Definition of Valid Proteomic Biomarkers: A Bayesian Solution

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

[1] Inference Group, Department of Computing Science, University of Glasgow, UK
{keithh,girolami}@dcs.gla.ac.uk
http://www.dcs.gla.ac.uk/inference
[2] Mosaiques Diagnostics and Therapeutics AG, Hannover, Germany

Abstract. Clinical proteomics is suffering from high hopes generated by reports on apparent biomarkers, most of which could not be later substantiated via validation. This has brought into focus the need for improved methods of finding a panel of clearly defined biomarkers. To examine this problem, urinary proteome data was collected from healthy adult males and females, and analysed to find biomarkers that differentiated between genders. We believe that models that incorporate sparsity in terms of variables are desirable for biomarker selection, as proteomics data typically contains a huge number of variables (peptides) and few samples making the selection process potentially unstable. This suggests the application of a two-level hierarchical Bayesian probit regression model for variable selection which assumes a prior that favours sparseness. The classification performance of this method is shown to improve that of the Probabilistic K-Nearest Neighbour model.

Keywords: Proteomic biomarkers, classification, sparsity, feature selection, Bayesian inference.

1 Introduction

Proteins and peptides in body fluids hold considerable information on the physiology of an organism and thus can serve as biomarkers for disease. However, the fields of biomarker discovery and clinical proteomics are suffering from high hopes generated by reports on potential biomarkers, most of which subsequently could not be substantiated via validation [1]. This development has resulted in much scepticism from both clinicians and regulatory agencies, which will make the application of valid biomarkers even more of a challenge. This vicious circle has to be broken by pinpointing the major errors made in earlier research and highlighting good practice that will enable the definition of valid biomarkers with a much higher probability than currently observed. While some of the initial issues have already been dealt with satisfactorily, others are still unresolved.

For example, it is now generally accepted that single biomarkers should not be applied, as the complexity of a disease is unlikely to be thoroughly displayed by just one marker, and that a panel of such biomarkers should be employed

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 137–149, 2009.

instead [1,2]. However, it is equally evident that such a panel must consist of clearly defined biomarkers, and not of an ill-defined signature, as reported in several of the original manuscripts, almost exclusively based on the Surface-Enhanced Laser Desorption/Ionization (SELDI) technology, that subsequently could not be validated [3,4].

This brings the issue of definition of a valid biomarker into focus. The fundamental question that should be asked is whether the change observed in the disease (frequency or abundance) of a certain molecule, based on data from a proteomics study, is in fact a result of the disease, or does it merely reflect an artefact due to technical variability in the pre-analytical steps, or in the analysis. Other likely suspects for suggesting an apparent but erroneous association with disease are biological variability or bias introduced in the study (for example, due to lifestyle, age and gender). In fact, these two problems are likely responsible for the majority of erroneous biomarkers.

The most appropriate answer to this challenge appears to be the application of stringent statistical analysis. Not only does good statistical practice need to be highlighted, but also more sophisticated multivariate selection methods need to be developed, so that valid biomarkers will be defined with a much higher probability than currently observed. To this end, we adopt a Bayesian approach to classification and feature selection, as this approach offers formal and well-calibrated probabilities for class prediction which is useful for medical decision making. We compare the Probabilistic K-Nearest Neighbour and hierarchic linear probit regression classifiers. Feature selection was incorporated in the latter method by assuming priors that favoured sparse solutions. It should be noted that other classification methods like support vector machines with recursive feature elimination, adaptive boosting and random forests could have been used, but that in this paper we decided to focus solely on highlighting possible Bayesian approaches for proteomic biomarker selection.

The rest of this paper is organised as follows: in Sect. 2 we discuss the illustrative experiment to find biomarkers that differentiate between males and females. Section 3 describes the classification and feature selection methods used in this paper in more detail. Section 4 presents the results of our experiments comparing the classification performance of the two methods and the feature selection performance of the three priors used to induce sparsity in the probit regression model. Finally, Sect. 5 discusses the conclusions that can be drawn from our experimental results.

2 Application

To avoid any uncertainty in the assignment of a physiological condition, we choose as an illustrative example defining proteomic differences between apparently healthy adult males and females. While clinical diagnosis or pathophysiological conditions are in general associated with a certain degree of uncertainty, gender can be assessed with almost 100% confidence. Furthermore, the differences between male and female, while quite obvious at first sight, are likely to be rather subtle at the proteomic level.

We chose urine to be the body fluid of interest, since urine has been found to be of much higher stability than blood-derived samples (serum or plasma), hence reducing pre-analytical variability [2,5]. Capillary electrophoresis-mass spectrometry (CE-MS) was used to analyse the urine samples, as this technology allows the routine analysis of a large number of samples and has been thoroughly validated as a platform technology [6,7].

The second urine of the morning was collected from a group of apparently healthy male and female volunteers (aged 21-40) during a routine medical checkup before recruitment at the Hannover Medical School. All samples were prepared and analysed using CE-MS as described in [6,7]. The goal of the analysis was to define biomarkers that would enable differentiation between male and female samples (based on the hypothesis that such biomarkers must exist).

3 Methods

3.1 Probabilistic K-Nearest Neighbour Classification

The Probabilistic K-Nearest Neighbour (PKNN) classification method (see [8] for an empirical analysis) adopts a fully Bayesian approach to obtaining posterior probabilities over the scaling parameter and the number of nearest neighbours to be employed. Markov chain Monte Carlo using the Metropolis-Hastings algorithm is employed to perform posterior sampling and Monte Carlo averaging provide the predicitive probabilities of class labels. Some more detail is provided below.

Consider a finite data sample $\{(t_1, \boldsymbol{x}_1), \cdots, (t_N, \boldsymbol{x}_N)\}$ where each $t_n \in \{1, \cdots, C\}$ denotes the class label associated with the D-dimensional feature vector $\boldsymbol{x}_n \in \mathbb{R}^D$ and the feature space $\mathbb{R}^D$ has an associated metric with parameters $\boldsymbol{\theta}$ denoted as $\mathcal{M}_{\boldsymbol{\theta}}$. To define a probabilistic representation of the KNN method an approximate conditional joint likelihood is defined in [9] such that

$$p(\boldsymbol{t}|\boldsymbol{X}, \beta, k, \boldsymbol{\theta}, \mathcal{M}) \approx \prod_{n=1}^{N} \frac{\exp\left\{\frac{\beta}{k} \sum_{j \sim n|k}^{\mathcal{M}_{\boldsymbol{\theta}}} \delta_{t_n t_j}\right\}}{\sum_{c=1}^{C} \exp\left\{\frac{\beta}{k} \sum_{j \sim n|k}^{\mathcal{M}_{\boldsymbol{\theta}}} \delta_{c t_n}\right\}} \tag{1}$$

where we define the $N \times 1$-dimensional vector $\boldsymbol{t}$ as $[t_1, \cdots t_N]^T$ and the $N \times D$-dimensional matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N]^T$, $\mathcal{M}$ denotes the metric employed in the feature space and $\boldsymbol{\theta}$ are the associated parameters. The number of nearest neighbours is k and β defines a scaling variable. The expression

$$\sum_{j \sim n|k}^{\mathcal{M}_{\boldsymbol{\theta}}} \delta_{t_n t_j} \tag{2}$$

denotes the number of the nearest k neighbours of $\boldsymbol{x}_n$, as measured under the metric $\mathcal{M}_{\boldsymbol{\theta}}$ within $N - 1$ samples from $\boldsymbol{X}$ remaining when $\boldsymbol{x}_n$ is removed which we denote as $\boldsymbol{X}_{-n}$, and have the class label value of t_n, whilst each of the terms in the summation of the denominator provides a count of the number of the k neighbours of $\boldsymbol{x}_n$ which have class label equaling c.

Full posterior inference will follow by obtaining the parameter posterior distribution $p(\beta, k, \boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{X}, \mathcal{M})$ and subsequent predictions of the target class label t_* of a new datum $\boldsymbol{x}_*$ are made by posterior averaging such that

$$p(t_*|\boldsymbol{x}_*, \boldsymbol{t}, \boldsymbol{X}, \mathcal{M}) = \sum_k \int p(t_*|\boldsymbol{x}_*, \boldsymbol{t}, \boldsymbol{X}, \beta, k, \boldsymbol{\theta}, \mathcal{M}) p(\beta, k, \boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{X}, \mathcal{M}) d\beta d\boldsymbol{\theta}. \quad (3)$$

As the required posterior takes an intractable form an MCMC procedure is proposed in [9] and extended in [10] to enable metric inference so that the following Monte-Carlo estimate is employed

$$\hat{p}(t_*|\boldsymbol{x}_*, \boldsymbol{t}, \boldsymbol{X}, \mathcal{M}) = \frac{1}{N_s} \sum_{s=1}^{N_s} p(t_*|\boldsymbol{x}_*, \boldsymbol{t}, \boldsymbol{X}, \beta^{(s)}, k^{(s)}, \boldsymbol{\theta}^{(s)}, \mathcal{M}) \quad (4)$$

where each $\beta^{(s)}, k^{(s)}, \boldsymbol{\theta}^{(s)}$ are samples obtained from the full parameter posterior $p(\beta, k, \boldsymbol{\theta}|\boldsymbol{t}, \boldsymbol{X}, \mathcal{M})$ using a Metropolis style sampler.

As the standard KNN method has no straightforward way to learn the metric we restrict this study to posterior inference over k and β and fix the metric to the standard Euclidean metric. We therefore adopt the Metropolis scheme detailed in [9] and obtain samples from the posterior $p(\beta, k, |\boldsymbol{t}, \boldsymbol{X}, \mathcal{M})$ and employ Monte-Carlo estimates $\hat{p}(t_*|\boldsymbol{x}_*, \boldsymbol{t}, \boldsymbol{X}, \mathcal{M}) = \frac{1}{N_s} \sum_{s=1}^{N_s} p(t_*|\boldsymbol{x}_*, \boldsymbol{t}, \boldsymbol{X}, \beta^{(s)}, k^{(s)}, \mathcal{M})$ in the following experimental section.

3.2 Hierarchic Linear Probit Regression Models

The fundamental problem of biomarker selection via CE-MS data is to identify which peptides best discriminate between different types of protein samples, in this case between male and female samples. CE-MS data contains a large 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. Bae and Mallick [11] proposed a two-level hierarchical Bayesian probit regression model for variable selection which used three different priors to incorporate different levels of sparsity in the model. Details of this model, the sparsity inducing priors and the Gibbs sampler used to perform posterior sampling are given below. This method is preferable to using support vector machines for performing variable selection as we can obtain predictive probabilities of the class labels for new observations by Monte Carlo averaging, similar to the Probabilistic K-Nearest Neighbour method mentioned earlier.

Model. Consider a finite data sample $\{(t_1, \boldsymbol{x}_1), \cdots, (t_N, \boldsymbol{x}_N)\}$ where each $t_n \in \{1, 2\}$ denotes the class label associated with the D-dimensional feature vector $\boldsymbol{x}_n \in \mathrm{I\!R}^D$. Define the binary regression model as $p_i = P(t_i = 2) = \Phi(\boldsymbol{x}_i^T \boldsymbol{\beta})$,

$i = 1, \ldots n$, where $\boldsymbol{\beta}$ is the $D \times 1$-dimensional vector of unknown regression parameters and Φ is the standard normal cumulative density function linking the probability p_i with the linear structure $\boldsymbol{x}_i^T\boldsymbol{\beta}$.

Albert and Chib [12] introduced n independent latent variables $\boldsymbol{z} = [z_1, \ldots, z_n]^T$ into the problem, where $z_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1)$ and define $t_i = 2$ if $z_i > 0$ and $t_i = 1$ if $z_i \leq 0$. This approach connects the probit binary regression model for t_i to a normal linear regression model for the latent variable z_i.

Bae and Mallick [11] considered different sparsity inducing priors for $\boldsymbol{\beta}$ in a two-level hierarchical Bayesian model. They placed a zero-mean Gaussian prior on $\boldsymbol{\beta}$ with unknown variances and assigned three different priors for the variances under the assumption that they were independent, i.e., $\boldsymbol{\beta}|\boldsymbol{\Lambda} \sim N(\mathbf{0}, \boldsymbol{\Lambda})$, where $\mathbf{0} = [0, \ldots, 0]^T$, $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$ and λ_i is the variance of β_i.

Prior distributions for $\boldsymbol{\Lambda}$. Model I - conjugate Inverse Gamma priors for each λ_i, i.e.,

$$\boldsymbol{\Lambda} \sim \prod_{i=1}^{D} \mathrm{IG}\left(\frac{a}{2}, \frac{2}{b}\right) \propto \prod_{i=1}^{D} \left(\frac{1}{\lambda_i}\right)^{(a/2)+1} \exp\left(-\frac{b}{2\lambda_i}\right). \tag{5}$$

Model II - exponential priors for each λ_i, i.e.,

$$\boldsymbol{\Lambda} \sim \prod_{i=1}^{D} \mathrm{Exponential}(\gamma) \propto \prod_{i=1}^{D} \exp\left(-\frac{\gamma\lambda_i}{2}\right). \tag{6}$$

Model III - non-informative Jeffreys priors for each λ_i, i.e.,

$$\boldsymbol{\Lambda} \sim \prod_{i=1}^{D} \frac{1}{\lambda_i}. \tag{7}$$

Note that Model III is the special case of Model I in which the hyperparameters a and b are both set to 0.

Gibbs sampler. 1. Sample z_i, for $i = 1, \ldots, n$, from its full conditional distribution

$$z_i|\boldsymbol{\beta}, t_i \propto \begin{cases} N(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1) \text{ truncated at the left by 0 if } t_i = 2, \\ N(\boldsymbol{x}_i^T\boldsymbol{\beta}, 1) \text{ truncated at the right by 0 if } t_i = 1. \end{cases} \tag{8}$$

2. Sample $\boldsymbol{\beta}$ from its full conditional distribution $p(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{\Lambda}) \propto N(\Sigma\boldsymbol{X}^T\boldsymbol{z}, \Sigma)$, where $\Sigma = (\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Lambda}^{-1})^{-1}$.
3. Sample $\boldsymbol{\Lambda}$ from its full conditional distribution. The full conditional distributions for Models I, II and III, respectively, are:

$$p(\boldsymbol{\Lambda}^{-1}|\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{\beta}) \propto \prod_{i=1}^{D} \mathrm{Gamma}\left(\frac{a+1}{2}, \frac{2}{b+\beta_i^2}\right), \tag{9}$$

$$p(\boldsymbol{\Lambda}^{-1}|\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{\beta}) \propto \prod_{i=1}^{D} \mathrm{InverseGaussian}\left(\frac{\sqrt{\gamma}}{|\beta_i|}, \gamma\right) \tag{10}$$

and

$$p(\boldsymbol{\Lambda}^{-1}|\boldsymbol{z}, \boldsymbol{t}, \boldsymbol{\beta}) \propto \prod_{i=1}^{D} \text{Gamma}\left(\frac{1}{2}, \frac{2}{\beta_i^2}\right). \tag{11}$$

Predictive classification. The predictive classification of the target class label t_* of a new datum $\boldsymbol{x}_*$ is given by the following Monte-Carlo estimate:

$$\hat{P}(t_* = 2|\boldsymbol{x}_*) = \frac{1}{N_s}\sum_{s=1}^{N_s} p(t_* = 2|\boldsymbol{x}_*, \boldsymbol{\beta}^{(s)}, \boldsymbol{z}^{(s)}, \boldsymbol{\Lambda}^{(s)}) = \frac{1}{N_s}\sum_{s=1}^{N_s} \Phi(\boldsymbol{x}_*^T \boldsymbol{\beta}^{(s)}), \tag{12}$$

where $\boldsymbol{\beta}^{(s)}$, $\boldsymbol{z}^{(s)}$ and $\boldsymbol{\Lambda}^{(s)}$ are the MCMC samples from the posterior distribution.

4 Experimental Results

4.1 PKNN Classifiers

To maintain consistency of data representation with other studies on this data the same arbitrary threshold for data sparsity (80%) was employed to reduce the number of covariates to 1524. However, instead of normalising samples with their sum of intensity values, we normalised features with their sum of intensity values, so as to not distort the original feature space before applying our models. A Wilcoxon Rank-Sum non-parametric test was used to provide a ranking of individual covariates based on p-value, and from this it is clear that a very small percentage of the 1524 peptides have any statistical evidence supporting their discrimination ability. Setting a p-value threshold of 2% the number of peptides was reduced further to 229 and these were used in devising a series of PKNN classifiers.

Starting with the full set of 229 peptides a PKNN classifier was devised by using Metropolis sampling with a burn-in of 5000 samples and a further 45000 post-burn-in samples retained for Monte-Carlo averaging. The proposal distribution was tuned to achieve acceptance rates between 35% to 50%. A randomised ten-fold cross validation (10-CV) was used to obtain estimates of predictive performance and only 0-1 error loss is reported here, however predictive probabilities are obtained from PKNN. These probabilities are used to make decisions based on the cost and threshold selected which in this case as the classes are balanced was set at 0.5.

The 10-CV score and associated standard error is reported when 229 peptides are used, then 228 are used where the peptide with the highest p-value is removed. This is done for fourteen peptides after which groups of 10 peptides were removed each time and the 10-CV score was measured. These results are shown in Fig. 1. A minimum mean 10-CV error of 8.68% is achieved, however this is rather meaningless taken on its own without considering the standard error, which would increase if multiple randomisations were employed. At around 220 to 210 peptides the range of error is minimal and this increases as the number of low p-value peptides are removed. It is conjectured that due to the relatively

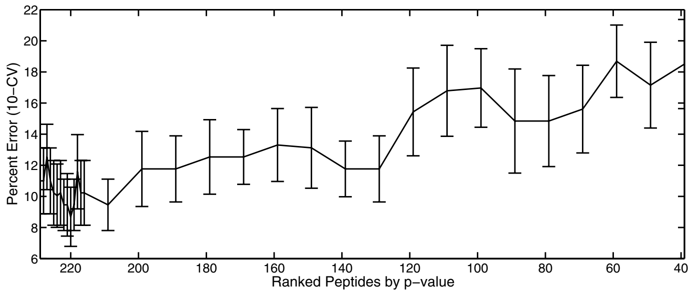

Fig. 1. This graph shows the estimated 10-CV prediction error (mean ± one standard error)

high levels of sparsity in the data which remains a large number of peptides are required to make reasonable predictions, this will be discussed further on in the paper.

4.2 Hierarchic Linear Probit Regression Models

As in Sect. 4.1, we remove all peptides that have more than 80% zero intensity values and normalise each feature with their sum of intensity values. We again use the ranking of covariates provided by the p-values of the Wilcoxon test to further reduce the number of peptides, but this time choose a p-value threshold of roughly 5% to reduce the number of peptides to 350. These peptides were then used to build the three classifiers discussed in Sect. 3.2.

Like Bae and Mallick [11], we fixed the hyperparameters for Models I and II so that $E(\lambda_i) = 10$ and $\mathrm{Var}(\lambda_i) = 100$. We ran the Gibbs sampler of Sect. 3.2 for 50,000 iterations and discarded the first 20,000 iterations as burn-in. As in Sect. 4.1, a randomised 10-CV was used to assess the predictive performance of the three models. Both Models I and II gave an average test error of $8.2\% \pm 2.1\%$, while Model III gave an average test error of $11.2\% \pm 2.0\%$. It should be noted that tuning the hyperparameters of Models I and II could potentially lead to improved performance. It is not surprising that Models I and II performed similarly, as although the form of the prior distribution for the variance of the regression coefficients was different, the mean and variance was set to be the same, and this result is consistent with the findings of Bae and Mallick [11]. We believe that the poorer performance of Model III is due to the Jeffreys prior inducing too much sparsity in the model, similar to the worsening performance of the PKNN classifier seen in Fig. 1 after the number of peptides in the model is reduced below 210.

We select potential biomarkers using the posterior variance of $\boldsymbol{\beta}$ with the idea being that the peptides with larger variance are more important in discriminating between the different types of protein samples than those with smaller variance. Figures 2, 3 and 4 show the variance of β_i for Models I, II and III, respectively.

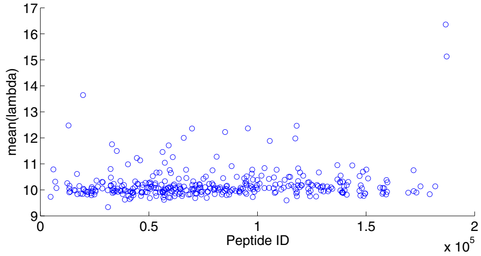

Fig. 2. Plot of the variance of β_i versus the peptide ID (Model I)

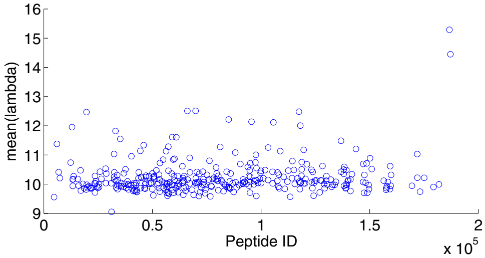

Fig. 3. Plot of the variance of β_i versus the peptide ID (Model II)

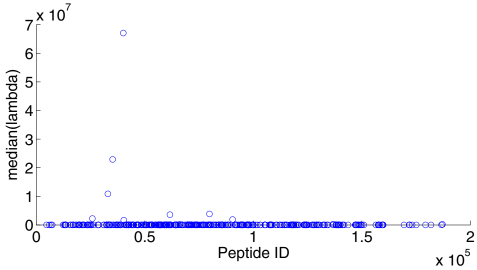

Fig. 4. Plot of the variance of β_i versus the peptide ID (Model III)

We see that Models I and II give very similar results and there are roughly 20 peptides that have significantly larger variance than the others. Comparing the top 20 peptides for Models I and II we see that Models I and II give very consistent selections, as 18 peptides are in both top 20s. In particular, both models rank peptides 186673 and 187114 as the two most important peptides by a comfortable margin. Of these 18 peptides, all but three of them had p-values less than 0.005 in the original Wilcoxon tests and of the top two peptides both had p-values less than 3×10^{-6}. We also checked the sensitivity of the peptide rankings with the 10-CV mentioned earlier and found that the rankings of Models I and II were broadly consistent between folds with the same peptides being ranked highly in the majority of folds.

We see from Fig. 4 that Model III induces sparseness much more strongly than Models I and II, as there are only 8 peptides that have significantly larger variance than the others. Only two of these peptides were selected by Models I and II, but both models ranked them outside the top 10. Unlike Models I and II, when we checked the sensitivity of the peptide rankings with the results for the 10-CV we found that the selected peptides were rarely consistent between folds. This suggests that the Jeffreys prior over-prunes the model and puts very little weight on many peptides useful for classifying the binary response, leading to its worse performance in terms of the average test error found earlier. We thus conclude that the peptides suggested by Models I and II are more likely to make a good set of biomarkers for this problem than those suggested by Model III.

4.3 Classification for the Blinded Test Data

Figures 5, 6, 7 and 8 show the posterior predictive probabilities obtained for each test sample from the PKNN classifier of Sect. 4.1 and Bae and Mallick's sparse probit regression Models I, II and III of Sect. 4.2, respectively. We see that the PKNN classifier and Bae and Mallick's Model III tend to give the most confident predictions, while the posterior predictive probabilities are very similar for Bae and Mallick's Models I and II and tend to give the least confident predictions. Note that it is easy to compare the predictive performance of two competing classifiers graphically by plotting the predictive probabilities of one method against the other. We also see that the four classifiers tend to allocate the test samples to the same class. In fact, all four classifiers are in agreement for 71 of the 92 test samples. The test samples where there is disagreement in the predictions of the four classifiers tend to happen when at least one classifier gives an unconfident prediction, that is, a posterior predictive probability close to 0.5. Even with the PKNN classifier there are four samples that have a posterior predictive probability of between 0.4 and 0.6. In such cases we would advocate

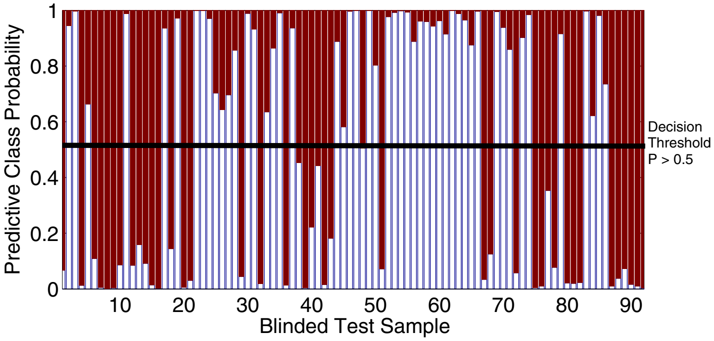

Fig. 5. Plot of the posterior predictive probabilities from the PKNN classifier

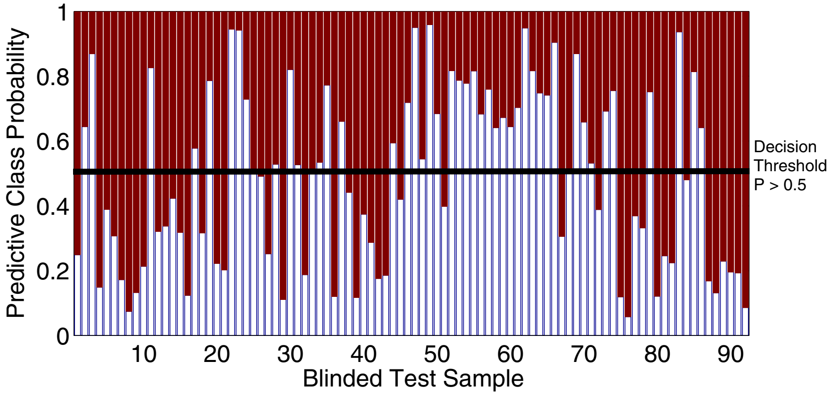

Fig. 6. Plot of the posterior predictive probabilities from Bae and Mallick Model I

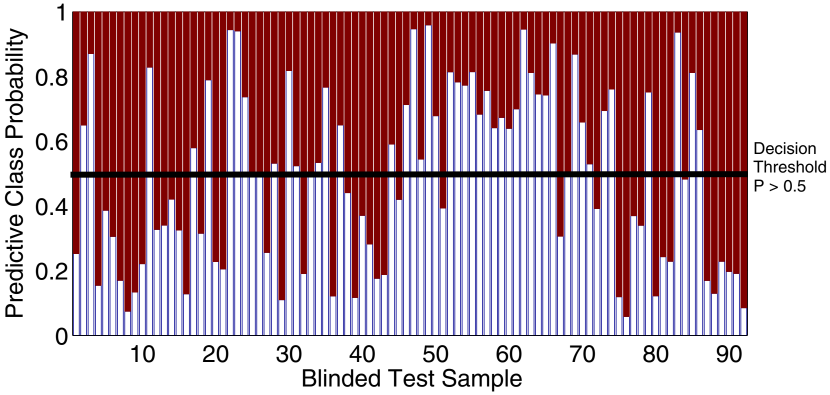

Fig. 7. Plot of the posterior predictive probabilities from Bae and Mallick Model II

not allocating the test sample to either class, as there is great uncertainty over the true class of the test sample. This transparency in the confidence of our class predictions is a huge advantage of the Bayesian approach over the more commonly used SVM techniques, which cannot provide such a formal and well-calibrated measure of the confidence of a class prediction.

The performance of the four classifiers on the blinded test set turned out to be very similar, as Bae and Mallick's Model II misclassified 14 out of the 92 samples, while both their Models I and III made 15 misclassifications, and the PKNN classifier performing slightly worse with 17 misclassifications. As we would expect, the test samples that were misclassified tended to have posterior predictive probabilities between 0.3 and 0.7, and thus had class predictions that were not very confident.

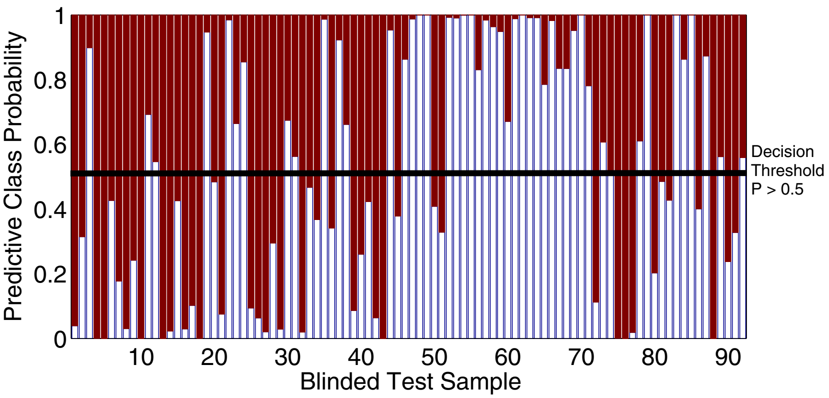

Fig. 8. Plot of the posterior predictive probabilities from Bae and Mallick Model III

We then trained Bae and Mallick's three models on three smaller data sets with 33, 20 and 7 cases and controls, in order to assess how model performance was affected by smaller training set sizes. We discovered that the confidence in our predictions declines significantly 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, which suggests that our predictive performance may be little better than guessing and that the biomarkers suggested by such a small data set would not be substantiated in practice. This suggestion of deteriorating predictive performance as the number of training samples is reduced was confirmed when we unblinded the test samples (see Table 1).

Table 1. 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%

5 Conclusions

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.

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 medical decision making.

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

Acknowledgements. K. Harris & M. Girolami are supported by an Engineering and Physical Sciences Research Council (EPSRC) grant EP/F009429/1 - Advancing Machine Learning Methodology for New Classes of Prediction Problems. M. Girolami is funded by an EPSRC Advanced Research Fellowship EP/E052029/1. H. Mischak is supported by EU funding through the InGenious HyperCare consortium, grant LSHM-CT-2006-037093.

References

1. Mischak, H., Apweiler, R., Banks, R.E., Conaway, M., Coon, J., Dominiczak, A., Ehrich, J.H.H., Fliser, D., Girolami, M., Hermjakob, H., Hochstrasser, D., Jankowski, J., Julian, B.A., Kolch, W., Massy, Z.A., Neusuess, C., Novak, J., Peter, K., Rossing, K., Schanstra, J., Semmes, O.J., Theodorescu, D., Thongboonkerd, V., Weissinger, E.M., Van Eyk, J.E., Yamamoto, T.: Clinical proteomics: A need to define the field and to begin to set adequate standards. Proteomics - Clinical Applications 1(2), 148–156 (2007)
2. Decramer, S., de Peredo, A.G., Breuil, B., Mischak, H., Monsarrat, B., Bascands, J.L., Schanstra, J.P.: Urine in clinical proteomics. Molecular and Cellular Proteomics 7(10), 1850–1862 (2008)
3. Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., Liotta, L.A.: Use of proteomic patterns in serum to identify ovarian cancer. The Lancet 359(9306), 572–577 (2002)
4. Check, E.: Proteomics and cancer - running before we can walk? Nature 429(6991), 496–497 (2004)
5. Mischak, H., Coon, J.J., Novak, J., Weissinger, E.M., Schanstra, J.P., Dominiczak, A.F.: Capillary electrophoresis-mass spectrometry as a powerful tool in biomarker discovery and clinical diagnosis: An update of recent developments. Mass Spectrometry Reviews (October 2008) (in press)
6. Coon, J.J., Zürbig, P., Dakna, M., Dominiczak, A.F., Decramer, S., Fliser, D., Frommberger, M., Golovko, I., Good, D.M., Herget-Rosenthal, S., Jankowski, J., Julian, B.A., Kellmann, M., Kolch, W., Massy, Z., Novak, J., Rossing, K., Schanstra, J.P., Schiffer, E., Theodorescu, D., Vanholder, R., Weissinger, E.M., Mischak, H., Schmitt-Kopplin, P.: CE-MS analysis of the human urinary proteome for biomarker discovery and disease diagnostics. Proteomics - Clinical Applications 2(7-8), 964–973 (2008)
7. Jantos-Siwy, J., Schiffer, E., Brand, K., Schumann, G., Rossing, K., Delles, C., Mischak, H., Metzger, J.: Quantitative urinary proteome analysis for biomarker evaluation in chronic kidney disease. Journal of Proteome Research 8(1), 268–281 (2009)
8. Manocha, S., Girolami, M.: An empirical analysis of the probabilistic k-nearest neighbour classifier. Pattern Recognition Letters 28(13), 1818–1824 (2007)
9. Holmes, C.C., Adams, N.M.: A probabilistic nearest neighbour method for statistical pattern recognition. J. R. Statist. Soc. B 64(2), 295–306 (2002)

10. Everson, R.M., Fieldsend, J.E.: A variable metric probabilistic k-nearest-neighbours classifier. In: Yang, Z.R., Yin, H., Everson, R.M. (eds.) IDEAL 2004. LNCS, vol. 3177, pp. 654–659. Springer, Heidelberg (2004)
11. Bae, K., Mallick, B.K.: Gene selection using a two-level hierarchical Bayesian model. Bioinformatics 20(18), 3423–3430 (2004)
12. Albert, J., Chib, S.: Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88, 669–679 (1993)