The Coxlogit model: feature selection from survival and classification data

Samuel Branders
ICTEAM - Machine Learning Group
Université catholique de Louvain
Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium.
Email: samuel.branders@uclouvain.be

Roberto D'Ambrosio
ICTEAM - Machine Learning Group
Université catholique de Louvain
Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium.
Email: roberto.dambrosio@uclouvain.be

Pierre Dupont
ICTEAM - Machine Learning Group
Université catholique de Louvain
Place Sainte Barbe 2, B-1348 Louvain-la-Neuve, Belgium.
Email: pierre.dupont@uclouvain.be

***Abstract*—**

This paper proposes a novel approach to select features that are jointly predictive of survival times and classification within subgroups. Both tasks are common but generally tackled independently in clinical data analysis. Here we propose an embedded feature selection to select common markers, i.e. genes, for both tasks seen as a multi-objective optimization. The Coxlogit model relies on a Cox proportional hazard model and a logistic regression that are constrained to share the same weights. Such model is further regularized through an elastic net penalty to enforce a common sparse support and to prevent overfitting. The model is estimated through a coordinate ascent algorithm maximizing a regularized log-likelihood.

This Coxlogit approach is validated on synthetic and real breast cancer data. Those experiments illustrate that the proposed approach offers similar predictive performances than a Cox model for survival times or a logistic regression for classification. Yet the proposed approach is shown to outperform those standard techniques at selecting discriminant features that are informative for both tasks simultaneously.

I. Introduction

An important objective in cancer research aims at developing guidelines to determine a personalized treatment strategy. A specific cancer is sometimes thought as a single disease, but increasing evidence suggests that even within the same typology of cancer (e.g. breast, colon,...) there are multiple subgroups that respond to different kinds of treatment [1]. These subgroups are more or less aggressive and they have variable long-term survival rates [2], [3]. Hence, in order to choose the right treatment for each population, one needs to understand the links between patient subgroups and the survival times, that is, the times at which specific events such as metastasis, relapse or death, occur. One typical approach towards this objective would find markers (e.g. genes) that could explain both the subgroup and the survival time observed. The list of these markers provides information about the biological processes and pathways that are involved [4], [5].

In general this task is not trivial as the cancer is a complex decease and distinct genes may be relevant for the survival time estimation [6], [7] or the classification in subgroups [8]. When the two problems are tackled independently, Cox proportional hazard models or logistic regression are typically used to model survival times, on one hand, and to classify samples, on the other [11]–[13].

When the two problems are faced jointly the traditional approach consists of a two step procedure [10]. Firstly, a subset of markers are identified according to their ability to differentiate between the subgroups of patients. Secondly, a model using those markers as input features is estimated to predict survival times. The features selected by such a two step procedure could however be inconsistent between both tasks. Indeed markers that are good predictors of the patient subgroups need not be good estimators for the survival times.

In this work we propose an original feature selection method that is able to detect markers relevant for fitting survival time and simultaneouly classifying samples in subgroups.

Section II briefly revisits the Cox proportional hazard model, logistic regression and the elastic-net regularization to perform an embedded feature selection. Section II-D specifically presents the Coxlogit model seen as a mixture of Cox and logistic regression. This novel model promotes the features selected both in subgroup classification and survival time prediction.

Practical experiments on synthetic and real breast cancer datasets are reported in Section III. They show that the Coxlogit approach offers similar predictive performances than a Cox model for survival times or a logistic regression for classification while it selects features jointly discriminant for both tasks.

II. Methods

One considers a typical survival analysis framework in addition to a class label denoting a specific subgroup. Each sample $i \in \{1, \ldots, n\}$ is characterized by a 4-tuple $(t_i, \delta_i, y_i, x_i)$ where t_i is the time of an event (such as metastasis or relapse) whenever δ_i equals 1 and the censoring time whenever δ_i equals 0. Furthermore y_i denotes a binary class label,

IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making, Orlando, Florida, Dec 9-12, 2014.

respectively -1 and 1 for the two subgroups. The vector x_i includes p covariates for the sample i, typically corresponding to a specific patient.

A. Cox proportional hazard model

The Cox model assumes the hazard $h_i(t)$ and $h_j(t)$ of any pair of samples (i, j) to be proportional [14]. The hazard function $h_i(t)$ is a time depending function that gives the probability of a patient i to have the event (death, relapse, etc) at time t knowing that he has not yet experienced the event just before time t.

With this Cox hypothesis of proportionality, the hazard of a patient can be rewritten as the product of a baseline hazard $h_0(t)$ and a positive function of the covariates :

$$h_i(t) = h_0(t) \exp\left(\beta^\top x_i\right) \tag{1}$$

The partial likelihood of the Cox model can be written as :

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\left(\beta^\top x_i\right)}{\sum_{j \in R(t_i)} \exp\left(\beta^\top x_j\right)} \right]^{\delta_i} \tag{2}$$

where $R(t_i) = \{j | t_j \geq t_i\}$ is the set of patients still at risk just before time t_i. A Breslow approximation is used in the partial likelihood for the ties (patients with the same time of event).

B. Logistic regression

The logistic regression is a very common generalized linear model for classification. The logistic model predicts from the vector x_i associated to patient i, the probability of this patient to be in a specific subgroup/class.

$$P(y_i = 1|x_i) = \frac{\exp(\beta^\top x_i)}{1 + \exp(\beta^\top x_i)} \tag{3}$$

$$P(y_i = -1|x_i) = \frac{1}{1 + \exp(\beta^\top x_i)} \tag{4}$$

$$= 1 - P(y_i = 1|x_i) \tag{5}$$

One typically looks for the parameters β to maximize the model likelihood $L(\beta)$ while assuming the samples to be independent.

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{1 + \exp(-y_i(\beta^\top x_i))} \tag{6}$$

C. Embedded feature selection

In this context, a popular way to perform feature selection is through an embedded method. Those methods use sparse regularization techniques such as the LASSO [15], [16]. Alternatively, a mixture of an l_1 (LASSO) and l_2 (ridge) penalties, referred to as the elastic-net regularization [17], is often considered. This regularization has been successfully applied to several biomedical problems with $p \gg n$ and correlated features. It can be used with any generalized linear model such as the logistic regression for classification or a Cox proportional hazard model for survival prediction. Hence the general form of the optimization problem to fit the model parameters is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left[\frac{2}{n} l(\beta) - \lambda P_\alpha(\beta) \right] \tag{7}$$

where $l(\beta)$ is the log-likelihood of the Cox or the logistic regression (respectively $l_{\text{cox}}(\beta)$ and $l_{\text{logit}}(\beta)$), and $P_\alpha(\beta)$ is the elastic-net penalty:

$$P_\alpha(\beta) = \left(\alpha \sum_{k=1}^{p} |\beta_k| + (1 - \alpha) \frac{1}{2} \sum_{k=1}^{p} \beta_k^2 \right) \tag{8}$$

$\alpha \in [0 - 1]$ is a meta-parameter of the regularization, which reduces to the LASSO (respectively to ridge regression) whenever $\alpha = 1$ (respectively $\alpha = 0$).

D. The Coxlogit mixed model

The Cox proportional hazard regression and logistic regression are standard models for their respective tasks: either fitting survival times or classifying samples. Here we make use of both types of supervision, i.e. the time to event (t_i, δ_i) and the class label y_i, to fit both models constrained to share the same set of parameters β. As both optimization problems have a similar form and can be efficiently solved by the same algorithm, one can maximize both likelihoods jointly. The log-likelihood of this model is a mixture of a Cox and a logistic model sharing the same sparse parameter vector β:

$$l(\beta) = (1 - \gamma)\ l_{\text{cox}}(\beta)\ +\ \gamma\ l_{\text{logi}}(\beta) \tag{9}$$

The meta-parameter $\gamma \in [0 - 1]$ controls the contribution of either log-likelihoods in the model, with $\gamma = 0$ (respectively $\gamma = 1$) corresponding to a pure Cox model (respectively a pure logistic regression).

In all cases, the same vector β is combined with the covariates x_i of a new sample to define a risk score $\beta^\top x_i$. This score can be either directly used in survival analysis or included in equation (3) to define a *posterior* class probability.

The mixed log-likelihood can be optimized under an elastic-net regularization using the same optimization framework as for equation (7):

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \left[\frac{2}{n} \Big((1-\gamma)\ l_{\text{cox}}(\beta) \ + \ \gamma\ l_{\text{logi}}(\beta) \Big) - \lambda P_\alpha(\beta) \right]$$

where $P_\alpha(\beta)$ is the regularization term (Equation 8).

In this paper, a coordinate ascent algorithm is used to solve the optimization problem for both the logistic and Cox model as presented in [18], [19]. This efficient algorithm computes the regularization path. It starts with the trivial optimal solution given by a high λ and follows the optimum when λ is decreased. Applied to feature selection, the algorithm follows the regularization path and stops when the model contains a desired number of features.

III. Experiments

In order to validate the proposed Coxlogit model in prediction and feature selection tasks, we perform experiments on synthetic and real datasets. The performances in classification and survival are assessed according to the classification accuracy and the concordance index, respectively. The concordance index (C-index) measures to which extent the risk groups are concordant with the time to event, that is, whether the patients in the high risk group actually experience the event before the patients in the low risk group [20].

The first experiments reported on a synthetic data set (section III-A) illustrate that the Coxlogit model is able to select features that are jointly informative for survival and subgroup classification. Those results are confirmed on real breast cancer prognosis studies (section III-C). In general, the Coxlogit approach offers similar classification results than a logistic regression and survival prediction results similar to a standard Cox model but Coxlogit is better at selecting genes predictive for both tasks.

A. Synthetic data

The synthetic data set is designed to have both supervisions in terms of survival times and subgroup classification. The data set is designed with four groups of features. The first 3 groups of k features are predictive of respectively

- both the survival and the class label.
- the survival only.
- the class label only.

while the $p - 3k$ remaining features are purely random and supposed to represent noise.

One would like to assess to which extent the Coxlogit approach is able to select features from the first group, as compared to a regularized Cox or logistic model alone. The data matrix $X \in \mathbb{R}^{n \times p}$ is drawn from a $\mathcal{N}(0,1)$ distribution to represent covariates that have been centered and normalized to unit variance, a common practice in our context.

The class assignments and hazards are generated from distinct linear combinations of the features. The weights of those predictors $\beta \in \mathbb{R}^{3k}$ are drawn from a uniform distribution over $[-1, 0.5] \cup [0.5, 1]$. Class labels $y \in \{-1, 1\}^n$ are generated from the first and third groups in the following way:

$$\begin{aligned} \beta_{class} &= (\ \beta_1 \dots \beta_k \quad 0 \dots 0 \quad \beta_{2k+1} \dots \beta_{3k} \quad 0 \dots 0\ ) \\ y &= \operatorname{sign}\left(X \beta_{class}^\top\right) \end{aligned}$$

The survival data (t_i, δ_i) are generated from two weibull distributions, for the time to event and the censoring respectively. The weibull distribution for the time to event is parametrized such that the hazard $h_i(t)$ depends on the features from the first and second groups :

$$\begin{aligned} \beta_{surv} &= (\ \beta_1 \dots \beta_k \quad \beta_{k+1} \dots \beta_{2k} \quad 0 \dots 0 \quad 0 \dots 0\ ) \\ h_i(t) &\propto \exp(X \beta_{surv}^\top) \end{aligned}$$

The two classes, -1 and 1, exhibit a difference in their survival, but these classes only account for part of the survival differences between patients, as expected in real data. In practice, we consider a data set of $n = 1000$ patients with $k = 10$ features in each of the 3 groups for a total of $p = 100$ features. 200 samples are used for training and 800 independent samples serve as validation.

B. Results on synthetic data

This section reports the results using the Coxlogit model on the synthetic data described above, as compared to a regularized Cox model ($(\gamma = 0)$ or a logistic regression ($\gamma = 1$). Figure 1 reports the model weights obtained while varying γ in $[0, 1]$.

For each experiment, the regularization path is followed till the model contains exactly 10 features. The absolute weight value assigned to each feature can be easily interpreted as the relevance of the features estimated by the model. Figure 1 shows a smooth transition between the features selected by the method while varying the value of γ. The Cox model only selects features that contains some survival informations (in green and red). Similarly, the regularized logistic model only selects features associated to class assignments (in green and blue), plus one random feature in this particular run. In contrast, the Coxlogit model (typically for $\gamma = 0.5$), tends to select only those features (in green) that are informative for both tasks.

Figure 2 reports the absolute weight values while repeating the above experiment 100 times and averaging those absolute values over the 100 runs. While the Cox model, respectively the logistic regression, always selects features related to the survival, respectively the classification in subgroups, the Coxlogit model clearly favors the selection of common features. This experiment also shows that no specific random feature is consistently selected over all runs.

Figures 3 and 4 report the prediction results, respectively in terms of classification accuracy and C-index, obtained using the features selected by the Coxlogit model and averaged over 100 runs.

The results presented in Figure 3 show that the best classification accuracy is obtained for a logistic regression or a Coxlogit model with $\gamma \geq 0.75$. Logically, a model fitted to maximize only a regularized Cox log-likelihood ($\gamma = 0$) performs poorly according to this metric.

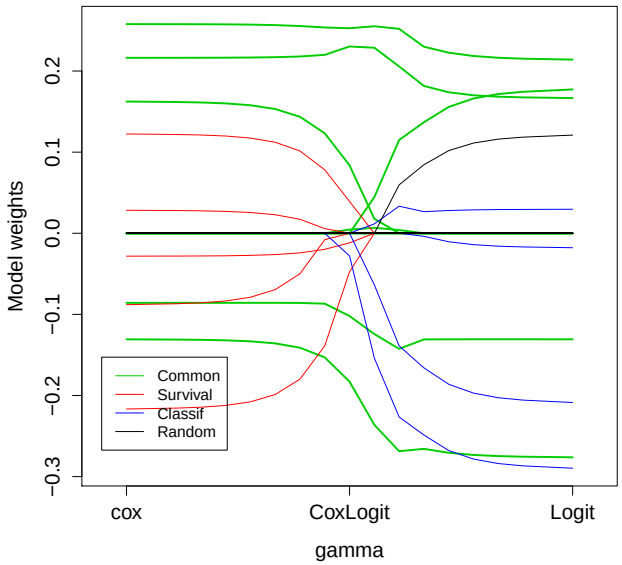

Fig. 1. Model weights obtained while varying γ from 0 (Cox model) to 1 (Logistic model). The absolute value of the weights represent the importance assigned by the model to each feature. Only 10 out of the 100 weights have a non-zero value as a consequence of the chosen working condition along the regularization path.

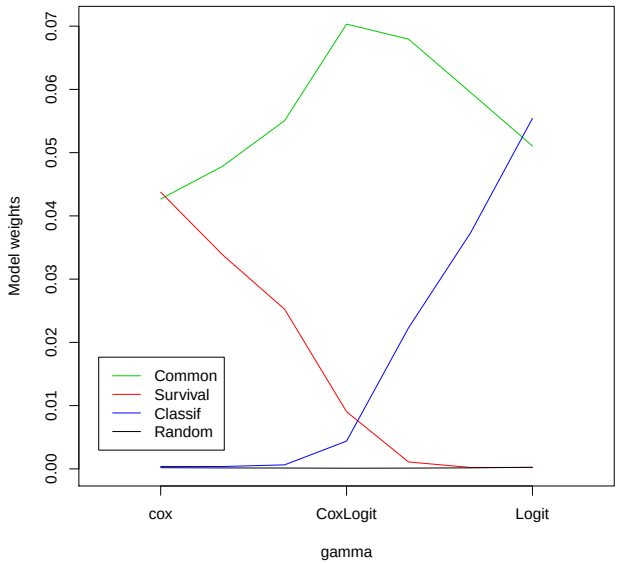

Fig. 2. Mean absolute weight values in each group of features (Common, Survival, Classification, Random) computed while varying γ within $[0, 1]$.

Similarly, as reported in Figure 4, the best C-index is obtained for a Cox model or a Coxlogit model with $\gamma \leq 0.5$ while a model fitted to maximize only a regularized logistic log-likelihood ($\gamma = 1$) is poor at predicting survival times.

In summary, the Coxlogit model is able to perform comparably at both tasks, provided an adequate choice of γ, while using only a third of the relevant features (those which are commonly informative for survival prediction and classification).

C. Breast cancer experiments

Further tests, reported here on 4 breast cancer datasets, validate the Coxlogit performances in a real scenario. These data sets are publicly available from the GEO database with accession numbers GSE2034, GSE5327, GSE7390 and GSE2990. In all those studies, distant metastasis serves as survival end point and gene expression data are measured on the Affymetrix HGU133A microarray platform. All data sets were summarized according to the MAS5.0 procedure and represented in $\log_2$ scale. A class independent pre-filtering is performed to keep only those features with highest variances.

Three histo-pathological markers, the tumor size, the number of invaded lymph nodes and the grade of the tumor, commonly predict the outcome for breast cancer patients [22]. In particular, the grade measures how much the tumor cells are differentiated from the physiological tissue, i.e. normal breast tissue. Understanding the link between the grade of the tumor and the probability of survival is an important task studied in the medical literature [10], [23].

We propose to revisit here this question with the Coxlogit

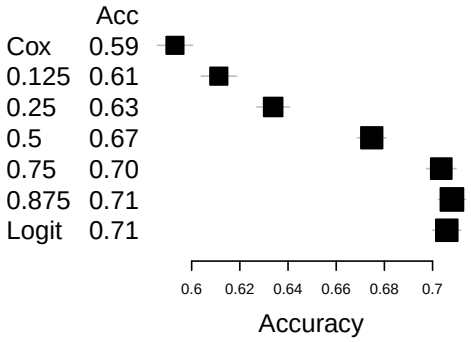

Fig. 3. Forest plot of the classification accuracy on 100 repeated experiments with synthetic data sets.

model. Our goal is to predict both the grade of the tumor and the survival probability of the patients. All the experiments were performed after pooling the 4 GEO datasets in a single dataset of 554 patients consisting of 1236 features, *i.e.* gene expression values. The two classes, low (≤ 2) and high (> 2) grades, of the binary classification problem approximately correspond to the same number of patients.

The predictive performances are computed in terms of accuracy and C-index, respectively for the classification and the survival prediction. Those performances are reliably estimated through a resampling protocol. The following steps are repeated 100 times:

- 80% of the patients are selected at random (without replacement) as training set. The remaining 20% is used as validation set.
- For each $\gamma \in \{0, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, \frac{7}{8}, 1\}$:

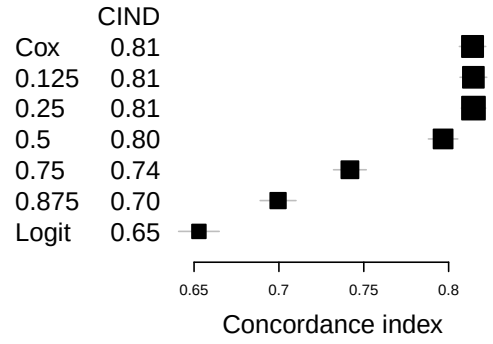

Fig. 4. Forest plot with the results in C-index on 100 repeated experiments with synthetic data sets.

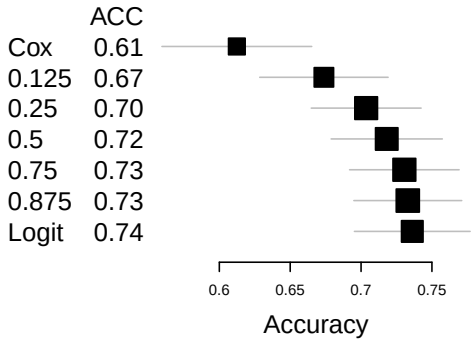

Fig. 5. Forest plot: grade classification results on breast cancer samples.

- Compute a Coxlogit model on the train.
- Predict the class and survival on the test.

Each model has been estimated by following the regularization path such that one selects exactly 10 features. The statistics recorded at each run of the above protocol are then aggregated to provide an average classification accuracy and average C-index for all models.

D. Breast cancer prognosis and classification results

Figures 5 and 6 report the average results over the 100 experiments generated according to the protocol presented in the section III-C.

Figure 5 shows the accuracy achieved with the Coxlogit model in the classification task while varying the value γ. These results confirm those obtained on synthetic data. In particular, the best classification results are obtained for a logistic regression or a Coxlogit model with $\gamma \geq 0.75$. Again, a model fitted to maximize only a regularized Cox log-likelihood ($\gamma = 0$) performs poorly according to this metric.

Figure 6 presents, in terms of C-index, the results of the survival prediction task. A Cox model or a Coxlogit model with $\gamma \leq 0.5$ offer the best performances while a model fitted to maximize only a regularized logistic log-likelihood ($\gamma = 1$) perform slightly worse. The differences are here less striking than those observed in classification accuracy. Globally, a Coxlogit model with $\gamma = 0.5$ offers a good performance trade-off for both tasks.

We further study the feature selection characteristics of the Coxlogit approach and use the 10 genes selected by a Coxlogit model with $\gamma = 0.5$ as reference. Figure 7 reports the size of the intersection between 10 gene signatures obtained while varying γ and the reference signature. By design, the overlap is maximum at the middle of the plot. Clearly, the signature content diverges quickly as one decreases or increases γ from its reference 0.5 value. The overlap is minimum at both extremes with respectively only 3 ($\gamma = 0$) or 4 genes ($\gamma = 1$) in common with the chosen reference.

We also consider a set of 7 genes, the so-called CGI signature, known to be predictive of the tumor grade [10]. Figure 8 reports the average overlap between the genes selected

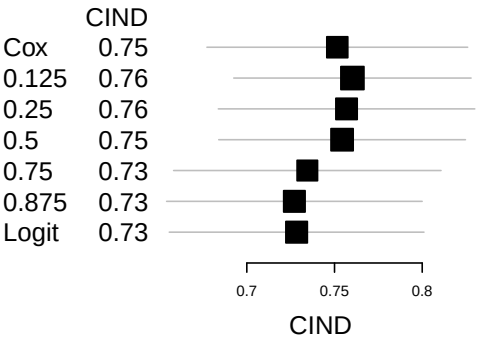

Fig. 6. Forest plot: survival prediction results (C-index) on breast cancer samples.

by a Coxlogit model while varying γ in $[0; 1]$ and this reference signature. Experiments were conducted here over 100 independent runs of the above protocol. The highest overlap is achieved with a Coxlogit model using $\gamma = .75$. Those results confirm that mixing of class and survival time supervisions brings informative gene signatures.

IV. Conclusion and perspectives

This work describes the Coxlogit method: a novel feature selection approach towards the combined objective of fitting survival times and classifying samples in subgroups. This method relies on a selection embedded into the fitting of a generalized linear model. The parameters of this model are estimated so as to maximize a mixture between a logistic and Cox log-likelihoods including an elastic net regularization.

Experiments reported both on synthetic and breast cancer data illustrate that the Coxlogit approach is indeed able at identifying features which are jointly predictive of survival times and classification within subgroups. As such this approach outperforms a Cox proportional hazard model or a logistic regression, each of those models being efficient essentially on their own task.

Several questions will be addressed in our future work on this topic. Firstly, an appropriate choice of the mixing coefficient γ deserves a further study, even though satisfactory results are reported here with a balanced contribution ($\gamma = 0.5$). Secondly, the number of selected features results

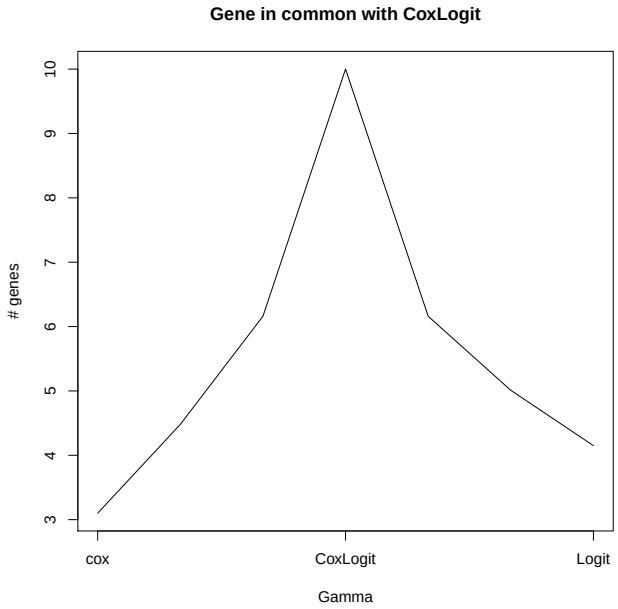

Fig. 7. Average number of genes selected for each model that are also selected by a Coxlogit model (at $\gamma = 0.5$)

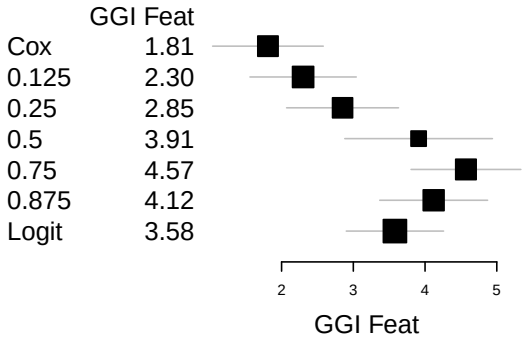

Fig. 8. Average number of genes selected for each model that come from the GGI signature.

from the user choice when to stop the regularization path while estimating this model. Such path is in turn controlled by the meta-parameters of the elastic net penalty. The above observations call for an appropriate model selection strategy or, at least, for studying the performances of the Coxlogit approach across various feature set sizes.

The supervision used at training time assumes that the samples are labeled into 2 specific subgroups. Generalization to more than 2 subgroups looks interesting and easy. It would essentially amount to replace the logistic part of this model by its multinomial extension using a softmax function. Finally, the specific subgroups are here assumed to be *a priori* known for all samples. Such assumption could be relaxed and unsupervised or semi-supervised learning of those groups could be considered.

ACKNOWLEDGMENT

a) Funding: The work of Samuel Branders is supported by the F.R.S. - FNRS - Télévie (Grant number FC 88088).

REFERENCES

[1] R. Rouzier, C. M. Perou, W. F. Symmans, N. Ibrahim, M. Cristofanilli, K. Anderson, K. R. Hess, J. Stec, M. Ayers, P. Wagner, P. Morandi, C. Fan, I. Rabiul, J. S. Ross, G. N. Hortobagyi, and L. Pusztai, "Breast cancer molecular subtypes respond differently to preoperative chemotherapy." *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 11, no. 16, pp. 5678–85, Aug. 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16115903

[2] L. Marisa, A. de Reyniès, A. Duval, J. Selves, M. P. Gaub, L. Vescovo, M.-C. Etienne-Grimaldi, R. Schiappa, D. Guenot, M. Ayadi, S. Kirzin, M. Chazal, J.-F. Fléjou, D. Benchimol, A. Berger, A. Lagarde, E. Pencreach, F. Piard, D. Elias, Y. Parc, S. Olschwang, G. Milano, P. Laurent-Puig, and V. Boige, "Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value." *PLoS medicine*, vol. 10, no. 5, p. e1001453, Jan. 2013. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3660251\&tool=pmcentrez\&rendertype=abstract

[3] B. Haibe-Kains, C. Desmedt, F. Rothé, M. Piccart, C. Sotiriou, and G. Bontempi, "A fuzzy gene expression-based computational approach improves breast cancer prognostication." *Genome biology*, vol. 11, no. 2, p. R18, Jan. 2010. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2872878\&tool=pmcentrez\&rendertype=abstract

[4] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[5] A. Essaghir and J.-B. Demoulin, "A minimal connected network of transcription factors regulated in human tumors and its application to the quest for universal cancer biomarkers." *PloS one*, vol. 7, no. 6, p. e39666, Jan. 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3382591\&tool=pmcentrez\&rendertype=abstract

[6] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. a. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer." *Nature*, vol. 415, no. 6871, pp. 530–6, Jan. 2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/11823860

[7] Y. Wang, J. G. M. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. J. J. Berns, D. Atkins, and J. a. Foekens, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer." *Lancet*, vol. 365, no. 9460, pp. 671–9, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/15721472

[8] A. V. Kapp, S. S. Jeffrey, A. Langerø d, A.-L. Bø rresen Dale, W. Han, D.-Y. Noh, I. R. K. Bukholm, M. Nicolau, P. O. Brown, and R. Tibshirani, "Discovery and validation of breast cancer subtypes." *BMC genomics*, vol. 7, no. 1, p. 231, Jan. 2006. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1574316\&tool=pmcentrez\&rendertype=abstract

[9] D. Venet, J. E. Dumont, and V. Detours, "Most random gene expression signatures are significantly associated with breast cancer outcome." *PLoS computational biology*, vol. 7, no. 10, p. e1002240, Oct. 2011. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3197658\&tool=pmcentrez\&rendertype=abstract

[10] C. Sotiriou, P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi, "Gene expression profiling in breast cancer: understanding the molecular basis

of histologic grade to improve prognosis." *Journal of the National Cancer Institute*, vol. 98, no. 4, pp. 262–72, Feb. 2006. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16478745

[11] K. Yoshihara, T. Tsunoda, D. Shigemizu, H. Fujiwara, M. Hatae, H. Fujiwara, H. Masuzaki, H. Katabuchi, Y. Kawakami, A. Okamoto, T. Nogawa, N. Matsumura, Y. Udagawa, T. Saito, H. Itamochi, M. Takano, E. Miyagi, T. Sudo, K. Ushijima, H. Iwase, H. Seki, Y. Terao, T. Enomoto, M. Mikami, K. Akazawa, H. Tsuda, T. Moriya, A. Tajima, I. Inoue, and K. Tanaka, "High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by downregulation of antigen presentation pathway." *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 18, no. 5, pp. 1374–85, Mar. 2012. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22241791

[12] J. R. Kratz, J. He, S. K. Van Den Eeden, Z.-H. Zhu, W. Gao, P. T. Pham, M. S. Mulvihill, F. Ziaei, H. Zhang, B. Su, X. Zhi, C. P. Quesenberry, L. a. Habel, Q. Deng, Z. Wang, J. Zhou, H. Li, M.-C. Huang, C.-C. Yeh, M. R. Segal, M. R. Ray, K. D. Jones, D. J. Raz, Z. Xu, T. M. Jahan, D. Berryman, B. He, M. J. Mann, and D. M. Jablons, "A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies." *Lancet*, vol. 379, no. 9818, pp. 823–32, Mar. 2012. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3294002\&tool=pmcentrez\&rendertype=abstract

[13] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by lasso penalized logistic regression." *Bioinformatics (Oxford, England)*, vol. 25, no. 6, pp. 714–21, Mar. 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2732298\&tool=pmcentrez\&rendertype=abstract

[14] D. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B*, vol. 34, no. 2, pp. 187–220, 1972. [Online]. Available: http://www.jstor.org/stable/10.2307/2985181

[15] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (* ..., vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[16] ——, "The lasso method for variable selection in the Cox model." *Statistics in medicine*, vol. 16, no. 4, pp. 385–95, Feb. 1997. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9044528

[17] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, Apr. 2005. [Online]. Available: http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x

[18] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2929880/

[19] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Coxs Proportional Hazards Model via Coordinate Descent," *Journal of statistical software*, vol. 39, no. 5, 2011. [Online]. Available: http://www.stanford.edu/~hastie/Papers/v39i05.pdf

[20] F. E. Harrell, K. L. Lee, and D. B. Mark, "Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors." *Statistics in medicine*, vol. 15, pp. 361–387, 1996.

[21] J. Paul, M. Verleysen, and P. Dupont, "Identification of Statistically Significant Features from Random Forests," *ECML workshop on Solving Complex Machine Learning Problems with Ensemble Methods*, 2013. [Online]. Available: http://perso.uclouvain.be/michel.verleysen/papers/ecmlpkdd13jp.pdf

[22] M. H. Galea, R. W. Blamey, C. E. Elston, and I. O. Ellis, "The Nottingham Prognostic Index in primary breast cancer." *Breast cancer research and treatment*, vol. 22, no. 3, pp. 207–19, Jan. 1992. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/1391987

[23] S. Loi, B. Haibe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, C. Gillet, P. Ellis, A. Harris, J. Bergh, J. a. Foekens, J. G. M. Klijn, D. Larsimont, M. Buyse, G. Bontempi, M. Delorenzi, M. J. Piccart, and C. Sotiriou, "Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade." *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 25, no. 10, pp. 1239–46, Apr. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17401012