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

Keith Harris[1], Mark Girolami[1] & Harald Mischak[2]

1. Inference Group, Department of Computing Science, University of Glasgow, UK
2. Mosaiques Diagnostics and Therapeutics AG, Hannover, Germany

April 1st 2009

University of Glasgow

Introduction

Proteins/peptides in body fluids hold information on the physiology of an organism and thus can serve as biomarkers for disease.

University of Glasgow

Introduction

Proteins/peptides in body fluids hold information on the physiology of an organism and thus can serve as biomarkers for disease.

However, clinical proteomics is suffering from high hopes generated by reports on potential biomarkers, most of which could not be later substantiated via validation.

This has resulted in much scepticism from clinicians and regulatory agencies, which will make the application of valid biomarkers even more challenging.

University of Glasgow

Introduction

Proteins/peptides in body fluids hold information on the physiology of an organism and thus can serve as biomarkers for disease.

However, clinical proteomics is suffering from high hopes generated by reports on potential biomarkers, most of which could not be later substantiated via validation.

This has resulted in much scepticism from clinicians and regulatory agencies, which will make the application of valid biomarkers even more challenging.

The cause of these erroneous biomarkers is often the inappropriate usage of basic statistics.

Good statistical practice thus needs to be highlighted and more sophisticated multivariate selection methods need to be developed, so that valid biomarkers will be defined with a much higher probability than currently observed.

University of Glasgow

Application

We chose as an illustrative example defining proteomic differences between healthy adult males and females.

University of Glasgow

Application

We chose as an illustrative example defining proteomic differences between healthy adult males and females.

We chose to collect urine samples, since urine has been found to be of much higher stability than blood-derived samples, hence reducing pre-analytical variability.

University of Glasgow

Application

We chose as an illustrative example defining proteomic differences between healthy adult males and females.

We chose to collect urine samples, since urine has been found to be of much higher stability than blood-derived samples, hence reducing pre-analytical variability.

Capillary electrophoresis-mass spectrometry (CE-MS) was used to analyse the urine samples, as this allows the routine analysis of a large number of samples and has been thoroughly validated as a platform technology.

University of Glasgow

Application

We chose as an illustrative example defining proteomic differences between healthy adult males and females.

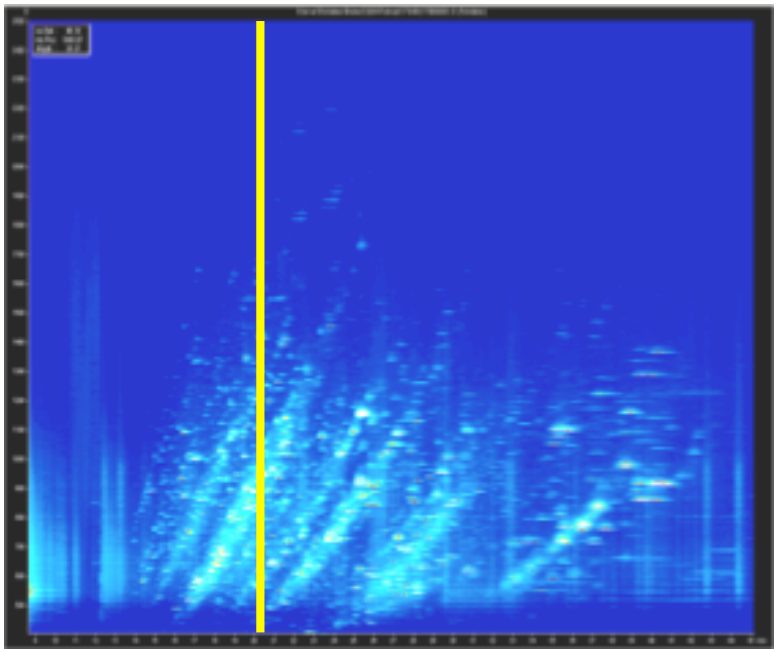
We chose to collect urine samples, since urine has been found to be of much higher stability than blood-derived samples, hence reducing pre-analytical variability.

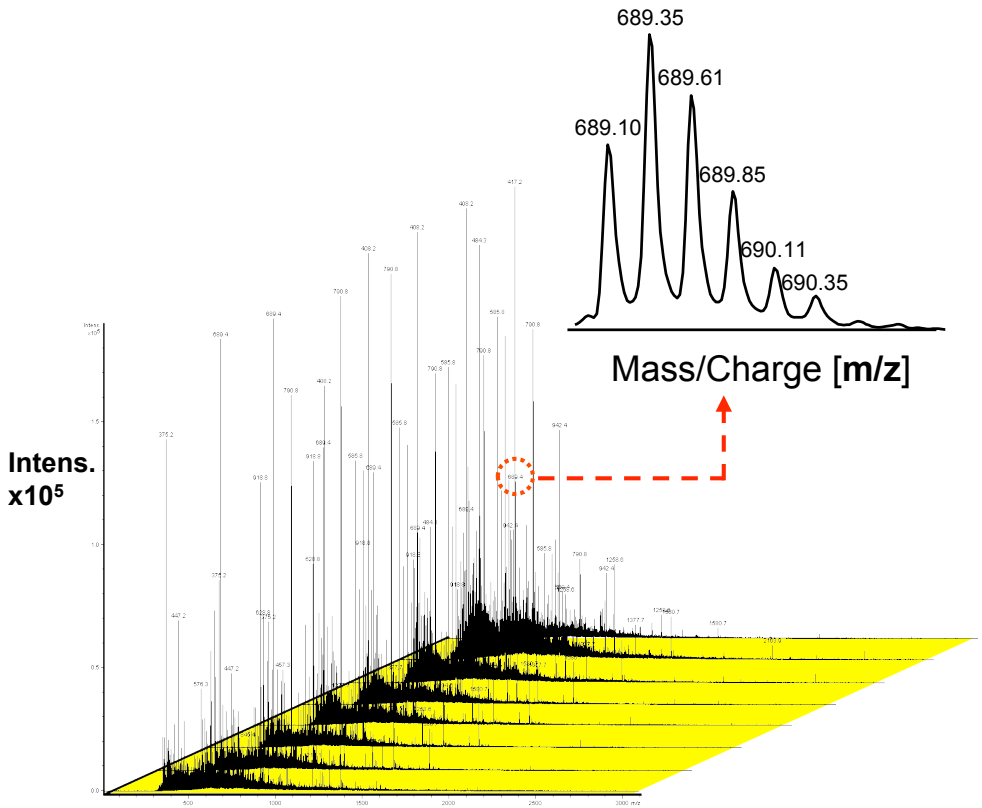
Capillary electrophoresis-mass spectrometry (CE-MS) was used to analyse the urine samples, as this allows the routine analysis of a large number of samples and has been thoroughly validated as a platform technology.

The goal of the analysis was to define biomarkers that would enable differentiation between male and female samples.

University of Glasgow

Raw data from CE-MS analysis





University of Glasgow

A Bayesian model for biomarker selection

CE-MS data contains a huge number of variables and the sample size is relatively small so the selection process can be unstable.

Hence, models which incorporate sparsity are desirable.

University of Glasgow

A Bayesian model for biomarker selection

CE-MS data contains a huge number of variables and the sample size is relatively small so the selection process can be unstable.

Hence, models which incorporate sparsity are desirable.

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

Sparsity was incorporated by choosing prior distributions for the variance of the regression coefficients (λ_i) that would shrink the coefficients of non-informative variables towards zero.

Three priors for λ_i were considered: Inverse Gamma (Model I), exponential (Model II) and non-informative Jeffreys (Model III).

University of Glasgow

A Bayesian model for biomarker selection

CE-MS data contains a huge number of variables and the sample size is relatively small so the selection process can be unstable.

Hence, models which incorporate sparsity are desirable.

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

Sparsity was incorporated by choosing prior distributions for the variance of the regression coefficients (λ_i) that would shrink the coefficients of non-informative variables towards zero.

Three priors for λ_i were considered: Inverse Gamma (Model I), exponential (Model II) and non-informative Jeffreys (Model III).

Posterior inference was conducted using an efficient Gibbs sampler.

University of Glasgow

Applying the Bae and Mallick model to our data

1. We removed all features that were present in less than 20% of the training samples.

2. We normalised features by dividing by their sums.

3. A Wilcoxon rank-sum test was used to provide a ranking of the features based on their p-value. Setting a threshold of 5% the number of features was reduced to 350.

University of Glasgow

Applying the Bae and Mallick model to our data

1. We removed all features that were present in less than 20% of the training samples.

2. We normalised features by dividing by their sums.

3. A Wilcoxon rank-sum test was used to provide a ranking of the features based on their p-value. Setting a threshold of 5% the number of features was reduced to 350.

4. The hyperparameters for Models I and II were fixed such that $E(\lambda_i) = 10$ and $\text{Var}(\lambda_i) = 100$.

University of Glasgow

Applying the Bae and Mallick model to our data

1. We removed all features that were present in less than 20% of the training samples.

2. We normalised features by dividing by their sums.

3. A Wilcoxon rank-sum test was used to provide a ranking of the features based on their p-value. Setting a threshold of 5% the number of features was reduced to 350.

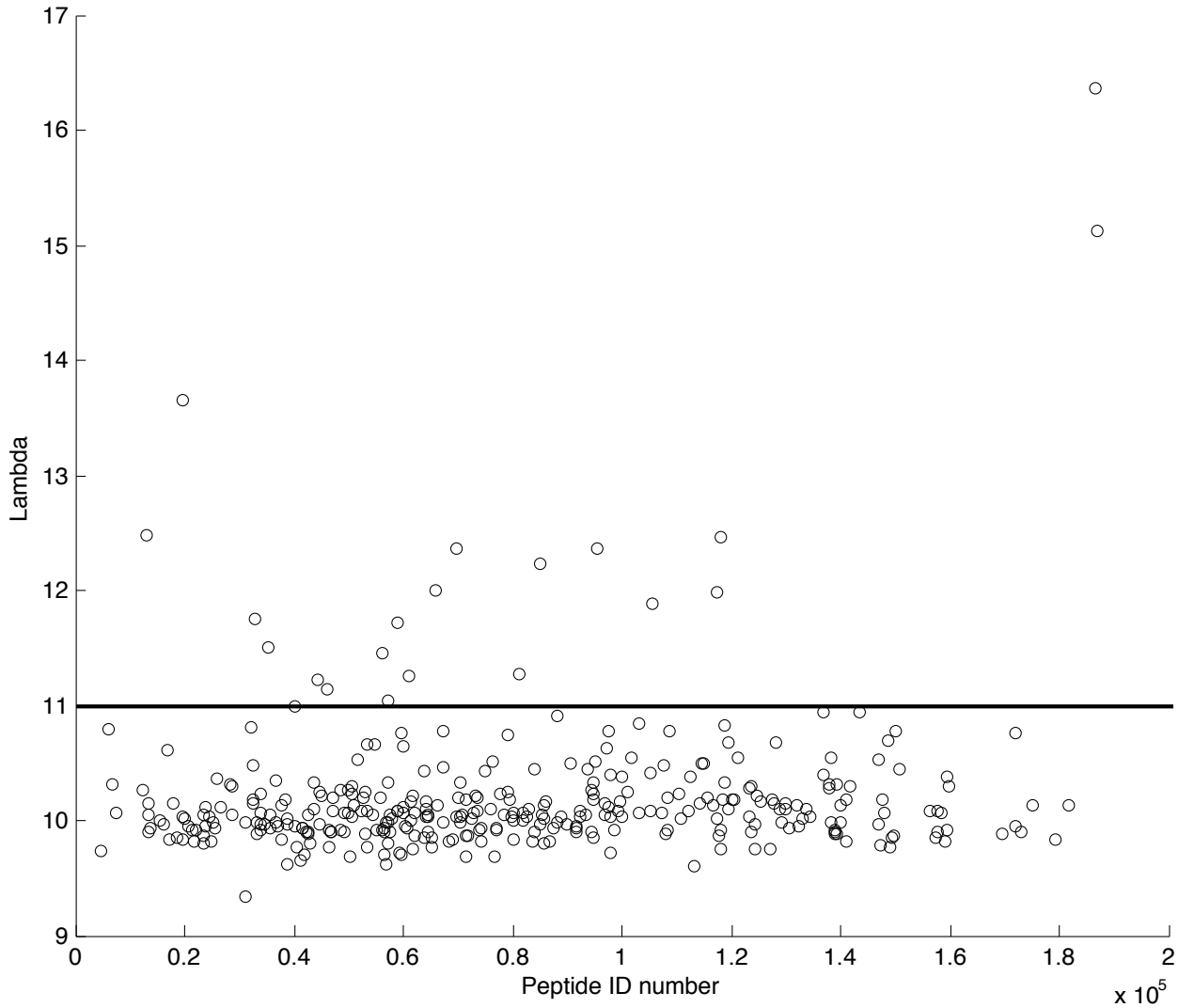
4. The hyperparameters for Models I and II were fixed such that $E(\lambda_i) = 10$ and $\text{Var}(\lambda_i) = 100$.

5. A randomised 10-CV was used to initially assess the performance of the 3 models.

Models I and II had an average test error of $8.2\% \pm 2.1\%$, while Model III's was $11.2\% \pm 2.0\%$.

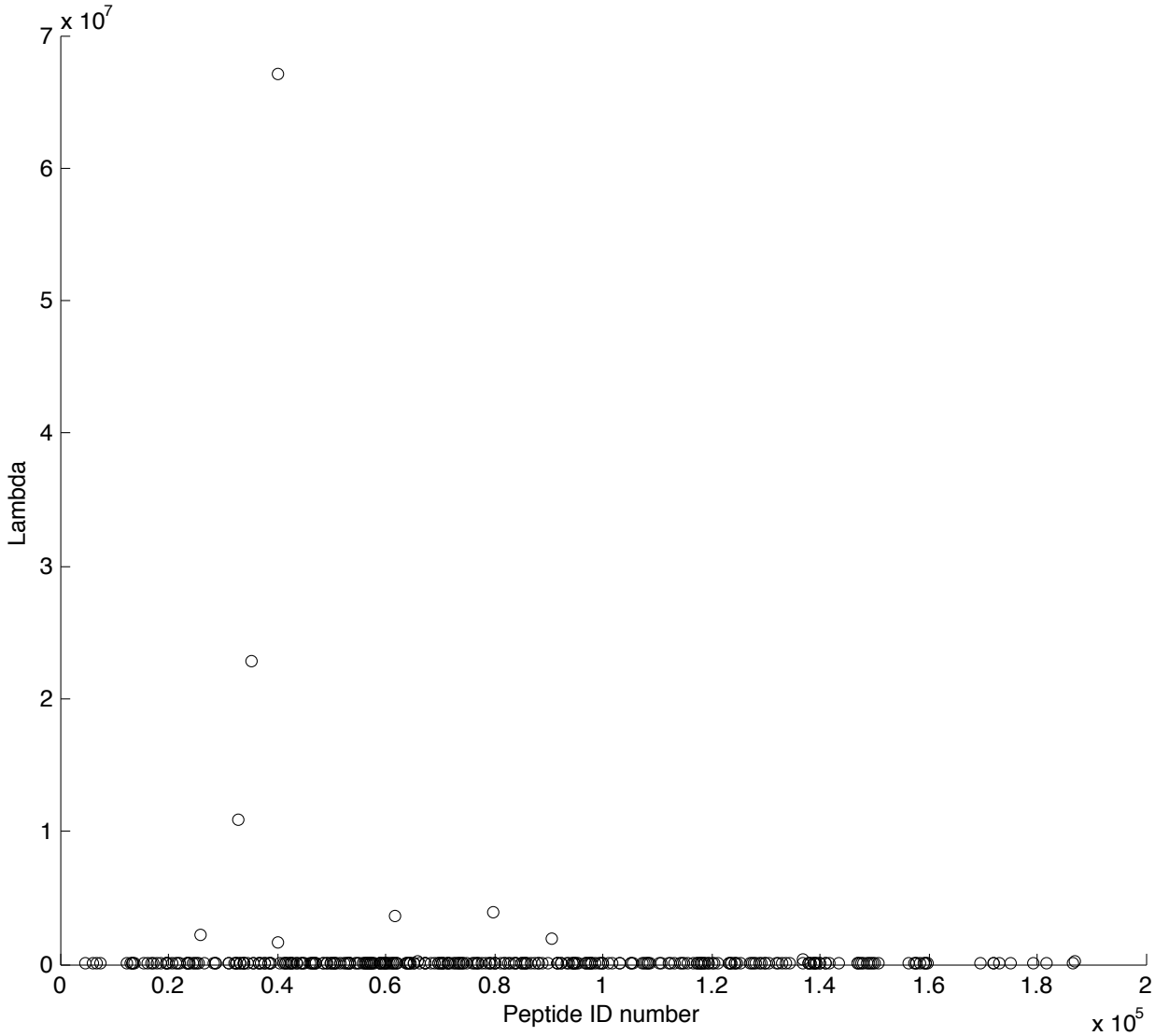
University of Glasgow

Plot of λ_i versus the peptide ID number (Model I)



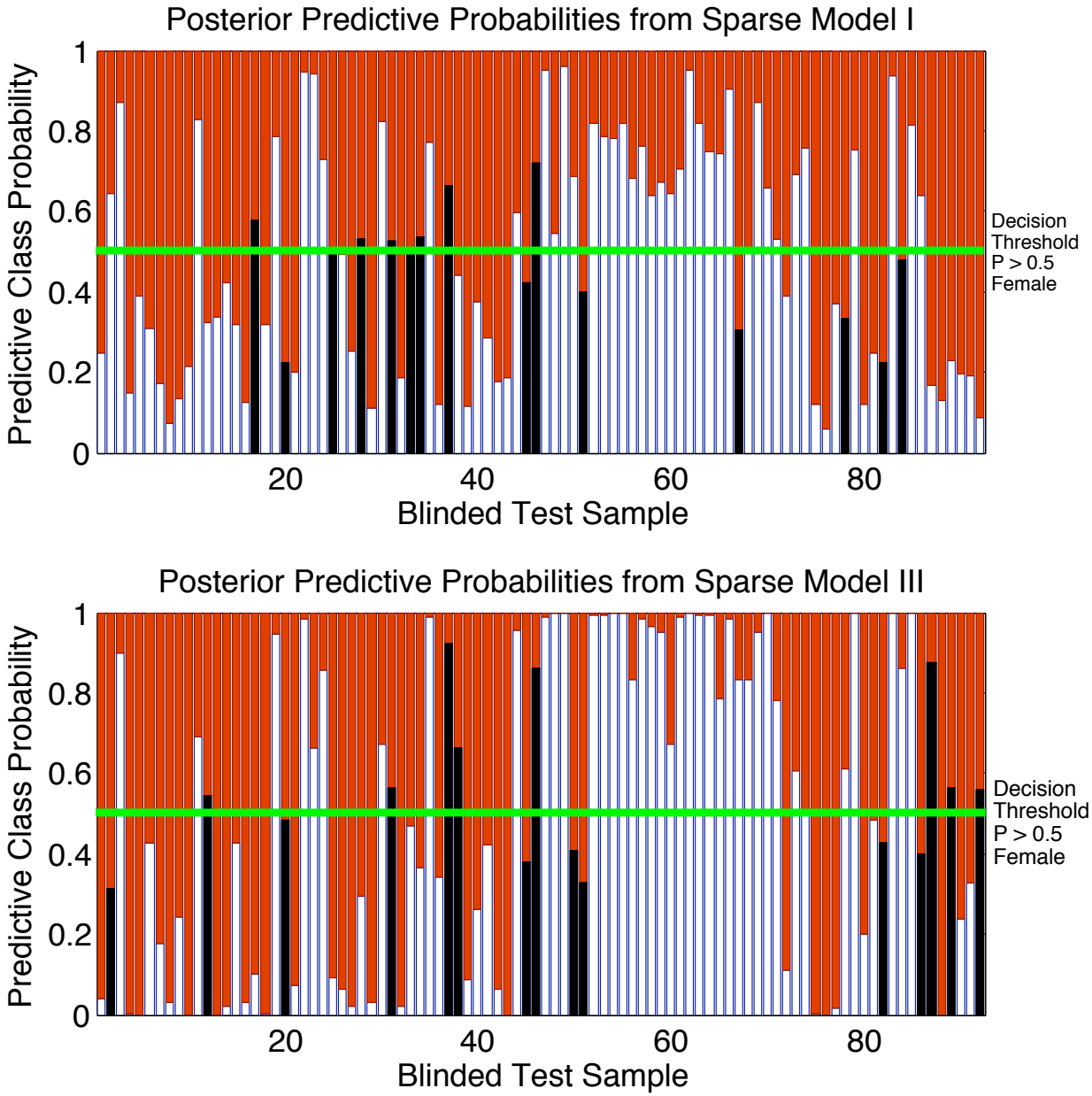
University of Glasgow

Plot of λ_i versus the peptide ID number (Model III)



University of Glasgow

Plots of the posterior predictive probabilities (Models I and III)



University of Glasgow

Test error for different training set sizes

Training set size	Model I	Model II	Model III
14	28.3%	27.2%	25%
40	27.2%	27.2%	23.9%
66	21.7%	21.7%	25%
134	16.3%	15.2%	16.3%

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.

University of Glasgow

Bayesian classification with averaged feature clusters

Scenario: A procedure that combines model based clustering and binary classification.

By averaging the features within the clusters obtained from model based clustering, we define “superfeatures” and use them in a classification model, thereby attaining concise interpretation and accuracy.

University of Glasgow

Bayesian classification with averaged feature clusters

Scenario: A procedure that combines model based clustering and binary classification.

By averaging the features within the clusters obtained from model based clustering, we define “superfeatures” and use them in a classification model, thereby attaining concise interpretation and accuracy.

Similar ideas, from a non-Bayesian two-step perspective, have been looked at by Hanczar et al. (2003) and Park et al. (2007).

With our simultaneous procedure, the clusters are formed considering the correlation of the predictors with the response in addition to the correlations among the predictors.

University of Glasgow

Bayesian classification with averaged feature clusters

Scenario: A procedure that combines model based clustering and binary classification.

By averaging the features within the clusters obtained from model based clustering, we define “superfeatures” and use them in a classification model, thereby attaining concise interpretation and accuracy.

Similar ideas, from a non-Bayesian two-step perspective, have been looked at by Hanczar et al. (2003) and Park et al. (2007).

With our simultaneous procedure, the clusters are formed considering the correlation of the predictors with the response in addition to the correlations among the predictors.

The proposed methodology should have wide applicability outside of proteomic biomarker selection.

University of Glasgow

Model details

Joint distribution:

$$p(\mathbf{t}, \mathbf{y}, X, \theta, \mathbf{w}) = p(\mathbf{t}, \mathbf{y}|\theta, \mathbf{w})p(X|\theta)p(\theta, \mathbf{w}).$$

University of Glasgow

Model details

Joint distribution:

$$p(\mathbf{t}, \mathbf{y}, X, \theta, \mathbf{w}) = p(\mathbf{t}, \mathbf{y}|\theta, \mathbf{w})p(X|\theta)p(\theta, \mathbf{w}).$$

Classification model:

$$t_n = \begin{cases} 1 & \text{if } y_n > 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$y_n = \mathbf{w}^T \boldsymbol{\theta}_n + \epsilon_n \text{ where } \epsilon_n \sim \mathcal{N}(0, 1).$$

University of Glasgow

Model details

Joint distribution:

$$p(\mathbf{t}, \mathbf{y}, X, \theta, \mathbf{w}) = p(\mathbf{t}, \mathbf{y}|\theta, \mathbf{w})p(X|\theta)p(\theta, \mathbf{w}).$$

Classification model:

$$t_n = \begin{cases} 1 & \text{if } y_n > 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$y_n = \mathbf{w}^T \boldsymbol{\theta}_n + \epsilon_n \text{ where } \epsilon_n \sim \mathcal{N}(0, 1).$$

Clustering model: Normal mixture model with equal weights and identity covariance matrices.

$$\Rightarrow p(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\theta}_k, I).$$

University of Glasgow

Model details

Joint distribution:

$$p(\mathbf{t}, \mathbf{y}, X, \theta, \mathbf{w}) = p(\mathbf{t}, \mathbf{y}|\theta, \mathbf{w})p(X|\theta)p(\theta, \mathbf{w}).$$

Classification model:

$$t_n = \begin{cases} 1 & \text{if } y_n > 0 \\ 0 & \text{otherwise.} \end{cases}$$

$$y_n = \mathbf{w}^T \boldsymbol{\theta}_n + \epsilon_n \text{ where } \epsilon_n \sim \mathcal{N}(0, 1).$$

Clustering model: Normal mixture model with equal weights and identity covariance matrices.

$$\Rightarrow p(\mathbf{x}) = \frac{1}{K} \sum_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\theta}_k, I).$$

Prior distributions:

$$p(\theta) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\theta}_k|\boldsymbol{\theta}_0, hI), \; p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, lI).$$

University of Glasgow

EM Algorithm

E-step:

$$\gamma(z_{dk}) = \frac{\exp\left\{-\frac{1}{2}\|\mathbf{x}_d - \boldsymbol{\theta}_k\|^2\right\}}{\sum_{j=1}^{K}\exp\left\{-\frac{1}{2}\|\mathbf{x}_d - \boldsymbol{\theta}_j\|^2\right\}},$$

$$E(y_n) = \begin{cases} \mathbf{w}^T\boldsymbol{\theta}_n + \frac{\phi(-\mathbf{w}^T\boldsymbol{\theta}_n)}{1-\Phi(-\mathbf{w}^T\boldsymbol{\theta}_n)} & \text{if } t_n = 1 \\ \mathbf{w}^T\boldsymbol{\theta}_n - \frac{\phi(-\mathbf{w}^T\boldsymbol{\theta}_n)}{\Phi(-\mathbf{w}^T\boldsymbol{\theta}_n)} & \text{otherwise.} \end{cases}$$

University of Glasgow

EM Algorithm

E-step:

$$\gamma(z_{dk}) = \frac{\exp\left\{-\frac{1}{2}\|\mathbf{x}_d - \boldsymbol{\theta}_k\|^2\right\}}{\sum_{j=1}^{K}\exp\left\{-\frac{1}{2}\|\mathbf{x}_d - \boldsymbol{\theta}_j\|^2\right\}},$$

$$E(y_n) = \begin{cases} \mathbf{w}^T\boldsymbol{\theta}_n + \frac{\phi(-\mathbf{w}^T\boldsymbol{\theta}_n)}{1-\Phi(-\mathbf{w}^T\boldsymbol{\theta}_n)} & \text{if } t_n = 1 \\ \mathbf{w}^T\boldsymbol{\theta}_n - \frac{\phi(-\mathbf{w}^T\boldsymbol{\theta}_n)}{\Phi(-\mathbf{w}^T\boldsymbol{\theta}_n)} & \text{otherwise.} \end{cases}$$

M-step:

$$\boldsymbol{\theta}_k = \frac{\left(E(\mathbf{y}) - \theta^T\mathbf{w}_{-k}\right)w_k + X\boldsymbol{\gamma}_k + \frac{1}{h}\boldsymbol{\theta}_0}{w_k^2 + \sum_{d=1}^{D}\gamma(z_{dk}) + \frac{1}{h}},$$

$$\mathbf{w} = \left(\theta\theta^T + \frac{1}{l}I\right)^{-1}\theta E(\mathbf{y}).$$

University of Glasgow

EM Algorithm

E-step:

$$\gamma(z_{dk}) = \frac{\exp\left\{-\frac{1}{2}\|\mathbf{x}_d - \boldsymbol{\theta}_k\|^2\right\}}{\sum_{j=1}^{K}\exp\left\{-\frac{1}{2}\|\mathbf{x}_d - \boldsymbol{\theta}_j\|^2\right\}},$$

$$E(y_n) = \begin{cases} \mathbf{w}^T\boldsymbol{\theta}_n + \frac{\phi(-\mathbf{w}^T\boldsymbol{\theta}_n)}{1-\Phi(-\mathbf{w}^T\boldsymbol{\theta}_n)} & \text{if } t_n = 1 \\ \mathbf{w}^T\boldsymbol{\theta}_n - \frac{\phi(-\mathbf{w}^T\boldsymbol{\theta}_n)}{\Phi(-\mathbf{w}^T\boldsymbol{\theta}_n)} & \text{otherwise.} \end{cases}$$

M-step:

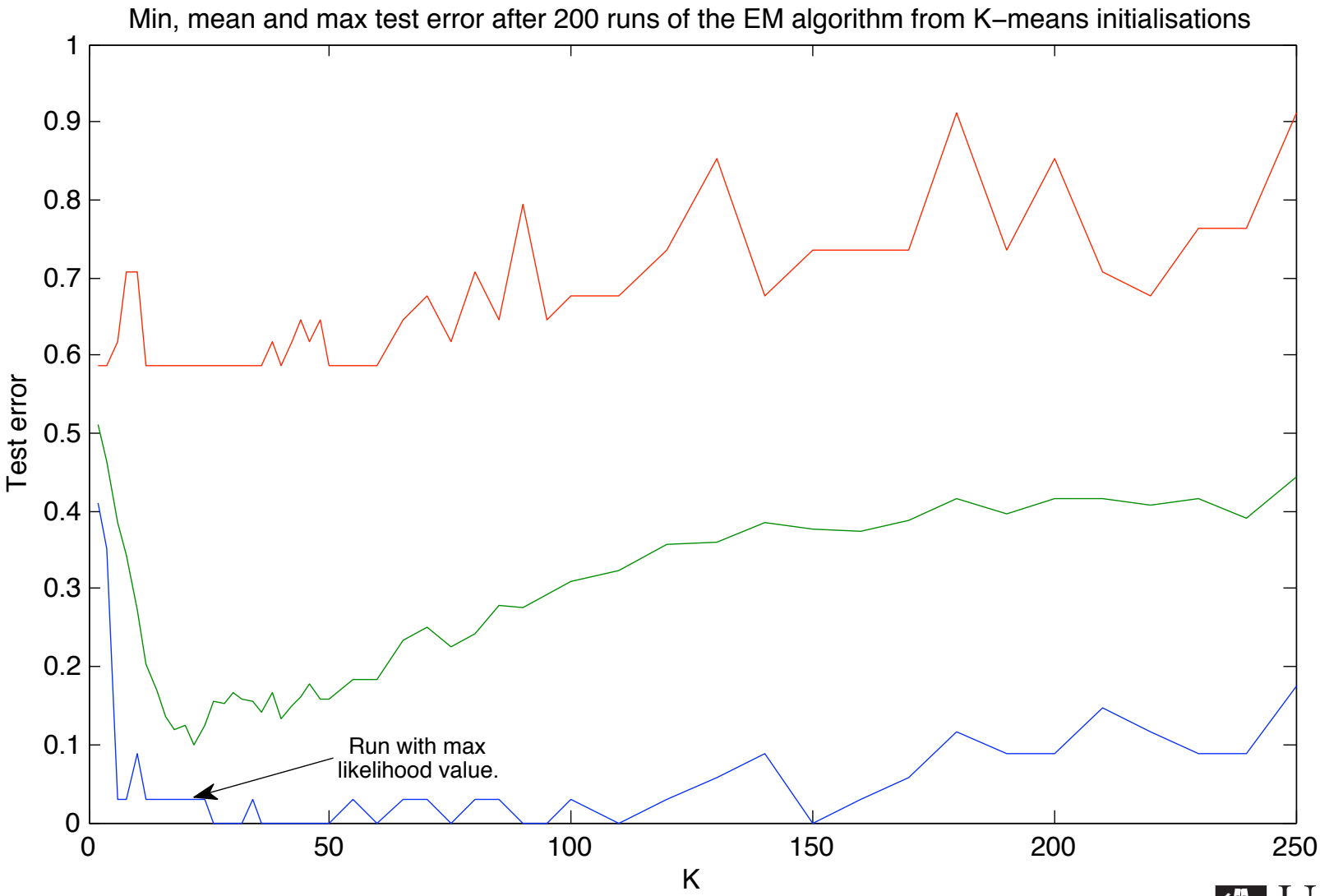
$$\boldsymbol{\theta}_k = \frac{\left(E(\mathbf{y}) - \theta^T\mathbf{w}_{-k}\right)w_k + X\boldsymbol{\gamma}_k + \frac{1}{h}\boldsymbol{\theta}_0}{w_k^2 + \sum_{d=1}^{D}\gamma(z_{dk}) + \frac{1}{h}},$$

$$\mathbf{w} = \left(\theta\theta^T + \frac{1}{l}I\right)^{-1}\theta E(\mathbf{y}).$$

Note that the first component of $\mathbf{w}$ is set to 1, so that the model is identifiable.

University of Glasgow

Results for Golub's leukemia data



University of Glasgow

Conclusions

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.

University of Glasgow

Conclusions

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.

University of Glasgow

Conclusions

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.

3. Meaningful results will only be obtained if the number of training samples collected is sufficient to allow the definition of statistically valid biomarkers.

University of Glasgow

Conclusions

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.

3. Meaningful results will only be obtained if the number of training samples collected is sufficient to allow the definition of statistically valid biomarkers.

4. Bayesian classification with averaged feature clusters is a promising new methodology with wide applicability.

University of Glasgow

Conclusions

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.

3. Meaningful results will only be obtained if the number of training samples collected is sufficient to allow the definition of statistically valid biomarkers.

4. Bayesian classification with averaged feature clusters is a promising new methodology with wide applicability.

5. The approach can be naturally extended to multiclass classification and to incorporate sparsity by employing an Inverse Gamma prior on the variance of the regression coefficients.

University of Glasgow

Acknowledgements

This work is supported by the EPSRC grant EP/F009429/1 - Advancing Machine Learning Methodology for New Classes of Prediction Problems.

University of Glasgow

References

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

Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D. and Lander E. S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439): 531–537.

Hanczar B., Courtine M., Benis A., Henegar C., Clément K. and Zucker J. D. (2003) Improving classification of microarray data using prototype-based feature selection. *SIGKDD Explorations*, 5(2): 23–30.

Park M. Y., Hastie T. and Tibshirani R. (2007) Averaged gene expressions for regression. *Biostatistics*, 8(2): 212–227.

University of Glasgow