# Partially supervised feature selection with regularized linear models

**Conference Paper** · January 2009

DOI: 10.1145/1553374.1553427 · Source: DBLP



**2 authors**, including:

Pierre Dupont
Université catholique de Louvain
**101** PUBLICATIONS **1,627** CITATIONS

# Partially Supervised Feature Selection with Regularized Linear Models

**Thibault Helleputte** THIBAULT.HELLEPUTTE@UCLOUVAIN.BE
**Pierre Dupont** PIERRE.DUPONT@UCLOUVAIN.BE
University of Louvain, Computing Science and Engineering Dept. & Machine Learning Group, Reaumur Building, Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium

## Abstract

This paper addresses feature selection techniques for classification of high dimensional data, such as those produced by microarray experiments. Some prior knowledge may be available in this context to bias the selection towards some dimensions (genes) *a priori* assumed to be more relevant. We propose a feature selection method making use of this partial supervision. It extends previous works on embedded feature selection with linear models including regularization to enforce sparsity. A practical approximation of this technique reduces to standard SVM learning with iterative rescaling of the inputs. The scaling factors depend here on the prior knowledge but the final selection may depart from it. Practical results on several microarray data sets show the benefits of the proposed approach in terms of the stability of the selected gene lists with improved classification performances.

## 1. Introduction

Classification of microarray data is a challenging problem as it typically relies on a few tens of samples but several thousand dimensions (genes). Feature selection techniques are commonly used in this context, both to increase the interpretability of the predictive model and possibly to reduce its cost (Guyon & Elisseef, 2003; Saeys et al., 2007). In some cases feature selection has also been shown to improve classification accuracy (Krishnapuram et al., 2004). *Biomarker selection* specifically refers to the identification of a small set of genes, also called a *signature*, related to a pathology or to an observed clinical outcome after a treatment.

Semi-supervised classification deals with problems for which only a fraction of the learning examples have known class labels, and semi-supervised feature selection methods have been recently proposed (Zhao & Liu, 2007; Cheng et al., 2008). We use here a different kind of partial supervision, namely on the *dimensions* of a feature selection procedure. For instance in the case of microarray data classification, a molecular biologist may know or guess that some genes are likely to be more discriminant. This knowledge is usually only partial, as the purpose of biomarker selection is often to discover new gene signatures, or even inaccurate, as gene expression may be influenced by several factors not related with the outcome. The technique presented in this paper makes use of such prior knowledge to guide feature selection while letting the final selection depart from it if necessary to optimize the classification objective.

Support vector machines (SVMs) are particularly convenient to classify high dimensional data with only a few samples. In their simplest form, SVMs simply reduce to maximal margin hyperplanes in the input space. Such models were shown to successfully classify microarray data either on the full input space (Mukherjee, 2003) or combined with feature selection (Weston et al., 2000; Chapelle et al., 2002; Guyon et al., 2002). The latter approaches are *embedded* as the selected features directly follow from the structure of the classifier. Our method extends the embedded AROM methods (Weston et al., 2003), by adding a partial supervision on the dimensions to be selected, in a simple yet efficient way.

A good set of features is ideally highly stable with respect to sampling variation. In the context of biomarker selection from microarray data, high stabil-

Appearing in *Proceedings of the 26$^{th}$ International Conference on Machine Learning*, Montreal, Canada, 2009. 

ity means that different sub-samples of patients lead to very similar sets of biomarkers. This is motivated by the assumption that the biological process explaining the outcome is common among different patients. We show in the present study that the use of prior knowledge on relevant genes effectively induces a large gain in stability with improved classification performances in most cases.

The rest of this paper is organized as follows. Section 2 briefly reviews the AROM methods. Section 3 details how to extend these methods with partial supervision on the selected features. Practical experiments on various microarray data sets are reported in section 4. We conclude and present our perspectives in section 5.

## 2. The AROM methods

Given $m$ examples $\mathbf{x}_i \in \mathbb{R}^n$ and the corresponding class labels $y_i \in \{\pm 1\}$ with $i = 1, ..., m$, a linear model $g(\mathbf{x})$ predicts the class of any point $\mathbf{x} \in \mathbb{R}^n$ as follows.

$$g(\mathbf{x}) = \text{sign}(\mathbf{w} \cdot \mathbf{x} + b) \tag{1}$$

Feature selection is closely related to a specific form of regularization of this decision function to enforce sparsity of the weight vector $\mathbf{w}$. Weston et al. (2003) study in particular the zero-norm minimization subject to linear margin constraints[1] :

$$\min_{\mathbf{w}} ||\mathbf{w}||_0 \text{ subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \tag{2}$$

where $||\mathbf{w}||_0 = card\{w_j | w_j \neq 0\}$ and $card$ is the set cardinality. Since problem (2) is NP-Hard, a log 1-norm minimization is proposed instead.

$$\min_{\mathbf{w}} \sum_{j=1}^{n} \ln(|w_j| + \epsilon) \text{ subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \tag{3}$$

where $0 < \epsilon \ll 1$ is added to smooth the objective when some $|w_j|$ vanishes. The natural logarithm in the objective facilitates parameter estimation with a simple gradient descent procedure (an extended version of this procedure is detailed in section 3). The resulting algorithm $l1$-AROM[2] simply optimizes the 1-norm of $\mathbf{w}$ with iterative rescaling of the inputs.

The $l2$-AROM method further approximates this objective by replacing the 1-norm by the 2-norm. Even though such an approximation may result in a less sparse solution, it is very efficient in practice when $m \ll n$. Indeed, a dual formulation may be used and the final algorithm boils down to a linear SVM estimation with iterative rescaling of the inputs. A standard SVM solver can be iteratively called on properly rescaled inputs. A smooth feature selection occurs during this iterative process since the weight coefficients along some dimensions progressively drop below the machine precision while other dimensions become more significant. A final ranking on the absolute values of each dimension can be used to obtain a fixed number of features.

[1] The constraints in problem (2) could be rewritten $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) > 0$ since the 0-norm is insensitive to the scale of $\mathbf{w}$. The use of margin constraints is motivated by the subsequent approximations to this problem.

[2] AROM stands for **A**pproximation of ze**RO**-norm **M**inimization.

## 3. Partially supervised AROM

Whenever some prior knowledge on the relative importance of each feature is available, the $l1$-AROM objective can be modified by adding a prior relevance vector $\boldsymbol{\beta} = [\beta_1, ..., \beta_n]^t$ defined over the input dimensions. Let $\beta_j \geq 1$ denote the relative prior relevance of the $j^{th}$ feature, the higher its value the more relevant the corresponding feature is *a priori* assumed. If no information is available about a given feature prior relevance, it is fixed to the default value $\beta_j = 1$. The optimization problem is modified to penalize less the dimensions which are assumed *a priori* more relevant:

$$\min_{\mathbf{w}} \sum_{j=1}^{n} \frac{1}{\beta_j} \ln(|w_j| + \epsilon) \text{ subject to } y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \tag{4}$$

Following the same line of reasoning as in (Weston et al., 2003), we derive below an iterative algorithm to solve problem (4). An equivalent problem is obtained after introducing auxiliary variables $v_j$'s:

$$\min_{\mathbf{v}} \sum_{j=1}^{n} \frac{1}{\beta_j} \ln(v_j + \epsilon) \text{ subject to } \begin{cases} y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \\ v_j \geq w_j \\ v_j \geq -w_j \end{cases} \tag{5}$$

Next, problem (5) is solved using an iterative constrained gradient descent technique due to (Franke & Wolfe, 1956):

**1.** Find the steepest descent direction of the objective function that is consistent with the constraints:

$$\min_{\mathbf{v},\mathbf{w}} \nabla h(\mathbf{v}_k) \cdot (\mathbf{v} - \mathbf{v}_k) \text{ subj.to } \begin{cases} y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \\ v_j \geq w_j \\ v_j \geq -w_j \end{cases} \tag{6}$$

where $h(\mathbf{v}_k)$ is the value of the objective function at step $k$. Let $(\bar{\mathbf{v}}, \bar{\mathbf{w}})$ be the optimum value of this problem.

**2.** Optimize along that steepest descent direction: compute $\lambda$ such that $h(\mathbf{v}_k + \lambda(\bar{\mathbf{v}} - \mathbf{v}_k))$ is minimal.

At step $k$, the objective function is approximately given by $h(\mathbf{v}_k) \approx \sum_{j=1}^{n} \frac{1}{\beta_j} \ln(v_{kj})$. It follows that:

$$\frac{\partial h(.)}{\partial v_{kj}} = \frac{1}{\beta_j v_{kj}}$$

Since the steepest descent is given by

$$\nabla h(\mathbf{v}_k) \cdot (\mathbf{v} - \mathbf{v}_k) = \sum_{j=1}^{n} \frac{v_j - v_{kj}}{\beta_j v_{kj}}$$

problem (6) becomes

$$\min_{\mathbf{v},\mathbf{w}} \sum_{j=1}^{n} \frac{v_j - v_{kj}}{\beta_j v_{kj}} \text{ subject to } \begin{cases} y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 \\ v_j \geq w_j \\ v_j \geq -w_j \end{cases} \tag{7}$$

By introducing new variables $v'_j = \frac{v_j}{\beta_j v_{kj}}$, it can be rewritten as:

$$\min_{\mathbf{v}',\mathbf{w}} \sum_{j=1}^{n} v'_j - \sum_{j=1}^{n} \frac{1}{\beta_j} = \min_{\mathbf{v}',\mathbf{w}} \sum_{j=1}^{n} v'_j \tag{8}$$

subject to $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$ ; $v'_j \geq \frac{w_j}{\beta_j v_{kj}}$ ; $v'_j \geq \frac{-w_j}{\beta_j v_{kj}}$.

By defining $w'_j = \frac{w_j}{\beta_j v_{kj}}$ and given that $|w_{kj}| = |v_{kj}|$, the two last constraints can be rewritten to obtain $w_j = w'_j w_{kj} \beta_j$. Hence problem (8) can be reformulated as a 1-norm optimization with margin constraints on rescaled inputs:

$$\min_{\mathbf{w}'} \sum_{j=1}^{n} |w'_j| \text{ subject to } y_i(\mathbf{w}' \cdot (\mathbf{x}_i * \mathbf{w}_k * \boldsymbol{\beta}) + b) \geq 1 \tag{9}$$

where $*$ denotes the component-wise product.

The second step of the iterative Franke and Wolfe's method aims at finding $\lambda$ such that $h(\mathbf{w}_k + \lambda(\bar{\mathbf{w}} - \mathbf{w}_k))$ is minimal (with $\bar{\mathbf{w}}$ being the optimal solution to problem (9)). Since $h$ is a weighted positive sum of logarithms, it is concave. Consequently, at each iteration, $\lambda$ is either 0, in which case $\mathbf{w}_{k+1} = \mathbf{w}_k$ and a local optimum is reached, or 1, in which case $\mathbf{w}_{k+1} = \bar{\mathbf{w}} * \mathbf{w}_k * \boldsymbol{\beta}$, and the process is iterated till convergence.

Similarly to the $l2$-AROM method presented in section 2, problem (9) can be approximated by replacing the 1-norm by the 2-norm. This formulation reduces to an iterative algorithm using hard-margin linear SVMs with rescaled margin constraints (a soft-margin variant is straightforward). The original $l2$-AROM method is obtained when $\beta_j = 1$ $(\forall j)$, in other words, without prior preference between the input features.

The RFE approach proposed by (Guyon et al., 2002) is an iterative procedure where a linear SVM is trained at each iteration and features corresponding to the smallest absolute weights are discarded. It has a different initial motivation but is algorithmically very similar to $l2$-AROM. RFE can be seen as an iterative thresholding of the vector $\mathbf{w}_k$, masking some features at each iteration: each dimension is either multiplied by 1 (kept) or 0 (discarded) resulting in a backward selection process. $l2$-AROM performs a smoother selection at each step. We show here how some prior knowledge $\boldsymbol{\beta}$ can weight the smooth selection mask $\mathbf{w}_k$.

Some *a priori* less relevant features may appear in the final solution to problem (9), or its 2-norm approximation, since all $\beta_j$'s are strictly positive. This observation, confirmed in our practical experiments reported in section 4, illustrates why our feature selection procedure is only *partially* (and softly) supervised.

## 4. Experiments

We report here practical experiments with the feature selection method proposed in section 3. These experiments are conducted with the partially supervised $l2$-AROM approach (PS-$l2$-AROM for short) because of its computational efficiency. This choice is also motivated by the results reported in (Weston et al., 2003) which show that classification performances of the original $l1$-AROM and $l2$-AROM methods do not significantly differ while the computational time is in favor of the latter. This method is applied to several microarray data sets described in section 4.1. Two evaluation metrics, respectively measuring the *stability* of the selected genes and the *classification performance*, are defined in section 4.2. The experiments reported in sections 4.3 and 4.4 illustrate that partial supervision leads to an increased stability with improved classification performance in most cases. Comparative results with random supervision also show the soundness of the proposed approach.

### 4.1. Microarray Data Sets

*Table 1.* Microarray data sets characteristics.

| Data Set | Samples | Features | Class Priors |
|---|---|---|---|
| DLBCL | 77 | 7129 | 75%/25% |
| Leukemia | 72 | 7129 | 65%/35% |
| Prostate | 102 | 6033 | 51%/49% |
| Colon | 62 | 2000 | 65%/35% |

Table 1 summarizes the main characteristics of the 4 data sets used in the present study, namely the number of samples, the initial dimension of the input space and the binary class priors. Each dimension corresponds to

the expression value of a particular gene. The classification task in DLBCL (standing for diffuse large B-cells) is the prediction of the tissue types (Shipp et al., 2002). The LEUKEMIA task distinguishes two subtypes of leukemia (Golub et al., 1999). The COLON cancer task discriminates between tumor and normal tissues (Alon et al., 1999). The PROSTATE cancer task discriminates between tumor and non-tumor samples (Singh et al., 2002).

### 4.2. Evaluation metrics

Stability measures to which extent $k$ sets $\mathbf{S}$ of $s$ selected features (gene signatures) share common features. Those sets can typically be produced by selecting features from different samplings of the data. Kuncheva (2007) proposed such a stability index:

$$K(\{\mathbf{S}_1,\ldots,\mathbf{S}_k\}) = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{|\mathbf{S}_i \cap \mathbf{S}_j| - \frac{s^2}{n}}{s - \frac{s^2}{n}}$$

where $n$ is the total number of features, and $\mathbf{S}_i$, $\mathbf{S}_j$ are two signatures built from different subsets of the training samples. The $\frac{s^2}{n}$ ratio in this formula corrects a bias due to the chance of selecting common features among two sets chosen at random. This correction motivates our use of this particular stability index. This index satisfies $-1 < K \leq 1$ and the greater its value the largest the number of commonly selected features in the various sets. A negative stability index means feature sets sharing common features mostly due to chance.

Stability alone cannot characterize the quality of a subset of features. Indeed, if a large randomly chosen set of features were purely forced in every signature, the stability would be very high, but the model built on those features would likely have a poor classification performance. This performance is assessed here with the *Balanced Classification Rate*:

$$BCR = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right)$$

where $TP$ (resp. $TN$) is the number of positive (resp. negative) test samples correctly predicted as positive (resp. negative) among the $P$ positive (resp. $N$ negative) test samples. BCR is preferred to accuracy because microarray data sets often have unequal class priors. BCR is the average between *specificity* and *sensitivity*, two very common measures in the medical domain. BCR can also be generalized to multi-class problems more easily than ROC analysis.

### 4.3. Partial supervision from prior biological knowledge

Shipp et al. (2002) mention two genes previously known as clinical markers to discriminate DLBCL tissues from Follicular Lymphomas: Transferrin Receptor (TR) and Lactate Dehydrogenase A (LDHA).

Our first experiment with PS-$l2$-AROM favors those two dimensions to build a signature of 30 genes (the same signature size as in (Shipp et al., 2002)). We define the prior relevance vector $\boldsymbol{\beta}$ as follows: $\beta_{j\in[\text{TR,LHDA}]} = 10$, $\beta_{j\notin[\text{TR,LHDA}]} = 1$. The relevance value for genes assumed more relevant is arbitrarily assigned to 10. Additional experiments (not detailed here) illustrate that results presented in this section depend only marginally on this choice.

We report here comparative results with no prior preference between genes ($\beta_j = 1\ , \forall j$, in which case PS-$l2$-AROM reduces to the original $l2$-AROM approach). This experiment is run on the whole DLBCL data set (77 samples). Without prior preference ($l2$-AROM), LDHA and TR are not ranked within the top 30 genes selected as signature. In contrast, they correspond to the two largest components of the weight vector $\mathbf{w}$ with non uniform prior relevance (PS-$l2$-AROM). The number of genes differing between signatures generated with and without prior relevance is greater than just the number of favored genes. Only 6 genes are shared between both signatures. This illustrates the multivariate nature of the selection.

Shipp et al. (2002) report a leave-one-out (LOO) accuracy of 91% on the 77 samples with their 30 genes signature. Their classifier is a linear model with weighted voting, where the weights measure the correlation with class labels. Their evaluation does not look completely sound. Firstly, because it relies on accuracy while class priors are unequal. More importantly, because it includes a selection bias as the signature was built on the whole data set before evaluating several classifiers with LOO. With the same biased protocol, a linear SVM built on the 30 genes produced by $l2$-AROM (respectively PS-$l2$-AROM) has 93 % (resp. 92 %) LOO-accuracy. Such a protocol includes an optimistic performance bias (Ambroise & McLachlan, 2002). The additional experiments detailed below avoid such a bias and aim at evaluating both stability and classification performances.

We consider ($k = 200$) independent sub-samplings without replacement of the DLBCL data set with arbitrary splits into 90% training - 10% test. Figure 1 reports the average stability and BCR results obtained with PS-$l2$-AROM and $l2$-AROM for several signature

sizes. The RANDOM approach refers to PS-$l2$-AROM when the partial supervision relies on two genes picked at random[3] instead of TR and LDHA. We also report results obtained with the related approach RFE (see section 3) and Golub's S/N ratio (Golub et al., 1999). This univariate filtering method measures the correlation with the class labels and ranks genes according to $\frac{|\mu_j^+ - \mu_j^-|}{\sigma_j^+ + \sigma_j^-}$, where $\mu_j^+$ (resp. $\mu_j^-$) is the mean expression value of the gene $j$ for positively (resp. negatively) labeled samples, and $\sigma_j^+, \sigma_j^-$ are the associated standard deviations. For all methods, soft-margin linear SVMs are built on selected features on the training sets and evaluated on the test sets[4].

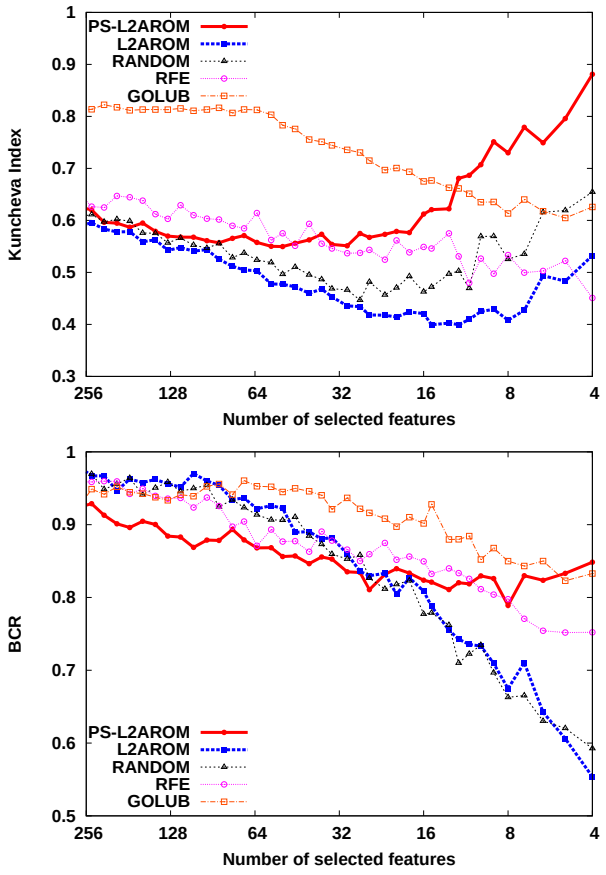


*Figure 1.* Signature stability (Kuncheva index) and classification performance (BCR) of PS-$l2$-AROM (with $\beta_{j\in[\text{TR,LHDA}]} = 10$), $l2$-AROM, RANDOM, RFE and Golub's S/N filtering on the DLBCL data set. Average results over 200 runs (90 % training - 10% test).

[3]We perform 10 independent random selection of genes and 20 random partition into 90% training and 10% test for a total of 200 independent runs on which average results are reported.

[4]Microarray data are usually normalized to make sure that each dimension has zero mean and unit variance across samples. In order to avoid another common bias, we estimate the normalization coefficients on the training sets only and apply those coefficients to normalize the test data.

The comparison between $l2$-AROM and PS-$l2$-AROM shows that a partial supervision on only 2 genes improves drastically the stability of gene signatures with 64 or fewer genes. There is also an important gain in classification performance for very small signatures ($\leq 8$ genes). It is expected that those effects would be even stronger, or also observed for larger signatures, if additional biological knowledge were available to favor more genes (see section 4.4). Partial supervision with randomly chosen genes increases the stability with respect to $l2$-AROM, because they are favored through PS-$l2$-AROM, but not at all the classification performance. This illustrates that, if the partial supervision is based on (likely) irrelevant dimensions, the PS-$l2$-AROM may depart from those but without improving prediction. RFE offers intermediate classification performance but a lower stability for small signatures, while S/N filtering offers good classification results but a drop in stability when fewer than 64 genes are selected.

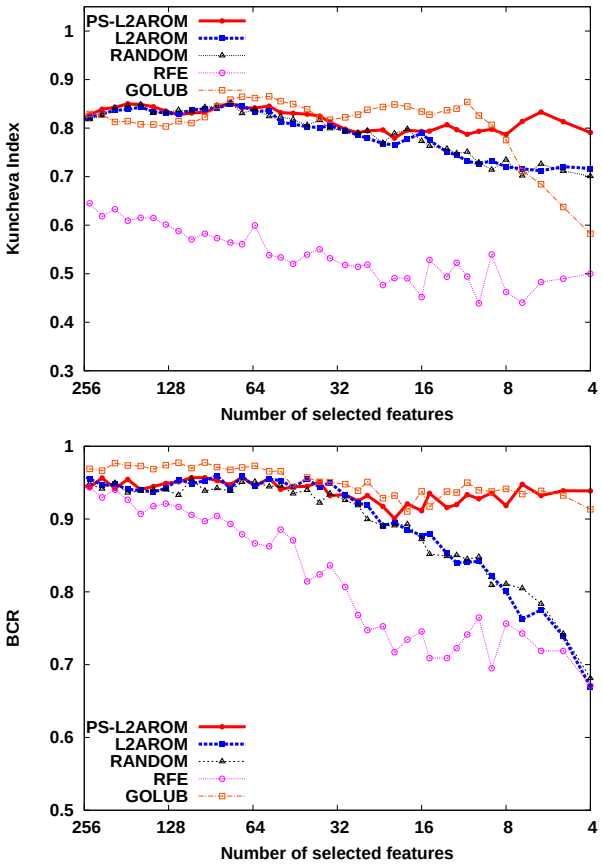


*Figure 2.* Signature stability (Kuncheva index) and classification performance (BCR) of PS-$l2$-AROM (with $\beta_{j\in[\text{CD11c,CD33,MB}-1]} = 10$), $l2$-AROM, RANDOM, RFE and Golub's S/N filtering on the LEUKEMIA data set. Average results over 200 runs (90 % training - 10% test).

Three genes, CD11c, CD33 and MB-1, are mentioned

in (Golub et al., 1999) as clinical markers to distinguish between AML and ALL leukemia subtypes. We repeat the same experiments on the LEUKEMIA data set with ($k = 200$) random splits in 90% training and 10% test. Figure 2 reports stability and classification performances. The conclusions are even stronger than those obtained on DLBCL, with substantial improvements both in stability and BCR for small signatures. Partial supervision with 3 randomly selected genes offers no benefit, neither in stability nor in BCR, as compared to no supervision. RFE performances are comparatively worse on this dataset both in stability and BCR, while S/N filtering offers BCR results equivalent to PS-L2-AROM with a drop in stability for small signatures.

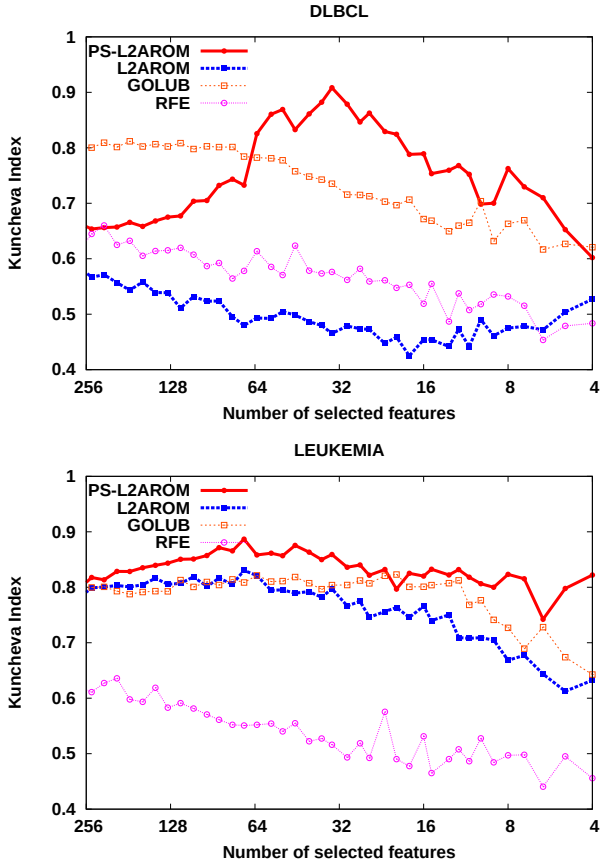


*Figure 3.* Signature stability (Kuncheva index) on DLBCL and LEUKEMIA. Partial supervision for PS-$l2$-AROM with 50 genes selected on an independent partition (20%) with Golub's S/N ratio.

### 4.4. Partial supervision from data partitioning

It is interesting to compare the various selection methods on additional data sets while checking the influence of a partial supervision on a larger number of genes. However we do not always have access to such a biological knowledge on public databases. Our experimental protocol is consequently adapted as follows. Each data set is first randomly split into two stratified folds representing respectively 20% and 80% of the whole data. Prior knowledge is simulated by computing on the 20% partition a signature $S_{prior}$ made of the 50 most differentially expressed genes according to Golub's S/N ratio. The 80% partition is subsequently randomly partitioned into 90% training and 10% test sets. The 50 dimensions in $S_{prior}$ are favored with PS-$l2$-AROM to select features on the 90% training set on which a linear SVM is built. BCR performances are estimated on the 10 % test set. We report average results over a total of 200 runs: 10 random external splits (20%-80%) and ($k = 20$) internal splits (90%-10%).

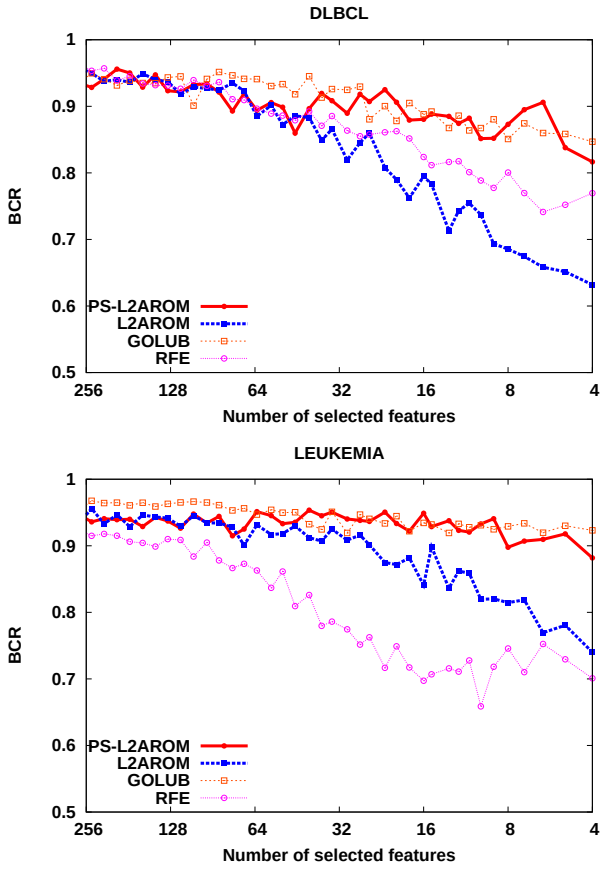


*Figure 4.* Classification performances (BCR) on DLBCL and LEUKEMIA. Partial supervision for PS-$l2$-AROM with 50 genes selected on an independent partition (20%) with Golub's S/N ratio.

Figures 3 and 4 report stability and classification performances on DLBCL and LEUKEMIA with $\beta_{j \in S_{prior}} = 10$, $\beta_{j \notin S_{prior}} = 1$ for PS-$l2$-AROM. The partial supervision greatly increases both the stability of the selected gene lists and BCR with respect to $l2$-AROM. Results with RFE are globally worse especially in terms of stability. BCR results are equivalent between Golub's filtering and PS-$l2$-AROM while the latter generally offers a better stability for small

signatures. Figures 5 and 6 show that significant stability improvements are observed with PS-$l2$-AROM on the PROSTATE and COLON datasets. If we combine stability and BCR performances in a single metric (for instance, by computing the geometric average between both measures), PS-$l2$-AROM offers improved results over $l2$-AROM in all our experiments. The closest competitor to PS-$l2$-AROM is Golub's S/N filtering combined with a linear SVM classifier but this is likely related with the fact that the prior knowledge was precisely simulated with Golub's S/N ratio.

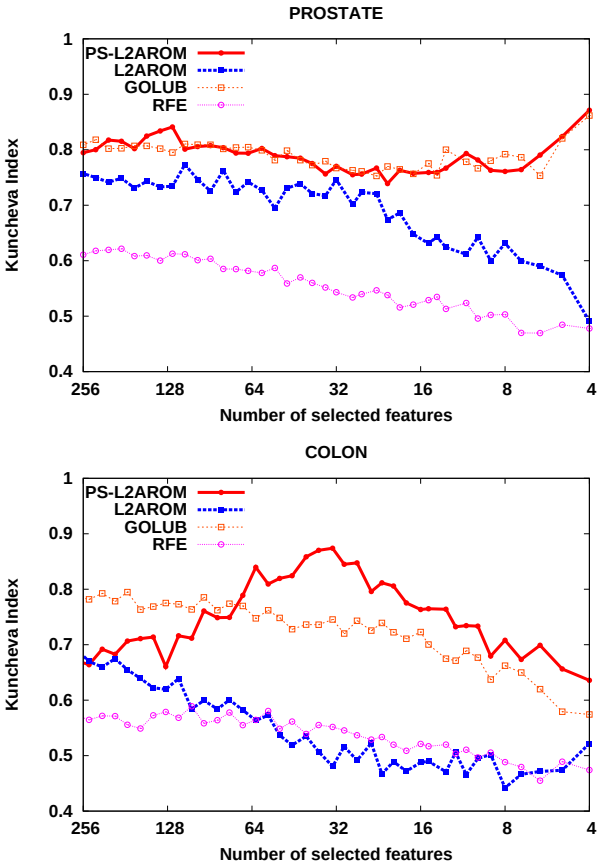


*Figure 5.* Signature stability (Kuncheva index) on PROSTATE and COLON. Partial supervision for PS-$l2$-AROM with 50 genes selected on an independent partition (20%) with Golub's S/N ratio.

## 5. Conclusion and perspectives

We propose a new feature selection method based on regularized linear models. This approach makes use of a partial supervision on the features *a priori* assumed to be more relevant. This method naturally extends the AROM methods due to (Weston et al., 2003). Several experiments on microarray data sets show that the partial supervision largely improves the stability of the selected gene lists, with respect to variation in data sampling. Classification performances are also improved in most cases. Since the selection algorithm is multivariate, partial supervision of a few dimensions may influence the other selected features. The dimensions *a priori* favored may also be discarded if necessary to optimize the classification objective.

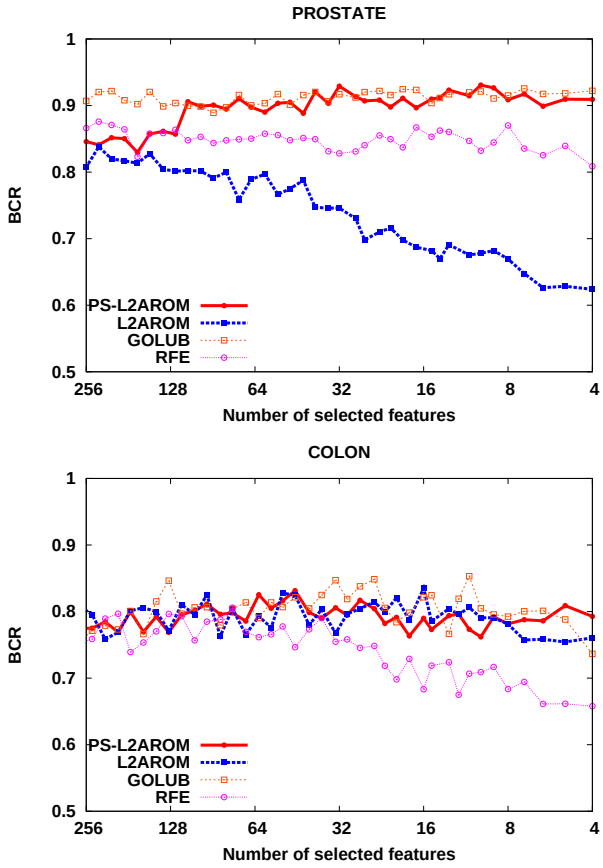


*Figure 6.* Classification performances (BCR) on PROSTATE and COLON. Partial supervisionfor PS-$l2$-AROM with 50 genes selected on an independent partition (20%) with Golub's S/N ratio.

The iterative learning algorithm uses a $l1$-norm regularization. This objective function can be subsequently approximated with a $l2$-norm. Even though such an approximation may result in a less sparse solution, it is very efficient in practice for high dimensional data with few samples. This algorithm then reduces to linear SVM learning with iterative rescaling of the inputs. The scaling factors directly depend on the prior relevance defined on each dimension. Approximating $l1$-AROM by $l2$-AROM had no significant influence on the practical results described in (Weston et al., 2003). It would be worthwhile to confirm this observation when partial supervision is added to the feature selection. We also plan to study to which extent partial supervision could be applied to other regularized models, such as the generalized LASSO (Roth, 2004) or Huberized SVMs (Wang et al., 2007).

Our selection approach was originally motivated by microarray data experiments. It is however a general feature selection technique that can be used in principle in any application domain with some prior preference on the relevant features. The weights to favor some dimensions could also depend on the degree of certainty of the prior knowledge. The proposed approach could also be used to perform transfer learning across tasks, while acquiring prior knowledge on one dataset and using it as partial supervision on others.

## Acknowledgements

T. Helleputte is funded by a FRIA grant. All computations were run on the INGRID cluster of the Center for Intensive Computation and Mass Storage (Louvain).

## References

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., & Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS*, *96*, 6745–6750.

Ambroise, C., & McLachlan, G. (2002). Selection bias in gene extraction on the basis of microarra gene-expression data. *PNAS*, *99*, 6562–6566.

Chapelle, O., Vapnik, V., Bousquet, O., & Mukherjee, S. (2002). Choosing multiple parameters for support vector machines. *Machine Learning*, *46*, 131–159.

Cheng, Y., Cai, Y., Sun, Y., & Li, J. (2008). Semi-supervised feature selection under logistic I-RELIEF framework. *19th International Conference on Pattern Recognition*.

Franke, M., & Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quaterly*, *3*, 95–110.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., & Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*, 531–537.

Guyon, I., & Elisseef, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*, 389–422.

Krishnapuram, B., Carin, L., & Hartemink, A. (2004). *Kernel methods in computational biology*, chapter 14: Gene Expression Analysis: Joint Feature Selection and Classifier Design, 299–317. Cambridge, MA: MIT Press.

Kuncheva, L. (2007). A stability index for feature selection. *Proceedings of the 25th International Multi-Conference: Artificial Intelligence and Applications* (pp. 390–395). Anaheim, CA, USA: ACTA Press.

Mukherjee, S. (2003). *A practical approach to microarray data analysis*, chapter 9: Classifying Microarray Data Using Support Vector Machines, 166–185. Springer.

Roth, V. (2004). The generalized LASSO. *IEEE Transactions on Neural Networks*, *15*, 16–28.

Saeys, Y., Inza, I., & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, *23*, 2507–2517.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., Ray, T. S., Koval, M. A., Last, K. W., Norton, A., Lister, T. A., Mesirov, J., Neuberg, D. S., Lander, E. S., Aster, J. C., & Golub, T. R. (2002). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, *8*, 68–74.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T., & Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, *1*, 203–209.

Wang, L., Zhu, J., & Zou, H. (2007). Hybrid huberized support vector machines for microarray classification. *Proceedings of the 24th International Conference on Machine Learning* (pp. 983–990).

Weston, J., Elisseef, A., Schlkopf, B., & Tipping, M. (2003). Use of the zero-norm with linear models and kernel methods. *Journal of Machine Learning Research*, *3*, 1439–1461.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., & Vapnik, V. (2000). Feature selection for SVMs. *Advances in Neural Information Processing Systems* (pp. 668–674).

Zhao, Z., & Liu, H. (2007). Semi-supervised feature selection via spectral analysis. *7th SIAM International Conference on Data Mining* (pp. 641–652).