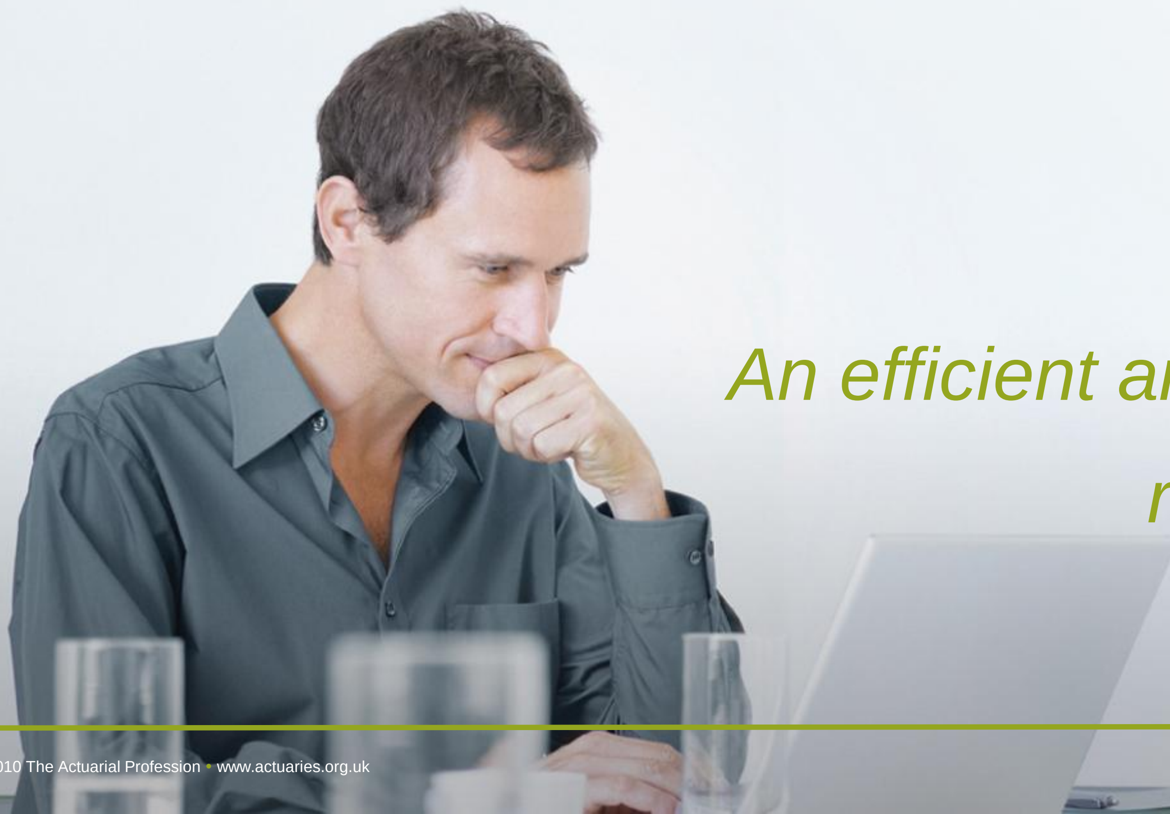
The Actuarial Profession

making financial sense of the future

GIRO conference and exhibition 2010

Pietro Parodi (Structured Risk Solutions, Willis Ltd)



Regularisation

An efficient and simple method for rating factor selection

12-15 October 2010

Agenda

I. Rating factors selections is best understood in the framework of statistical learning theory
II. The industry standard approach to rating factors selection is GLM
III. The machine learning community would solve the same problem quite differently... (A look at regularisation)
IV. A comparison between GLM and regularisation
V. Questions?

I. Rating factor selection is best understood in the context of statistical learning theory

The appropriate framework for rating factor selection is *statistical learning theory*

Rating factor selection

Find the combination of rating factors $X_1,\ldots X_n$ which best predicts future losses Y

Supervised learning

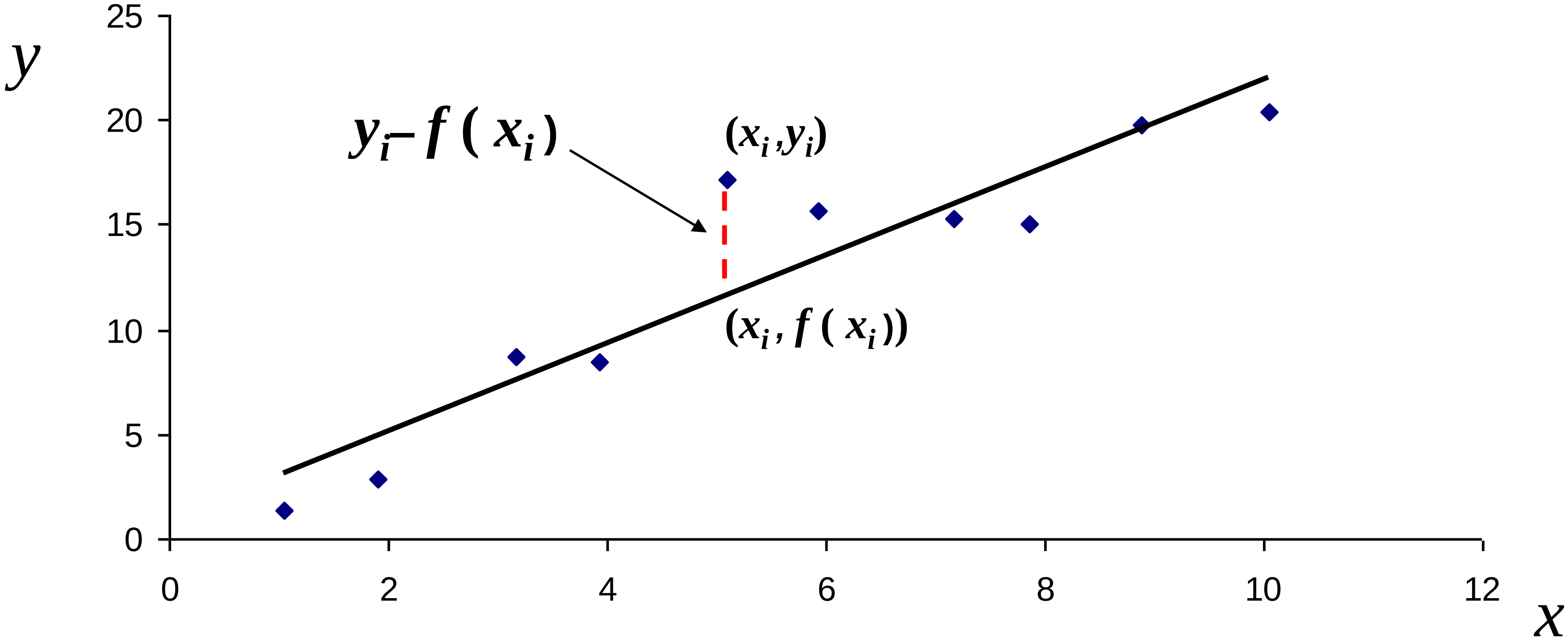
Given X (inputs), Y (outputs) with joint unknown distribution $\Pr(X,Y)$, find the model $f(X)$ of Y that minimises the expected prediction error

$$\text{EPE}(f)=\mathcal{E}(L(f(X),Y))$$

The **loss function** $L(f(X),Y)$ is the distance between the model and the data, e.g.

$$L(Y,f(X)) = \|Y-f(X)\|^2$$

The basic idea is the same as that of least squares regression...



$$L(Y, f(X)) = \|Y - f(X)\|^2 = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

… but with some complications

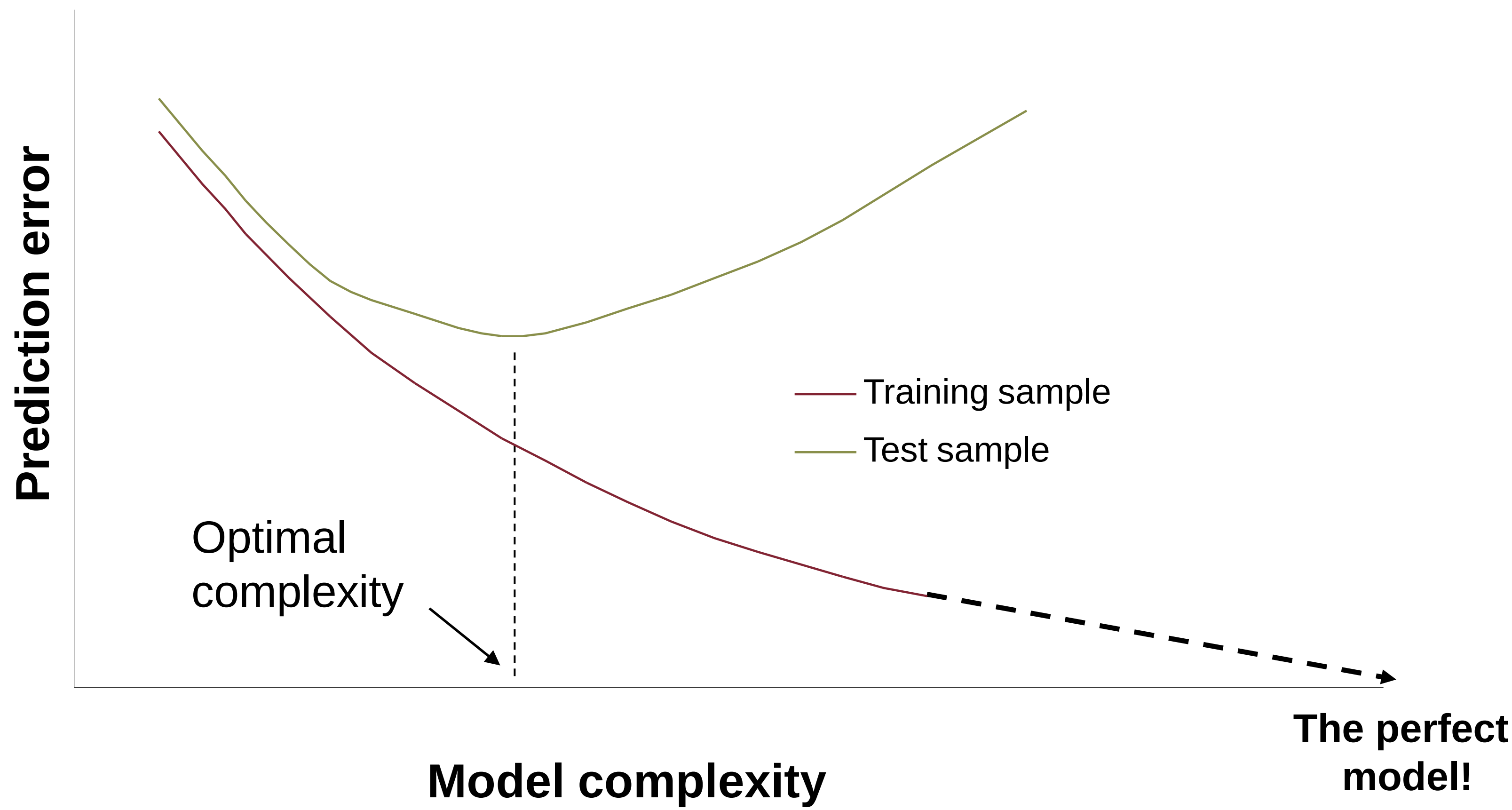
- We don’t know what the right model is
- We don’t even know how many variable it has
- We need a way to validate any model we produce

We speak of “learning” because there is **always** a training stage and a testing stage

We say “supervised” because there is a “teacher” – in the training stage we can see both the inputs *and* the outputs!

How do we choose $f(X)$?

The crucial problem: goodness of fit vs complexity



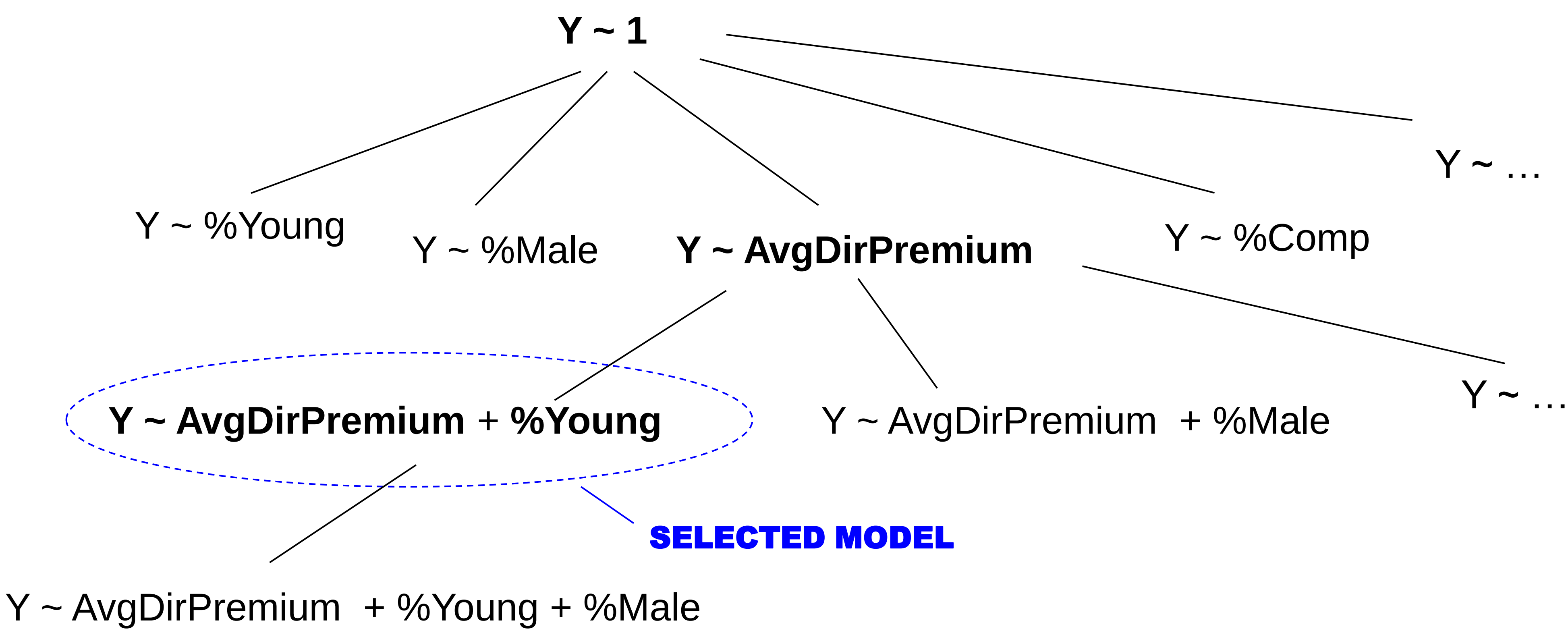
II. The industry standard approach to rating factors selection is GLM

The industry standard for rating factors selection is GLM

	Linear model	*Generalised linear model*
The model	$Y = \Sigma a_j X_j$	$Y \quad = g^{-1}(\Sigma\, a_j \psi_j(X_1, X_2, \ldots X_n)) =$ (eg) = $\exp(a_1 X_1 + a_2 X_2 + a_3 X_1 X_2)$
The loss function	$L(Y, f(X)) = \lVert Y - f(X) \rVert^2$	$L(Y, f(X)) = -2 \log \Pr_{f(X)}(Y)$
The noise	Gaussian	Exponential family (Gaussian, Poisson, Gamma…)
Model selection and validation	Greedy approach with penalty: AIC = – 2 loglik +2 d	Greedy approach with penalty: AIC = – 2 loglik +2 d

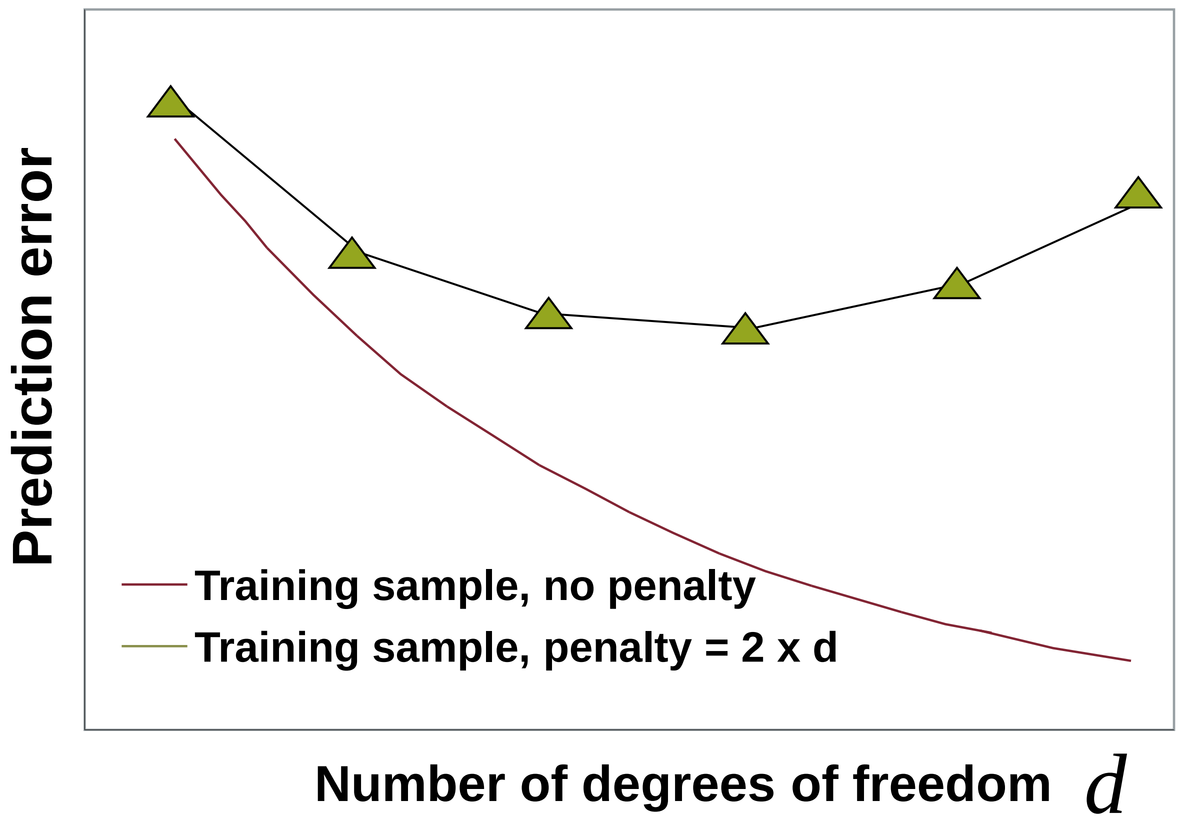
The greedy approach for GLM
A practical example

Consider the problem of predicting the reinsurance premium for a motor policy, based on the characteristics of the insurer's portfolio



An interpretation of the GLM selection scheme in terms of the error v complexity graph

The test sample error is (roughly) approximated by the AIC criterion



Shortcomings *of the textbook approach* to GLM

The GLM “core”

The usual limitations of GLM (linearity, exponential family, etc.)

Model selection

There is no guarantee that the solution found by the greedy approach (forward/backward selection) is optimal

Model validation

The model validation process is not rigorous

II. The machine learning community would solve the same problem quite differently...

A look at regularisation

Rating factors selection can be addressed by regularised regression

The main idea: to minimise the distance between the data and the model on a test set:

$$\mathrm{EPE}(f) = \left\| Y - f_\beta(X) \right\|_{l_2}^2, \text{ where } \quad f_\beta(X) = \sum_{j=1}^{\infty} \beta_j \psi_j(X_1, \ldots X_n)$$

minimise a *regularised* functional, *such as (Tychonov regul.)*:

$$\mathrm{EPE}(f) = \left\| Y - f_\beta(X) \right\|_{l_2}^2 + \lambda \left\| \beta \right\|_{l_2}^2$$

on the training set. Why does Tychonov regularisation work?

Some regularisation schemes also do variable selection!

The lasso (Tibshirani, 1996):

$$\mathrm{EPE}(\beta) = \left\| Y - f_{\beta}(X) \right\|_{l_2}^{2} + \lambda \left\| \beta \right\|_{l_1}$$

$$\left(\left\| \beta \right\|_{l_1} = |\beta_1| + |\beta_2| + \ldots + |\beta_n| \right)$$

- Performs automatic variable selection
- Can be solved as fast as least square regression

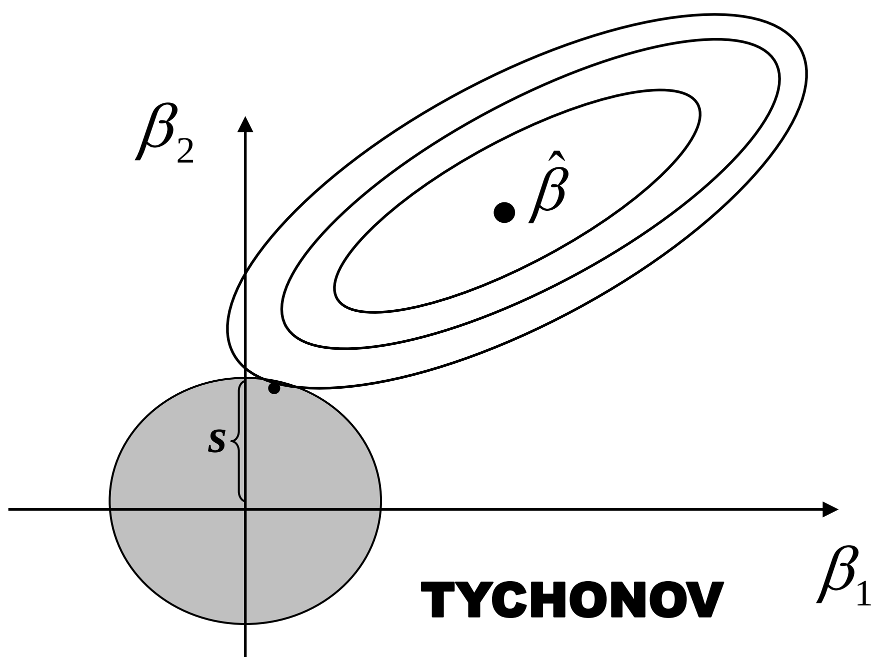
but

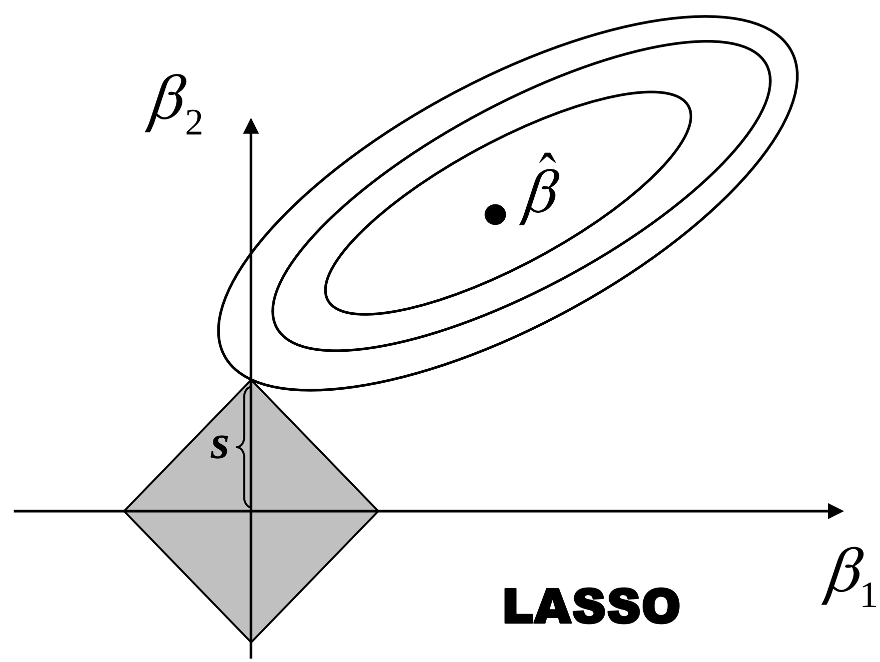
- Breaks down when no of factors > no of data points
- Is over-zealous in eliminating correlated features

How does the lasso achieve variable selection?

Tychonov regularisation
Minimise $\| Y - f(X) \|_{l_2}$
subject to $\|\beta\|_{l_2} < s$

Lasso regularisation
Minimise $\| Y - f(X) \|_{l_2}$
subject to $\|\beta\|_{l_1} < s$

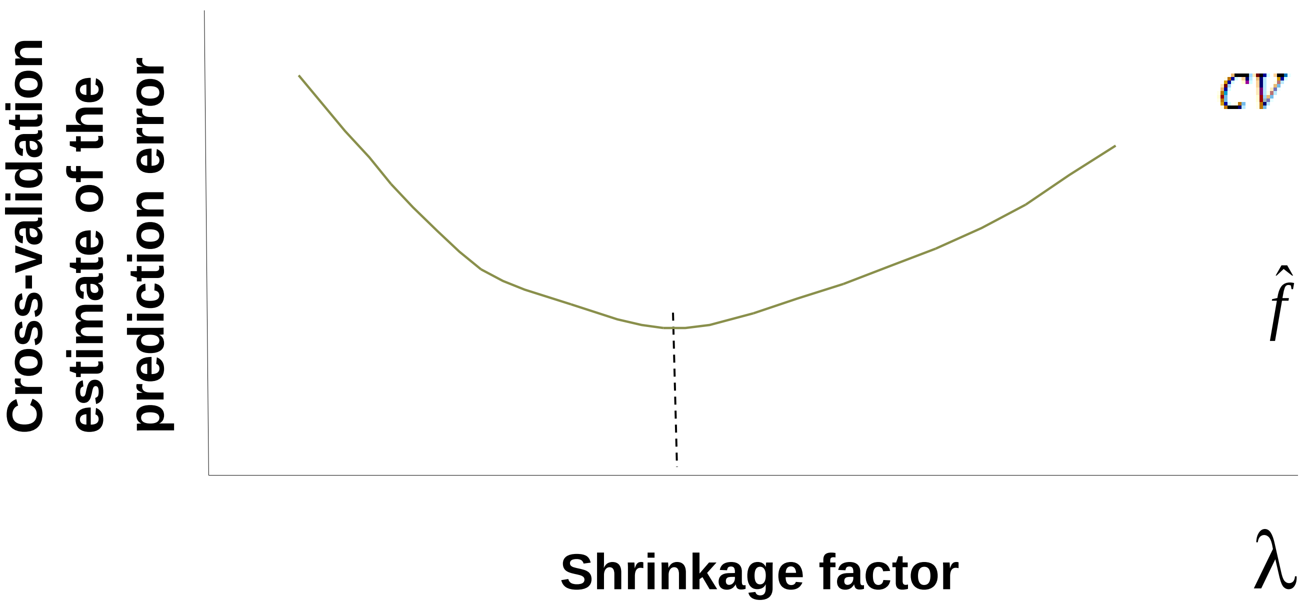




Estimating the expected prediction error for regularisation – *Cross validation*

Data set is split into K segments

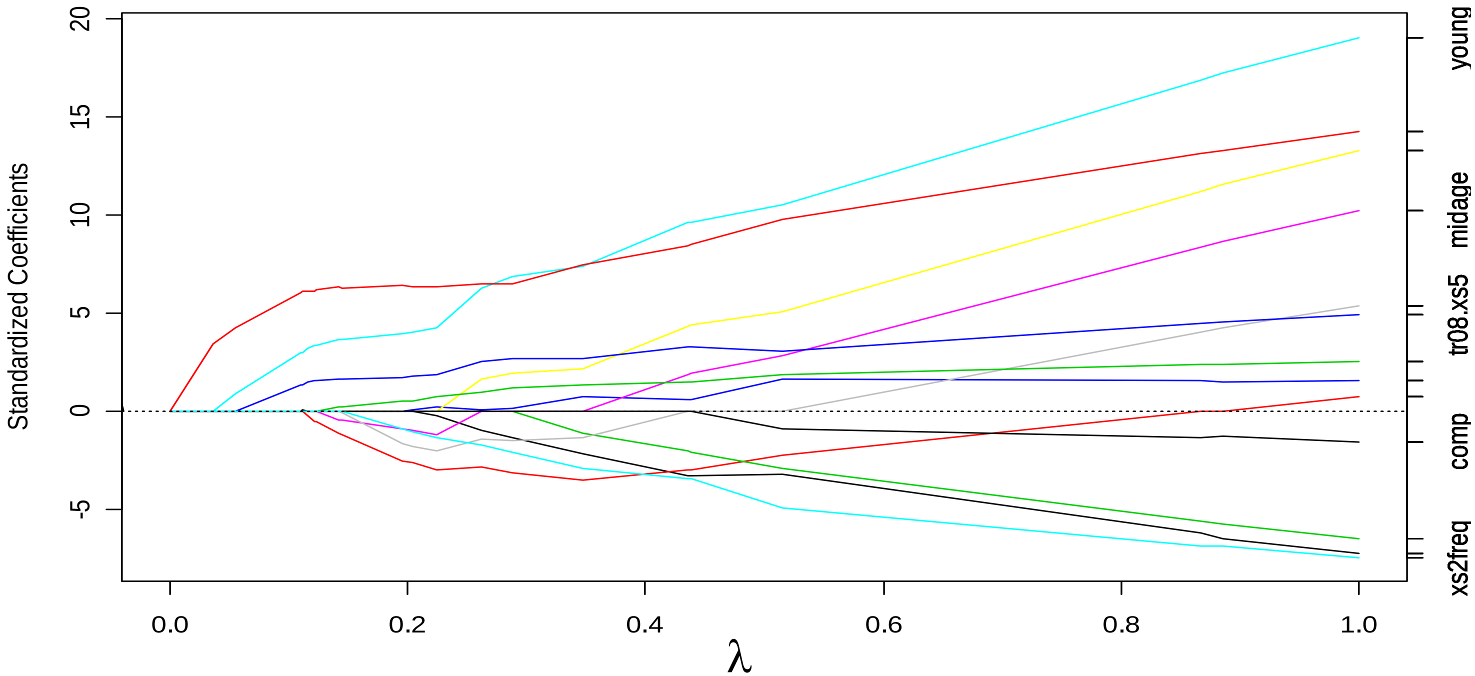
1 - Training	2 - Training	3 - Training	4 - Testing	5 - Training



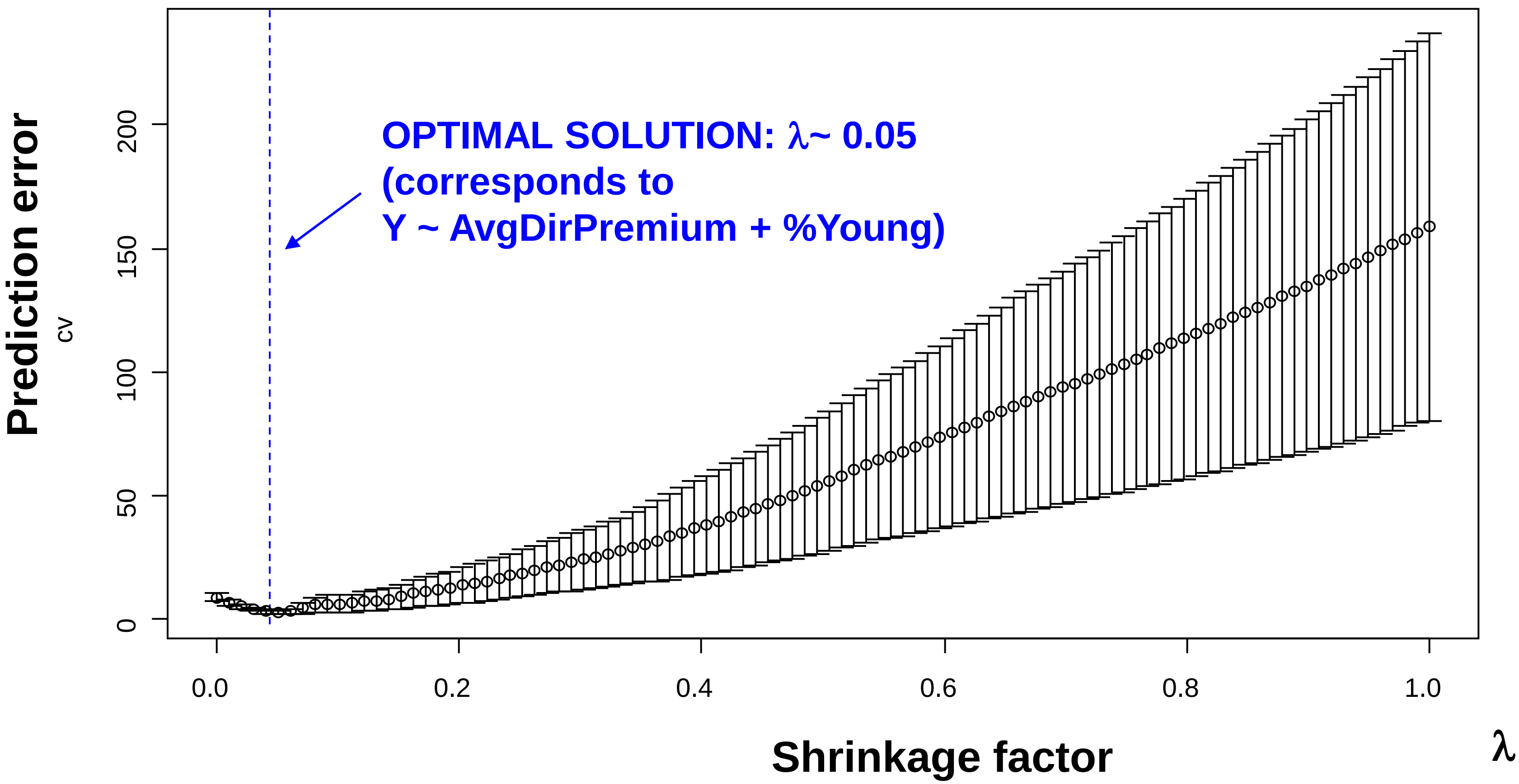
$$CV = \frac{1}{N}\sum_{i=1}^{N} L\left(y_i, \hat{f}^{\kappa(i)}(x_i)\right)$$

$\hat{f}^{\kappa(i)}(x_i)$ = fitted function with $\kappa(i)$ -th set removed

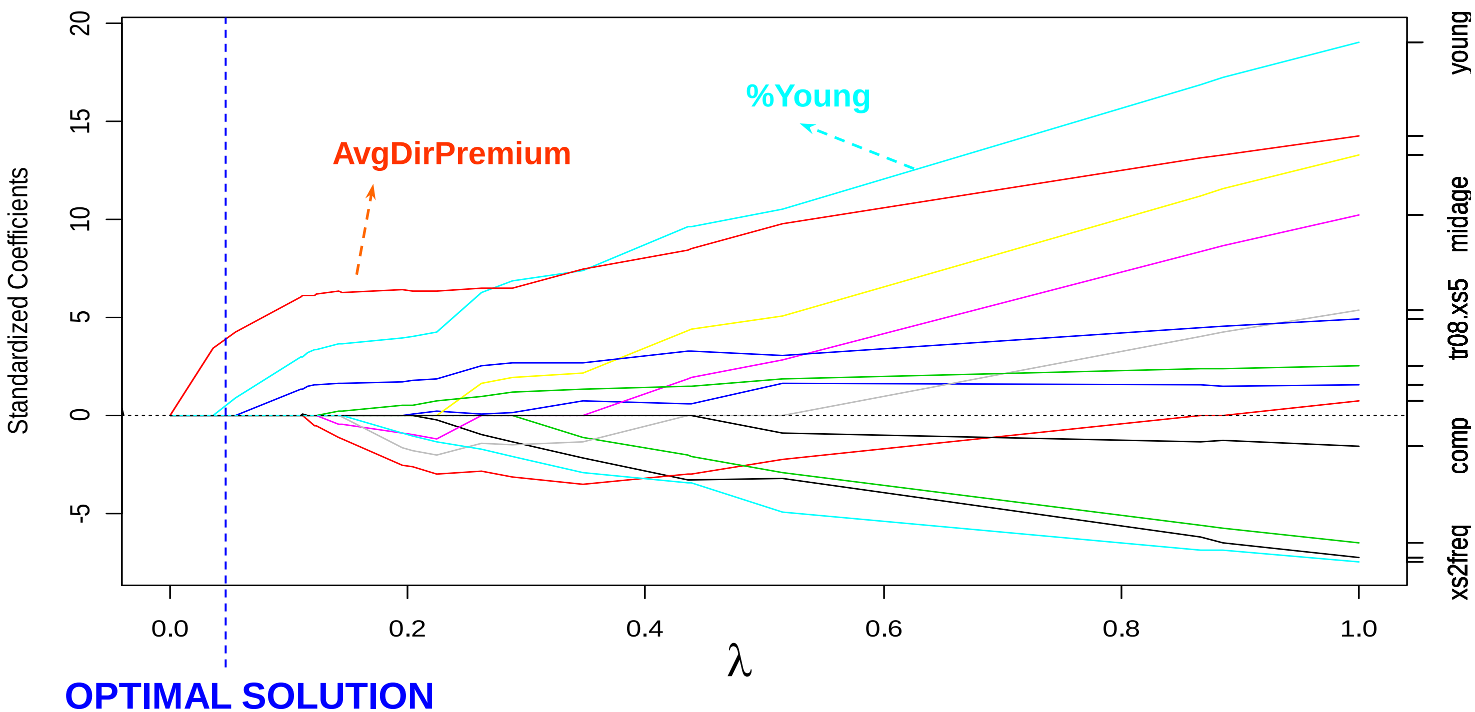
Lasso – Reinsurance example
Variables selected for different values of λ



Lasso – Reinsurance example
Model selection

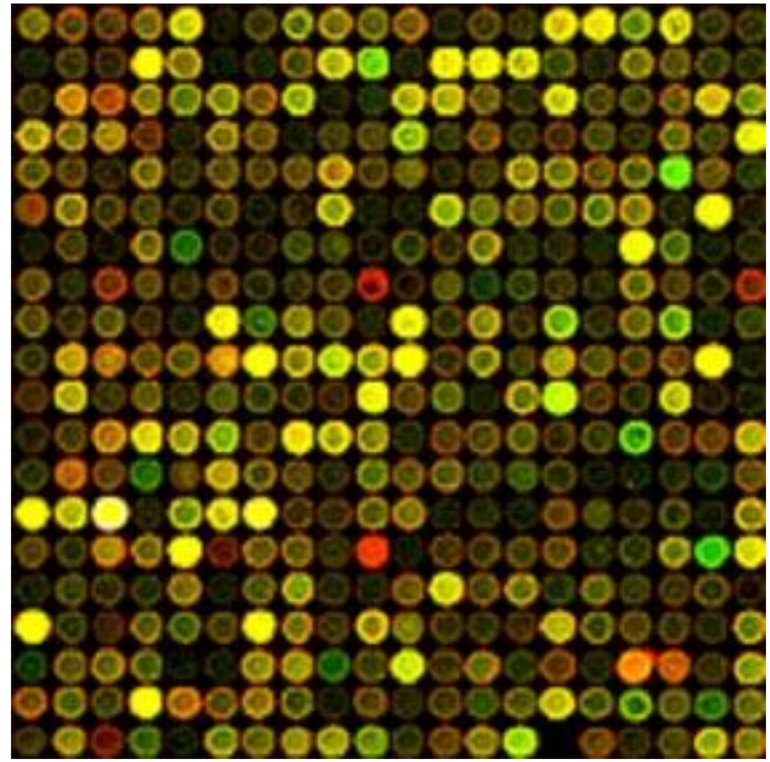


Lasso – Reinsurance example
Results



Where the lasso breaks down
Example: microarray data analysis

A microarray



Microarray technology

A tool to monitor genome-wide expression levels of genes in a given organism, as measured by the fluorescence level of spots on a glass slide (microarray)

The problem

Select the features (genes) that are responsible for a given disease, given DNA samples from a number of patients

The issues with the lasso

No of data points: ~ 100 (patients), no of genes ~ 10,000

Groups of highly correlated genes, need to capture them all

Beyond the lasso: Elastic net regularisation (Zou and Hastie, 2005)

$$\mathrm{E}_n^{\lambda}(\beta) = \left\| Y - f_{\beta}(X) \right\|_{l_2}^2 + \lambda \left\| \beta \right\|_{l_1} + \mu \left\| \beta \right\|_{l_2}^2$$

Improvements over the lasso

- Allows **variable selection** but avoids the excesses of lasso
- Deals successfully with **data sparsity**
- Deals with **groups of correlated features**

How is this relevant to insurance?

- Data sparsity is ubiquitous, especially in reinsurance and commercial insurance
- Many rating factors are strongly correlated (e.g. choice of comprehensive motor policies and driver's age)

III. A comparison between GLM and regularisation

Comparison of GLM and regularisation

GLM

- “log P” loss function more general than squared loss
- Greedy algorithms may get stuck in local minima
- Limited by linearity (but a large dictionary of functions is possible)

Regularised regression

- Guaranteed minimum and very efficient
- Can address cases where there # variables » # data points
- Use of quadratic loss function is a limit when data are sparse and the process is non-Gaussian: the Poisson example

Comparison of GLM and regularisation, using artificial Poisson data

$\mathbf{E}[Y] = c \cdot \exp(0.2 \cdot \text{Sex} - 0.3 \cdot \text{Age} + 0.15 \cdot \text{Region} - 0.4 \cdot \text{NCB} + 0.1 \cdot \text{Profession})$

(Y = number of motor losses; $Y \sim \text{Poi}$)

GLM performs well when the average Poisson rate decreases. What about the lasso?

Lasso performance as a function of overall exposure/frequency

	Sex	Age	Region	Colour	NCB	Profession	Garden	Dumb1	Dumb2	Dumb3
True model	0.20	-0.30	0.15	0.00	-0.40	0.10	0.00	0.00	0.00	0.00
Lasso										
Exp = 10m	0.21	-0.30	0.15	0.00	-0.41	0.10	0.00	0.01	-0.01	0.00
Exp = 1m	0.20	-0.28	0.16	0.04	-0.40	0.09	0.00	0.00	-0.01	0.00
Exp = 100k	0.09	-0.18	0.14	0.09	-0.18	0.07	0.04	0.00	-0.01	-0.06

The best of both worlds?

We have compared the textbook approach of GLM to a textbook approach to regularisation. However, hybrid approaches are possible:

- **Rigorous model selection/validation methods** of machine learning can be used in GLM without modifications
- The **limitations of the quadratic loss function** can be overcome by, e.g., using a regularised version of GLM:

  Park and Hastie, 2006: “*L1-regularized path algorithm for generalized linear models*”

Questions or comments?

Expressions of individual views by members of The Actuarial Profession and its staff are encouraged.

The views expressed in this presentation are those of the presenter.