

Definition of Valid Proteomic Biomarkers: A Bayesian Solution to a Currently Unmet Challenge

Keith Harris & Mark Girolami

Inference Group, Department of Computing Science, University of Glasgow

March 29th 2009

Sir Ronald Fisher - Likelihood Principle



Data X , Model parameters θ , Likelihood function $P(X|\theta)$, Maximum Likelihood $P(X|\hat{\theta})$, Predictions $P(X^*|\hat{\theta})$.

Note that predictions are made based on a point estimate for θ and that no allowance is made for our uncertainty about the value of θ .



Posterior distribution $P(\theta|X)$,

Posterior inference (posterior \propto prior \times likelihood) $P(\theta|X) \propto P(\theta)P(X|\theta)$,

Predictions $P(X^*|X) = \int P(X^*|\theta)P(\theta|X)d\theta$.

Note that the Bayesian approach allows us to take into account our uncertainty about θ when making predictions, unlike the classical approach.

A Bayesian model for biomarker selection

The fundamental problem of biomarker selection via CE-MS data is to identify which peptides best discriminate between different types of protein samples (e.g. male and female).

A Bayesian model for biomarker selection

The fundamental problem of biomarker selection via CE-MS data is to identify which peptides best discriminate between different types of protein samples (e.g. male and female).

CE-MS data contains a huge number of variables (peptides) and the sample size tends to be relatively small so the selection process can be unstable. Hence, models which incorporate sparsity in terms of variables are desirable for this kind of problem.

A Bayesian model for biomarker selection

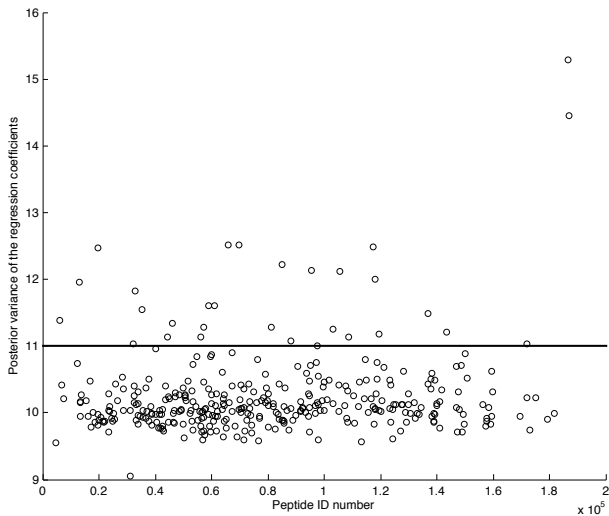
The fundamental problem of biomarker selection via CE-MS data is to identify which peptides best discriminate between different types of protein samples (e.g. male and female).

CE-MS data contains a huge number of variables (peptides) and the sample size tends to be relatively small so the selection process can be unstable. Hence, models which incorporate sparsity in terms of variables are desirable for this kind of problem.

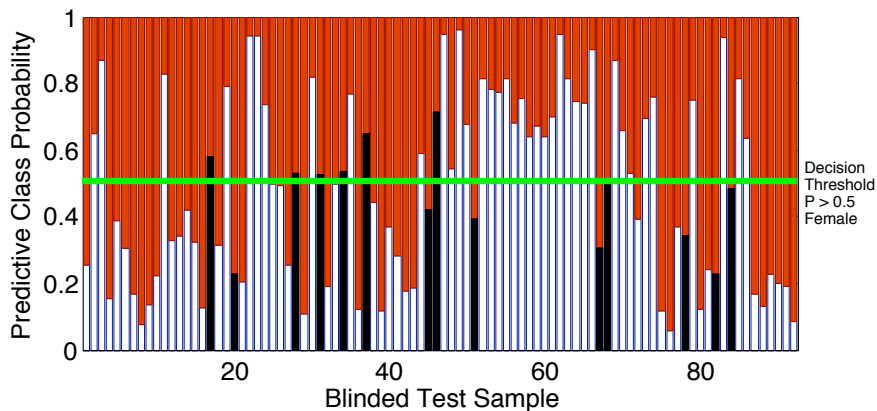
One such “sparse” model for binary classification was proposed by Bae and Mallick (2004).

Sparsity was incorporated by choosing a prior distribution that would shrink the regression coefficients of non-informative variables towards zero.

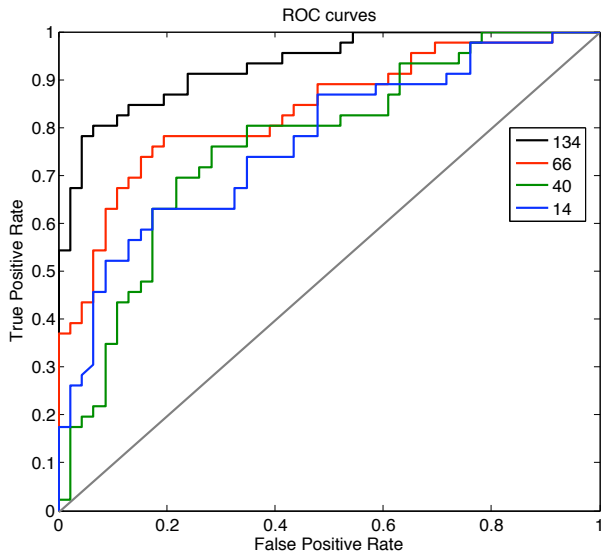
Plot of the posterior variance of β_i versus the peptide ID number



Plot of the posterior predictive probabilities



Classification performance using different size training sets



Performance on the test set for different training set sizes

Training set size	Test error
14	27.2%
40	27.2%
66	21.7%
134	15.2%

As we would expect, the confidence in our predictions also declines as the number of training samples decreases.

Indeed, when the number of training samples is only 14, almost all the predictive probabilities are between 0.3 and 0.7.

This suggests that the biomarkers selected by such a small data set would not be substantiated in practice.

1. Sparse models enable us to identify a small number of peptides having the greatest discriminating power, thereby allowing researchers to quickly focus on the most promising candidates for diagnostics and prognostics.

1. Sparse models enable us to identify a small number of peptides having the greatest discriminating power, thereby allowing researchers to quickly focus on the most promising candidates for diagnostics and prognostics.
2. The Bayesian approach yields a coherent way to assign new samples to particular classes. Rather than hard rules of assignment, we can evaluate the probability that the new sample will be of a certain type which is more helpful for decision making. Other non-statistical approaches (like SVMs) cannot provide such formal and well-calibrated probabilities.

1. Sparse models enable us to identify a small number of peptides having the greatest discriminating power, thereby allowing researchers to quickly focus on the most promising candidates for diagnostics and prognostics.
2. The Bayesian approach yields a coherent way to assign new samples to particular classes. Rather than hard rules of assignment, we can evaluate the probability that the new sample will be of a certain type which is more helpful for decision making. Other non-statistical approaches (like SVMs) cannot provide such formal and well-calibrated probabilities.
3. Meaningful results will only be obtained if the number of training samples collected is sufficient to allow the definition of statistically valid biomarkers.

Bae K. and Mallick B. K. (2004) Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20(18): 3423–3430.