Bias-corrected Hierarchical Bayesian Classification with a Subset of Selected Features

Longhai Li

longhai@math.usask.ca

Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, Saskatchewan, S7N 5E6 Canada

Presented on ICSA Canada Chapter Annual Meeting,
University of Calgary, 6 August, 2015

What is Feature Selection Bias?

- **High-throughput Data**

  Today, many biotechnologies, for example Microarrays, can gather high-dimensional profiles of a huge number (eg, hundreds of thousands) features with pretty low costs.

- **Classification**

  We are interested in building a classification mechanism for predicting a categorical response, for example disease/normal, types of tumors, from these high-dimensional profiles. The classification mechanism may be used for practical diagnosis of disease, or for evaluating the predictive goodness of the features (eg genes) under investigation.

- **Classification after Feature Selection**

  We use a certain univariate screening method (such as, t test) to select a small number (such as 10) of top features out of a very large number (such as 5000) of candidate features. Then we build a classification mechanism with only the top features.

What is Feature Selection Bias?

- **Feature Selection Bias**

  The classification mechanism built by treating the selected subset of features as ordinary features measured without feature selection will give over-confident (too extreme) prediction probabilities for future cases. For example, we predict that a set of test cases have class label equal to 1 with a probability between 0.9 to 1, but actually the fraction of class label 1 for them in the whole population is only 0.7. This is called the feature selection bias. The bias arises because many "falsely relevant" features are considered in making the prediction.

  An extreme example: All features are irrelevant, but the top selected features would appear very predictive to the response, which is made wholly by the feature selection, not by the biological signals.

- **Why do we care about the feature selection bias in classification?**
  - We need more accurate prediction tools in practice, for example, in personalized drug recommendation.
  - A better guidances for determining the number of features that should be retained for further more expensive investigation.

Methods for Correcting for Feature Selection Bias

- **Predictive analysis with microarrays (PAM), Tibshirani et al., 2002)**

  Correcting for feature selection bias in classification problems seemingly has not received much attention. One solution is PAM:

  PAM corrects for this bias by imposing *stronger* shrinkage for (ie., cut more) the signals of retained features when the number of retained features is *smaller*.

  The PAM cut corrects for the feature selection bias, but also hurts the signals of really relevant features. For example, the top feature selected by the t test is likely a real signal, but PAM punishes it. Therefore, the predictive probabilities returned by PAM become over-conservative.

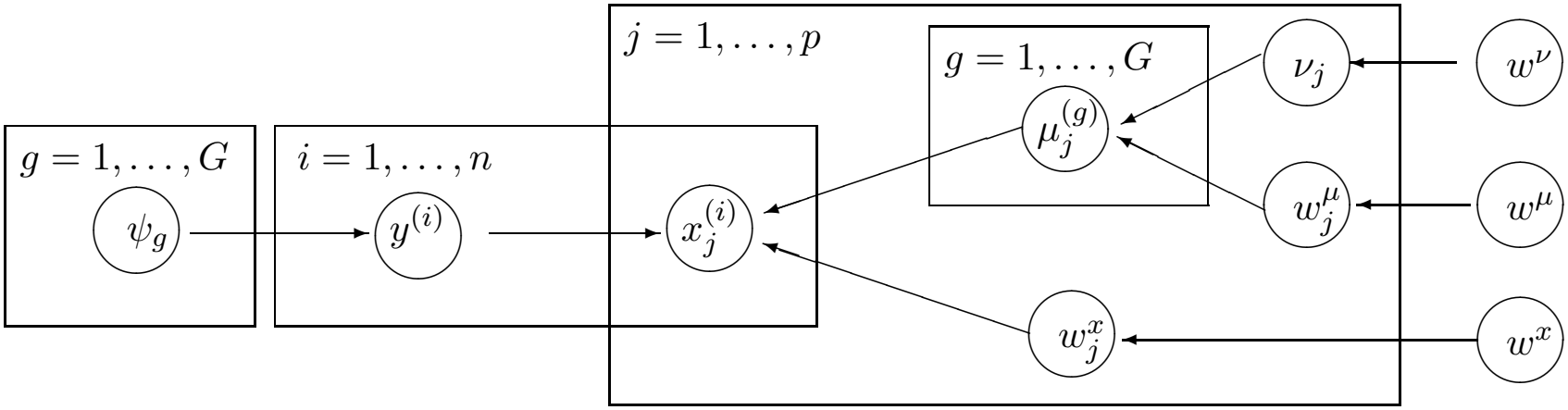
- **Our Bayesian solution**

  Adjusting posterior distribution of hyperparameters that control the overall signal-to-noise ratio of the whole data set.

- **How is our solution different from PAM?**

  Cut less the top large signals (which are likely real) while shrinking the small signals (which are likely made by the feature selection, ie., *by the chance*).

A Bayesian Model for High-dimensional Data



- `parameters`: $\mu_j^{(g)}$ and w_j^x are the mean and variance of of $x_j^{(i)}$ *within* class g. ψ_g is the prior label probability of class g.
- `hyperparameters`: ν_j and w_j^μ are the mean and the variance of signals $\mu_j^{(1:G)}$ across G classes.
- `hyperparameters`: w^μ is the scale of inverse-χ^2 prior for $w_1^\mu, \ldots, w_p^\mu$, representing the overall signal level, and w^x is the scale of inverse-χ^2 prior for $w_1^x, \ldots, w_p^x$, representing overall noise level.
- Overall signal-to-noise ratio: $\dfrac{w^\mu}{w^x}$.

Correction for Feature Selection Bias

When p is very large, for pragmatic reasons, we will select a small subset of features, $\boldsymbol{x}_{1:k}^{(1:n)}$, by some univariate score $R(\boldsymbol{x}_j^{(1:n)}, \boldsymbol{y}^{(1:n)})$. The posterior of w^μ/w^x given only retained features will be upwardly biased.

To correct for the bias, we should condition on all available information to form our posterior of parameters and hyperparameters, in particular, of w^μ and w^x. All the available information is:

$$\boldsymbol{y}^{(1:n)}, \boldsymbol{x}_{1:k}^{(1:n)}, \text{and}$$
$$\boldsymbol{x}_j^{(1:n)} \in \mathcal{S} = \{\boldsymbol{x}^{(1:n)} \mid R(\boldsymbol{x}^{(1:n)}, \boldsymbol{y}^{(1:n)}) \leq \gamma\}, \text{ for } j = k+1, \ldots, p$$

where where γ is the score value of the last retained feature x_k (or a threshold that is actually used in determining k).

These set statements for omitted features contain information about the overall signal-noise ratio $\dfrac{w^\mu}{w^x}$: a likelihood of $\dfrac{w^\mu}{w^x}$ based on them favors small values, and therefore can be used to correct for the feature selection bias.

Bias-corrected Posterior

We will base our posterior distribution on the following joint distribution:

$$\prod_{j=1}^{k} P(\boldsymbol{x}_j^{(1:n)} | \boldsymbol{\mu}_j^{(1:G)}, w_j^x, \boldsymbol{y}^{(1:n)}) \times$$

$$\prod_{j=1}^{k} \left[P(\boldsymbol{\mu}_j^{(1:G)} | \nu_j, w_j^\mu) P(\nu_j | w^\nu) P(w_j^\mu | w^\mu) P(w_j^x | w^x) \right] \times$$

$$P(w^\mu)\, P(w^\nu)\, P(w^x) \times C(w^\mu, w^x)^{p-k},$$

where $C(w^\mu, w^x)$ is the correction factor:

$$C(w^\mu, w^x) = P(\boldsymbol{x}_j^{(1:n)} \in \mathcal{S} | w^\mu, w^x, \boldsymbol{y}^{(1:n)}).$$

Note that $C(w^\mu, w^x)$ is the same for all $j = k+1, \ldots, p$. We need to approximate this value only once no matter how many features are omitted. Particularly, we have found a fast Monte Carlo method when F-statistic is used to select features, based on knowledges on non-central F distribution.

Application to Real Lymphoma Microarray Data

Lymphoma data set contains expression levels of $p = 4026$ genes from $n = 62$ patients with most prevalent adult lymphoid malignancies:

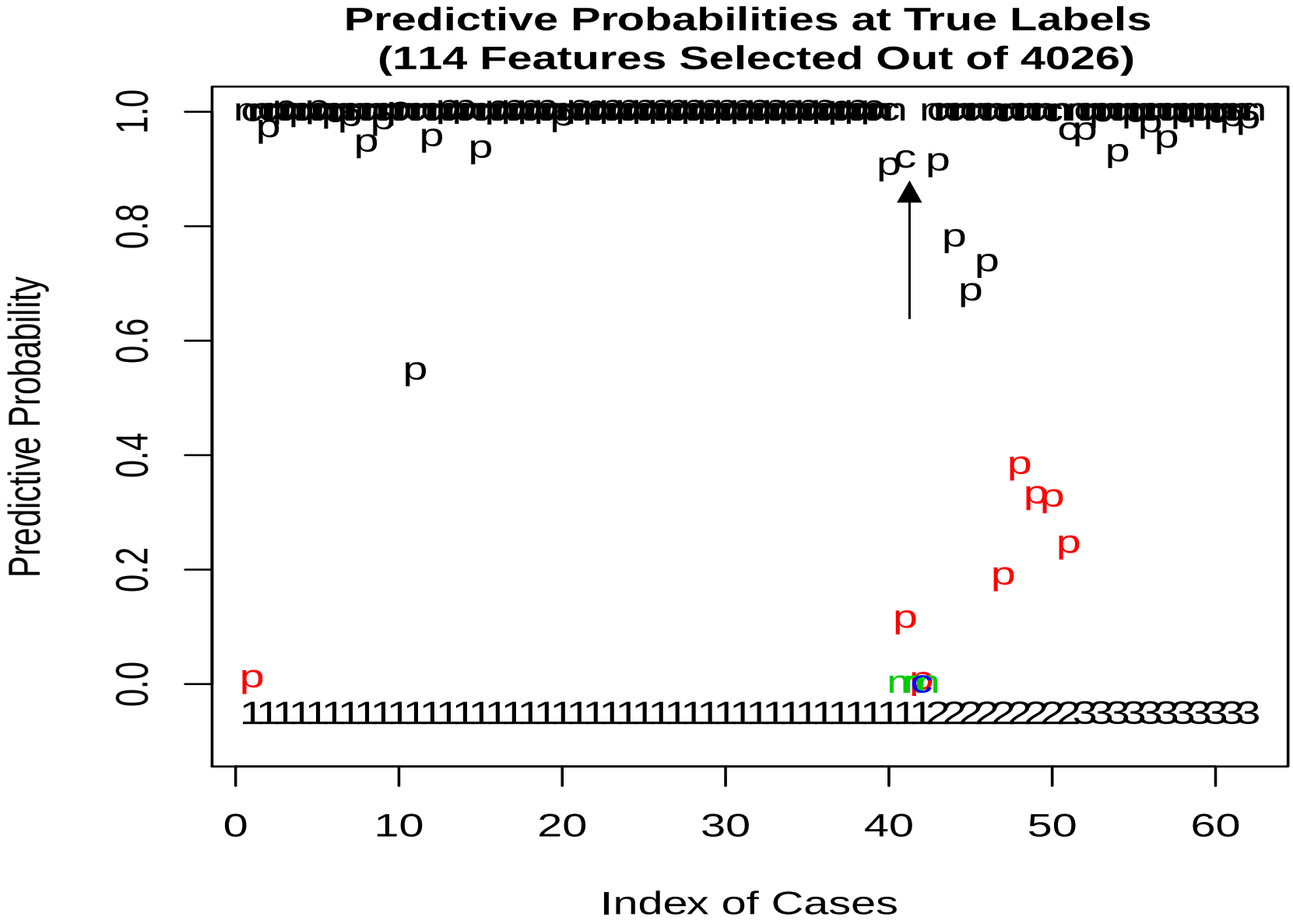
- 42 cases of diffuse large B-cell lymphoma (coded by 1)
- 9 cases of follicular lymphoma (coded by 2)
- 11 cases of chronic lymphocytic leukemia (coded by 3)

The data set was originally published by Alizadeh et al. (2000). I used a data set pre-processed by Dettling (2004).

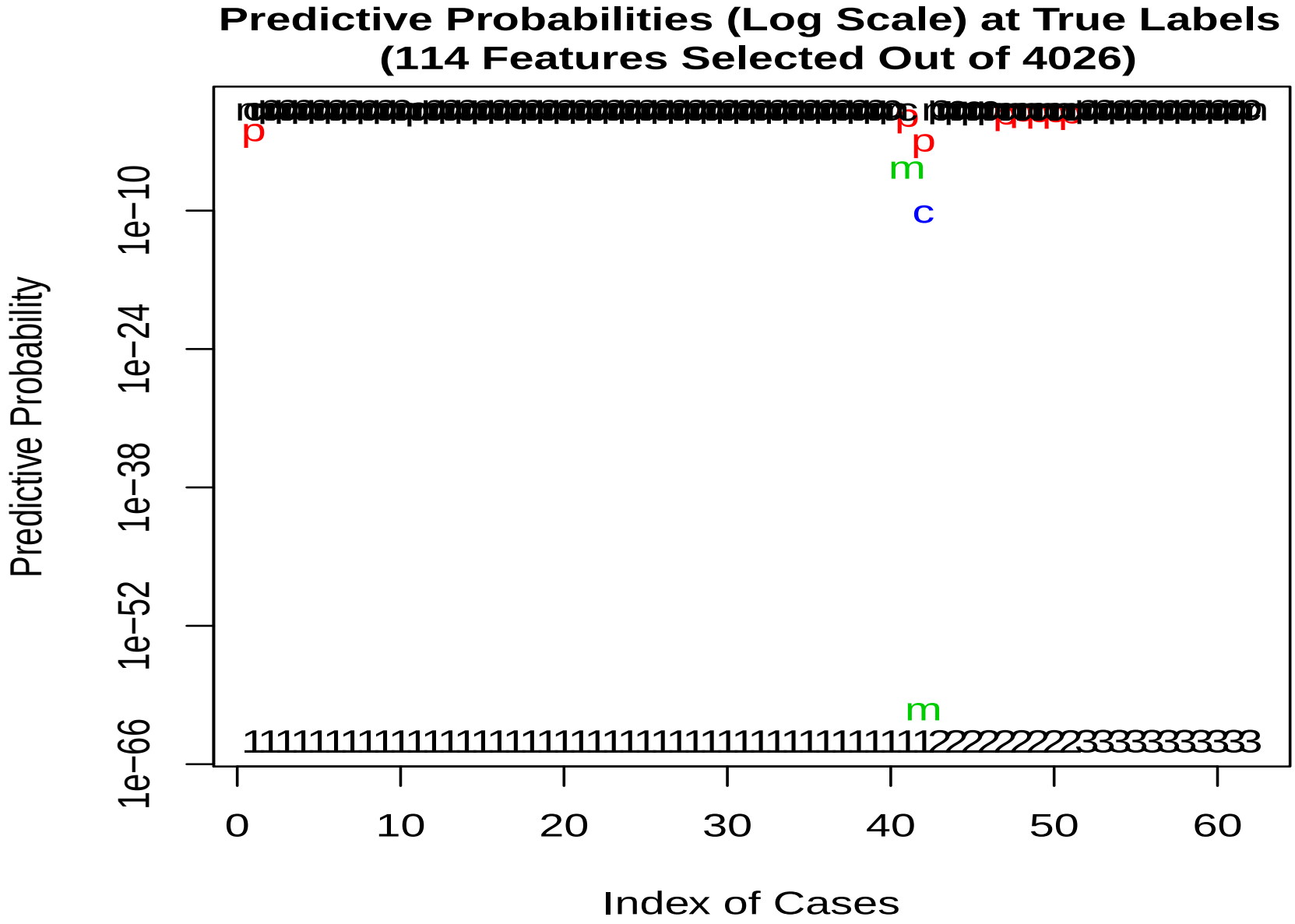
I used 10-fold cross-validation to compare three methods:

- m — DLDA (MLE) by Dudoit, Fridlyand, and Speed (2002), without correction for feature selection bias
- p — PAM by Tibshirani et al. (2002)
- c — BCBCSF, the method introduced here.

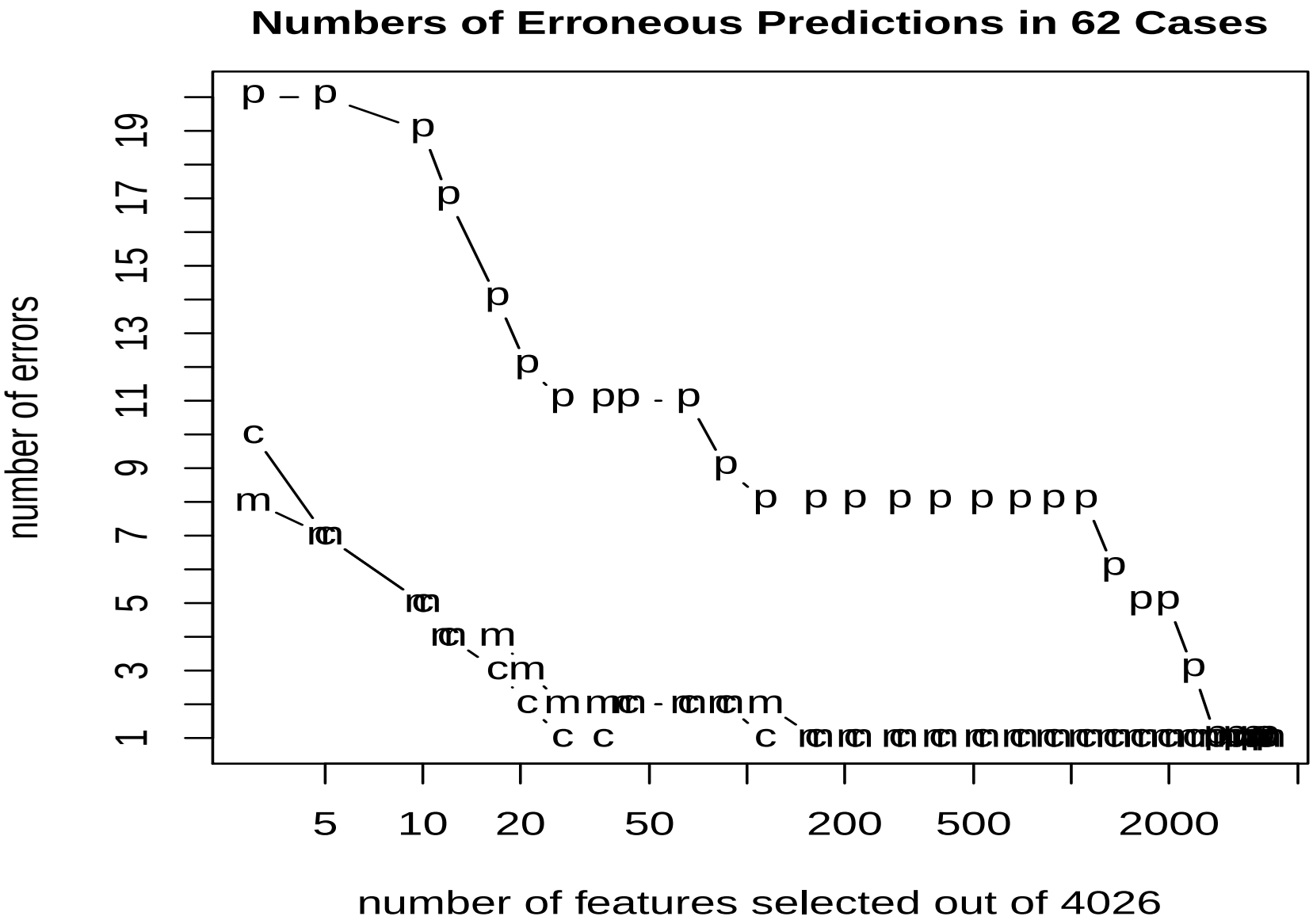
Predictive Probabilities with 114 Features Selected



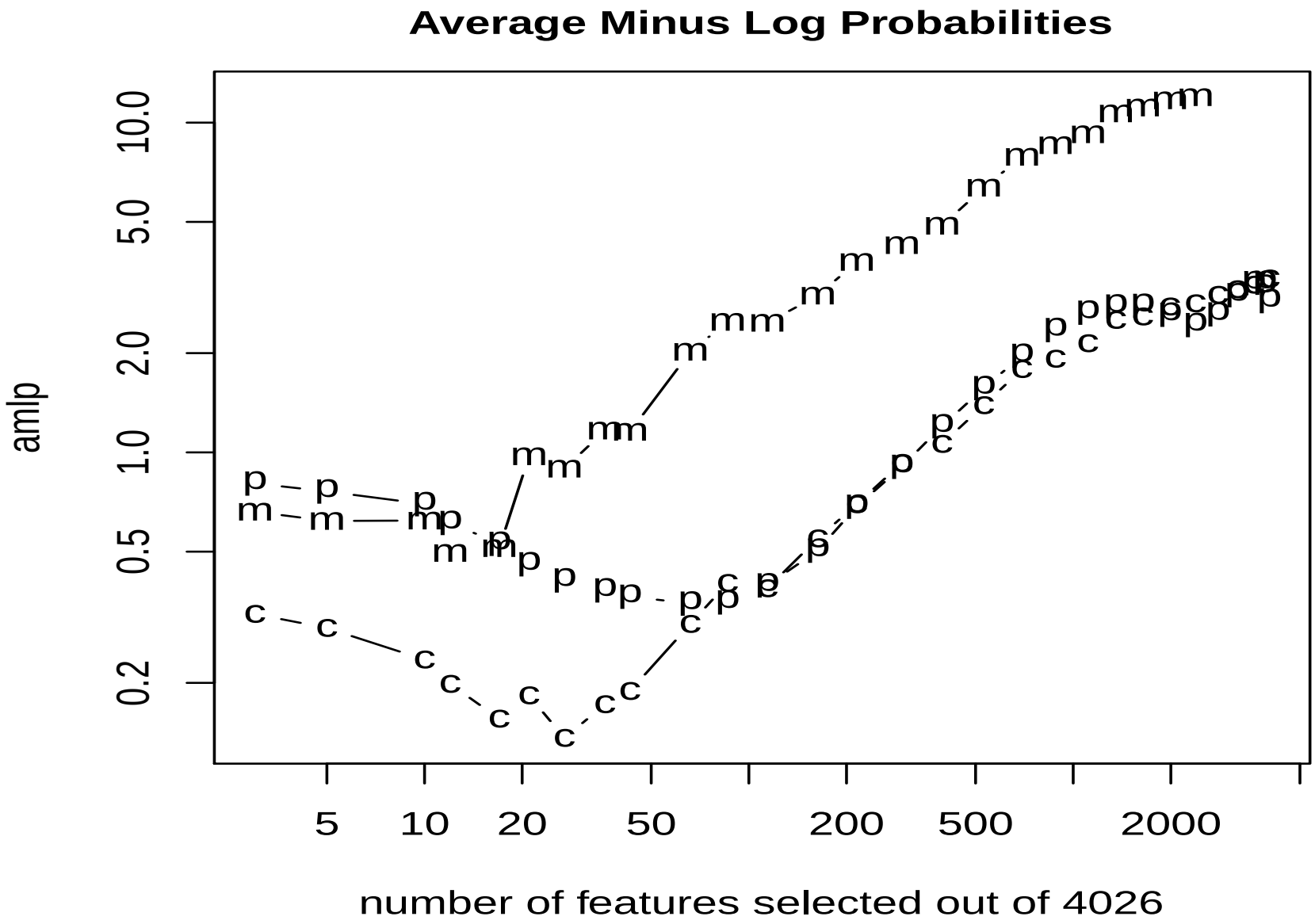
Log Predictive Probabilities with 114 Features Selected



Comparing Classification Error Rates



Comparing Average Minus Log Probabilities (AMLP)



A Simulation Study

Data Generation

Using the following fixed top level hyperparameters and degrees of freedom for IG distributions:

$$\alpha_1^{\mu} = 3,\ w^{\mu} = 0.003,\ w^{x} = 1,\ \alpha_1^{x} = 10,\ w^{\nu} = 1,$$

we generated a data set of $n = 2100$ cases that are evenly distributed in $G = 4$ classes, with $p = 5000$ features.

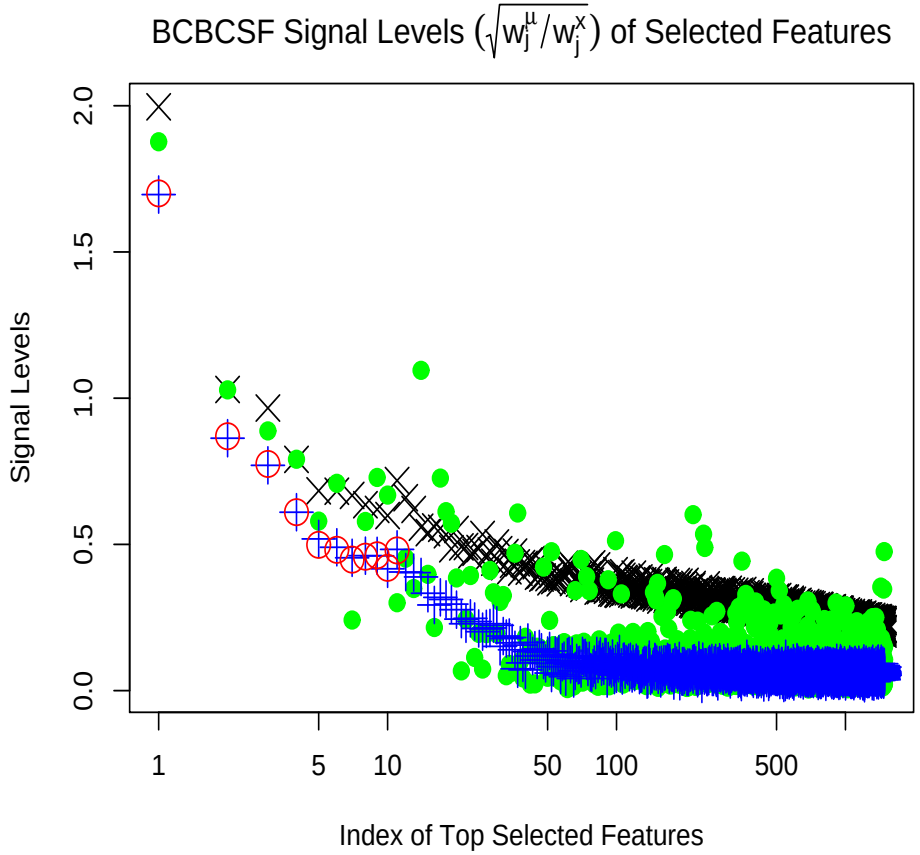
- 100 cases are used as training set, and
- 2000 cases are used as test cases.

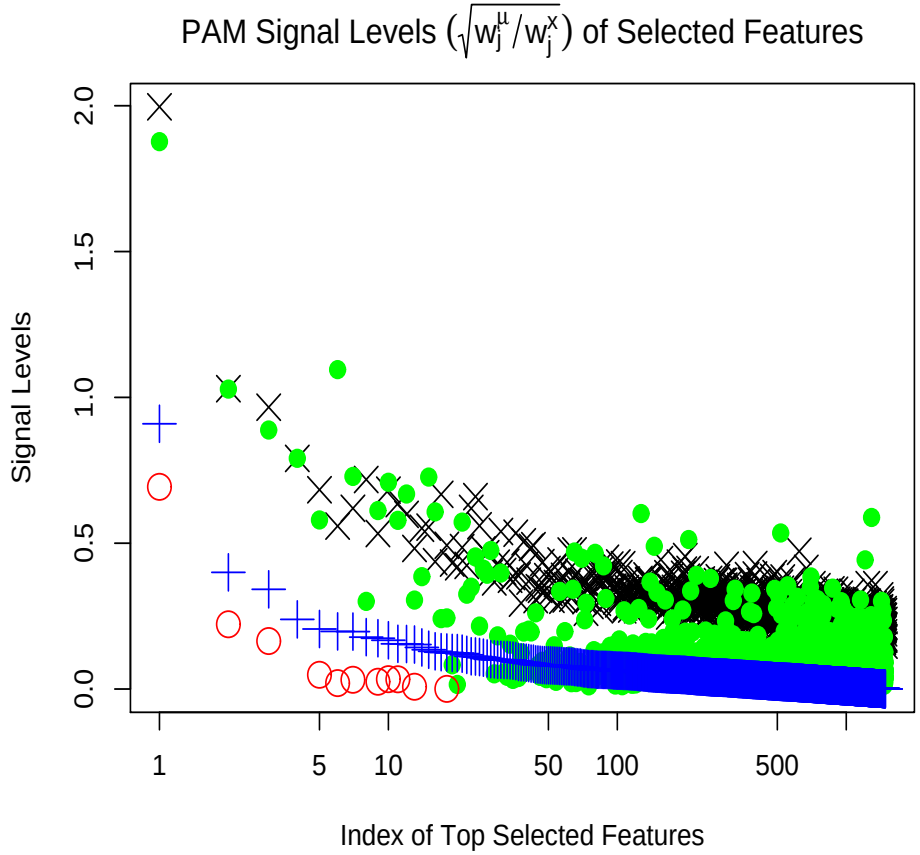
A Simulation Study

●: True signals based on 2000 test cases.

×: MLE estimates based on 100 training cases.

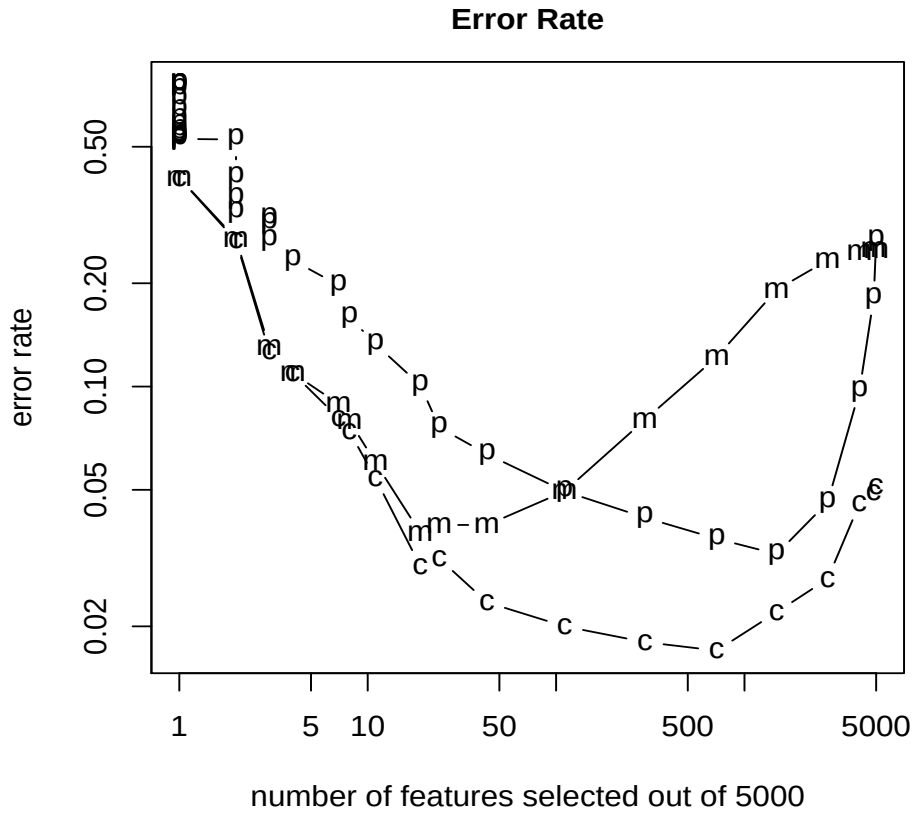
○ and +: Two signals estimates based on 11 and 1482 top features.

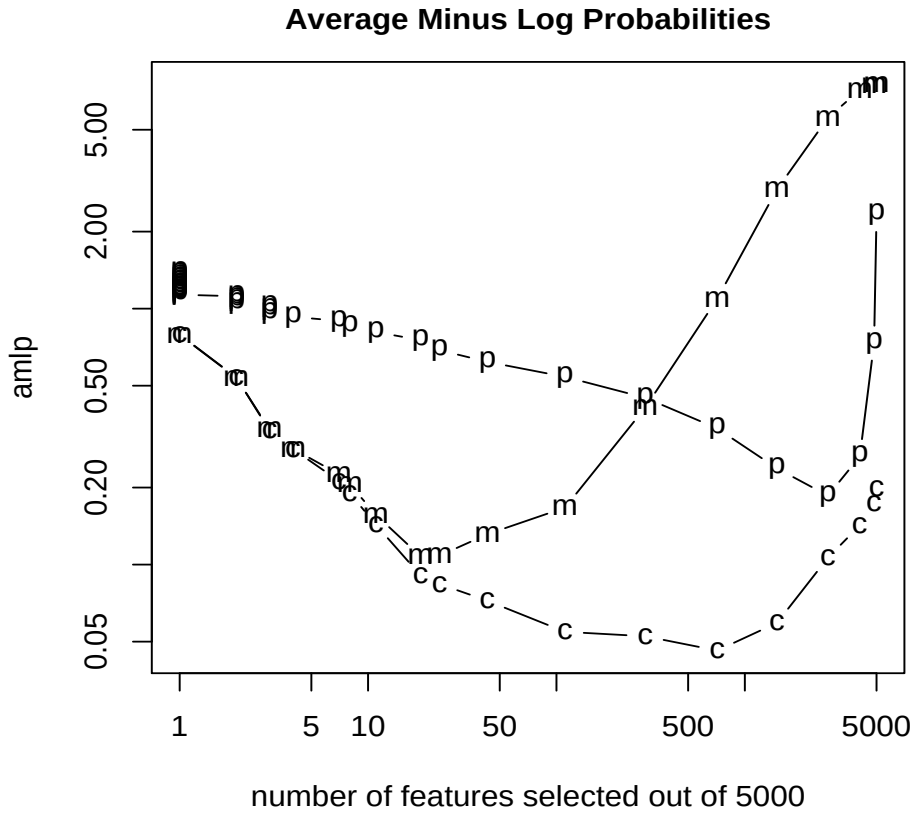




A Simulation Study

Comparisons of predictive probabilities produced by three classification methods ("p" — PAM, "m" — DLDA, "c" — BCBCSF) in terms of error rate and AMLP (average of minus log probabilities at the actually observed class labels).





Conclusions and Future Work

- DLDA (MLE) without correction for feature selection bias is over-confident. PAM is over-conservative. BCBCSF is in the middle. It can correct for feature selection bias, but doesn't over-shrink strong signals.
- In the future, we could extend BCBCSF to other more complicated models, more complicated feature selection schemes, and more inferences problems (such as interval estimates and hypothesis testing).
- How to correct for feature selection bias in classification and other inference problems when the features are selected by fitting a linear model, such as many variants of LASSO, or other more complicated selection procedures?

  Recently, many researchers, including R. Tibshirani (Stanford), and J. Taylor (Stanford), R. Lockhart (SFU), and may others, have proposed to do inference conditional on that the response values y are in a subset:

$$\{y|Ay < b\}$$

  They call this topic **selective inference** or **post-selection inference**. Read Tibshirani (2015) and the references therein.

References

If you are interested in this topic, you can start from reading:

- Li, Longhai (2012). "Bias-Corrected Hierarchical Bayesian Classification With a Selected Subset of High-Dimensional Features." *Journal of the American Statistical Association*, 107(497): 120-134. doi:10.1198/jasa.2011.ap10446.
- Tibshirani, Robert (2015). "Two novel applications of selective inference (talk slides)". Workshop on Statistical inference for large scale data (Simons Fraser University, Canada), April. 2015. url: `http://www-stat.stanford.edu/~tibs/ftp/lockharttalk.pdf`

Thank you for your attention.

Questions are welcomed!