genome-wide monitoring of wine yeast gene expression during alcoholic fermentation. Yeast 2003;20(16):1369–85.

[23] Rodriguez-Colman MJ, Reverter-Branchat G, Sorolla MA, Tamarit J, Ros J, Cabiscol E. The forkhead transcription factor Hcm1 promotes mitochondrial biogenesis and stress resistance in yeast. J Biol Chem 2010;285(47):37092–101.

[24] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25(1):25–9.

[25] Beißbarth T, Speed TP. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics 2004;20(9):1464–5.

[26] Dhanasekaran SM, Barrette TR, Ghosh D, Shah R, Varambally S, Kurachi K, et al. Delineation of prognostic biomarkers in prostate cancer. Nature 2001;412(6849):822–6.

[27] Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001;98(24):13790–5.

[28] Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 2002;415(6870):436–42.

[29] Manning CD, Raghavan P, Schütze H. Introduction to information retrieval, vol. 1. Cambridge: Cambridge University Press; 2008.

[30] Kuncheva LI. A stability index for feature selection. In: Proceedings of the 25th IASTED artificial intelligence and applications, AIAP'07. 2007. p. 390–5.

[31] Jun J-I, Lau LF. Taking aim at the extracellular matrix: CCN proteins as emerging therapeutic targets. Nat Rev Drug Discov 2011;10(12):945–63.

[32] Chen C-C, Lau L. Functions and mechanisms of action of CCN matricellular proteins. Int J Biochem Cell Biol 2009;41(4):771–83.

[33] Bennewith KL, Huang X, Ham CM, Graves EE, Erler JT, Kambham N, et al. The role of tumor cell-derived connective tissue growth factor (CTGF/CCN2) in pancreatic tumor growth. Cancer Res 2009;69(3):775–84.

[34] Xie D, Yin D, Wang H-J, Liu G-T, Elashoff R, Black K, et al. Levels of expression of CYR61 and CTGF are prognostic for tumor progression and survival of individuals with gliomas. Clin Cancer Res 2004;10(6):2072–81.

[35] Chen P-P, Li W-J, Wang Y, Zhao S, Li D-Y, Feng L-Y, et al. Expression of cyr61, CTGF, and WISP-1 correlates with clinical features of lung cancer. PLoS ONE 2007;2(6):e534.

[36] Fahmy RG, Dass CR, Sun L-Q, Chesterman CN, Khachigian LM. Transcription factor Egr-1 supports FGF-dependent angiogenesis during neovascularization and tumor growth. Nat Med 2003;9(8):1026–32.

[37] Lu D, Wolfgang CD, Hai T. Activating transcription factor 3, a stress-inducible gene, suppresses ras-stimulated tumorigenesis. J Biol Chem 2006;281(15):10473–81.

[38] Troup S, Njue C, Kliewer EV, Parisien M, Roskelley C, Chakravarti S, et al. Reduced expression of the small leucine-rich proteoglycans, lumican, and decorin is associated with poor outcome in node-negative invasive breast cancer. Clin Cancer Res 2003;9(1):207–14.

[39] Shahzad MMK, Arevalo JM, Armaiz-Pena GN, Lu C, Stone RL, Moreno-Smith M, et al. Stress effects on FosB- and interleukin-8 (IL8)-driven ovarian cancer growth and metastasis. J Biol Chem 2010;285(46):35462–70.

[40] Kataoka F, Tsuda H, Arao T, Nishimura S, Tanaka H, Nomura H, et al. EGRI and FOSB gene expressions in cancer stroma are independent prognostic indicators for epithelial ovarian cancer receiving standard therapy. Genes Chromosomes Cancer 2012;51(3):300–12.

[3] Please note that in this set of experiments we have not been able to compute statistical significance as we did not have access to the individual state of the art results for the different cross-validation folds.

[41] Wielockx B, Libert C, Wilson C. Matrilysin (matrix metalloproteinase-7): a new promising drug target in cancer and inflammation? Cytokine Growth Factor Rev 2004;15(2–3):111–5.

[42] Tokunaga K, Nakamura Y, Sakata K, Fujimori K, Ohkubo M, Sawada K, et al. Enhanced expression of a glyceraldehyde-3-phosphate dehydrogenase gene in human lung cancers. Cancer Res 1987;47(21):5616–9.

[43] Revillion F, Pawlowski V, Hornez L, Peyrat J. Glyceraldehyde-3-phosphate dehydrogenase gene expression in human breast cancer. Eur J Cancer 2000;36(8):1038–42.

[44] Minn A, Gupta G, Siegel P, Bos P, Shu W, Giri D, et al. Genes that mediate breast cancer metastasis to lung. Nature 2005;436(7050):518–24.

[45] D'Amico G, Korhonen EA, Anisimov A, Zarkada G, Holopainen T, Hagerling R, et al. Tie1 deletion inhibits tumor growth and improves angiopoietin antagonist therapy. J Clin Investig 2014;124(2):824–34.

[46] Shankar J, Messenberg A, Chan J, Underhill TM, Foster LJ, Nabi IR. Pseudopodial actin dynamics control epithelial–mesenchymal transition in metastatic cancer cells. Cancer Res 2010;70(9):3780–90.

[47] Kim J, Lee S, Chae Y, Kang B, Lee Y, Oh S, et al. Association between phosphorylated amp-activated protein kinase and mapk3/1 expression and prognosis for patients with gastric cancer. Oncology 2013;85(2):78–85.

[48] Balkwill F. Cancer and the chemokine network. Nat Rev Cancer 2004;4(7):540–50.

[49] Lin W-C, Li AF-Y, Chi C-W, Chung W-W, Huang CL, Lui W-Y, et al. Tie-1 protein tyrosine kinase: a novel independent prognostic marker for gastric cancer. Clin Cancer Res 1999;5(7):1745–51.

[50] Moncho-Amor V, Ibanez de Caceres I, Bandres E, Martinez-Poveda B, Orgaz J, Sanchez-Perez I, et al. Dusp1/mkp1 promotes angiogenesis, invasion and metastasis in non-small-cell lung cancer. Oncogene 2011;30(6):668–78.

[51] Bieche I, Lerebours F, Tozlu S, Espie M, Marty M, Lidereau R. Molecular profiling of inflammatory breast cancer: identification of a poor-prognosis gene expression signature. Clin Cancer Res 2004;10(20):6789–95.

[52] Zhang Y-J, Li H, Wu H-C, Shen J, Wang L, Yu M-W, et al. Silencing of hint1, a novel tumor suppressor gene, by promoter hypermethylation in hepatocellular carcinoma. Cancer Lett 2009;275(2):277–84.

[53] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 2002;1(2):203–9.

[54] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999;96(12):6745–50.

[55] Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, et al. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 2002;8(1):68–74.

[56] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286(5439):531–7.

[57] Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530–6.

[58] Yu L, Han Y, Berens M. Stable gene selection from microarray data via sample weighting. IEEE/ACM Trans Comput Biol Bioinform 2012;9:262–72.

[59] Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1226–38.

[60] Weston J, Elisseeff A, Schölkopf B, Tipping M. Use of the zero norm with linear models and kernel methods. J Mach Learning Res 2003;3:1439–61.

[61] Perina A, Cristani M, Castellani U, Murino V, Jojic N. Free energy score spaces: using generative information in discriminative classifiers. IEEE Trans Pattern Anal Mach Intell 2012;34(7):1249–62.

[62] Perina A, Bicego M, Castellani U, Murino V. Exploiting geometry in counting grids. SIMBAD; 2013. p. 250–64.

[63] Jaakkola T, Haussler D. Exploiting generative models in discriminative classifiers. In: Advances in neural information processing systems. 1999. p. 487–93.

[64] Wang X, Gotoh O. A robust gene selection method for microarray-based cancer classification. Cancer Inform 2010;9:15–30.

[65] Chen P-C, Huang S-Y, Chen W, Hsiao C. A new regularized least squares support vector regression for gene selection. BMC Bioinformatics 2009;10(1):44.

[66] Osareh A, Shadgar B. Classification and diagnostic prediction of cancers using gene microarray data analysis. J Appl Sci 2009;9(3):459–68.

[67] Liu H, Liu L, Zhang H. Ensemble gene selection by grouping for microarray data classification. J Biomed Inform 2010;43(1):81–7.

[68] Bolón-Canedo V, Sánchez-Maro no N, Alonso-Betanzos A. An ensemble of filters and classifiers for microarray data classification. Pattern Recogn 2012;45(1):531–9.

[69] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 2005;21(5):631–43.