An Integrated Generative and Discriminative Bayesian Model for Binary Classification

Keith Harris[1] & Mark Girolami[2]

1. Water and Environment Research Group, School of Engineering, University of Glasgow, UK
2. Inference Group, School of Computing Science, University of Glasgow, UK

October 16th 2010

Introduction

High dimensional data sets typically consist of several thousand covariates and a much smaller number of samples.

Introduction

High dimensional data sets typically consist of several thousand covariates and a much smaller number of samples.

Analysing such data is statistically challenging, as the covariates are highly correlated, which results in unstable parameter estimates and inaccurate prediction.

Introduction

High dimensional data sets typically consist of several thousand covariates and a much smaller number of samples.

Analysing such data is statistically challenging, as the covariates are highly correlated, which results in unstable parameter estimates and inaccurate prediction.

To alleviate this problem, we have developed a statistical model that uses a small number of meta-covariates inferred from the data through a Gaussian mixture model, rather than all the original covariates, to classify samples.

Introduction

High dimensional data sets typically consist of several thousand covariates and a much smaller number of samples.

Analysing such data is statistically challenging, as the covariates are highly correlated, which results in unstable parameter estimates and inaccurate prediction.

To alleviate this problem, we have developed a statistical model that uses a small number of meta-covariates inferred from the data through a Gaussian mixture model, rather than all the original covariates, to classify samples.

The novelty of our approach is that our meta-covariates are formed considering predictor-outcome correlations as well as inter-predictor correlations.

Possible benefits

This idea was partly inspired by recent empirical research that has shown that optimum predictive performance often corresponds to an intermediate trade-off between the purely generative and purely discriminative approaches to classification.

Possible benefits

This idea was partly inspired by recent empirical research that has shown that optimum predictive performance often corresponds to an intermediate trade-off between the purely generative and purely discriminative approaches to classification.

The main advantage over using a sparse classification model is that we can extract a much larger subset of covariates with essential predictive power and partition this subset into groups, within which the covariates are similar.

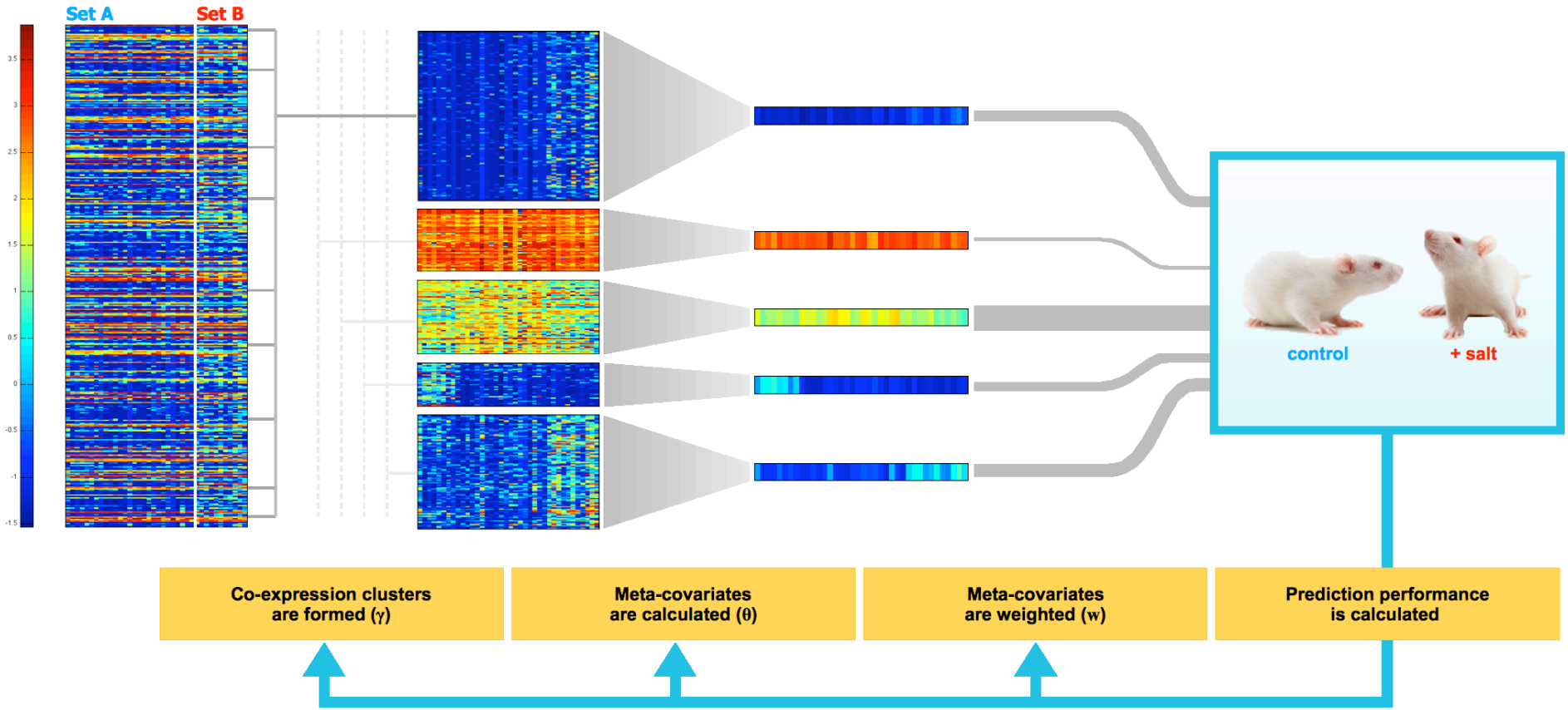
Possible benefits

This idea was partly inspired by recent empirical research that has shown that optimum predictive performance often corresponds to an intermediate trade-off between the purely generative and purely discriminative approaches to classification.

The main advantage over using a sparse classification model is that we can extract a much larger subset of covariates with essential predictive power and partition this subset into groups, within which the covariates are similar.

Moreover, our meta-covariates have a natural ordering and interpretation as increasingly predictive response-relevant clusters.

Model overview

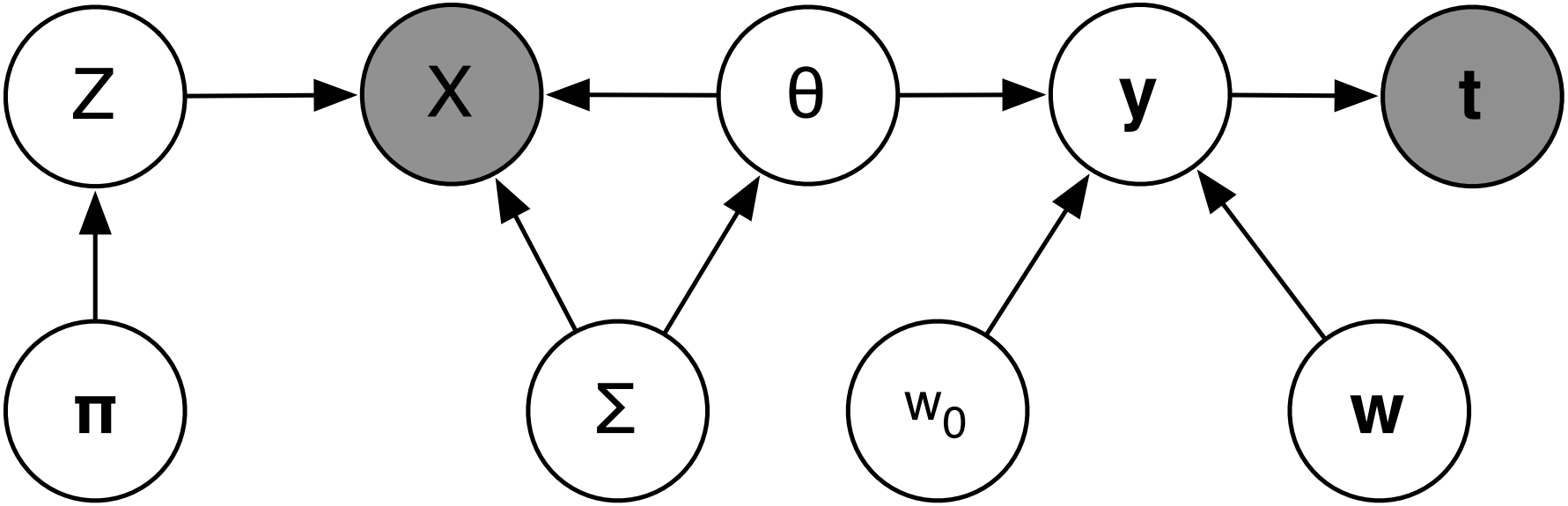


Here, the meta-covariate method is applied to gene expression data. Co-expression clusters are identified and represented by its mean. Each cluster mean is assigned a weight according to its ability to distinguish between set A and set B data.

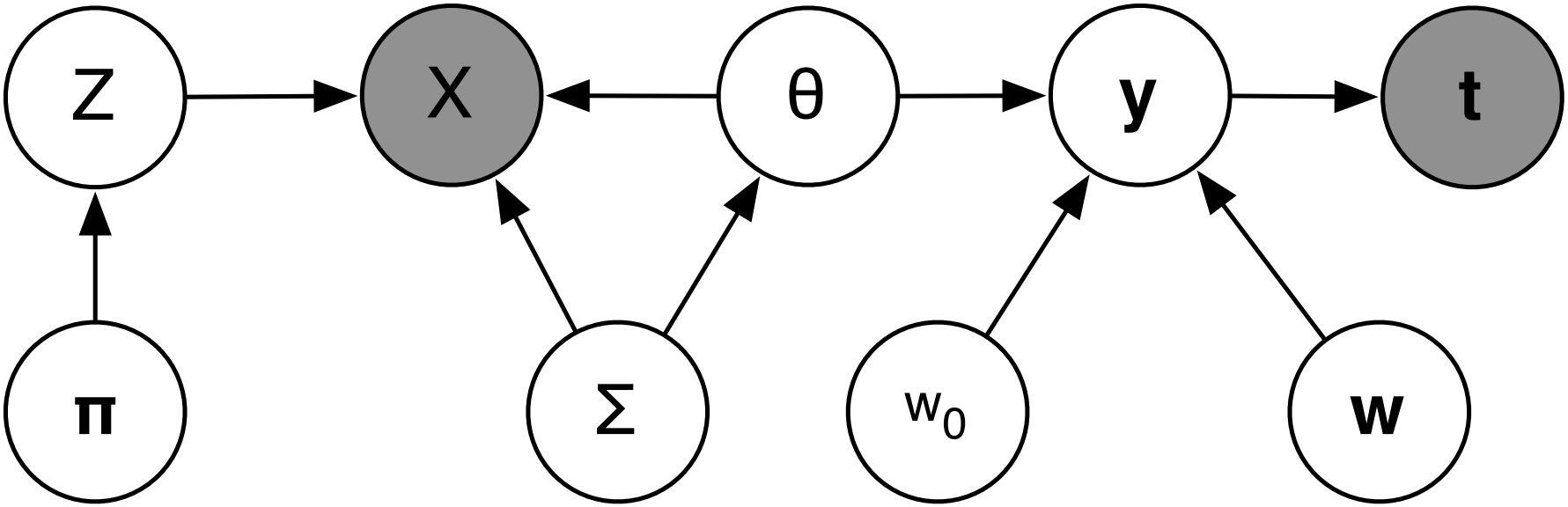
Model notation

X	Design matrix	$N \times D$	$X = [\mathbf{x}_1, \ldots, \mathbf{x}_D]$
$\mathbf{t}$	Response vector	$N \times 1$	$t_n \in \{0, 1\}$
θ	Matrix of clustering mean parameters θ_{kn}	$K \times N$	K *meta-covariates* $\boldsymbol{\theta}_k$
Σ	Matrix of clustering variance parameters	$K \times N$	σ^2_{kn}
$\boldsymbol{\pi}$	Vector of mixing coefficients	$K \times 1$	π_k
Z	Matrix of clustering latent variables	$D \times K$	$z_{dk} \in \{0, 1\}$
w_0	Regression bias parameter	1×1	*Scalar intercept*
$\mathbf{w}$	Vector of regression coefficients	$K \times 1$	w_k
$\mathbf{y}$	Vector of classification auxiliary variables	$N \times 1$	y_n

Conditional dependency structure



Conditional dependency structure



Joint distribution:

$$p(\mathbf{t}, \mathbf{y}, X, Z, \pi, \theta, \Sigma, w_0, \mathbf{w}) = p(\mathbf{t}, \mathbf{y}|\theta, w_0, \mathbf{w})p(X, Z|\pi, \theta, \Sigma)p(\pi)p(\theta|\Sigma)p(\Sigma)p(w_0)p(\mathbf{w}).$$

Model components

Generative component:

$$p(X, Z|\boldsymbol{\pi}, \theta, \Sigma) = \prod_{d=1}^{D} \prod_{k=1}^{K} \pi_k^{z_{dk}} \mathcal{N}(\mathbf{x}_d|\boldsymbol{\theta}_k, \Sigma_k)^{z_{dk}}, \text{ where } \Sigma_k = \text{diag}(\sigma_{k1}^2, \ldots, \sigma_{kN}^2).$$

Model components

Generative component:

$$p(X, Z|\boldsymbol{\pi}, \theta, \Sigma) = \prod_{d=1}^{D} \prod_{k=1}^{K} \pi_k^{z_{dk}} \mathcal{N}(\mathbf{x}_d|\boldsymbol{\theta}_k, \Sigma_k)^{z_{dk}}, \text{ where } \Sigma_k = \text{diag}(\sigma_{k1}^2, \ldots, \sigma_{kN}^2).$$

Discriminative component:

$$p(\mathbf{t}, \mathbf{y}|\theta, w_0, \mathbf{w}) = \prod_{n=1}^{N} p(t_n|y_n) p(y_n|\boldsymbol{\theta}_n, w_0, \mathbf{w}), \text{ where}$$

$$p(t_n|y_n) = \begin{cases} \delta(y_n > 0) & \text{if } t_n = 1 \\ \delta(y_n \leq 0) & \text{otherwise} \end{cases} \text{ and } p(y_n|\boldsymbol{\theta}_n, w_0, \mathbf{w}) = \mathcal{N}(y_n|w_0 + \mathbf{w}^T \boldsymbol{\theta}_n, 1).$$

Model components

Generative component:

$$p(X, Z|\pi, \theta, \Sigma) = \prod_{d=1}^{D} \prod_{k=1}^{K} \pi_k^{z_{dk}} \mathcal{N}(\mathbf{x}_d|\boldsymbol{\theta}_k, \Sigma_k)^{z_{dk}}, \text{ where } \Sigma_k = \text{diag}(\sigma_{k1}^2, \ldots, \sigma_{kN}^2).$$

Discriminative component:

$$p(\mathbf{t}, \mathbf{y}|\theta, w_0, \mathbf{w}) = \prod_{n=1}^{N} p(t_n|y_n) p(y_n|\boldsymbol{\theta}_n, w_0, \mathbf{w}), \text{ where}$$

$$p(t_n|y_n) = \begin{cases} \delta(y_n > 0) & \text{if } t_n = 1 \\ \delta(y_n \leq 0) & \text{otherwise} \end{cases} \text{ and } p(y_n|\boldsymbol{\theta}_n, w_0, \mathbf{w}) = \mathcal{N}(y_n|w_0 + \mathbf{w}^T \boldsymbol{\theta}_n, 1).$$

Prior distributions:

$$p(\pi) = \text{const}, \; p(\theta|\Sigma) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{\theta}_k|\boldsymbol{\theta}_0, h\Sigma_k),$$

$$p(\Sigma) = \prod_{k=1}^{K} \prod_{n=1}^{N} \text{Inv-Gamma}\left(\sigma_{kn}^2 \middle| \nu, \xi\right), \; p(w_0) = \mathcal{N}(w_0|0, l_0) \text{ and } \; p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, l\mathbf{I}).$$

EM algorithm: the E-step

$$\gamma(z_{dk}) = \frac{\pi_k \left(\prod_n \sigma_{kn}^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\sum_n \frac{(x_{nd}-\theta_{kn})^2}{\sigma_{kn}^2}\right\}}{\sum_j \pi_j \left(\prod_n \sigma_{jn}^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\sum_n \frac{(x_{nd}-\theta_{jn})^2}{\sigma_{jn}^2}\right\}},$$

EM algorithm: the E-step

$$\gamma(z_{dk}) = \frac{\pi_k \left(\prod_n \sigma_{kn}^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\sum_n \frac{(x_{nd}-\theta_{kn})^2}{\sigma_{kn}^2}\right\}}{\sum_j \pi_j \left(\prod_n \sigma_{jn}^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\sum_n \frac{(x_{nd}-\theta_{jn})^2}{\sigma_{jn}^2}\right\}},$$

$$E(y_n) = \begin{cases} w_0 + \mathbf{w}^T\boldsymbol{\theta}_n + \frac{\phi(-w_0-\mathbf{w}^T\boldsymbol{\theta}_n)}{1-\Phi(-w_0-\mathbf{w}^T\boldsymbol{\theta}_n)} & \text{if } t_n = 1 \\ w_0 + \mathbf{w}^T\boldsymbol{\theta}_n - \frac{\phi(-w_0-\mathbf{w}^T\boldsymbol{\theta}_n)}{\Phi(-w_0-\mathbf{w}^T\boldsymbol{\theta}_n)} & \text{otherwise.} \end{cases}$$

EM algorithm: the M-step

$$\theta_{kn} = \frac{\left(E(y_n) - w_0 - \sum_{k' \neq k} w_{k'} \theta_{k'n}\right) w_k + \frac{1}{\sigma_{kn}^2} \left(\sum_{d=1}^{D} \gamma(z_{dk}) x_{nd} + \frac{\theta_{0n}}{h}\right)}{w_k^2 + \frac{1}{\sigma_{kn}^2} \left(\sum_{d=1}^{D} \gamma(z_{dk}) + \frac{1}{h}\right)},$$

EM algorithm: the M-step

$$\theta_{kn} = \frac{\left(E(y_n) - w_0 - \sum_{k' \neq k} w_{k'} \theta_{k'n}\right) w_k + \frac{1}{\sigma_{kn}^2} \left(\sum_{d=1}^{D} \gamma(z_{dk}) x_{nd} + \frac{\theta_{0n}}{h}\right)}{w_k^2 + \frac{1}{\sigma_{kn}^2} \left(\sum_{d=1}^{D} \gamma(z_{dk}) + \frac{1}{h}\right)},$$

$$\sigma_{kn}^2 = \frac{\sum_{d=1}^{D} \gamma(z_{dk})(x_{nd} - \theta_{kn})^2 + \frac{1}{h}(\theta_{kn} - \theta_{0n})^2 + 2\xi}{\sum_{d=1}^{D} \gamma(z_{dk}) + 2\nu + 3},$$

EM algorithm: the M-step

$$\theta_{kn} = \frac{\left(E(y_n) - w_0 - \sum_{k' \neq k} w_{k'} \theta_{k'n}\right) w_k + \frac{1}{\sigma_{kn}^2}\left(\sum_{d=1}^{D} \gamma(z_{dk}) x_{nd} + \frac{\theta_{0n}}{h}\right)}{w_k^2 + \frac{1}{\sigma_{kn}^2}\left(\sum_{d=1}^{D} \gamma(z_{dk}) + \frac{1}{h}\right)},$$

$$\sigma_{kn}^2 = \frac{\sum_{d=1}^{D} \gamma(z_{dk})(x_{nd} - \theta_{kn})^2 + \frac{1}{h}(\theta_{kn} - \theta_{0n})^2 + 2\xi}{\sum_{d=1}^{D} \gamma(z_{dk}) + 2\nu + 3},$$

$$\pi_k = \frac{1}{D} \sum_{d=1}^{D} \gamma(z_{dk}),$$

EM algorithm: the M-step

$$\theta_{kn} = \frac{\left(E(y_n) - w_0 - \sum_{k' \neq k} w_{k'} \theta_{k'n}\right) w_k + \frac{1}{\sigma_{kn}^2} \left(\sum_{d=1}^{D} \gamma(z_{dk}) x_{nd} + \frac{\theta_{0n}}{h}\right)}{w_k^2 + \frac{1}{\sigma_{kn}^2} \left(\sum_{d=1}^{D} \gamma(z_{dk}) + \frac{1}{h}\right)},$$

$$\sigma_{kn}^2 = \frac{\sum_{d=1}^{D} \gamma(z_{dk})(x_{nd} - \theta_{kn})^2 + \frac{1}{h}(\theta_{kn} - \theta_{0n})^2 + 2\xi}{\sum_{d=1}^{D} \gamma(z_{dk}) + 2\nu + 3},$$

$$\pi_k = \frac{1}{D} \sum_{d=1}^{D} \gamma(z_{dk}),$$

$$w_0 = \frac{\sum_{n=1}^{N} \left(E(y_n) - \sum_{k'=1}^{K} w_{k'} \theta_{k'n}\right)}{N + \frac{1}{l_0}},$$

EM algorithm: the M-step

$$\theta_{kn} = \frac{\left(E(y_n) - w_0 - \sum_{k' \neq k} w_{k'} \theta_{k'n}\right) w_k + \frac{1}{\sigma_{kn}^2} \left(\sum_{d=1}^{D} \gamma(z_{dk}) x_{nd} + \frac{\theta_{0n}}{h}\right)}{w_k^2 + \frac{1}{\sigma_{kn}^2} \left(\sum_{d=1}^{D} \gamma(z_{dk}) + \frac{1}{h}\right)},$$

$$\sigma_{kn}^2 = \frac{\sum_{d=1}^{D} \gamma(z_{dk})(x_{nd} - \theta_{kn})^2 + \frac{1}{h}(\theta_{kn} - \theta_{0n})^2 + 2\xi}{\sum_{d=1}^{D} \gamma(z_{dk}) + 2\nu + 3},$$

$$\pi_k = \frac{1}{D} \sum_{d=1}^{D} \gamma(z_{dk}),$$

$$w_0 = \frac{\sum_{n=1}^{N} \left(E(y_n) - \sum_{k'=1}^{K} w_{k'} \theta_{k'n}\right)}{N + \frac{1}{l_0}},$$

$$\mathbf{w} = \left(\theta\theta^T + \frac{1}{l}\mathrm{I}\right)^{-1} \theta \left(E(\mathbf{y}) - w_0 \mathbf{1}\right).$$

EM algorithm: the M-step

$$\theta_{kn} = \frac{\left(E(y_n) - w_0 - \sum_{k' \neq k} w_{k'} \theta_{k'n}\right) w_k + \frac{1}{\sigma_{kn}^2}\left(\sum_{d=1}^{D} \gamma(z_{dk}) x_{nd} + \frac{\theta_{0n}}{h}\right)}{w_k^2 + \frac{1}{\sigma_{kn}^2}\left(\sum_{d=1}^{D} \gamma(z_{dk}) + \frac{1}{h}\right)},$$

$$\sigma_{kn}^2 = \frac{\sum_{d=1}^{D} \gamma(z_{dk})(x_{nd} - \theta_{kn})^2 + \frac{1}{h}(\theta_{kn} - \theta_{0n})^2 + 2\xi}{\sum_{d=1}^{D} \gamma(z_{dk}) + 2\nu + 3},$$

$$\pi_k = \frac{1}{D} \sum_{d=1}^{D} \gamma(z_{dk}),$$

$$w_0 = \frac{\sum_{n=1}^{N}\left(E(y_n) - \sum_{k'=1}^{K} w_{k'} \theta_{k'n}\right)}{N + \frac{1}{l_0}},$$

$$\mathbf{w} = \left(\theta\theta^T + \frac{1}{l}\mathrm{I}\right)^{-1} \theta \left(E(\mathbf{y}) - w_0 \mathbf{1}\right).$$

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

Insights into salt-sensitive hypertension

Hypertension is a common precursor to cardiovascular disease and is often exacerbated by increased dietary intake of sodium.

Insights into salt-sensitive hypertension

Hypertension is a common precursor to cardiovascular disease and is often exacerbated by increased dietary intake of sodium.

The stroke prone spontaneously hypertensive rat (SP) is an excellent model of human essential hypertension that exhibits salt sensitivity.

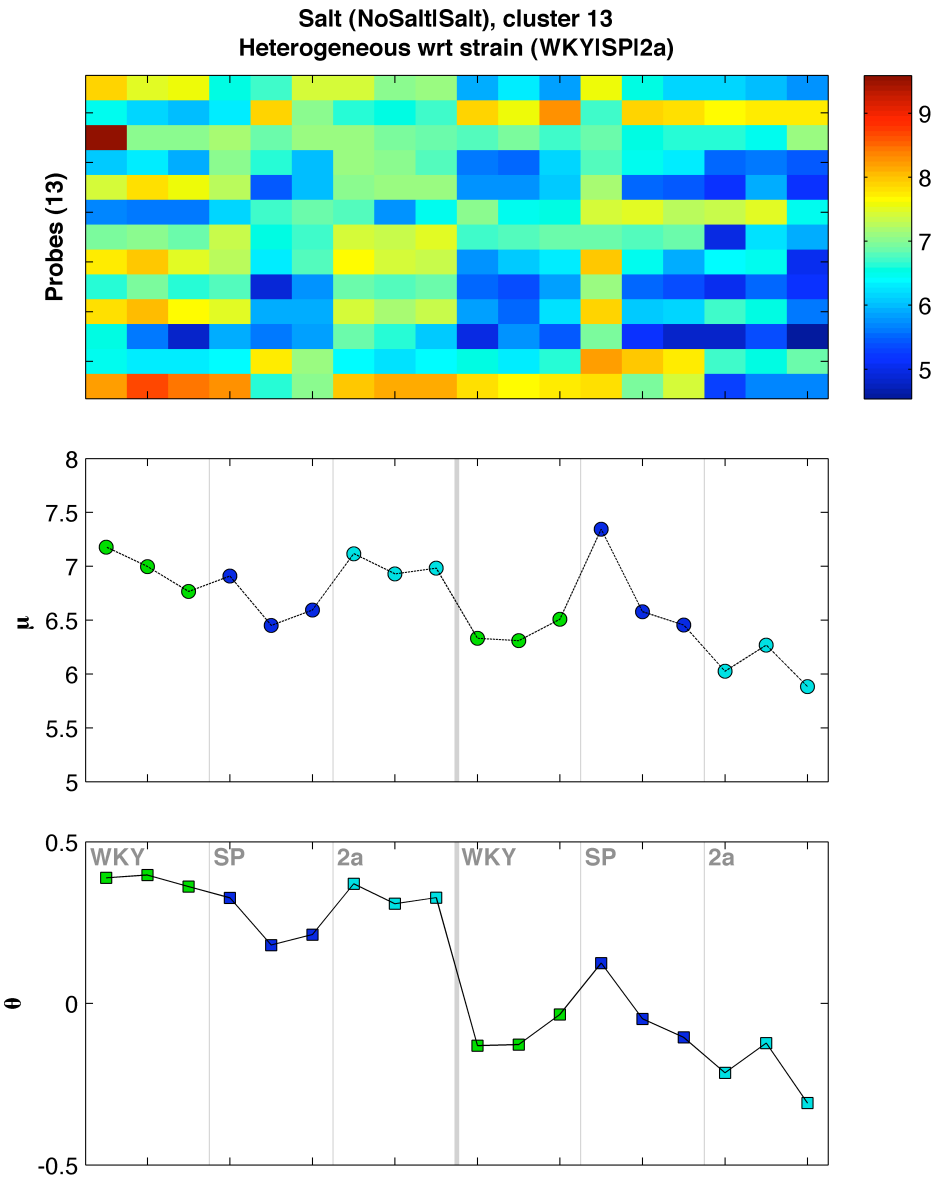
Insights into salt-sensitive hypertension

Hypertension is a common precursor to cardiovascular disease and is often exacerbated by increased dietary intake of sodium.

The stroke prone spontaneously hypertensive rat (SP) is an excellent model of human essential hypertension that exhibits salt sensitivity.

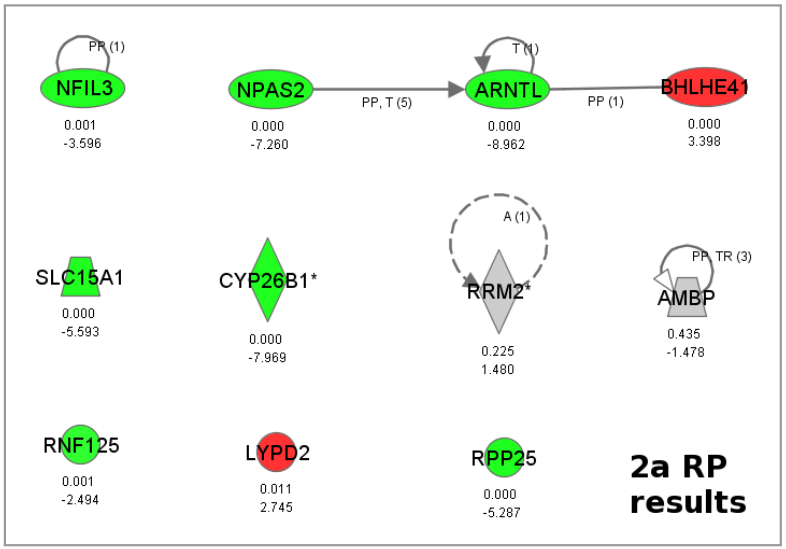
By analysing microarray data of the SP, a salt-insensitive strain (WKY) and an intermediate congenic strain (2a), the genes and pathways that influence salt-sensitive hypertension can be elucidated.

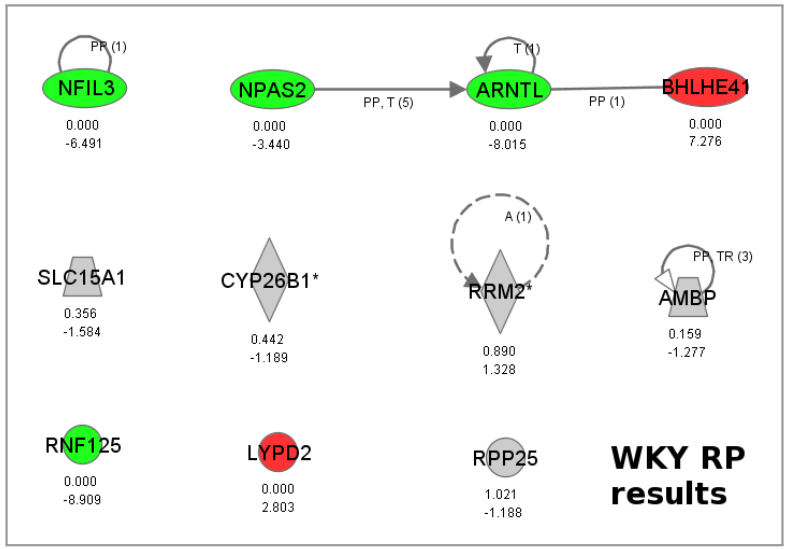
A highly influential cluster of 13 genes

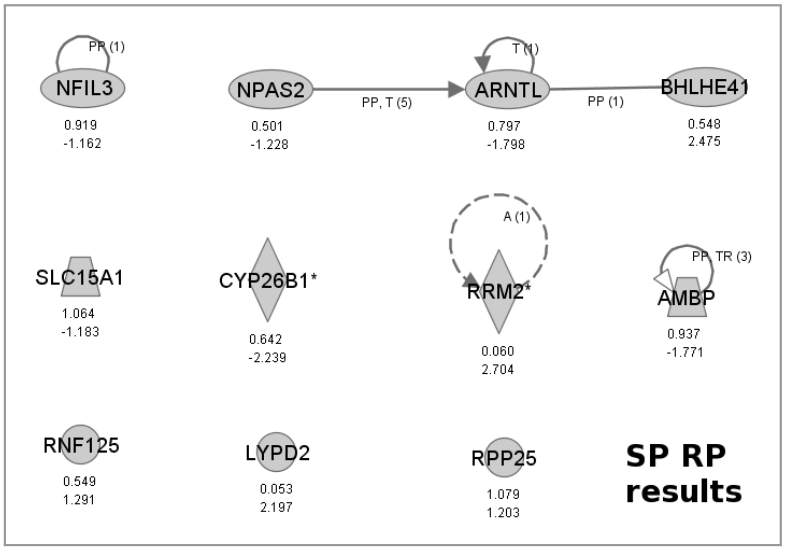


Characterising cluster 13 of the meta-covariate model with $K = 20$ clusters suggested by BIC. The expression of all 13 genes (left top); the mean expression (left middle) and alternative θ (meta-covariate) representation (left bottom).

RP analysis of this cluster







The genes in cluster 13, overlaid with 2a (left top), WKY (right top) and SP (left bottom) rank products results: red indicates over-expression on salt, while green indicates under-expression on salt.

Circadian rhythm genes are implicated

Canonical pathway analysis of cluster 13 shows that this cluster is enriched for circadian rhythm genes, implicating circadian rhythm genes as important in differentiating between salt-loaded and non salt-loaded animals.

Circadian rhythm genes are implicated

Canonical pathway analysis of cluster 13 shows that this cluster is enriched for circadian rhythm genes, implicating circadian rhythm genes as important in differentiating between salt-loaded and non salt-loaded animals.

This is relevant, given that these noctural animals exhibit increased hypertension during the night, and this difference is exacerbated on salt-loading.

Circadian rhythm genes are implicated

Canonical pathway analysis of cluster 13 shows that this cluster is enriched for circadian rhythm genes, implicating circadian rhythm genes as important in differentiating between salt-loaded and non salt-loaded animals.

This is relevant, given that these noctural animals exhibit increased hypertension during the night, and this difference is exacerbated on salt-loading.

Now we consider how these genes are related to the strains. We saw from the RP analysis of cluster 13 that there are differences on salt-loading in the 2a and WKY strains, but not in the SP strain.

Circadian rhythm genes are implicated

Canonical pathway analysis of cluster 13 shows that this cluster is enriched for circadian rhythm genes, implicating circadian rhythm genes as important in differentiating between salt-loaded and non salt-loaded animals.

This is relevant, given that these noctural animals exhibit increased hypertension during the night, and this difference is exacerbated on salt-loading.

Now we consider how these genes are related to the strains. We saw from the RP analysis of cluster 13 that there are differences on salt-loading in the 2a and WKY strains, but not in the SP strain.

We can therefore hypothesize that the genes in the most influential meta-covariate cluster are protective against hypertension in response to an increase in dietary sodium.

Extension to Gibbs sampling

The full conditional distribution for π:

$$\text{Dirichlet}\left(\sum_{d=1}^{D} z_{d1} + 1, \cdots, \sum_{d=1}^{D} z_{dK} + 1\right).$$

Extension to Gibbs sampling

The full conditional distribution for π:

$$\text{Dirichlet}\left(\sum_{d=1}^{D} z_{d1} + 1, \cdots, \sum_{d=1}^{D} z_{dK} + 1\right).$$

The full conditional distribution for θ_{kn}:

$$\mathcal{N}((e_n w_k + m_{kn})v_{kn}, v_{kn}), \text{ where } e_n = y_n - w_0 - \sum_{k' \neq k} w_{k'}\theta_{k'n},$$

$$m_{kn} = \frac{1}{\sigma_{kn}^2}\left(\sum_{d=1}^{D} z_{dk}x_{nd} + \frac{\theta_{0n}}{h}\right) \text{ and } v_{kn} = \left[w_k^2 + \frac{1}{\sigma_{kn}^2}\left(\sum_{d=1}^{D} z_{dk} + \frac{1}{h}\right)\right]^{-1}.$$

Extension to Gibbs sampling

The full conditional distribution for π:

$$\text{Dirichlet}\left(\sum_{d=1}^{D} z_{d1} + 1, \cdots, \sum_{d=1}^{D} z_{dK} + 1\right).$$

The full conditional distribution for θ_{kn}:

$$\mathcal{N}((e_n w_k + m_{kn})v_{kn}, v_{kn}), \text{ where } e_n = y_n - w_0 - \sum_{k' \neq k} w_{k'}\theta_{k'n},$$

$$m_{kn} = \frac{1}{\sigma_{kn}^2}\left(\sum_{d=1}^{D} z_{dk}x_{nd} + \frac{\theta_{0n}}{h}\right) \text{ and } v_{kn} = \left[w_k^2 + \frac{1}{\sigma_{kn}^2}\left(\sum_{d=1}^{D} z_{dk} + \frac{1}{h}\right)\right]^{-1}.$$

The full conditional distribution for σ_{kn}^2:

$$\text{Inv-Gamma}\left(\frac{1}{2}\sum_{d=1}^{D} z_{dk} + \nu + \frac{1}{2}, \frac{1}{2}\sum_{d=1}^{D} z_{dk}(x_{nd} - \theta_{kn})^2 + \frac{1}{2h}(\theta_{kn} - \theta_{0n})^2 + \xi\right).$$

Extension to Gibbs sampling continued

The full conditional distribution for w_0:

$$\mathcal{N}\left(\frac{\sum_{n=1}^{N}\left(y_n - \sum_{k'=1}^{K} w_{k'}\theta_{k'n}\right)}{N + \frac{1}{l_0}}, \left(N + \frac{1}{l_0}\right)^{-1}\right).$$

Extension to Gibbs sampling continued

The full conditional distribution for w_0:

$$\mathcal{N}\left(\frac{\sum_{n=1}^{N}\left(y_n - \sum_{k'=1}^{K} w_{k'}\theta_{k'n}\right)}{N + \frac{1}{l_0}}, \left(N + \frac{1}{l_0}\right)^{-1}\right).$$

The full conditional distribution for $\mathbf{w}$:

$$\mathcal{N}\left(\left(\theta\theta^T + \frac{1}{l}\mathrm{I}\right)^{-1}\theta\left(\mathbf{y} - w_0\mathbf{1}\right), \left(\theta\theta^T + \frac{1}{l}\mathrm{I}\right)^{-1}\right).$$

Extension to Gibbs sampling continued

The full conditional distribution for w_0:

$$\mathcal{N}\left(\frac{\sum_{n=1}^{N}\left(y_n - \sum_{k'=1}^{K} w_{k'}\theta_{k'n}\right)}{N + \frac{1}{l_0}}, \left(N + \frac{1}{l_0}\right)^{-1}\right).$$

The full conditional distribution for $\mathbf{w}$:

$$\mathcal{N}\left(\left(\theta\theta^T + \frac{1}{l}\mathrm{I}\right)^{-1}\theta\left(\mathbf{y} - w_0\mathbf{1}\right), \left(\theta\theta^T + \frac{1}{l}\mathrm{I}\right)^{-1}\right).$$

The full conditional distribution for $\mathbf{z}_d$: Multinomial($n_{\text{trials}}, p_1, \ldots, p_K$), where $n_{\text{trials}} = 1$ and $p_k = \gamma(z_{dk})$.

Extension to Gibbs sampling continued

The full conditional distribution for w_0:

$$\mathcal{N}\left(\frac{\sum_{n=1}^{N}\left(y_n - \sum_{k'=1}^{K} w_{k'}\theta_{k'n}\right)}{N + \frac{1}{l_0}}, \left(N + \frac{1}{l_0}\right)^{-1}\right).$$

The full conditional distribution for $\mathbf{w}$:

$$\mathcal{N}\left(\left(\theta\theta^T + \frac{1}{l}\mathrm{I}\right)^{-1}\theta\left(\mathbf{y} - w_0\mathbf{1}\right), \left(\theta\theta^T + \frac{1}{l}\mathrm{I}\right)^{-1}\right).$$

The full conditional distribution for $\mathbf{z}_d$: Multinomial$(n_{\text{trials}}, p_1, \ldots, p_K)$, where $n_{\text{trials}} = 1$ and $p_k = \gamma(z_{dk})$.

The full conditional distribution for y_n:

$$p(y_n|\mathbf{y}_{-n}, \pi, \theta, \Sigma, w_0, \mathbf{w}, \mathbf{t}, X, Z) \propto \begin{cases} \delta(y_n > 0)\mathcal{N}\left(y_n|w_0 + \mathbf{w}^T\boldsymbol{\theta}_n, 1\right) & \text{if } t_n = 1 \\ \delta(y_n \leq 0)\mathcal{N}\left(y_n|w_0 + \mathbf{w}^T\boldsymbol{\theta}_n, 1\right) & \text{otherwise.} \end{cases}$$

Posterior predictive distribution

We obtain the predictive classification of a new observation t^*, conditioning on the test point $\mathbf{x}^*$, using the Monte-Carlo estimate:

$$P(t^* = 1|\mathbf{x}^*, \mathbf{t}, X) \approx \frac{1}{m} \sum_{t=1}^{m} \Phi \left(w_0^{(t)} + \mathbf{w}^{(t)^T} \boldsymbol{\theta}^{*(t)} \right),$$

where $w_0^{(t)}$, $\mathbf{w}^{(t)}$ and $\boldsymbol{\theta}^{*(t)}$ are the MCMC samples of the parameters w_0, $\mathbf{w}$ and $\boldsymbol{\theta}^*$.

Posterior predictive distribution

We obtain the predictive classification of a new observation t^*, conditioning on the test point $\mathbf{x}^*$, using the Monte-Carlo estimate:

$$P(t^* = 1|\mathbf{x}^*, \mathbf{t}, X) \approx \frac{1}{m} \sum_{t=1}^{m} \Phi\left(w_0^{(t)} + \mathbf{w}^{(t)^T} \boldsymbol{\theta}^{*(t)}\right),$$

where $w_0^{(t)}$, $\mathbf{w}^{(t)}$ and $\boldsymbol{\theta}^{*(t)}$ are the MCMC samples of the parameters w_0, $\mathbf{w}$ and $\boldsymbol{\theta}^*$.

Thus, we also need to sample θ_k^* from:

$$\mathcal{N}\left(\frac{\sum_{d=1}^{D} z_{dk} x_d^* + \frac{\theta_0^*}{h}}{\sum_{d=1}^{D} z_{dk} + \frac{1}{h}}, \left[\frac{1}{\sigma_k^{*2}}\left(\sum_{d=1}^{D} z_{dk} + \frac{1}{h}\right)\right]^{-1}\right),$$

and σ_k^{*2} from:

$$\text{Inv-Gamma}\left(\frac{1}{2}\sum_{d=1}^{D} z_{dk} + \nu + \frac{1}{2}, \frac{1}{2}\sum_{d=1}^{D} z_{dk}(x_d^* - \theta_k^*)^2 + \frac{1}{2h}(\theta_k^* - \theta_0^*)^2 + \xi\right).$$

Hedenfalk's breast cancer data

We apply our method to a publicly available breast cancer dataset from patients carrying mutations in the predisposing genes, BRCA1 or BRCA2, and from patients not expected to carry either of these hereditary predisposing mutations.

$D = 3226$, $N = 22$ (7 BRCA1, 8 BRCA2 and 7 sporadic).

A Wilcoxon rank-sum test was used to provide a ranking of the features based on their p-value. Setting a threshold of 10% the number of features was reduced to 626.

Hedenfalk's breast cancer data

We apply our method to a publicly available breast cancer dataset from patients carrying mutations in the predisposing genes, BRCA1 or BRCA2, and from patients not expected to carry either of these hereditary predisposing mutations.

$D = 3226$, $N = 22$ (7 BRCA1, 8 BRCA2 and 7 sporadic).

A Wilcoxon rank-sum test was used to provide a ranking of the features based on their p-value. Setting a threshold of 10% the number of features was reduced to 626.

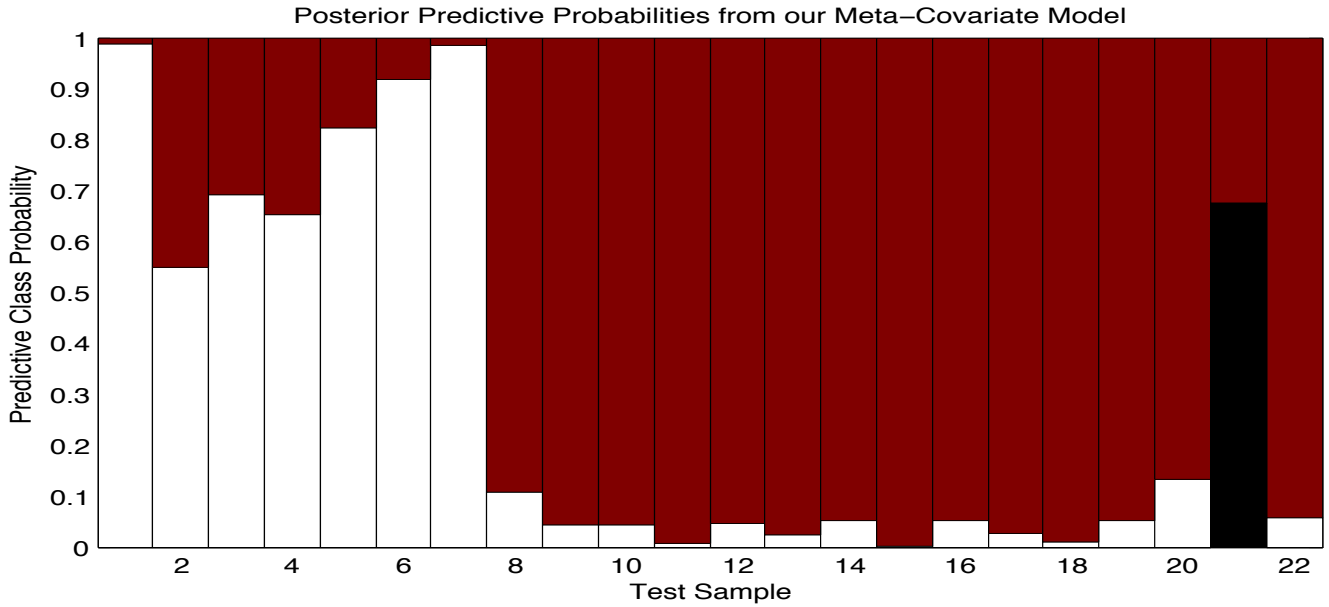
We use our method to classify BRCA1 versus the others and compare our method to a Bayesian sparse probit regression model.

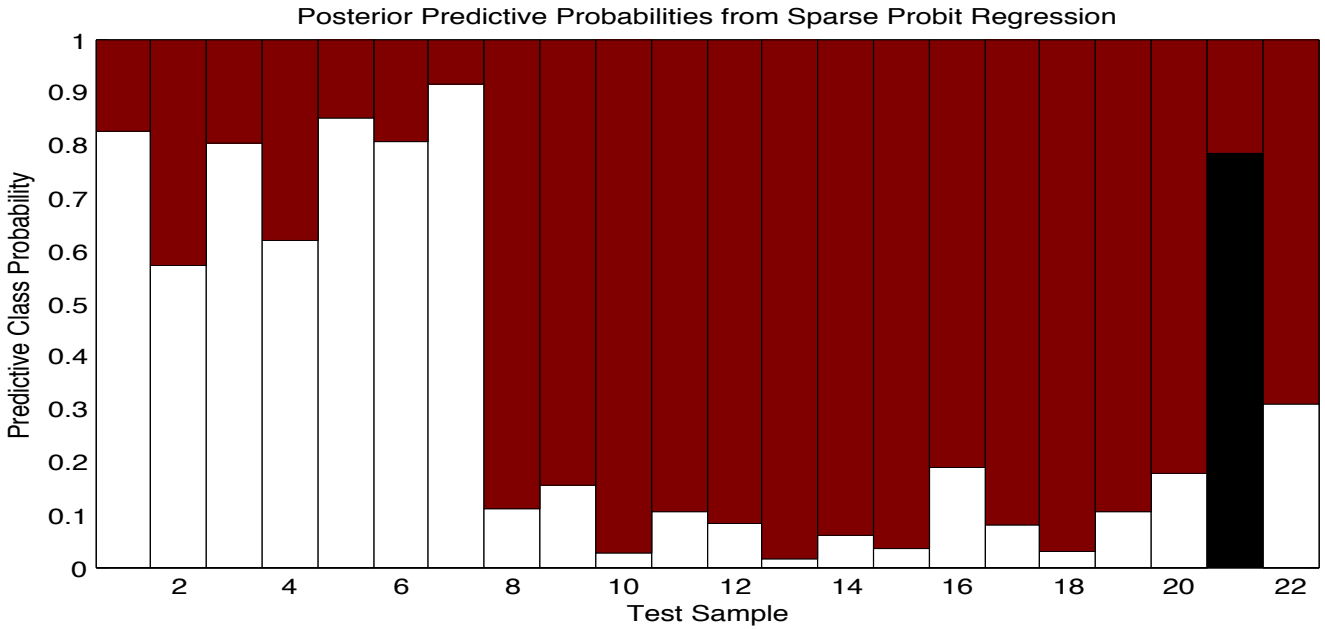
We initialised our Gibbs sampler using the EM algorithm.

We ran the Gibbs samplers of both methods for 100000 iterations and discard the first half of each chain as burn-in.

We compared the methods using leave-one-out cross validation.

Plots of the posterior predictive probabilities





Conclusions

Our experimental results thus indicate that our Gibbs sampling approach of inferring meta-covariates in classification has competitive performance with Bayesian sparse probit regression.

Conclusions

Our experimental results thus indicate that our Gibbs sampling approach of inferring meta-covariates in classification has competitive performance with Bayesian sparse probit regression.

Moreover, our approach can be naturally extended to multiclass classification.

Conclusions

Our experimental results thus indicate that our Gibbs sampling approach of inferring meta-covariates in classification has competitive performance with Bayesian sparse probit regression.

Moreover, our approach can be naturally extended to multiclass classification.

Future research will focus on applying our methodology to functional magnetic resonance imaging data and developing a Bayesian sampler that can infer directly from the data the optimal number of clusters in our model via an infinite mixture model.

References

Bishop C. M. and Lasserre J. (2007) Generative or discriminative? Getting the best of both worlds. In *Bayesian Statistics*, volume 8, pages 3-24. Oxford University Press.

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

Hedenfalk I., Duggan, D. , Chen Y. D., Radmacher M., Bittner M., Simon R., Meltzer P., Gusterson B., Esteller M., Kallioniemi O. P., Wilfond B., Borg A., and Trent J. (2001) Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine*, 344(8):539-548.

Hopcroft L. E. M., McBride M. W., Harris K. J., Sampson A., McClure J. D., Graham D., Young G., Holyoake T. L., Girolami M. and Dominiczak A. F. (2010) Predictive response-relevant clustering of expression data provides insights into disease processes. *Nucleic Acids Research*, Published online.