# Robust methods and model selection

Garth Tarr

September 2015

# Outline

1. The past: robust statistics

2. The present: model selection

3. The future: protein data, meat science, joint modelling, data visualisation…

The past:

# Robust statistics

# PhD and postdoc at Sydney University

- Inference in quantile regression models

- Robust scale estimator
- Robust covariance and autocovariance (short and long range dependence)
- Robust precision matrix estimation (with regularisation for sparsity)

# $P_n$: our robust scale estimator

- Given data $\mathbf{X} = (X_1, \ldots, X_n)$, consider the $U$-statistic based on the pairwise mean kernel,

$$U_n(\mathbf{X}) = \binom{n}{2}^{-1} \sum_{i<j} \frac{X_i + X_j}{2}.$$

- Let $H(t) = P((X_i + X_j)/2 \leq t)$ be the cdf of the kernels with corresponding empirical distribution function,

$$H_n(t) = \binom{n}{2}^{-1} \sum_{i<j} \mathbb{1}\left\{ \frac{X_i + X_j}{2} \leq t \right\}, \qquad \text{for } t \in \mathbb{R}.$$

For $0 < p < 1$, let $H_n^{-1}(p) := \inf\{t : H_n(t) \geq p\}$.

- We define $P_n$ as the interquartile range of the pairwise means:

$$P_n = H_n^{-1}(3/4) - H_n^{-1}(1/4).$$

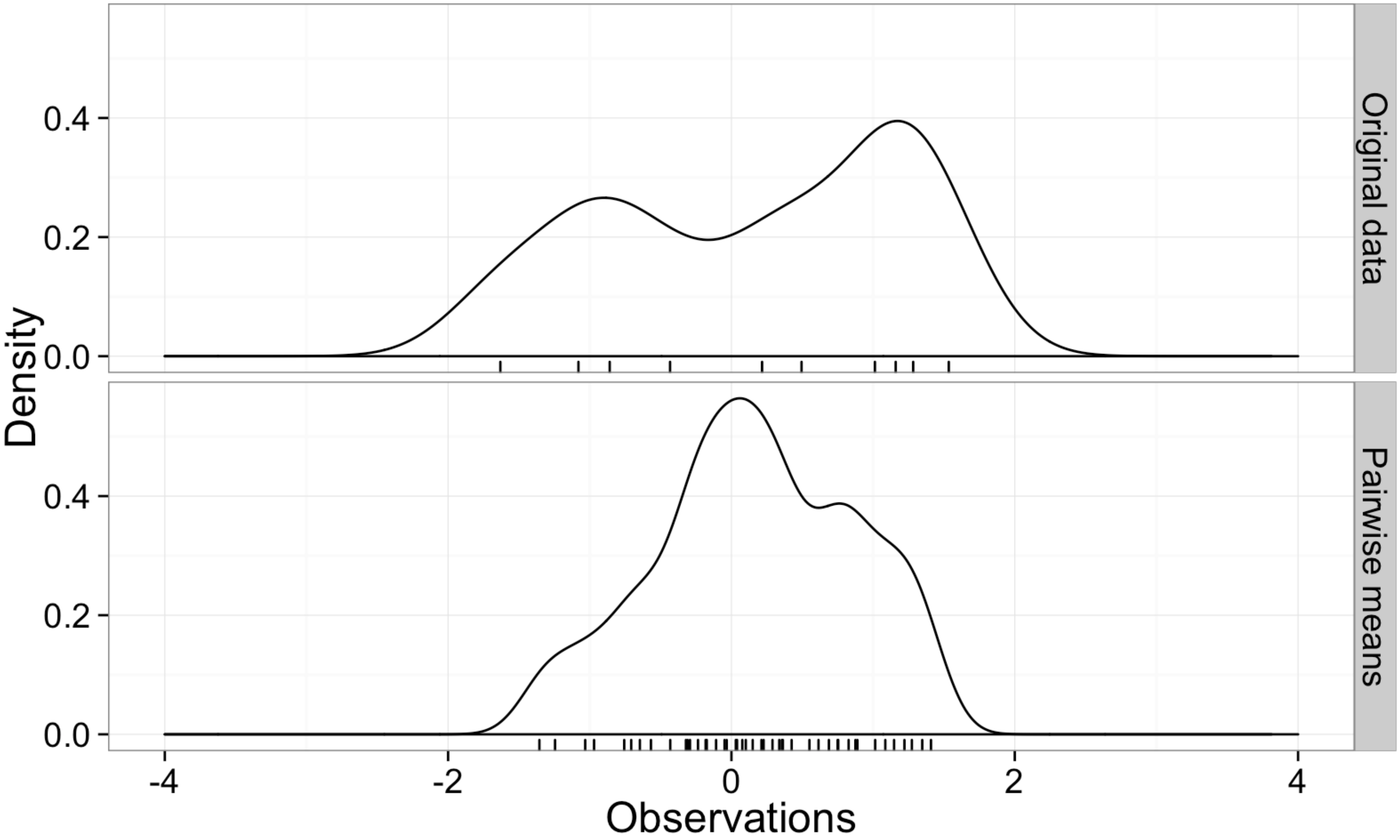
# Why another scale estimator?

| Location | Scale | Properties |
|---|---|---|
| Mean | Standard deviation | Efficient at normal but not robust |
| Median | Interquartile Range | Robust but not efficient |
| Hodges-Lehmann estimator | $P_n$ | Good robustness and efficiency properties |

- The Hodges-Lehmann estimator of **location** is the median of the pairwise means.
- $P_n$ is the interquartile range of the pairwise means.

# Why pairwise means?

Consider 10 observations drawn from $\mathcal{N}(0,1)$.



# Bounded influence function

The influence curve for a functional $T$ at distribution $F$ is

$$\mathrm{IF}(x; T, F) = \lim_{\epsilon \downarrow 0} \frac{T((1-\epsilon)F + \epsilon\delta_x) - T(F)}{\epsilon}$$

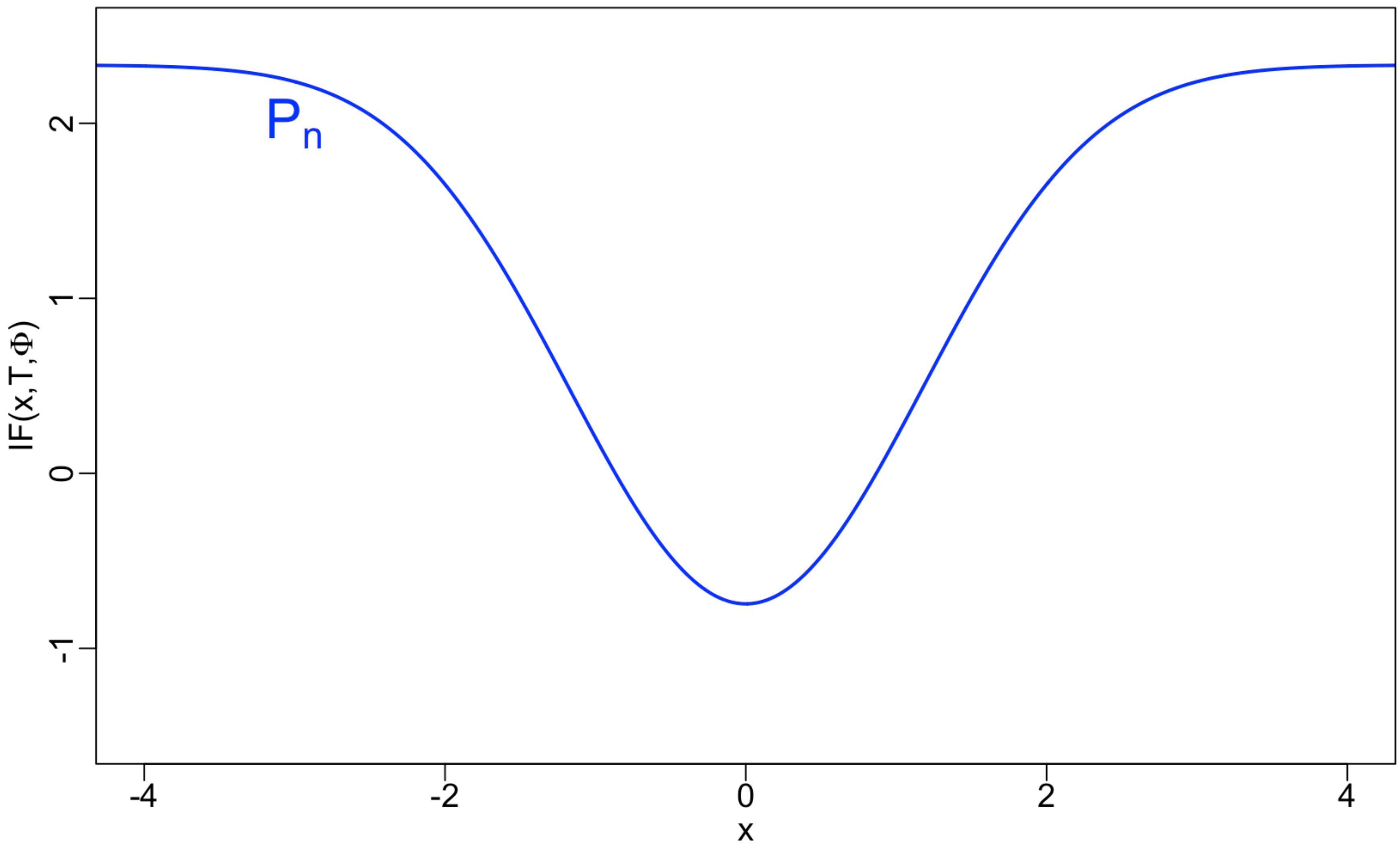
where $\delta_x$ has all its mass at $x$.

## Influence curve for $P_n$

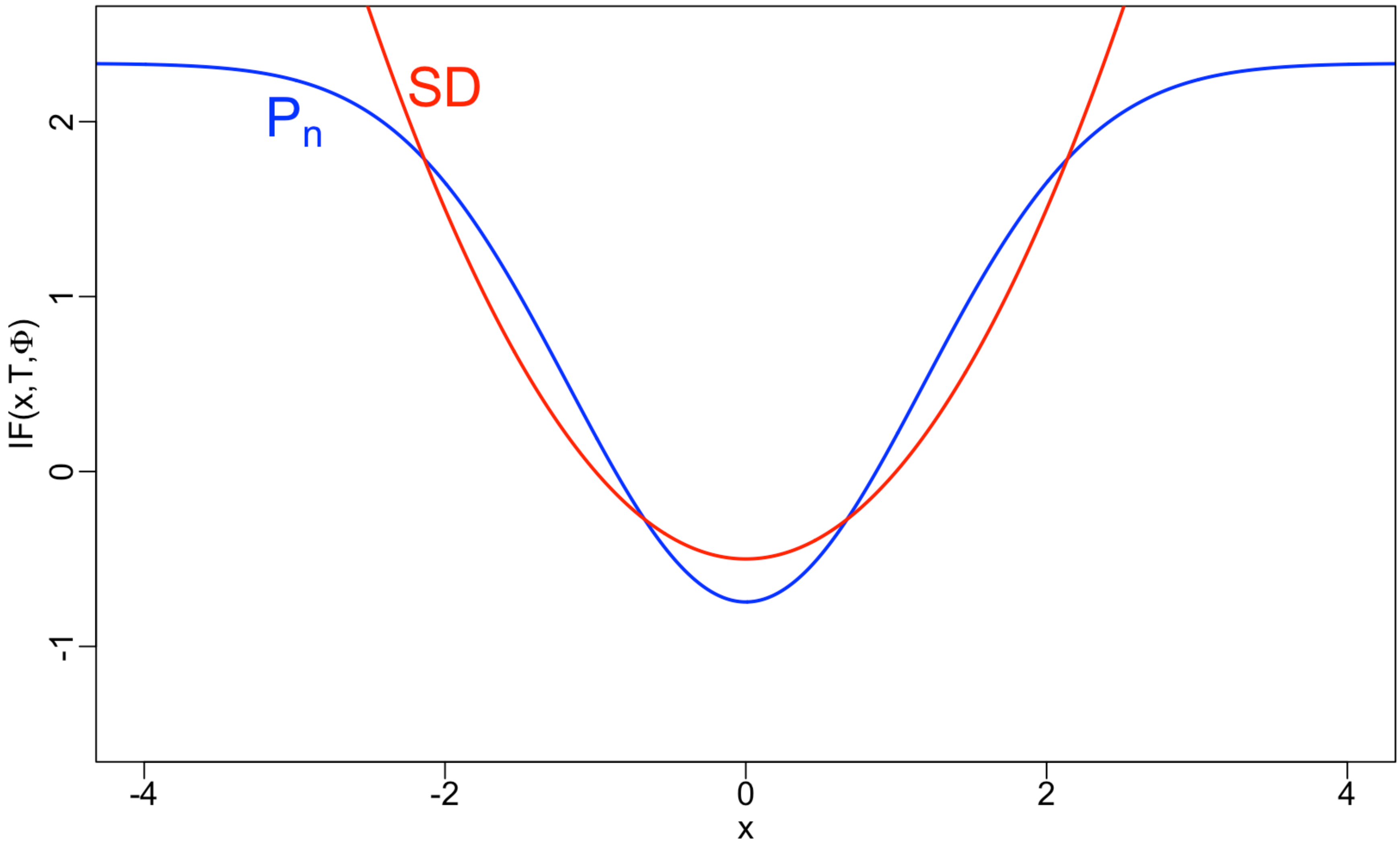
Assuming that $F$ has derivative $f > 0$ on $[F^{-1}(\epsilon), F^{-1}(1-\epsilon)]$ for all $\epsilon > 0$,

$$\begin{aligned} IF(x; P_n, F) = \Bigg[ & \frac{0.75 - F(2H_F^{-1}(0.75) - x)}{\int f(2H_F^{-1}(0.75) - x) f(x) dx} \\ & - \frac{0.25 - F(2H_F^{-1}(0.25) - x)}{\int f(2H_F^{-1}(0.25) - x) f(x) dx} \Bigg]. \end{aligned}$$

# Bounded influence function



# Bounded influence function



# Properties

- Bounded influence function
- When the underlying observations are independent, $P_n$ is asymptotically normal with variance given by the expected square of the influence function.
- When the underlying data are independent Gaussian, $P_n$ has an asymptotic efficiency of 86%.
- Breakdown value of 13%.

Tarr, Müller, and Weber (2012)

# Properties

- Bounded influence function
- When the underlying observations are independent, $P_n$ is asymptotically normal with varaince given by the expected square of the influence function.
- When the underlying data are independent Gaussian, $P_n$ has an asymptotic efficiency of 86%.
- Breakdown value of 13%.

Tarr, Müller, and Weber (2012)

- Also looked at the distribution of the estimator under short and long range dependence
- LRD turned out to be very complicated
- Took a step back and looked at the interquartile range

Tarr, Weber, and Müller (2015)

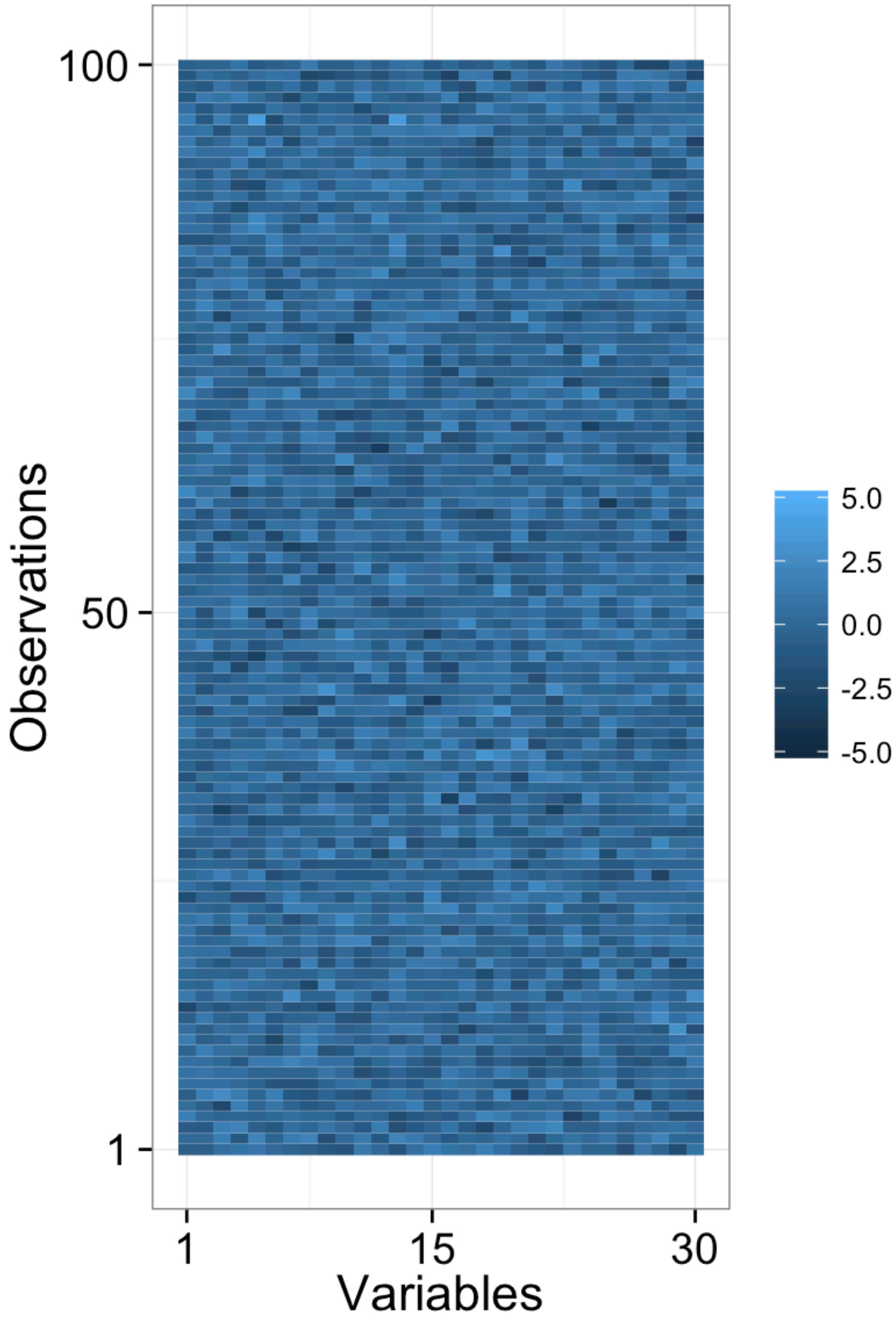
# Cellwise contamination

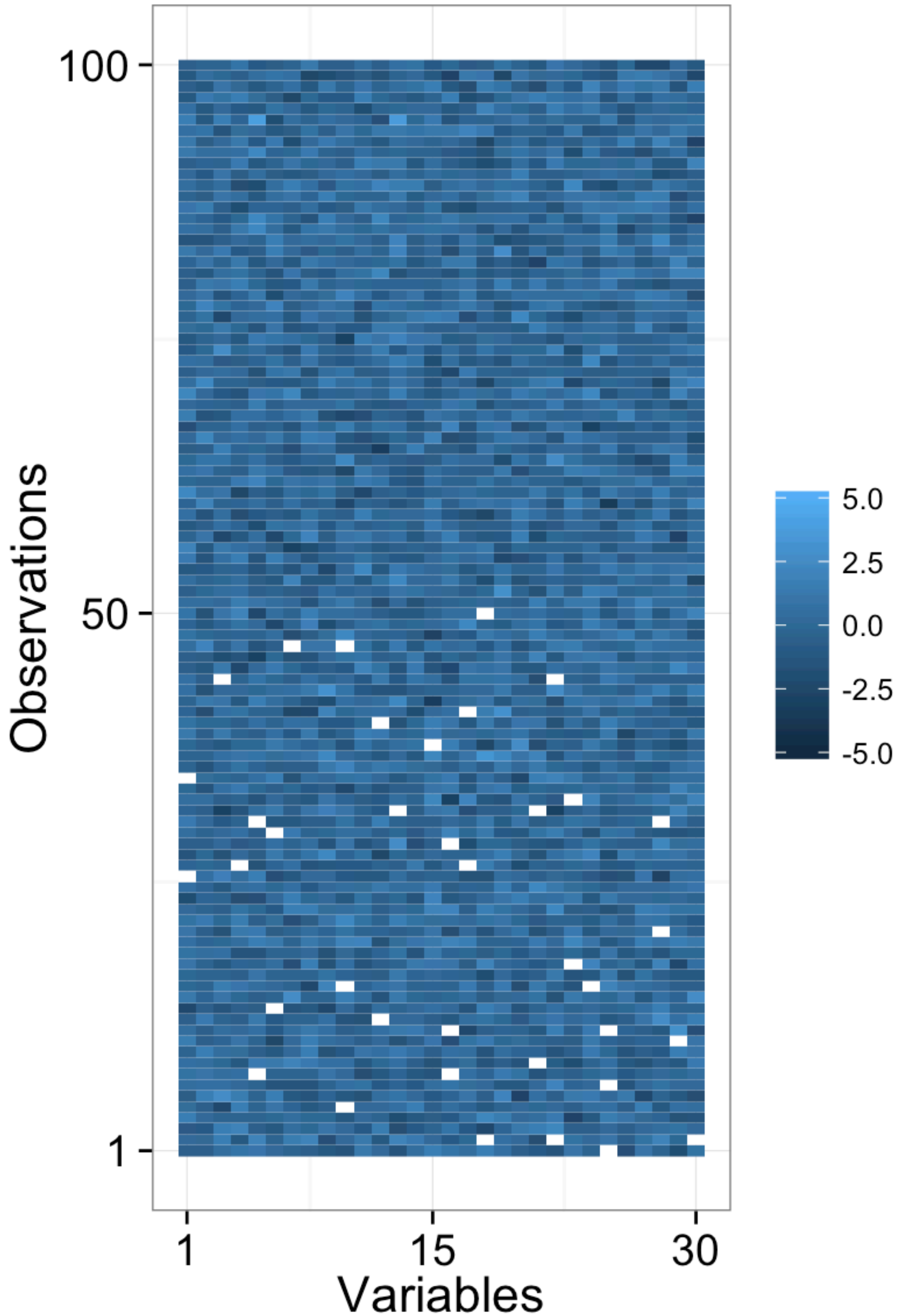
A key component of my PhD looked at estimating **precision matrices** for data contaminated in a *cellwise* manner.

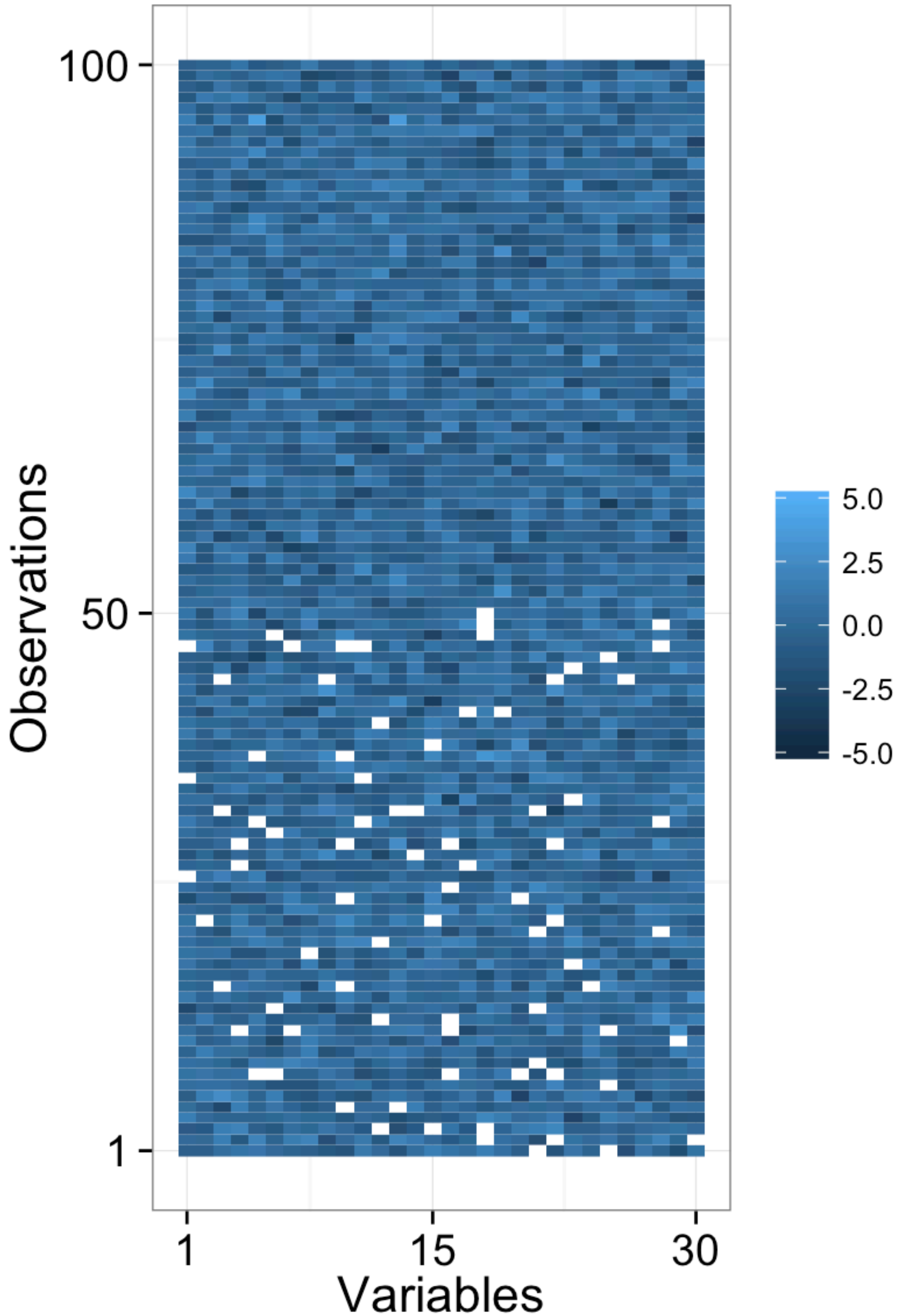
Important for:

- high dimensional data
- automated data collection and analysis methods
- e.g. -omics type data

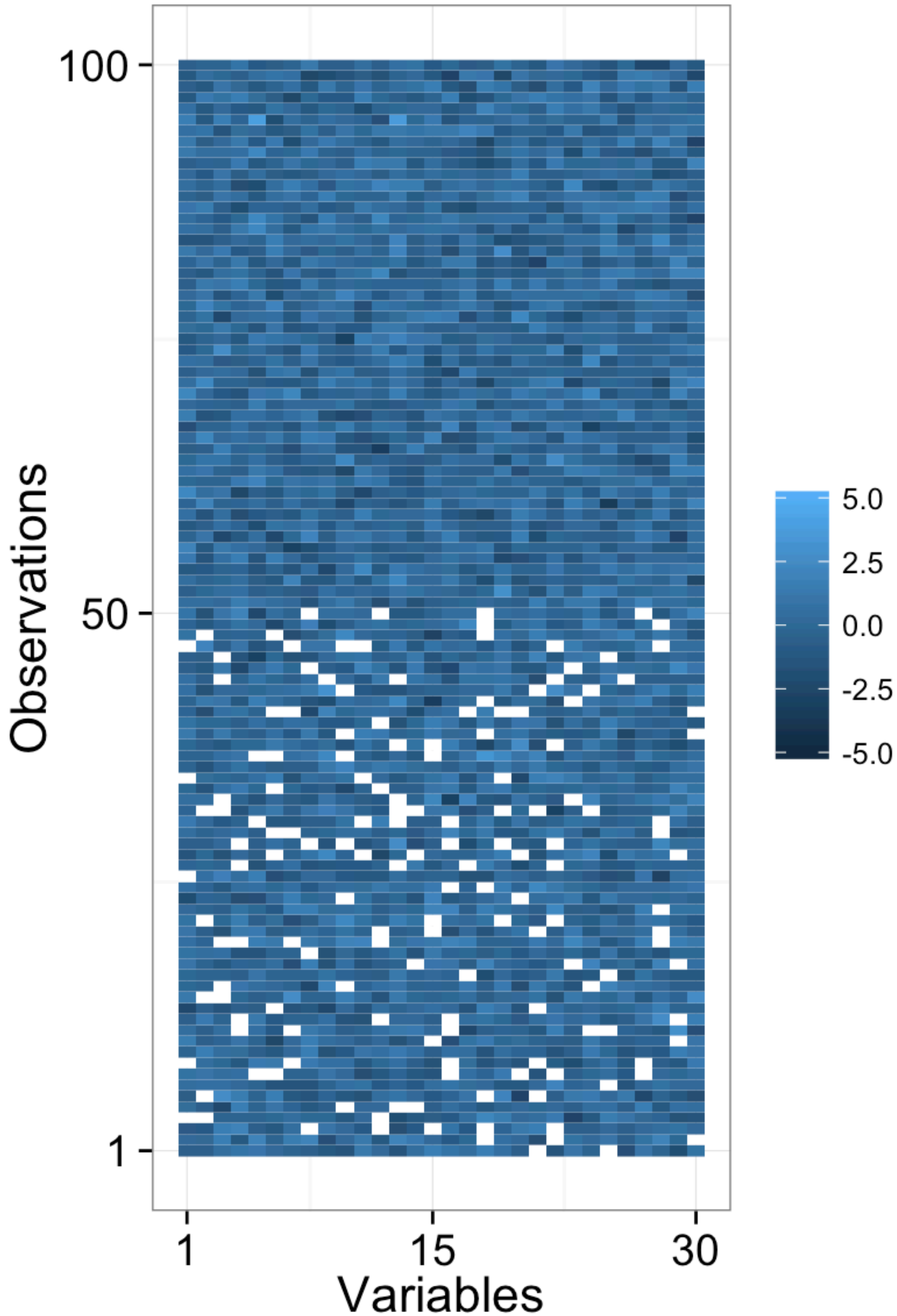
Often **sparsity** is assumed, i.e. the precision matrix will have many zero entries.


100
50
1
Observations
1
15
30
Variables
5.0
2.5
0.0
-2.5
-5.0

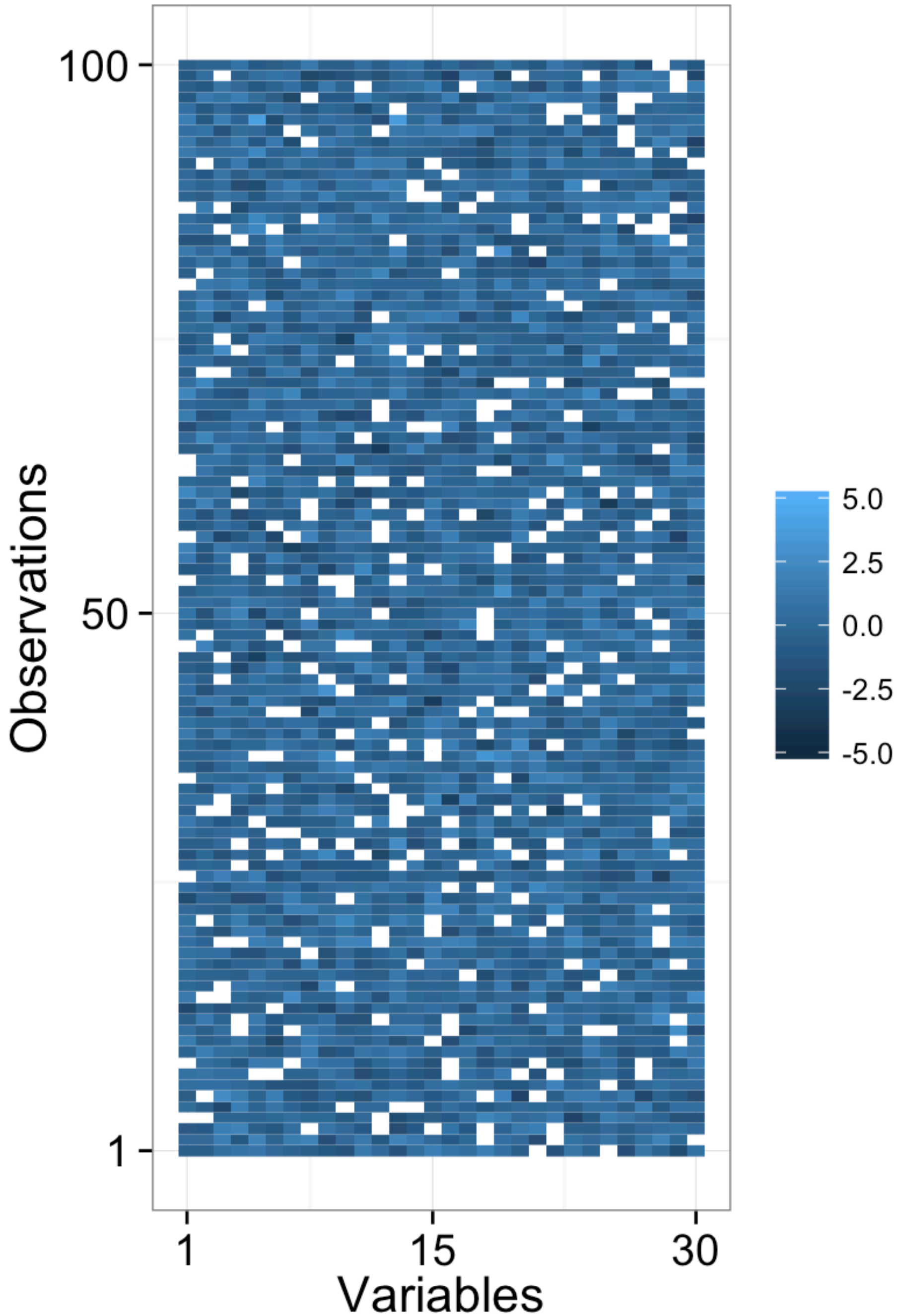

100
50
1
Observations
1
15
30
Variables
5.0
2.5
0.0
-2.5
-5.0



100
50
1
Observations
1
15
30
Variables
5.0
2.5
0.0
-2.5
-5.0



100
50
1
Observations
1
15
30
Variables
5.0
2.5
0.0
-2.5
-5.0



100
50
1
Observations
1
15
30
Variables
5.0
2.5
0.0
-2.5
-5.0

# Financial example

**Aim:** to estimate the dependence structure with S&P 500 stocks over the period 01/01/2003 to 01/01/2008 (before the GFC).

- We have $n = 1258$ obervations (trading days) over $p = 452$ dimensions (stocks).
- Observe $S_{t,j}$ the closing price of stock $j$ on day $t$ for $j = 1, \ldots, p$ and $t = 1, \ldots, n$.
- Look at the return series $X_{t,j} = \log\left(\frac{S_{t,j}}{S_{t-1,j}}\right)$.
- We want to estimate a sparse **precision matrix** where the zero entries correspond to (conditional) independence between the stocks.

**How:** using the **graphical lasso** with a robust covariance matrix as the input.

Tarr, Müller, and Weber (2015)

# What's the graphical lasso?

The graphical lasso minimises the penalised negative Gaussian log-likelihood: over non-negative definite matrices $\mathbf{\Theta}$:

$$f(\mathbf{\Theta}) = \mathrm{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) - \log|\mathbf{\Theta}| + \lambda||\mathbf{\Theta}||_1,$$

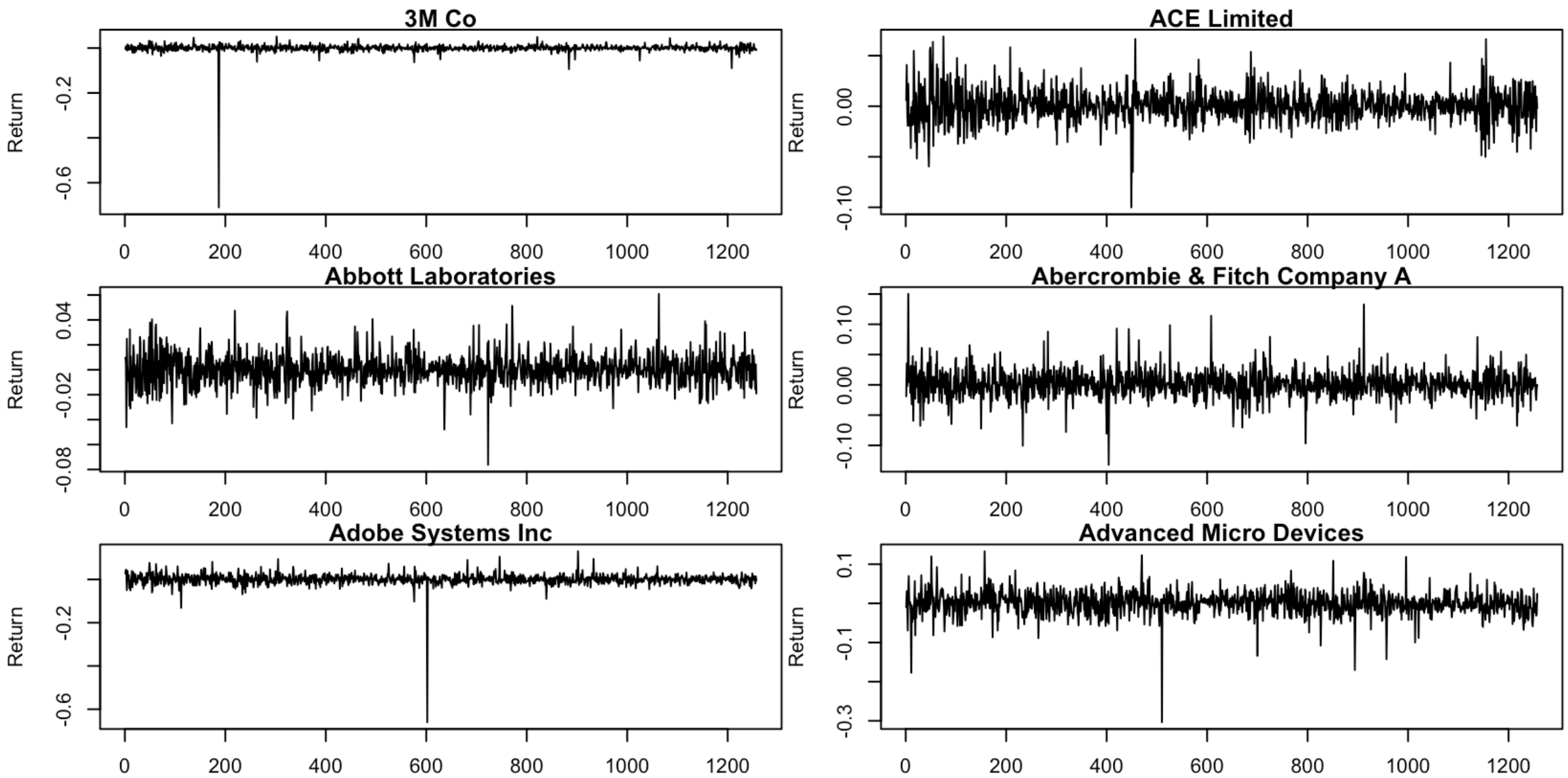
where $||\mathbf{\Theta}||_1$ is the $L_1$ norm, $\lambda$ is a tuning parameter for the amount of shrinkage and $\hat{\mathbf{\Sigma}}$ is a sample covariance matrix.

Friedman, Hastie, and Tibshirani (2008)

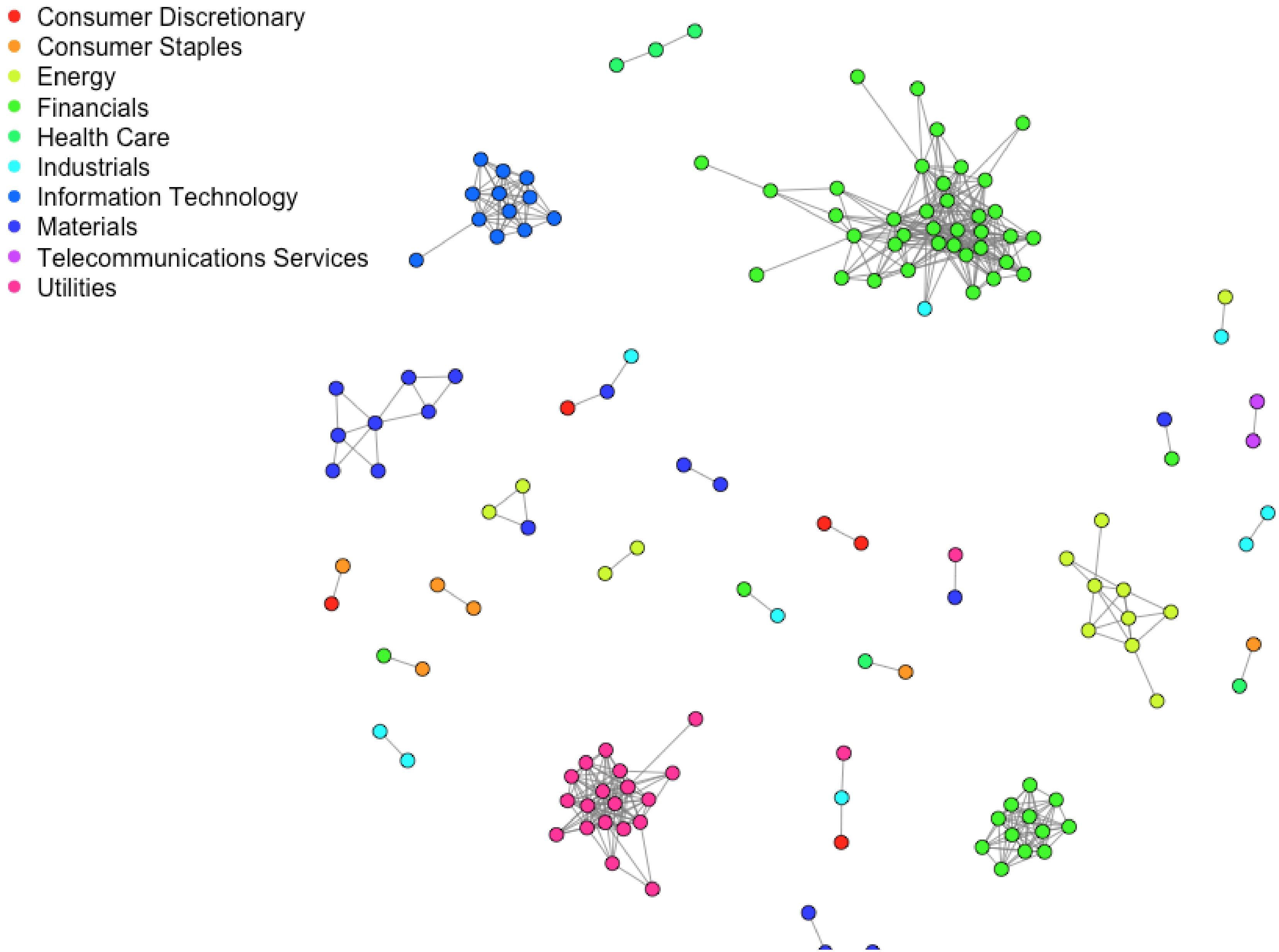
Why? **Sparsity!**

# Financial example

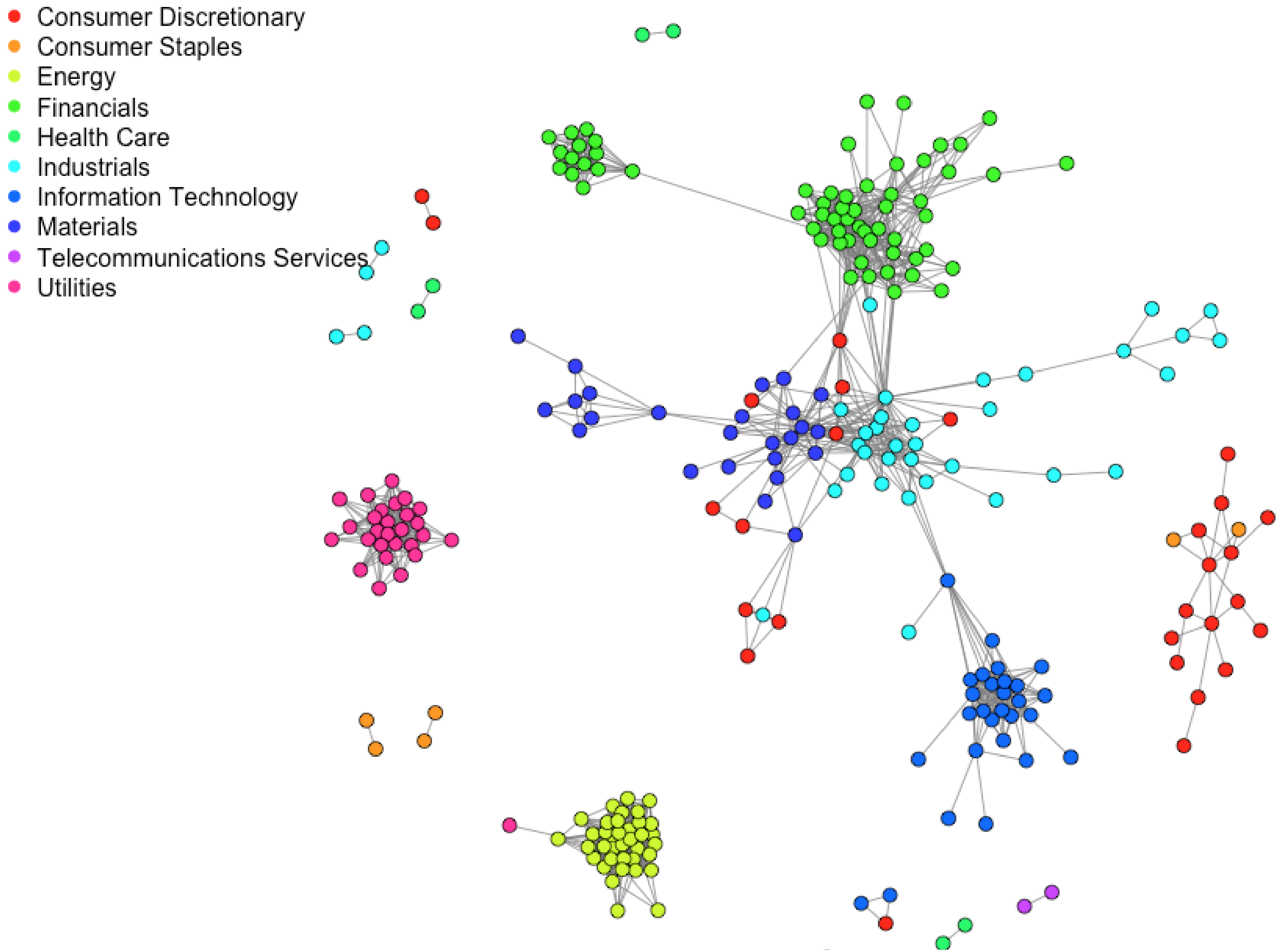
```
require(huge)
data(stockdata)
X = log(stockdata$data[2:1258,]/stockdata$data[1:1257,])
par(mfrow=c(3,2),mar=c(2,4,1,0.1))
for(i in 1:6) ts.plot(X[,i],main=stockdata$info[i,3],ylab="Return")
```



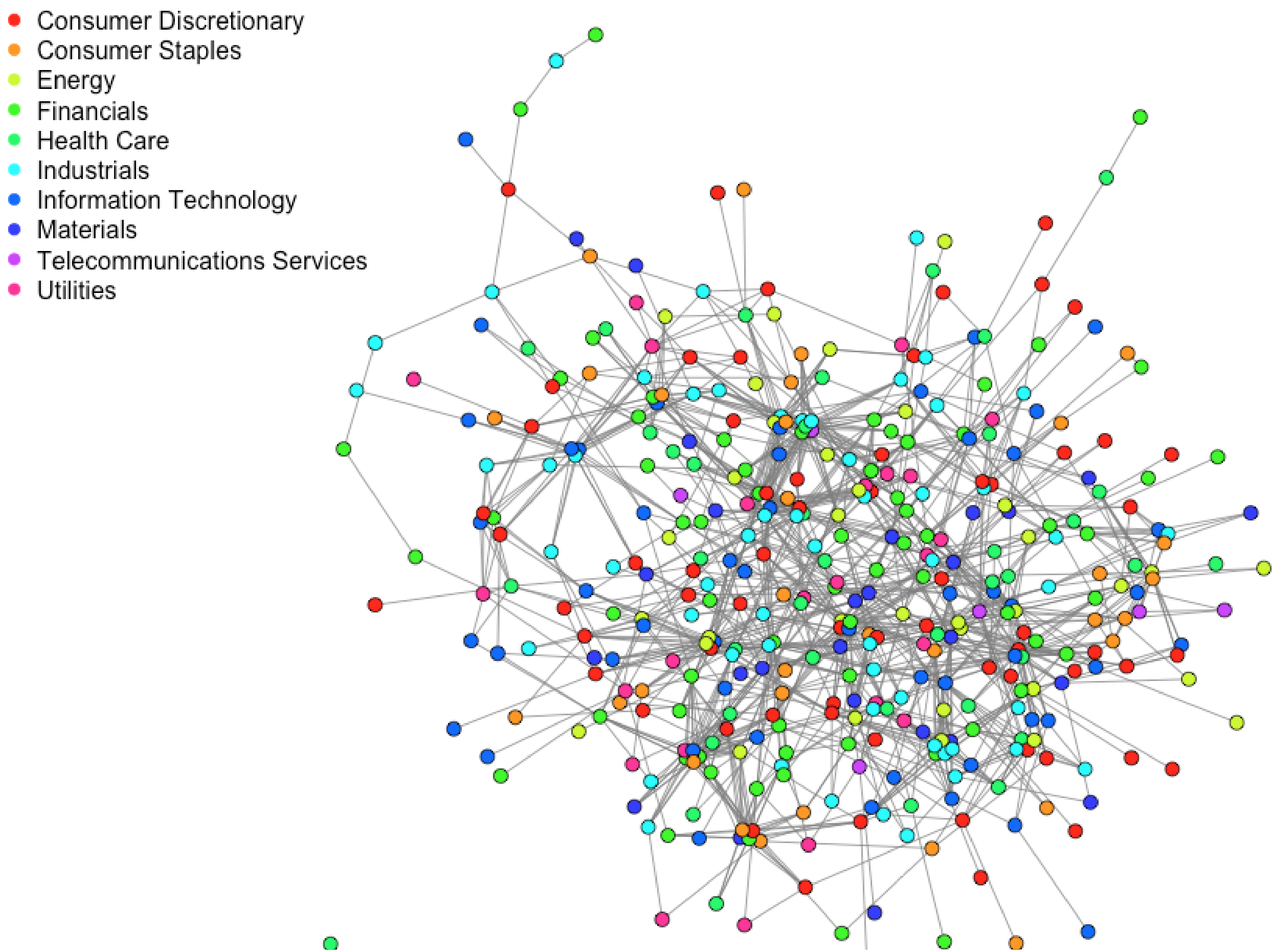
# Classical approach



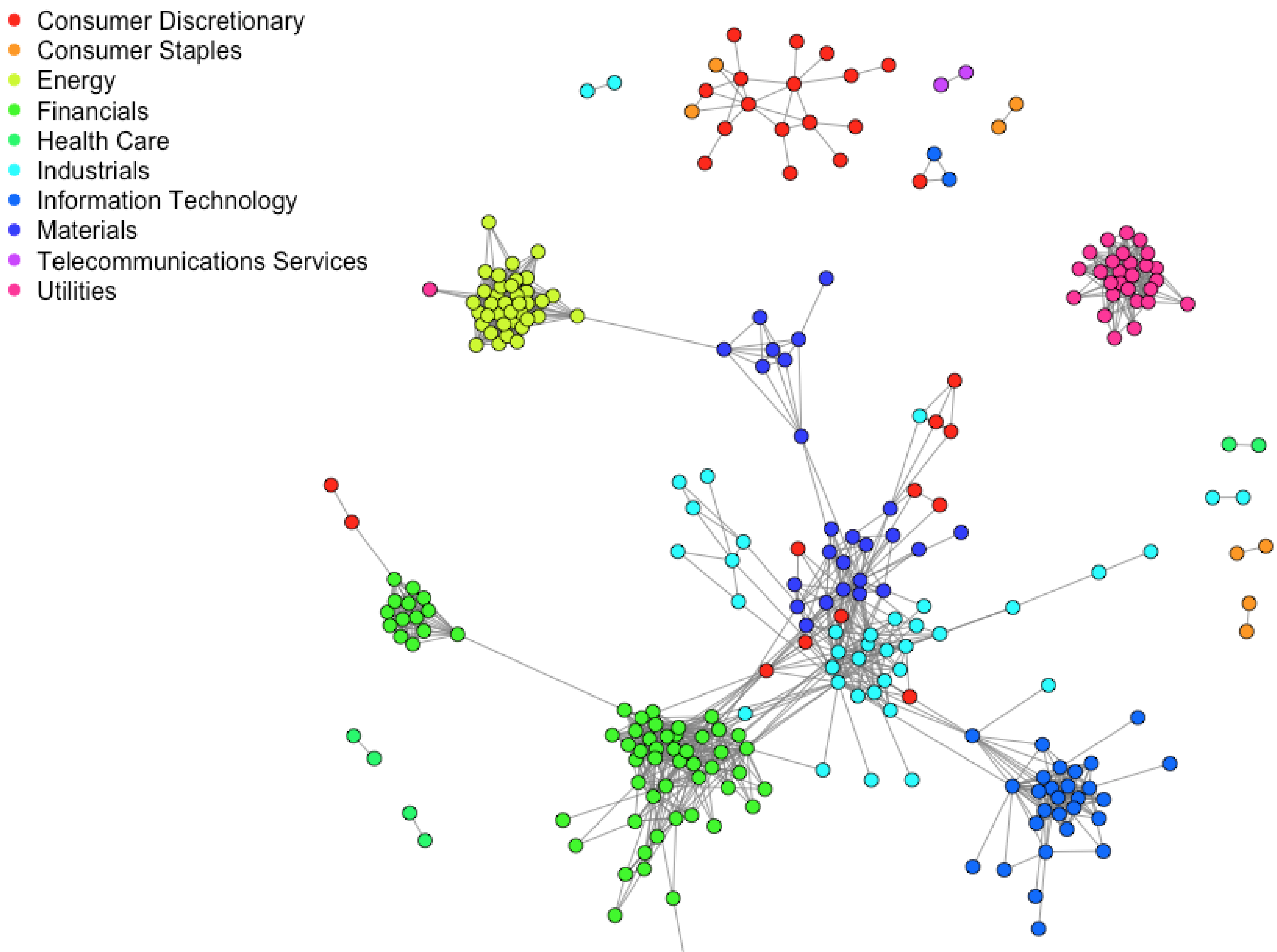
# Robust approach



# Classical approach (extra contamination)



# Robust approach (extra contamination)



# Take home messages

## Robust methods

- Robustness has always been quite niche, but it deserves more attention
- Analysing real data means dealing with errant observations
- Having reliable methods to deal with these observations is important

## Cellwise contamination

- With big data comes big problems
- Traditional robust methods can fail
- Downweighting rows is no longer appropriate

The present:

# Model selection

# Model selection

- Started working in this area last year during a post doc at ANU.

## Some notation

- Say we have a *full model* with an $n \times p$ design matrix $\mathbf{X}$.
- Let $\alpha$ be any subset of $p_\alpha$ distinct elements from $\{1, \ldots, p\}$.
- We can define a $n \times p_\alpha$ submodel with design matrix $\mathbf{X}_\alpha$ subset from $\mathbf{X}$ by the elements of $\alpha$.
- Denote the set of all possible models as $\mathcal{A} = \{\{1\}, \ldots, \alpha_f\}$.

# A smörgåsbord of tuning parameters…

## Information Criterion

- Generalised IC: $\mathrm{GIC}(\alpha;\lambda) = -2 \times \mathrm{LogLik}(\alpha) + \lambda p_\alpha$

With important special cases:

- AIC: $\lambda = 2$
- BIC: $\lambda = \log(n)$
- HQIC: $\lambda = 2\log(\log(n))$

## Regularisation routines

- Lasso: minimises $-\mathrm{LogLik}(\alpha) + \lambda\, ||\beta_\alpha||_1$
- Many variants of the Lasso, SCAD,…

# A stability based approach

## Aim

To provide scientists and researchers with tools that give them more information about the model selection choices that they are making.

## Method

- interactive graphical tools
- exhaustive searches (where feasible)
- bootstrapping to assess selection stability

Concept of **model stability** independently introduced by Meinshausen and Bühlmann (2010) and Müller and Welsh (2010) for different models.

# Diabetes example

| Variable | Description |
|---|---|
| age | Age |
| sex | Gender |
| bmi | Body mass index |
| map | Mean arterial pressure (average blood pressure) |
| tc | Total cholesterol (mg/dL) |
| ldl | Low-density lipoprotein ("bad" cholesterol) |
| hdl | High-density lipoprotein ("good" cholesterol) |
| tch | Blood serum measurement |
| ltg | Blood serum measurement |
| glu | Blood serum measurement (glucose?) |
| y | A quantitative measure of disease progression one year after baseline |

Source: Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., (2004). "Least angle regression. *The Annals of Statistics* 32 (2): 407-499. doi:10.1214/009053604000000067

# Variable inclusion plots

## Aim

To visualise **inclusion probabilities** as a function of the penalty multiplier $\lambda \in [0, 2\log(n)]$.

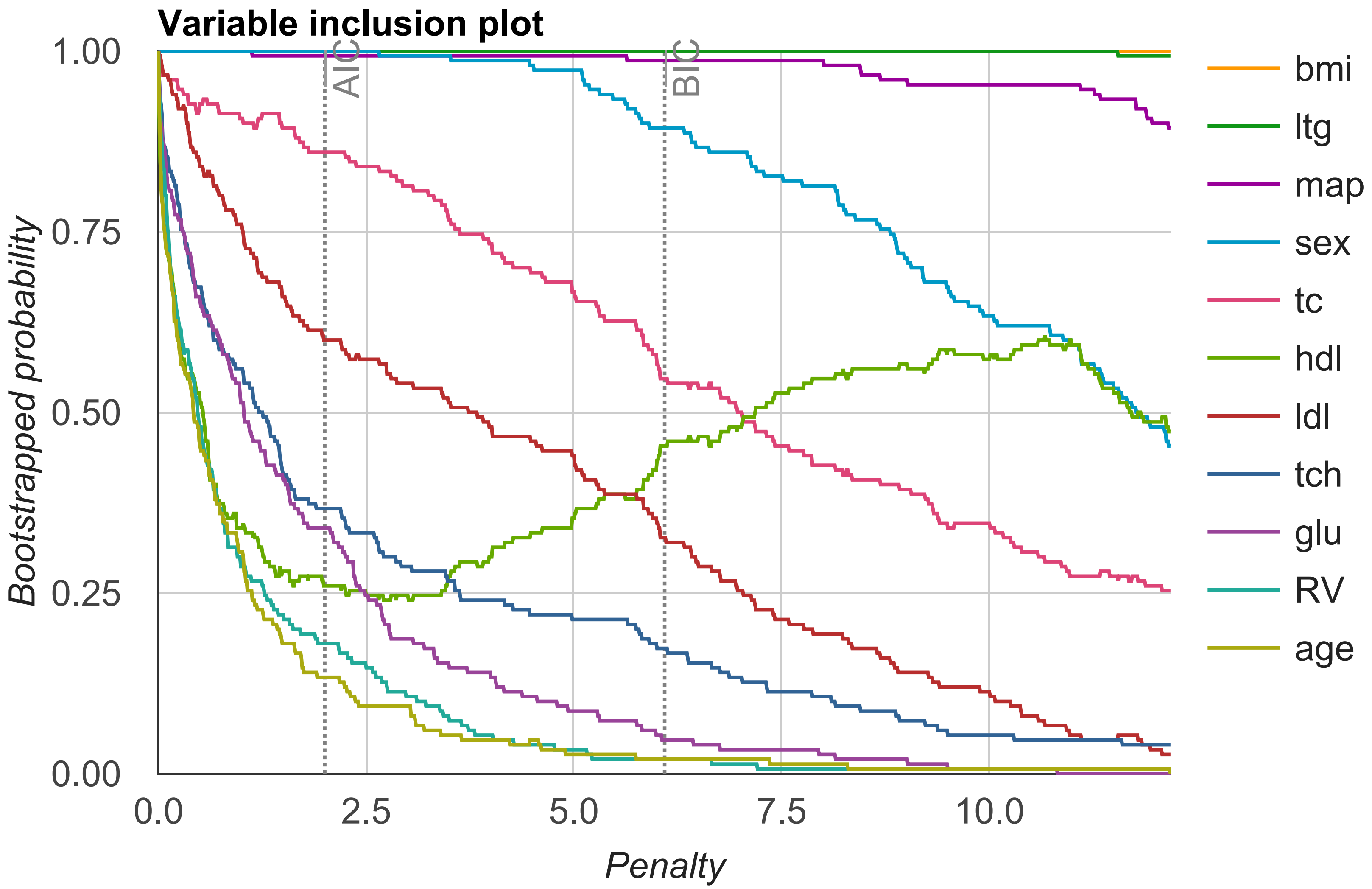
## Procedure

1. Calculate (weighted) bootstrap samples $b = 1, \dots, B$.
2. For each bootstrap sample, at each $\lambda$ value, find $\hat{\alpha}_\lambda^{(b)} \in \mathcal{A}$ as the model with smallest $\mathrm{GIC}(\alpha; \lambda) = -2 \times \mathrm{LogLik}(\alpha) + \lambda p_\alpha$.
3. The inclusion probability for variable $x_j$ is estimated as $\frac{1}{B}\sum_{b=1}^{B} 1\{j \in \hat{\alpha}_\lambda^{(b)}\}$.

## References

- Müller and Welsh (2010) for linear regression models
- Murray, Heritier, and Müller (2013) for generalised linear models

# Diabetes example – VIP



# Model stability plots

## Aim

To add value to the loss against size plots by choosing a symbol size proportional to a measure of stability.

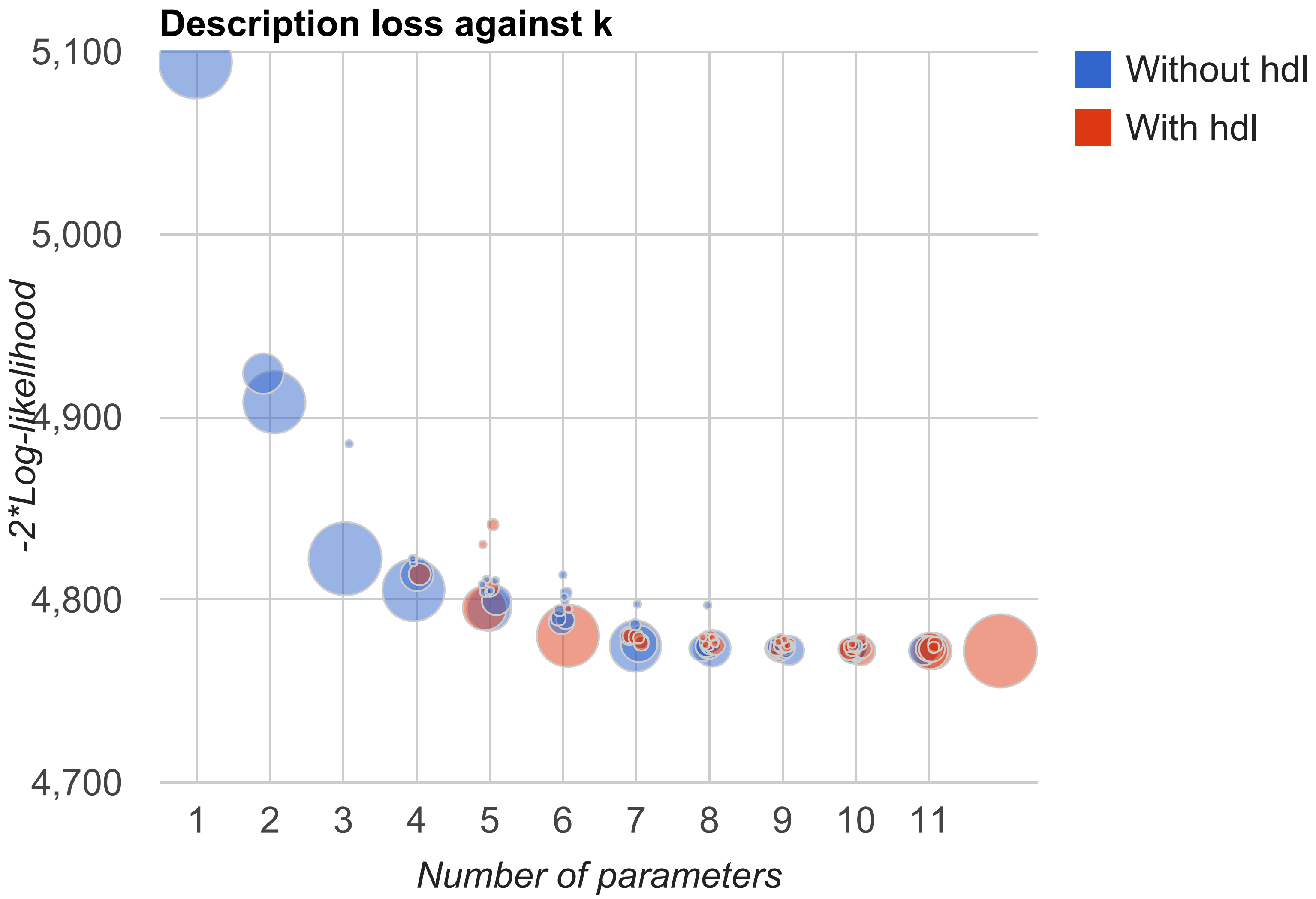
## Procedure

1. Calculate (weighted) bootstrap samples $b = 1, \ldots, B$.
2. For each bootstrap sample, identify the *best* model at each dimension.
3. Add this information to the loss against size plot using model identifiers that are proportional to the frequency with which a model was identified as being *best* at each model size.

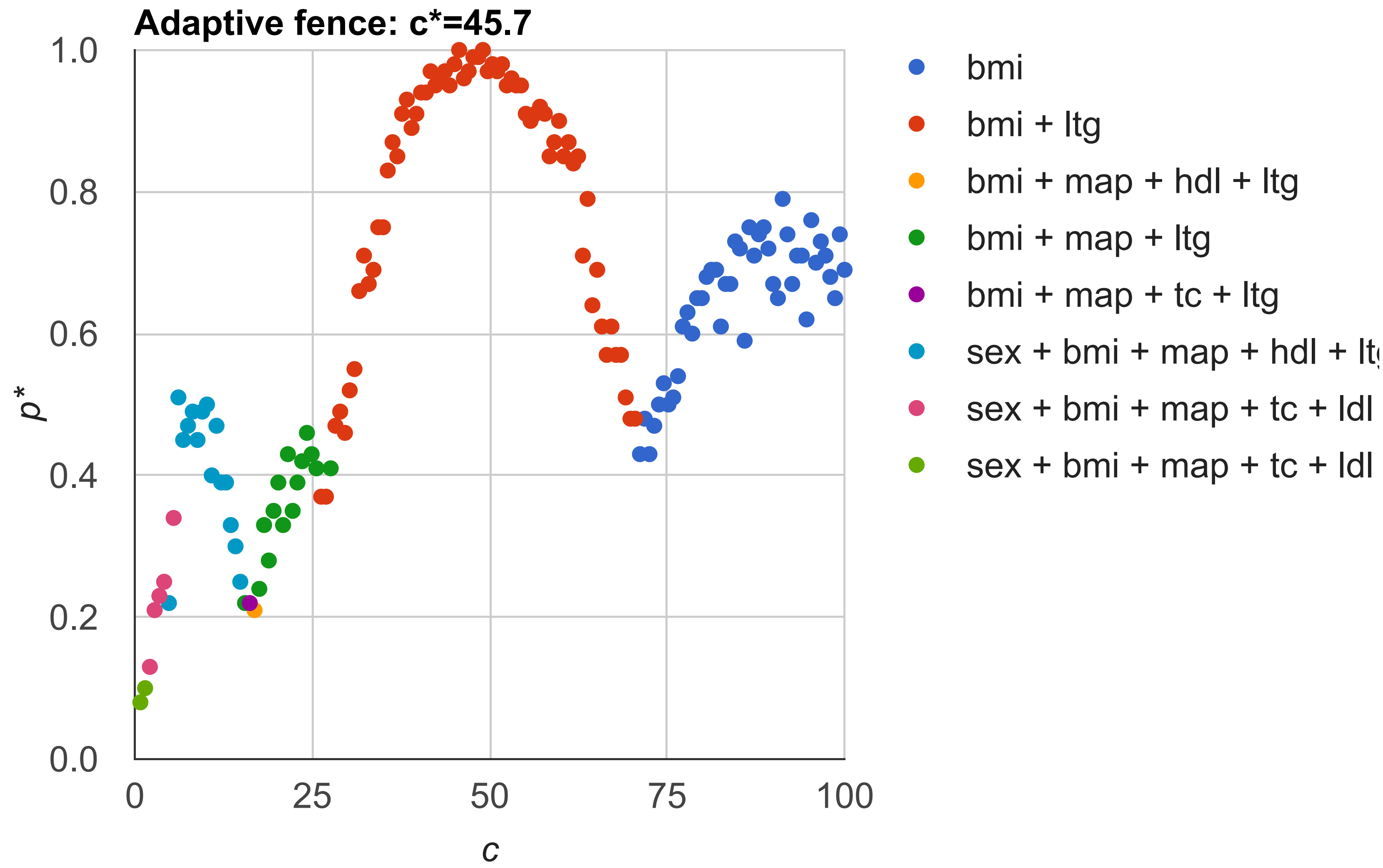
## References

- Murray, Heritier, and Müller (2013) for generalised linear models

# Artificial example – Model stability plot



# Adaptive fence



# Get it on Github

```
install.packages("devtools")
require(devtools)
install_github("garthtarr/mplot",quick=TRUE)
require(mplot)
```

### Main functions

- `af()` for the adaptive fence
- `vis()` for VIP and model stability plots
- `bglmnet()` bootstrapping glmnet
- `mplot()` for an interactive shiny interface

### Diabetes example

- Interact with it online at garthtarr.com/apps/mplot/

Tarr, Müller, and Welsh (2015)

# Take home messages

## Concept of "model stability"

- Relatively new
- Should be used more often

## Still to do:

- Approximating linear mixed models by linear models (with Alan Welsh, ANU)
- Approximating generalised linear models by linear models (with Samuel Mueller, USYD)
- Implement other models, e.g. Cox type models
- The role of robust analysis in model selection

# The future

# Projects underway (or soon to be)

- Melanoma **prognosis prediction** using protein data (with Jean Yang, USYD)
- **Predicting the eating quality** of beef and lamb (with Meat and Livestock Australia + international collaborators)
- Model selection in **joint models** (with Irene Hudson, UON)
- R packages for interactive **data visualisation** - bringing the power of D3 to R (edgebundleR, pairsD3)

# References

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. "Sparse Inverse Covariance Estimation with the Graphical Lasso." *Biostatistics* 9 (3): 432–41. doi:10.1093/biostatistics/kxm045.

Meinshausen, N, and P Bühlmann. 2010. "Stability Selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (4): 417–73. doi:10.1111/j.1467-9868.2010.00740.x.

Murray, K, S Heritier, and S Müller. 2013. "Graphical Tools for Model Selection in Generalized Linear Models." *Statistics in Medicine* 32 (25): 4438–51. doi:10.1002/sim.5855.

Müller, S, and AH Welsh. 2010. "On Model Selection Curves." *International Statistical Review* 78 (2): 240–56. doi:10.1111/j.1751-5823.2010.00108.x.

Tarr, G, S Müller, and NC Weber. 2012. "A Robust Scale Estimator Based on Pairwise Means." *Journal of Nonparametric Statistics* 24 (1): 187–99. doi:10.1080/10485252.2011.621424.

———. 2015. "Robust Estimation of Precision Matrices Under Cellwise Contamination." *Computational Statistics & Data Analysis* to appear. doi:10.1016/j.csda.2015.02.005.

Tarr, G, S Müller, and AH Welsh. 2015. *mplot: Graphical Model Stability and Model Selection Procedures*. https://github.com/garthtarr/mplot.

Tarr, G, NC Weber, and S Müller. 2015. "The Difference of Symmetric Quantiles Under Long Range Dependence." *Statistics & Probability Letters* 98: 144–50. doi:10.1016/j.spl.2014.12.022.