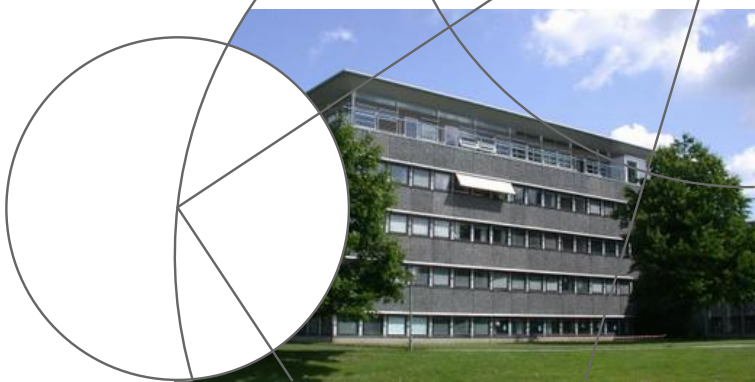
Modeling Tissue Heterogeneity of Test Samples to Improve Class Prediction

Niels Richard Hansen & Martin Vincent
Department of Mathematical Sciences



March 14, 2014

Overview

- Present qPCR data set on miRNA expression from primary cancers and liver biopsies.
- A brief detour around the multinomial group lasso predictor.
- Present a computational method for dealing with heterogeneous tissue composition in biopsy samples.
- Present a general modeling framework for class prediction based on heterogeneous tissue and some preliminary methods and results.

Prediction of primary site

Class description	Resections (primaries)	Liver core biopsies
Breast cancer	17	7 (5/2)
Colorectal cancer	20	12 (8/4)
Gastric/Cardia cancer	18	12 (8/4)
Pancreatic cancer	20	10 (5/5)
Squamous cell cancers (of different origins)	16	12 (6/6)
Hepatocellular carcinoma	17	3
Cholangiocarcinoma	20	4
Subtotal	**128**	**60**
Cirrhotic liver	17	8
Normal liver	20	7
Total	**165**	**75**

Objective: Predict site of primary tumor from liver biopsy.

Misclassification for biopsies from metastases

Principal training data	Number of core biopsies	ANOVA+PAM		Multinomial group lasso	
		Number of miRNAs			
		50	100	50	100
Primaries	0 (0)	81%[a]	77%[a]	77%	74%
	2 (10)	74%	71%	59%	54%
	4 (20)	64%	64%	48%	45%
Artificial	0 (0)	60%	57%	45%	43%
	2 (10)	-[b]	-[b]	39%	41%
	4 (20)	-[b]	-[b]	34%	39%

[a]Constructed as in *Ferracin et al.* J. Pathol., 255, 4353, 2011.

[b]Sample weights not directly supported by ANOVA+PAM.

Multinomial regression

Class variable $Y \in \{1, \ldots, K\}$, $X \in \mathbb{R}^p$

$$P(Y = y \mid X) \propto \exp\left(\sum_i X_i \beta_{iy}\right).$$

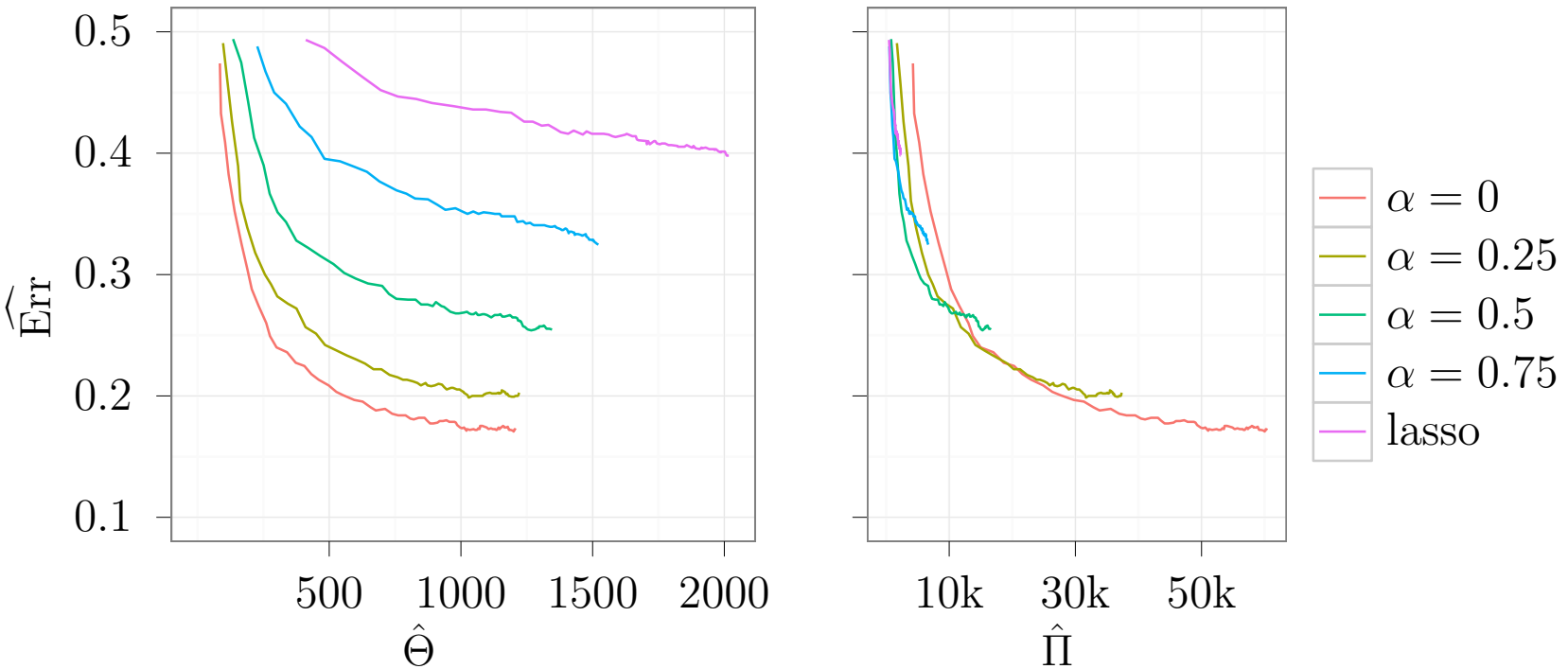
Ordinary lasso objective:

$$\underbrace{\ell(\beta)}_{\text{neg. log-like}} + \lambda \sum_{iy} |\beta_{iy}|.$$

Sparse group lasso objective:

$$\ell(\beta) + \lambda \left((1-\alpha) \sum_i ||\beta_i||_2 + \alpha \sum_{iy} |\beta_{iy}| \right).$$

Multinomial regression - test example



Classification of Amazon reviewers. Group lasso clearly outperforms lasso.

Sparse group lasso implementation in R package `msgl`.

The heterogeneity model

The "standard" model of molecular signatures from heterogenous tissue:

$$\alpha \times \text{primary tumor signature} + (1 - \alpha) \times \text{normal liver signature}$$

Our model, conditionally on class $Y = y$, allows for a non-linear transformation due to qPCR:

$$Z_y = f\left(\alpha f^{-1}(X_y) + (1 - \alpha) f^{-1}(X_0)\right)$$

Model assumption:

$$\alpha \perp\!\!\!\perp X \perp\!\!\!\perp X_0 \mid Y$$

Artificial training data

Based on

$$Z_y = f\left(\alpha f^{-1}(X_y) + (1-\alpha) f^{-1}(X_0)\right)$$

and

- sampling of X_y with replacement from primary signatures for class y
- sampling of X_0 with replacement from liver signatures
- and sampling of α from the Beta(2,2)-distribution

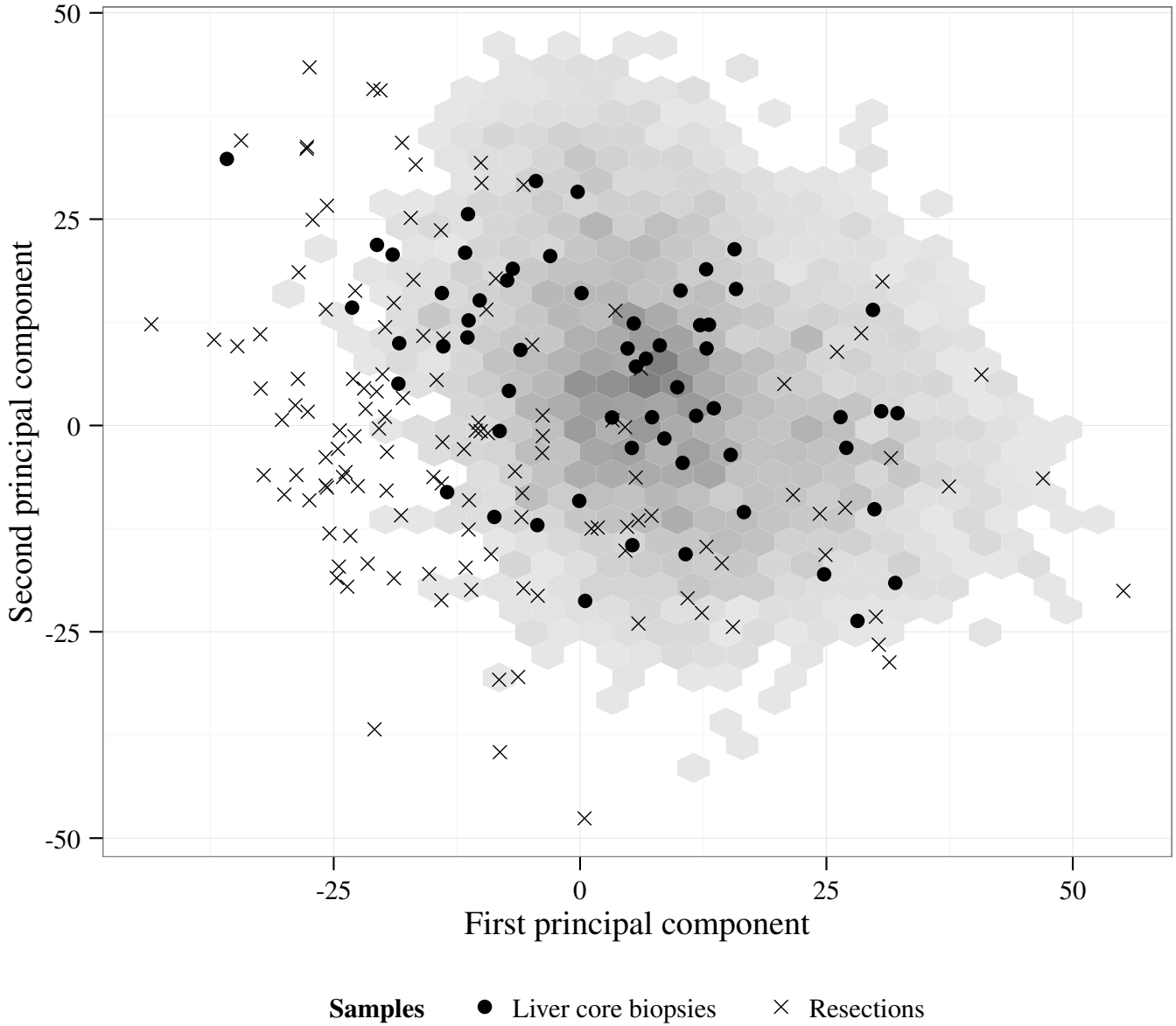
we artificially sampled Z_y used to train the multinomial predictor.

In the paper we considered two choices of f: the identity or

$$f_i(x_i) = -1.7 \log x_i$$

corresponding to a PCR amplification efficiency of 80%.

Comparison of real and artificial data



Misclassification for biopsies from metastases

Principal training data	Number of core biopsies	ANOVA+PAM		Multinomial group lasso	
		Number of miRNAs			
		50	100	50	100
Primaries	0 (0)	81%[a]	77%[a]	77%	74%
	2 (10)	74%	71%	59%	54%
	4 (20)	64%	64%	48%	45%
Artificial	0 (0)	60%	57%	45%	43%
	2 (10)	-[b]	-[b]	39%	41%
	4 (20)	-[b]	-[b]	34%	39%

[a]Constructed as in *Ferracin et al.* J. Pathol., 255, 4353, 2011.

[b]Sample weights not directly supported by ANOVA+PAM.

A general modeling approach

Consider a triple of variables (X, Z, Y) with $X, Z \in \mathbb{R}^p$ and $Y \in \{1, \ldots, K\}$ the class label.

- Observations of (X, Y) are available for construction of a predictor,
- but observations of Z are available for prediction.

With π_0 the joint distribution of (Z, Y), then if $Y \perp\!\!\!\perp Z \mid X$

$$\begin{aligned} \pi_0(z, y) &= \int p(z|x)\pi(x, y)\mathrm{d}x \\ \pi_0(y|z) &= \int \pi(y|x)q(x|z)\mathrm{d}x. \end{aligned}$$

Our previous solution

$$\pi_0(z, y) = \int p(z|x)\pi(x, y)\mathrm{d}x$$

Effectively, we computed estimates $\hat{\pi}(x, y)$ (the empirical distribution), and $\hat{p}(z|x)$ to make a forward simulation from $\hat{\pi}_0(z, y)$.

The forward simulated data were used to fit a model of $\hat{\pi}_0(y|z)$.

An alternative solution

$$\pi_0(y|z) = \int \pi(y|x)q(x|z)\mathrm{d}x$$

Alternatively, we can compute the estimate $\hat{\pi}(y|x)$ and use a Monte Carlo method to compute

$$\hat{\pi}_0(y|z) = \frac{1}{B}\sum_{i=1}^{B} \hat{\pi}(y|x_i)$$

with x_i from a Markov Chain with invariant distribution $q(x|z)$.

This is a backward simulation solution.

Latent Gaussian model

If

$$Z = [X \ \mathbf{X}_{-1}]\alpha + \varepsilon$$

with $\mathbf{X} = [X \ \mathbf{X}_{-1}]$ being a $p \times k$ matrix, and α, ε, $\mathbf{X}$ are independent Gaussian, then

$$\mathbf{X} \mid Z, \alpha \sim \mathcal{N}(\cdot, \cdot)$$

and

$$\alpha \mid Z, \mathbf{X} \sim \mathcal{N}(\cdot, \cdot).$$

This is what is needed to implement the Gibbs sampler.

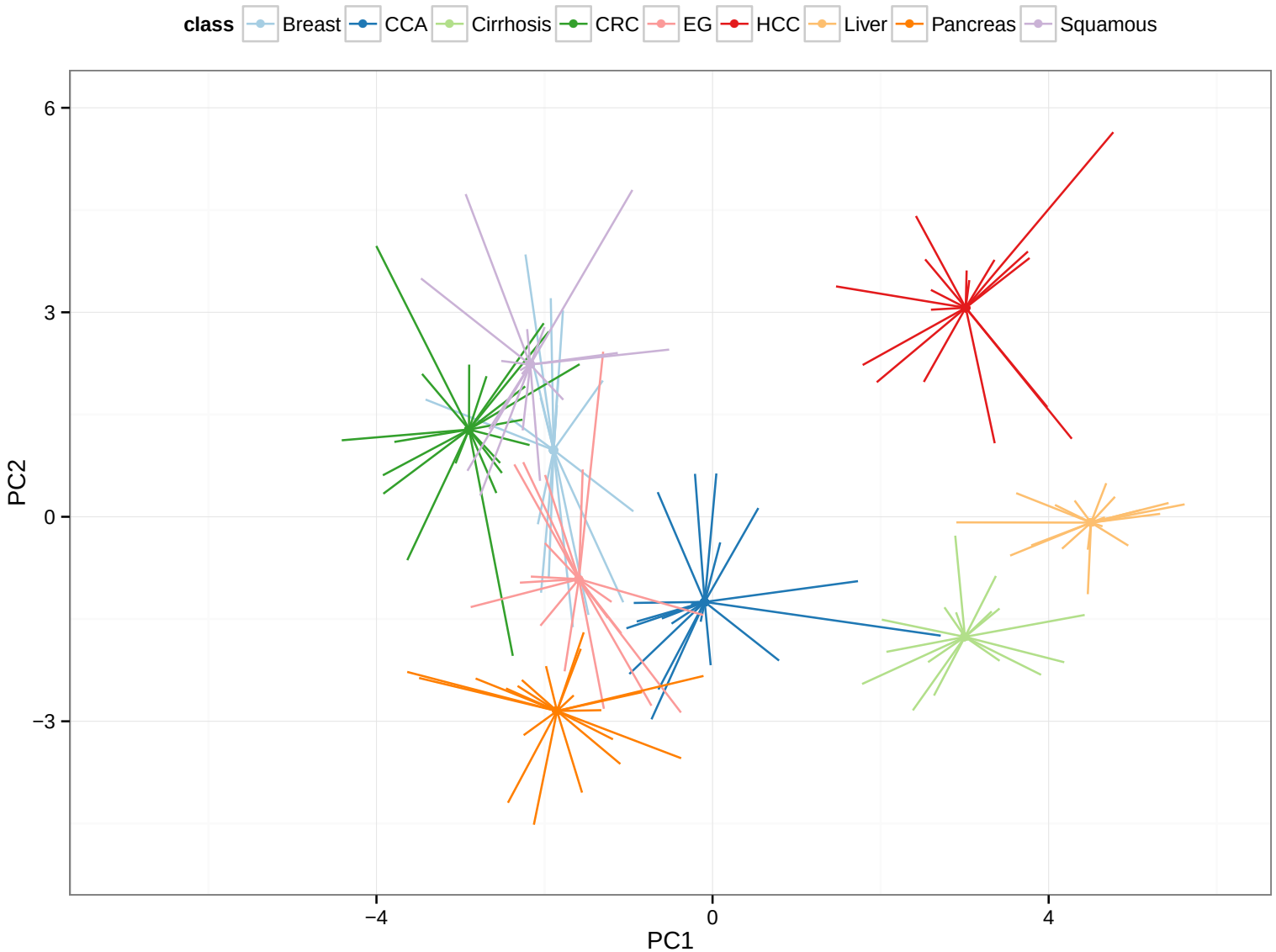
Parameters used

$$\alpha \sim \left(\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}, 0.05 \begin{bmatrix} 1 & -0.95 \\ -0.95 & 1 \end{bmatrix} \right)$$

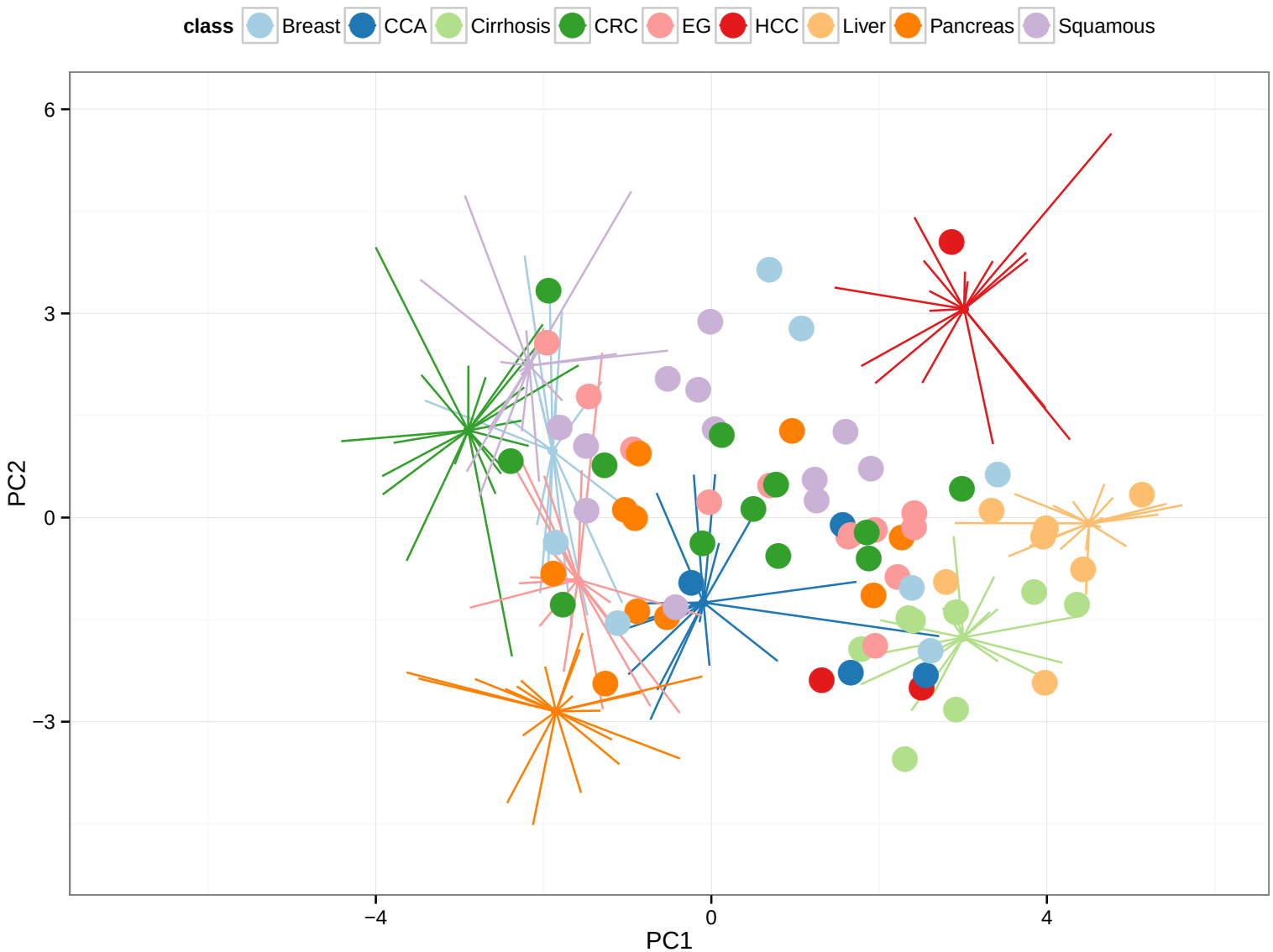
$$\mathbf{X} \sim \left(\begin{bmatrix} \mathbf{0} \\ \hat{\xi}_{\text{liver}} \end{bmatrix}, \begin{bmatrix} \hat{\boldsymbol{\xi}}\Sigma_{\text{class}}\hat{\boldsymbol{\xi}} + \begin{bmatrix} \hat{\sigma}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\sigma}_p \end{bmatrix} & \mathbf{0} \\ \mathbf{0} & \begin{bmatrix} \hat{\sigma}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{\sigma}_p \end{bmatrix} \end{bmatrix} \right)$$

$$\varepsilon \sim (\mathbf{0}, 0.2 I_p)$$

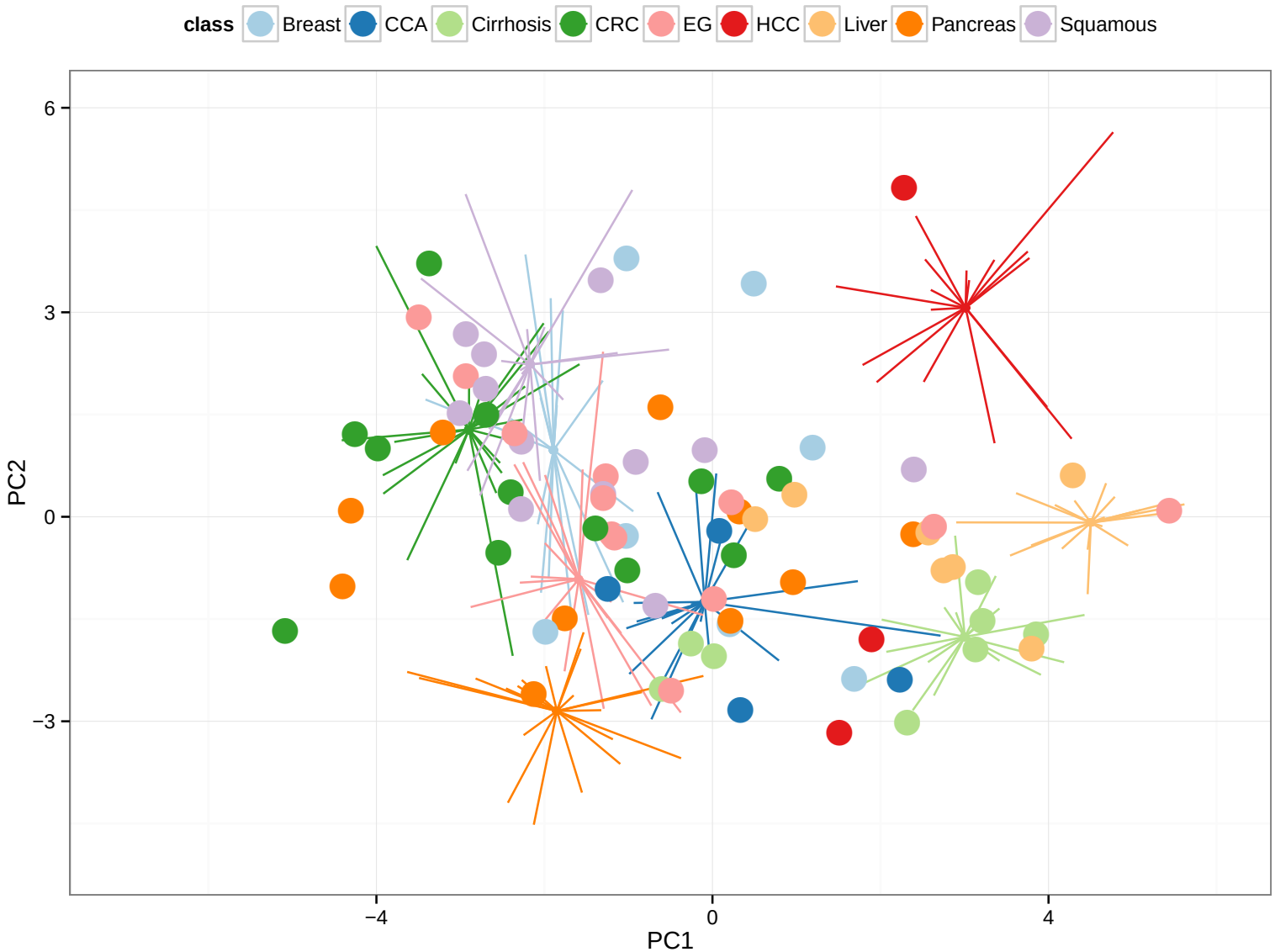
Projections of primary samples



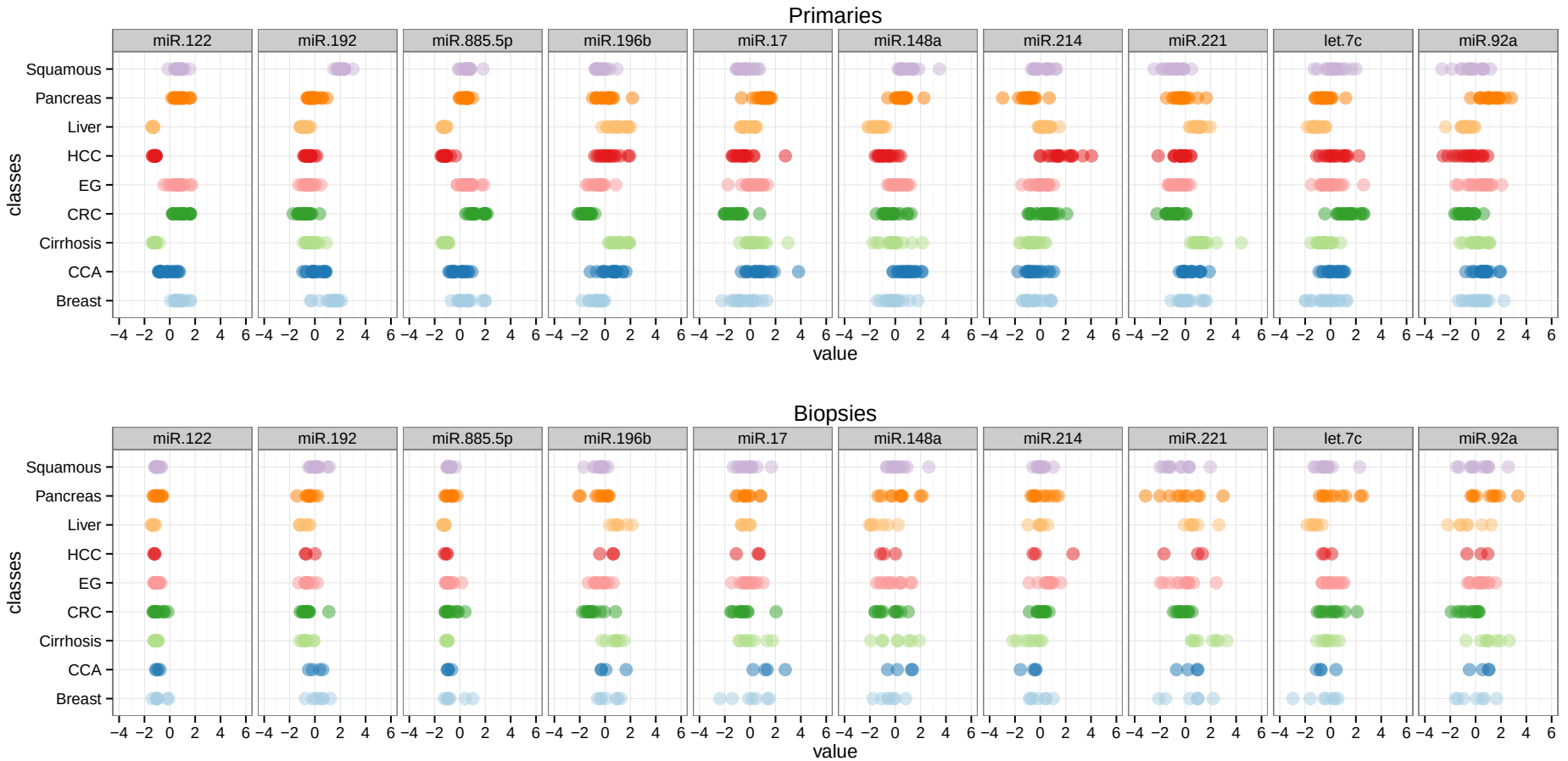
Projections of primary and biopsy samples



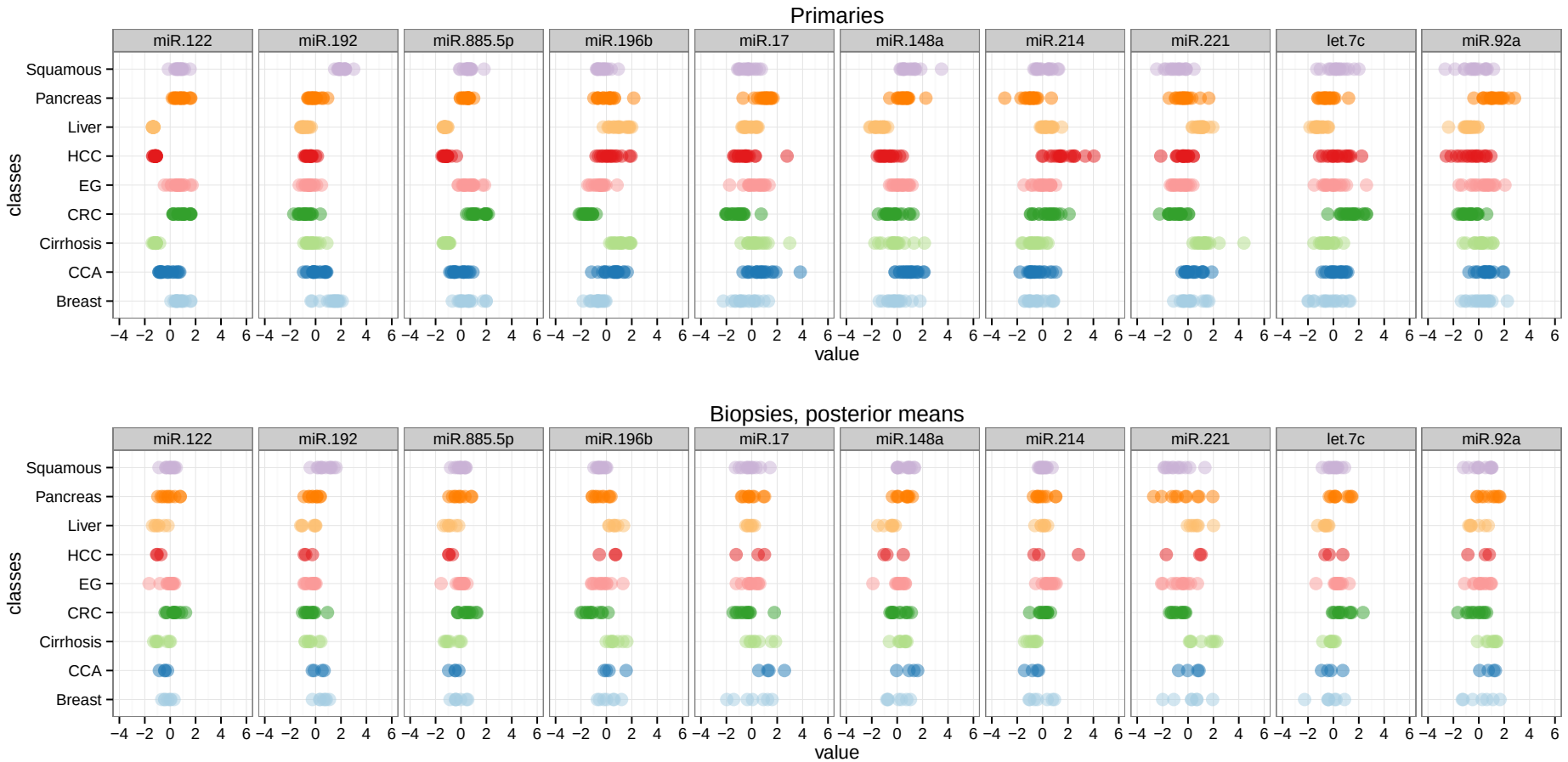
Projections of primary and biopsy posterior means



Top 10 miRNAs for class prediction



Top 10 miRNAs for class prediction



Preliminary results

Principal training data	Number of core biopsies	Backward multinomial		Forward multinomial	
		Method		Number of miRNAs	
		$\hat{\pi}_0(y\|z)$	$\hat{\pi}(y\|\hat{x})$	50	100
Primaries	0 (0)	48%	50%	77%	74%
	2 (10)	-	-	59%	54%
	4 (20)	-	-	48%	45%
Artificial	0 (0)	-	-	45%	43%
	2 (10)	-	-	39%	41%
	4 (20)	-	-	34%	39%

Conclusions

- Tissue heterogeneity can be a big problem for prediction based on molecular signatures.
- A forward or backward simulation can decrease but not solve the problem.
- The forward solution was first understood in the Machine Learning lingo as domain adaptation.
- Backward simulation is closely related to deconvolution of the molecular signature.

M. Vincent, N. R. Hansen. *Sparse group lasso and high dimensional multinomial classification*, Comp. Stat. Data Anal. 2014

M. Vincent, K. Perell, F. C. Nielsen, G. Daugaard and N. R. Hansen *Modeling tissue contamination to improve molecular identification of the primary tumor site of metastases*, Bioinformatics, 2014.

Proportion posterior means