Curve selection for predicting breast cancer metastasis from prospective gene expression in blood

Einar Holsbø[1] Vittorio Perduca[2] Lars Ailo Bongo[1]
Eiliv Lund[3] Etienne Birmelé[2]

[1]Department of Computer Science, University of Tromsø
[2] Laboratoire MAP5, Université Paris Descartes
[3]Department of Community Medicine, University of Tromsø

May 22, 2017

Abstract

In this article we use gene expression measurements from blood samples to predict breast cancer metastasis. We compare several predictive models and propose a biologically motivated variable selection scheme. Curve selection is based on the assumption that gene expression intensity as a function of time should diverge between cases and controls: there should be a larger difference between case and control closer to diagnosis than years before. We obtain better predictions and more stable predictive signatures by using curve selection and show some evidence that metastasis can be detected in blood samples.

1 Introduction

About one in ten women will at some point develop breast cancer. About 25% have an aggressive cancer at the time of diagnosis, with spread to axillary lymph nodes.[1] The tool to detect this spread is a surgical procedure known

[1]http://oncolex.org/Breast-cancer/

as a sentinel node biopsy. According to the Norwegian Cancer Registry, out of 1000 women who attend all ten screenings they're normally invited to, 200 will experience at least one false positive. Out of these 200, 40 will have to do a biopsy. This biopsy is an invasive procedure. If we could use blood samples to predict metastasis, we could reduce the number of unnecessary biopsies. Several recent articles develop this idea of *liquid biopsies* [3]. Different relevant signals appear in blood for already-diagnosed breast cancer. For instance: circulating tumor cells [10], serum microRNA [22], or tumor-educated platelets [1]. A recent review in *Cancer and Metastasis Reviews* [16] lists liquid biopsies and large data analysis tools as important challenges in metastatic breast cancer research.

Norwegian Women and Cancer (NOWAC) [4] is a prospective study containing blood samples. Prospective blood samples provide gene expression trajectories over time. The hope is that such trajectories diverge between cases and controls as the tumor grows. Lund et al. [18] show a significant difference in trajectories for groups of genes. In this paper we aim to show that we can go one step further and find information even about sentinel node status.

The main difficulty here is high dimensionality. There are about ten to fifteen thousand potential predictor genes. It's very easy to over-fit such data. The number of observations needed to fill some region of p-dimensional space grows more than exponentially fast with p. But there are often lower-dimensional structures in the data. For instance, we expect genes to work together in pathways. We don't expect all genes to be relevant in all processes. The analysis of high-dimensional data is an active research area of statistics and machine learning [8]. Usually we try to discover these low-dimensional structures by projection approaches like PLS-methods [17], or by variable selection.

Variable selection approaches highlight the most discriminative variables, which has a straight-forward interpretation. There is a variety of variable selection schemes, for a review of which see [7]. If we are working with gene expression, we can rank genes for example based on genewise t-tests for differential expression. The top k of these provide a lower-dimensional space where we can apply any classifier. Haury et al [14] show that such a ranking coupled with a simple classification method compares favorably to more sophisticated methods. There are also integrated methods that do simultaneous selection and

statistical learning. A popular choice is the penalized maximum likelihood family of generalized linear models. They optimize the likelihood plus a penalty term that encourages sparse solutions. These include the popular lasso and elastic net methods [13].

Regardless of variable selection method, the chosen predictor set can be unstable. Ein-Dor et al [6] examined the effect of using different subsets of the same data to choose a predictive gene set. They show that predictor gene sets depend strongly on the subset of patients used for analysis. So stability criteria is a complimentary feature to predictive power. These can be integrated in the model selection, as the stability selection for penalized regression [19]. Or they can be used as an a posteriori evaluation criterion [14].

In this paper we compare several learning methods to predict metastasis in breast cancer. We use blood gene expression data from NOWAC taken no more than one year before diagnosis. If we take all genes into account in a desultory manner, we do no better than random guess. In fact, we tend to do worse than random if we don't account for stratification in the data. Hence we propose variable selection based on a gene's prediagnostic trajectory. We call this biologically motivated approach curve selection. Curve selection improves both predictive power and signature stability. We see some evidence that there is a signal of sentinel node status already present before diagnosis. This gives some hope for the pursuit of liquid sentinel node biopsies as a cheaper and less invasive option to surgery.

2 Material and methods

2.1 Data

Our dataset is 88 pairs of breast cancer cases and age-matched controls from the NOWAC Post-genome cohort. The cohort profile by Dumeaux et al. describes the details [4]. In brief, women were recruited by random draw from the Central Person Register by Statistics Norway. They were invited to fill out a questionnaire and provide a blood sample. The Cancer Registry of Norway provided followup information on cancer diagnoses and lymph node status. The women received a diagnosis at most one year after providing a blood sample.

The NTNU genomics core facility processed the blood samples on Illumina microarray chips of either the HumanWG-6 v. 3 or the HumanHT-12 v. 4 type. A case and its matching control are together for the entire processing pipeline. Eg. they lie next to one another on chip, and so on. Afterwards we checked the data for technical outliers. These are observations that get distorted in the lab. We have removed low-signal probes, ie probes that lie below a certain detection threshold. We quantile-normalize the data before analysis. The preprocessing for these particular data is described in detail in Lund et al. [18]. Günther, Holden, and Holden's report from the Norwegian Computing Centre [11] provides more technical details.

In practice we have a 88×12404 gene expression fold change matrix X on the $\log_2$ scale. For each gene, g, and each case–control pair, i, we have the measurement $\log_2 x_{ig} - \log_2 x'_{ig}$. Here x_{ig} is the g expression level for the ith case, and x'_{ig} is the corresponding control. For each case we have the number of days between the blood sample and the cancer diagnosis. Call this the followup time. Note that although there is a time component to this, we don't have time series data. Each observation is a different woman. We also have a detection stratum variable. This takes one of the following values:

- **Screening** denotes a cancer that was detected in the regular screening program.
- **Interval** denotes a cancer that was detected between two screening sessions. The interval between screenings is two years.
- **Clinical** denotes a cancer that was detected outside of the screening program. These women either never took part in the screening program, or did not attended a screening in at least two years.

Finally, our response variable, metastasis ($\in \{0, 1\}$), indicates whether a sentinel node biopsy showed evidence of metastasis.

Table 1 shows the incidence of metastasis in the different strata. We see a certain heterogeneity between strata.

	Screening	**Interval**	**Clinical**
No spread	43	10	13
Spread	6	6	10

Table 1: Incidence of metastasis across detection strata. The incidence seems to vary across strata. Nearly half of the cancers in the clinical stratum show metastasis.

2.2 Predictive models

2.2.1 Penalized regression

We fit penalized logistic regression models. These models take the form

$$\text{logit}(p) = \beta_0 + \beta_1 x_1, + \ldots + \beta_p x_p + \epsilon,$$

$$\text{subject to } c(\hat{\beta}) \leq t,$$

where ϵ is iid mean-zero noise, and $c(\cdot) < t$ is a constraint on the magnitude $c(\cdot)$ of the coefficients β_i. The parameter t controls how severe this constraint is.

We investigate ridge penalty, $c_r(\hat{\beta}) = \sum \hat{\beta}_i^2$, lasso penalty, $c_l(\hat{\beta}) = \sum |\hat{\beta}_i|$, and the elastic net [23] penalty, which is a linear combination of ridge and lasso, $\alpha c_l(\hat{\beta}) + (1-\alpha)c_r(\hat{\beta})$, with $\alpha \in [0, 1]$. Ridge and lasso penalties are special cases of the elastic net. Lasso is well-known to encourage sparse solutions, where many coefficients are set to exactly zero. It's expected to be the better model if there are few relevant predictors. Ridge on the other hand, never shrinks coefficients to exactly zero and as such lets all predictors contribute to some extent. Hence ridge can be expected to do better of most predictors are relevant.

Logistic regression allows us to correct for strata by adding interactions between gene expression and stratum. In the case of genome-wide association studies, it's been shown to be one of the best methods to take stratification into account [2].

2.2.2 Stability selection

The set of predictors picked by regularization are often unstable with correlated predictors. It is also hard to choose the correct amount of regularization, the result is often over-regularized models. Stability selection [19] is a method to

deal with this. Basically: i) make a bootstrap estimate of the probability of a sample's being chosen by your regularized method, ii) define some probability threshold above which you you use all predictors, iii) fit your favorite model using these predictors.

We examine stability selection for both the lasso and the elastic net. Instead of setting a threshold, we simply pick the top 50 predictors in each case (this because we actually don't see much stability for the lasso). We then use Bayesian logistic regression with a weakly informative prior as described by Gelman et al. [9].

2.2.3 Nearest centroids

We also consider the purely geometrical algorithm of nearest centroids (NC). A class C_i is represented by it's centroid point c_i in p-dimensional space, eg. the class mean $c_i = \mu(x|x \in C_i)$. Then $p(C_i|x) \propto d(x, c_i)$, which is to say the probability x belonging to C_i is proportional to the distance between x and the centroid c_i. We normalize all features for this model. Being a distance-based classification algorithm, NC shouldn't be expected to do very well in thousands of dimensions. Hence we use the top 50 genes ranked by simple genewise t-tests.

2.2.4 Stratification

We account for stratification in the regression models by adding an interaction between all genes and the stratum variable. In simplified notation this is the model logit(spread) $= \beta$(expression + expression $\times$ stratum). In practice this leads to a three times as large design matrix of roughly 88 by 36 000 entries. In the nearest centroids model we include the stratum indicators as extra dimensions.

2.3 Curve selection

We would like bring some biology into this model and to take a cancer's potential evolution over time into account in our modeling. Our idea is that, for the relevant genes, cases and controls either have constant differential expression over time, or that they diverge in expression levels over time as a cancer grows and spreads. To detect this we propose to do genewise regression of fold change,

e, on time, t, and metastasis, $M \in \{0, 1\}$, in the following model:

$$e = \beta_0 + \beta_1 t + \beta_2 M + \beta_3 tM + \epsilon, \tag{1}$$

where ϵ is iid noise. For models with stratification we add the stratum variable as another interaction:

$$e = \beta_0 + \ldots + \beta_4 S + \beta_5 tS + \beta_6 SM + \beta_7 tSM + \epsilon, \tag{2}$$

with S the stratum variable. For a ranking score on the genes use the largest Wald statistic $\hat{\beta}_i / \widehat{s.e}(\hat{\beta}_i)$ of any coefficient corresponding to a term with the metastasis variable M as a factor. Ie in equation 1, this is β_2 or β_3. In equation 2 it's one of $\beta_2, \beta_3, \beta_6$, or β_7. This ranking restricts the predictive models to a smaller predictor space. The ranking should favor genes for which metastasized cases diverge from their controls as time progresses. We call this filtering method curve selection.

Figure 1 shows the curve selection model. The top row contains the top three genes in our data as ranked by curve selection, the bottom row contains three random genes for comparison.

2.3.1 Application to models

We use curve selection to filter out uninformative genes with all the models above. In all cases but one we do curve selection as a preselection step to narrow down our predictor space to the 200 best genes. We then apply the models in the usual way. The exception is nearest centroids, where we replace the t-test ranking with curve selection to obtain the 50 genes to compute centroids for.

When we account for stratification in the predictive models, we also account for it in the curve selection as in equation 2.

2.4 Cross-validation

We estimate performance generalization by repeated cross-validation. We have found that simple cross-validation in our setting produces point estimates and confidence bands variable enough to not be of any use. A possible fix to this is to use the bootstrap [5], but there are situations where the bootstrap estimates are biased [15, 20]. Repeated cross-validation puts cross-validation on an equal

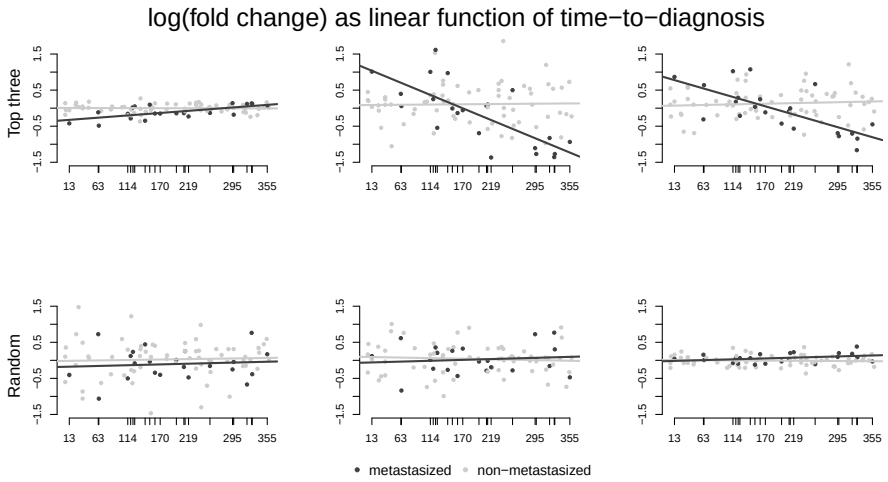

Figure 1: Examples of ranking by curve selection. The light points are non-metastasized cases, while the dark ones are metastasized cases. The slopes in metastasized cases in the top three suggest that gene expression levels between case and control diverge as we get closer to the time of diagnosis. The common y-axis is log_2(fold change). The common x-axis is followup in days; the x-axis ticks are located at samples with metastasis. The top row shows the top three genes ranked by this model. The bottom row shows three randomly chosen genes for comparison.

footing with bootstrapping in terms of computation. It also has comparatively low bias and variance in the 2009 study of Kim [15]. The process is simply to do regular cross-validation, compute the average error statistic, $\overline{\text{err}}$, and repeat as many times as feasible to get a set of error estimates $\{\overline{\text{err}}_i\}$. We can use these to construct quantile intervals in the same way that we would have with the bootstrap. We do 1500 cross-validations for each experiment.

We do any parameter tuning by cross-validation nested in the repeated procedure. This is not repeated, but simply done once per model fitting.

2.5 Metrics: AUC and stability

We're doing two-class prediction: metastasis vs. no metastasis. We measure the predictive performance of our models with the area under the receiver operating characteristic curve (AUC) [13]. AUC measures the probability that

two randomly chosen samples are ranked correctly, ie a positive sample has higher predicted probability of being positive than a negative sample. It is an equivalent statistic to the Mann-Whitney-Wilcoxon U [12]. Hence a simpler interpretation of AUC is that it's the probability of ranking a randomly chosen metastatic sample higher than a randomly chosen non-metastatic sample.

All the models we evaluate do some sort of feature selection to find the set of genes that best predict the outcome. The question is whether the predictors selected by each model change substantially on different data sets. We measure gene set stability as Haury et al. [14]. We use the Jaccard index, $\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$ to measure stability.

In cross-validation, the degree of overlap of observations in two of k folds is $\frac{|k-2|}{|k-1|}$. We use $k = 5$, which leads to 0.75 overlap. To get as many stability measures AUC measures, ie one statistic per fold, we calculate stability between fold one and fold two, fold two and fold three, and so on, wrapping around when we come to the kth fold.

3 Results

3.1 Danger of missing stratum

In figure 2 we see that fitting the models without regard for stratification can lead to worse-than-random predictive performance. Stratifying ameliorates this, and predictions from the stratified elastic net stability selection look promising.

We see that detection method is an important factor in predicting node status, or at the very least calibrating predictions so that they aren't outright wrong. It makes sense for the stratification to be important. The cancers are likely to have different character in different strata. You can expect the clinical cancers to be older, as they are large enough that the women suspected something on their own. Hence they have had a lot of time to metastasize. The screening and interval cancers haven't had much time to grow. The interval cancers are likely to be more aggressive as they were not detectable at the last screening, which was at most two years ago.

The AUC $< .5$ problem looks a lot like a Simpson's paradox [21], we suspect that there is some contradictory information between strata. A toy example of

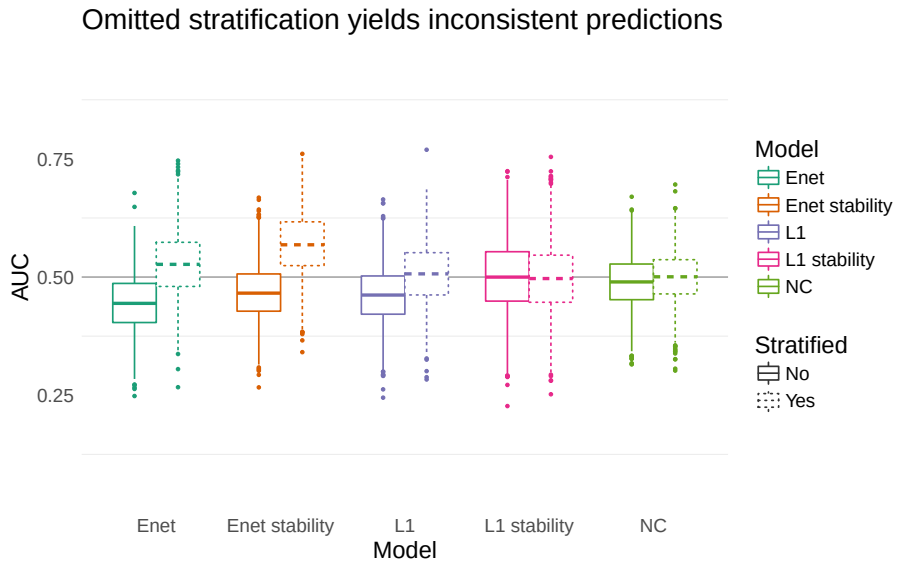

Figure 2: The effect of leaving out an important stratification variable across models. The strange behavior of getting AUC mostly $< .5$ (random guess) vanishes when we take stratum into account.

such an effect:

- Let $x_i = \mu_i + e_i$, where e_i is iid, mean-zero noise.
- Draw an outcome y_i and a stratum, s_i, both $\in \{0, 1\}$
- Let $\mu_i = 1$ if $s_i = y_i$, 0 otherwise

In this example, ignoring strata there is basically no information. Whether the outcome is 0 or 1 the predictor is distributed as a mixture two normals with modes at 0 and 1. Taking strata into account, you have in stratum 1: $E[X|y = 0] < E[X|y = 1]$ In stratum 0, the opposite: $E[X|y = 0] > E[X|y = 1]$ If the proportions of stratum 0 and stratum 1 in training and test data are sufficiently different and the stratum variable is missing, the estimated effect is opposite of what's happening in test data.

3.2 Stratification vs. curve filtering

Accounting for stratification fixes the AUC $< .5$ problem, and for the stratified elastic net model yields better-than-random predictions. But it does require

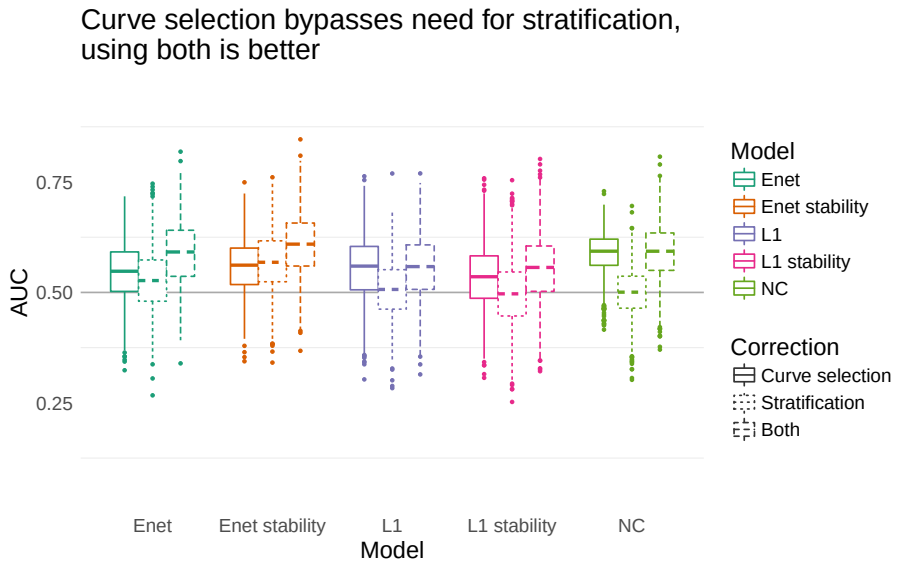

Figure 3: Effect of curve filtering and stratification. We see that preselecting genes by curve filtering also fixes the AUC $< .5$ issue, bypassing the need to consider the detection method stratification. Using both preselection and stratification is better for each model, but the nearest centroids model with preselection only does as well as the best models and with less variance.

us to actually know the detection method of a cancer at the time of modeling. Such a model would not work in for eg a screening setting.

In figure 3 we see that the use of curve selection avoids the need for explicitly modeling detection method. This suggests that the followup variable contains some compensating information and that metastatic cancers behave differently to non-metastatic ones over time. This is not something that simple gene-wise t-tests pick up, as we can see by comparing the performance of nearest centroids between figure 2 and figure 3. To confirm that it's not simple differential expression that's picked up by curve selection, we have investigated how often a gene gets selected based on the time interaction. If it's always the constant terms that contribute to selection, t-tests are good enough. By choosing gene sets of size 50 for 1500 bootstrap samples of our data, we get a distribution over selection frequency for the two candidate coefficients in the non-stratified model:

Curve selection improves predictions for all models, which is not the case

	First quartile	Median	Third Quartile
Spread	0.08	0.16	0.32
Spread × Time	0.68	0.84	0.92

Table 2: The frequency at which spread or spread×time is the term contributing to selection

for stratification alone. Using both together is in most cases better. But the nearest centroids with curve selection alone does as well as the best combined model, and does so with less variance.

3.3 Stability

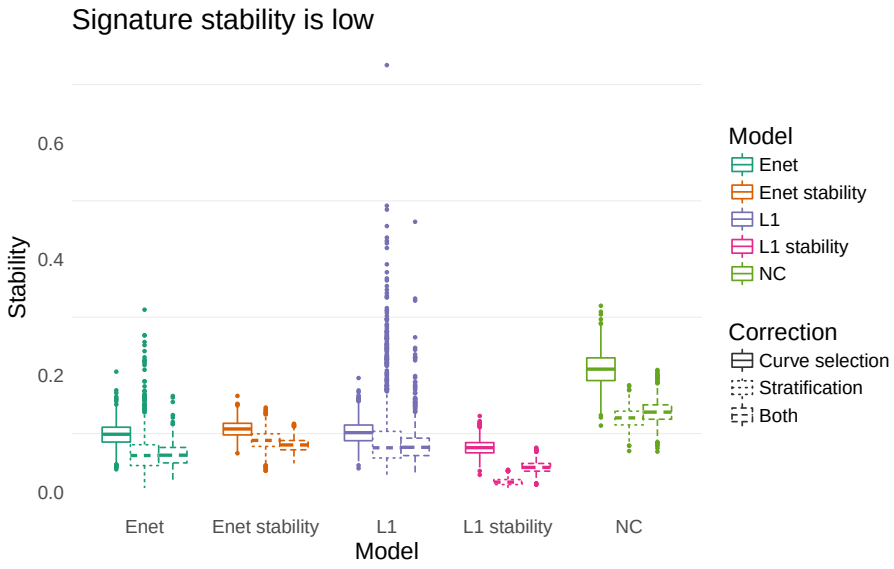

Figure 4: Signature stability. It's low across the board, with the nearest centroids models coming out clearly ahead. Stability selection doesn't improve stability much in these data.

Figure 4 shows that the selected gene sets are quite unstable. In the best case a 75% overlap in data yields a stability of about .2. This means that when we pick a 50 gene signature for the centroids model twice on mostly overlapping data, we can expect an overlap of ten genes between the two signatures. Interestingly stability selection doesn't seem to improve neither predictions nor

stability for the lasso penalized models in our data.

4 Conclusion

Curve selection is biologically motivated. We see that using the biology to select likely predictor genes and then fitting a very simple predictive model can outperform very clever, mathematically motivated models.

By doing curve selection we improve predictions and obtain more stable predictive signatures. As it is the dataset is quite small, so there remains a question of statistical power. There also seems to be very low signal to noise, something that is probably made worse by the fact that we don't have repeated measurements for any of the women.

However, there is some promise to these data. Further work is needed, but it does look as there is some predictive signal of breast cancer metastasis in prospective blood samples. It is a small step toward liquid biopsies for lymph node metastasis.

References

[1] Myron G. Best, Nik Sol, Irsan Kooi, Jihane Tannous, Bart A. Westerman, François Rustenburg, Pepijn Schellen, Heleen Verschueren, Edward Post, Jan Koster, Bauke Ylstra, Najim Ameziane, Josephine Dorsman, Egbert F. Smit, Henk M. Verheul, David P. Noske, Jaap C. Reijneveld, R. Jonas A. Nilsson, Bakhos A. Tannous, Pieter Wesseling, and Thomas Wurdinger. Rna-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell*, 28(5):666–676, 2017/05/22.

[2] Matthieu Bouaziz, Christophe Ambroise, and Michael Guedj. Accounting for population stratification in practice: A comparison of the main strategies dedicated to genome-wide association studies. *Plos One*, 6(12):e28845, 2011.

[3] Kelly Rae Chi. The tumour trail left in blood. *Nature*, 532:269 – 271, 2016.

[4] Vanessa Dumeaux, Anne-Lise Børresen-Dale, Jan-Ole Frantzen, Merethe Kumle, Vessela N. Kristensen, and Eiliv Lund. Gene expression analyses in breast cancer epidemiology: the norwegian women and cancer postgenome cohort study. *Breast Cancer Research*, 10(1):R13, 2008.

[5] B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26, jan 1979.

[6] Liat Ein-Dor, Itai Kela, Gad Getz, David Givol, and Eytan Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21(2):171–178, 2005.

[7] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high-dimensional feature space. *Statistica Sinica*, 20:101 – 148, 2010.

[8] A. Frigessi, P. Bühlmann, I.K. Glad, M. Langaas, S. Richardson, and M. (Eds.) Vannucci, editors. *Statistical Analysis for High-Dimensional Data*. Springer International Publishing, 2016.

[9] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Stat.*, 2(4):1360–1383, 12 2008.

[10] Mario Giuliano, Antonio Giordano, Summer Jackson, Ugo De Giorgi, Michal Mego, Evan N. Cohen, Hui Gao, Simone Anfossi, Beverly C. Handy, Naoto T. Ueno, Ricardo H. Alvarez, Sabino De Placido, Vicente Valero, Gabriel N. Hortobagyi, James M. Reuben, and Massimo Cristofanilli. Circulating tumor cells as early predictors of metastatic spread in breast cancer patients with limited metastatic dissemination. *Breast Cancer Research*, 16(5):440, 2014.

[11] Clara-Cecilie Günther, Marit Holden, and Lars Holden. Preprocessing of gene-expression data related to breast cancer diagnosis. https://www.nr.no/files/samba/smbi/note2015SAMBA3514preprocessing.pdf, 2014.

[12] J A Hanley and B J McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, apr 1982.

[13] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, New York, NY, 2009.

[14] Anne-Claire Haury, Pierre Gestraud, and Jean-Philippe Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *Plos One*, 6(12):e28210, 2011.

[15] Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745, 2009.

[16] Bora Lim and Gabriel N. Hortobagyi. Current challenges of metastatic breast cancer. *Cancer and Metastasis Reviews*, pages 1–20, 2016.

[17] Benoît Liquet, Pierre Lafaye de Micheaux, Boris P. Hejblum, and Rodolphe Thiébaut. Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*, 2015.

[18] Eiliv Lund, Lars Holden, Hege Bøvelstad, Sandra Plancade, Nicolle Mode, Clara-Cecilie Günther, Gregory Nuel, Jean-Christophe Thalabard, and Marit Holden. A new statistical method for curve group analysis of longitudinal gene expression data illustrated for breast cancer in the nowac postgenome cohort as a proof of principle. *BMC Medical Research Methodology*, 16(1):28, 2016.

[19] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[20] Annette M Molinaro, Richard Simon, and Ruth M Pfeiffer. Prediction error estimation: a comparison of resampling methods. *Bioinformatics (Oxford, England)*, 21(15):3301–7, aug 2005.

[21] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 238–241, 1951.

[22] Eleni van Schooneveld, Maartje CA Wouters, Ilse Van der Auwera, Dieter J. Peeters, Hans Wildiers, Peter A. Van Dam, Ignace Vergote, Peter B. Vermeulen, Luc Y. Dirix, and Steven J. Van Laere. Expression profiling of cancerous and normal breast tissues identifies micrornas that are differentially expressed in serum from patients with (metastatic) breast cancer and healthy volunteers. *Breast Cancer Research*, 14(1):R34, 2012.

[23] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.