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

Joint Statistical Meeting, Montreal
7 August 2013

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 usefulness of the features (eg genes) under investigation.

- **Feature Selection**: We often build a classification mechanism by using only a small subset of features selected by an univariate screening method, eg t-test.

- **Feature Selection Bias**: The classification mechanism built by treating the selected subset of features as ordinary features will give over-confident (too extreme) prediction probabilities for future cases.

Methods for Correcting for Feature Selection Bias

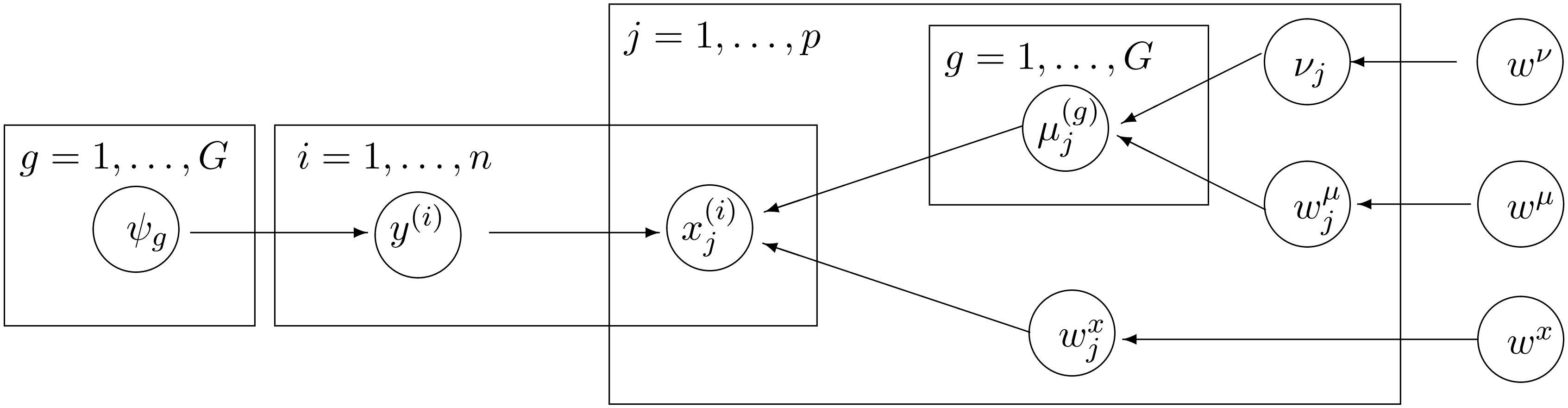
We are interested in finding a bias-corrected and better-calibrated classification mechanism built from a selected subset of features, because

- We need more accurate diagnosis tool in practice.
- Classification error estiamtes returned by a better-calibrated classification mechanism are better guildances for determining the number of features that should be retained for further more expensive investigation.

Correcting for feature selection bias seemingly hasn't received much attention. **Predictive Analysis with Microarrays** (Tibshirani et al., 2002) corrects for this bias by imposing stronger shrinkage for signals of retained features when the number of retained features is smaller.

In this talk, I present an alternative Bayesian method that attempts to correct for the bias by **adjusting posterior distribution of hyperparameters** that control signal-to-noise ratio of the whole data set.

A Bayesian Model for High-dimensional Data



- `data`: $y^{(i)}$ is the label for the ith case, $x_j^{(i)}$ is the value of the jth feature, for $j = 1, \ldots, p$.

- `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 $\mu_j^{(1:G)}$ 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.

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 — indeed a likelihood based on them favors small signal-noise ratio, and therefore can be used to correct for the lifted overall signal-noise ratio.

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 Lymphoma Microarray Data

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

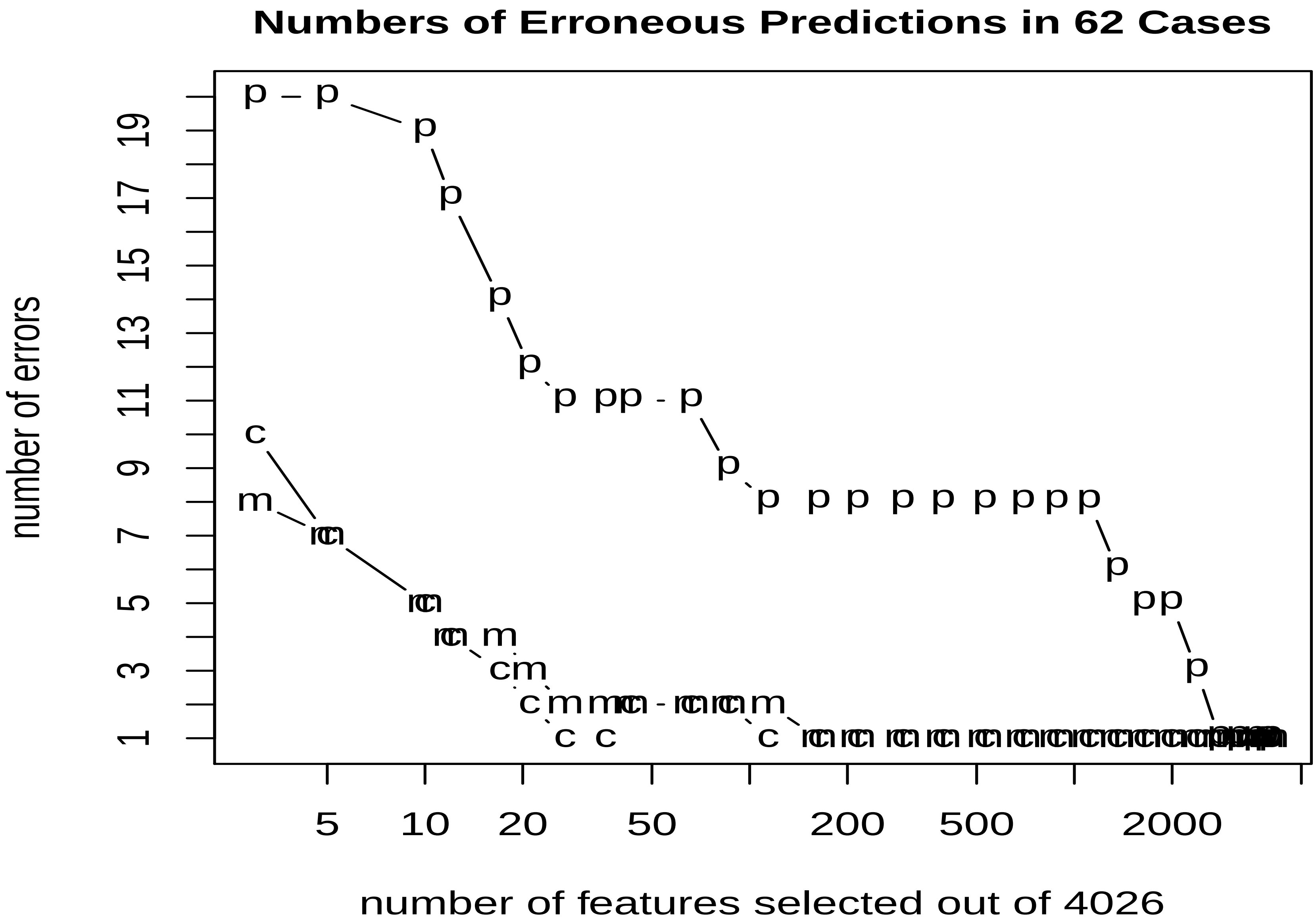
- 42 cases of diffuse large B-cell lymphoma (coded by 1)
- 9 cases of follicular lymphoma (coded by 2)
- 11 cases of chronic lymphocytic leukemia (coded by 3)

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

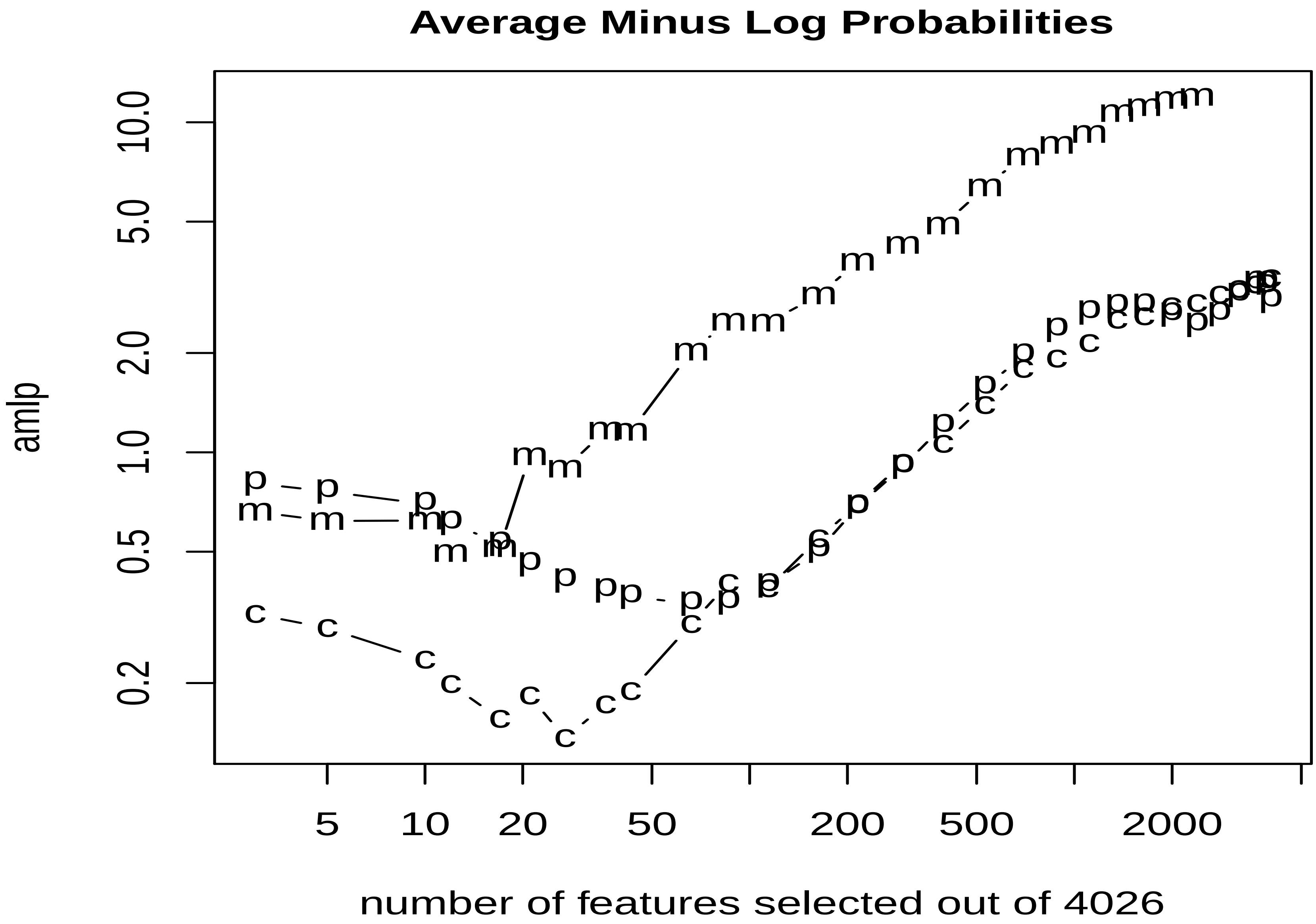
I used 10-fold cross-validation to compare three methods:

- 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.

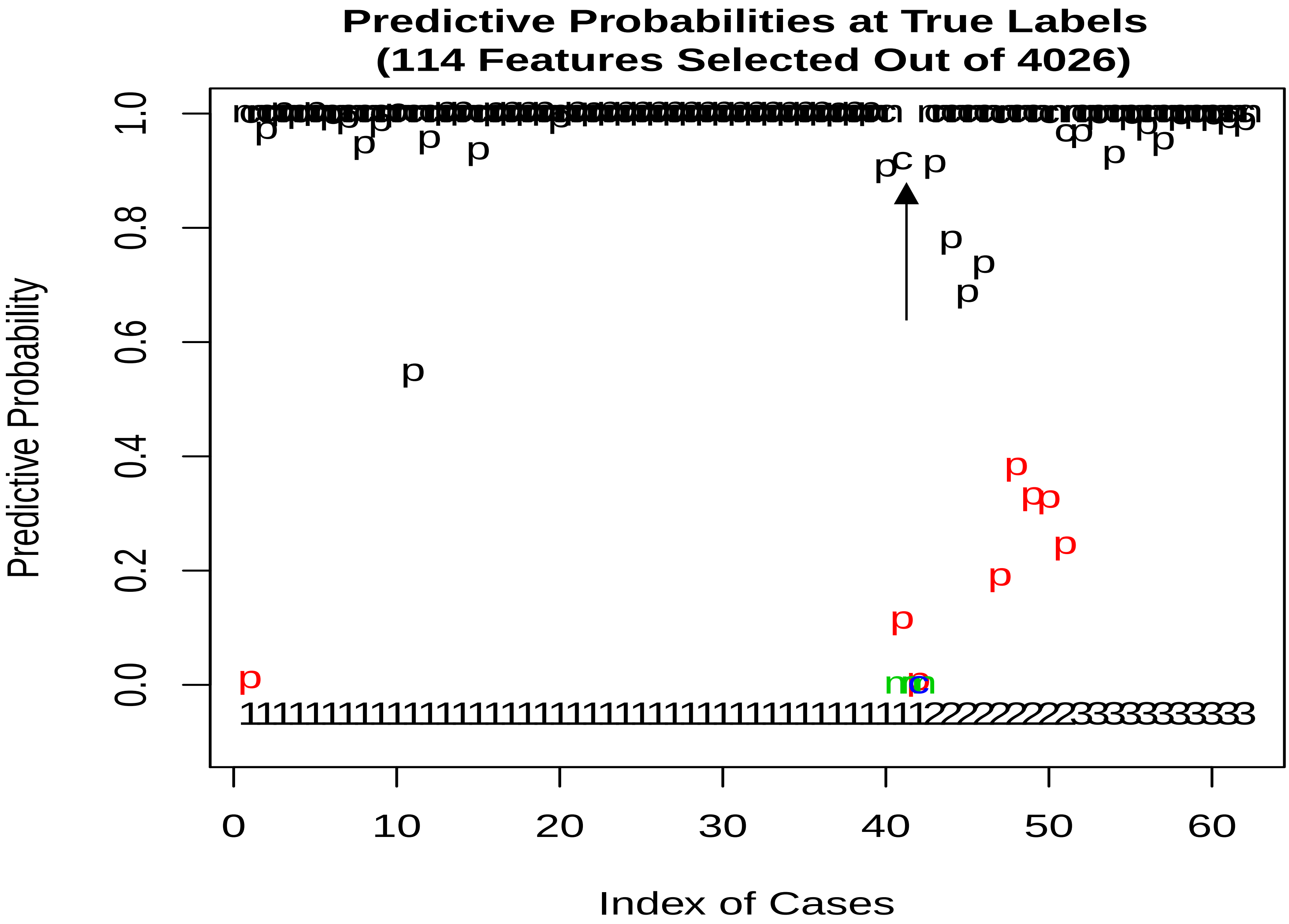
Comparing Classification Error Rates



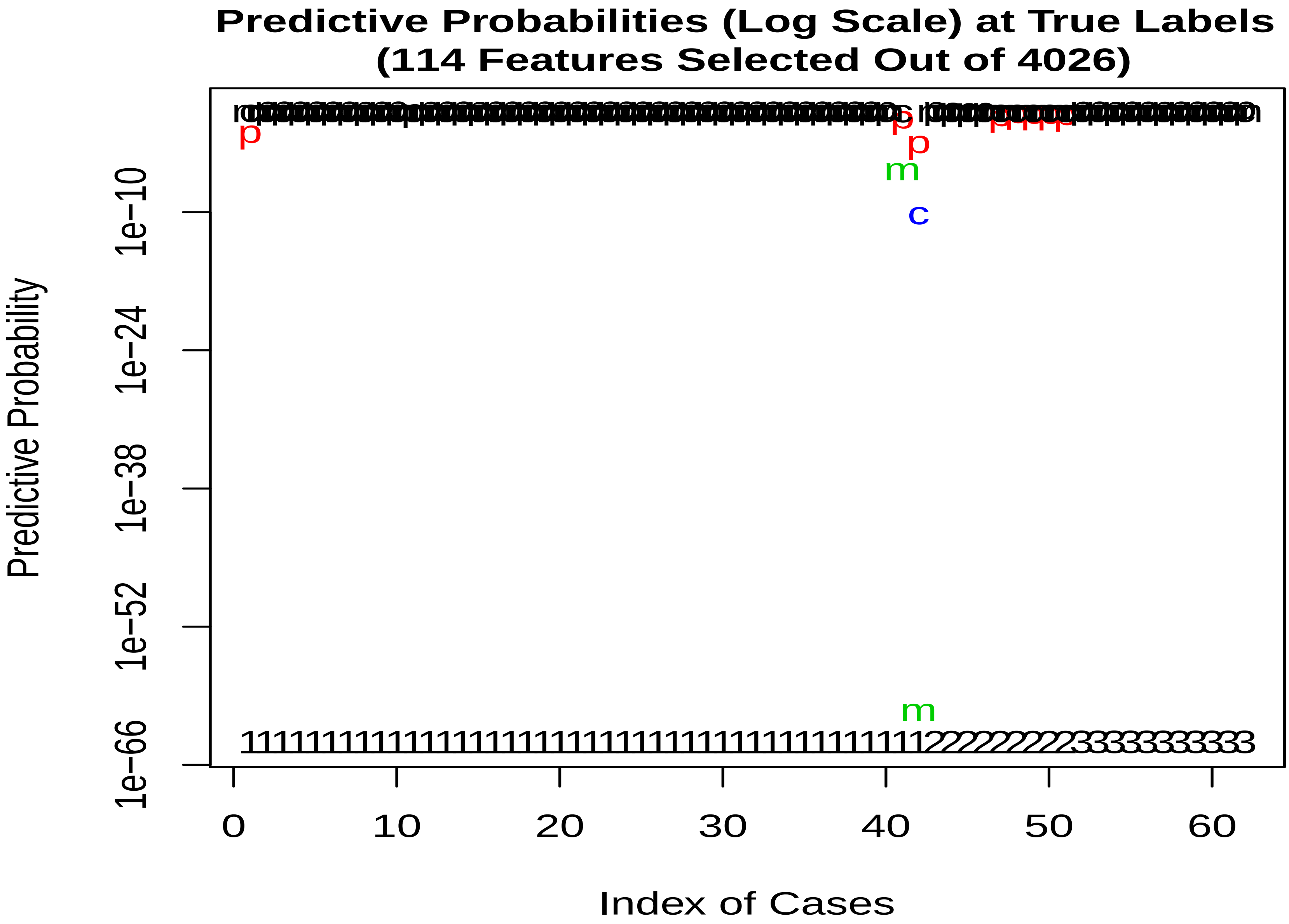
Comparing Average Minus Log Probabilities (AMLP)



Predictive Probabilities with 114 Features Selected



Log Predictive Probabilities with 114 Features Selected



Conclusions and Future Work

- DLDA without correction for feature selection bias is over-confident. It may give very low probability on the true class label and very high probability on the wrong class label. The overall measure of classification error in terms of AMLP is generally worse than BCBCSF and PAM.

- PAM is over-conservative. Its classification errors are generally very high when a very small number of features are retained, because the signals are over-shrunken. As a tool for determining how many features are to be retained for further investigation, it will result in a large feature subset.

- 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 models, and other univariate feature selection method.

Thank you for your attention.

Questions are welcomed!