ASTIN, AFIR/ERM and IACA
Colloquia
Innovation & Invention
23-27 August 2015 | Amora Hotel Jamison Sydney

Actuaries Institute

A discussion on credibility and penalised regression, with implications for actuarial work

Prepared by Hugh Miller

Presented to the Actuaries Institute
ASTIN, AFIR/ERM and IACA Colloquia
23-27 August 2015
Sydney

This paper has been prepared for the Actuaries Institute 2015 ASTIN, AFIR/ERM and IACA Colloquia.

The Institute's Council wishes it to be understood that opinions put forward herein are not necessarily those of the Institute and the Council is not responsible for those opinions.

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 2, 50 Carrington Street, Sydney NSW Australia 2000
t +61 (0) 2 9233 3466 **f** +61 (0) 2 9233 3446
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

Abstract

This paper explores the use of credibility and penalised regression for actuarial work, with particular application to reserving work. We survey recent developments in the statistical literature in penalised regression and boosted learning, and compare them to actuarial credibility approaches. We demonstrate the strong theoretical links between the different approaches. Further, we illustrate applications to actuarial problems such as reserving and pricing, where approaches based on penalised regression can provide an intuitive way to impose desired structure.

Keywords: Pricing, reserving, credibility, penalisation, boosting

1 Background

1.1 Introduction

Penalised regression approaches have proven to be a fruitful area of research in statistics. The key idea is to add a 'penalty' to a regression to encourage desirable behaviour in the model; often this is done to reduce variability in estimated parameters. A web search of papers with 'penalised' (and variants) in the title of six leading statistics journals uncovers 133 such papers over the past 20 years. However, these ideas have not permeated the actuarial landscape; an identical search of six leading actuarial journals yields zero such papers.

In contrast, credibility theory is a long-established actuarial discipline and continues to be an active area of research. Performing a similar search for 'credibility' in the title finds 109 papers in actuarial journals, compared with two papers in the statistics journals.

Intriguingly, this separation in emphasis between the actuarial and statistical disciplines masks deep connections between the two approaches. While they have different viewpoints and emphasis, the solutions produced are often identical. Further, penalised regression approaches allow actuaries to tackle a wider range of credibility setups. Knowledge of both approaches has the potential to greatly enhance actuarial assumption setting.

This paper provides a gentle introduction to penalised regression, and points readers to relevant literature. Section 2 introduces a basic credibility problem and then solves it using both Bayesian credibility and penalised regression. Section 3 explores applications to different actuarial problems. Section 4 discusses some (relatively easy) theoretical results and Section 5 explores further links to boosting and mixed models.

1.2 Relevant literature

The Bayesian approach to credibility in actuarial literature dates back to Bühlmann, H. (1967). The theory developed over the following decades, with important contributions by Bühlmann & Straub (1970), Jewell (1974), Taylor (1979), Heilmann (1989), Verral (1990), Nelder & Verrall (1997) and De Alba (2002). Makov (2001) provides a survey of approaches to different actuarial problems. Ohlsson (2008) presents (standard and hierarchical) credibility for multiplicative models, appropriate to the generalised linear models typically used in rating problems.

The textbook by Bühlmann & Gisler (2006) gives a comprehensive treatment of credibility theory, including the popular Bühlmann-Straub model. The text by Klugman (1992) also covers Bayesian credibility in actuarial applications in depth.

The origins of penalised regression can be traced back to ridge regression which was developed by Hoerl & Kennard (1970). Its popularity has grown dramatically since the development of the LASSO (Tibshirani, 1996). Important contributions include those by Fan & Li (2001), Tibshirani et al. (2002), Efron et al. (2004), Zou & Hastie (2005), Zou (2006), Candes, E., & Tao (2007) and Yuan & Lin (2007). Some uses of penalised models for premium rate setting were explored by Semenovich (2013).

The first popular boosting algorithm was developed by Freund & Schapire (1996) although related ideas can be traced back to Schapire (1990). Theoretical developments include those by Schapire et al. (1998), Friedman et al., (2000) and Friedman, (2001). The last introduced boosting with decision trees, a popular application. Interest has grown due to its links with other penalised procedures (see Friedman et al, 2010, for instance).

The textbook by Hastie et al. (2009) provides further background on penalised regression and boosting procedures.

2 Credibility and Penalised regression

2.1 Notation

We start by introducing the notation for a typical predictive modelling setup. Let $Y_1, Y_2, \dots, Y_n$ be a series of response variables, and let each of $X_1, X_2, \dots, X_n$ be a p-vector of predictor variables. We aim to find a function of X_i such that it is a good estimate of Y_i. This is usually done with reference to a loss function L, so that f is chosen to minimise the average loss between the $f(X_i)$ and Y_i:

$$\frac{1}{n}\sum_{i=1}^{n} L\{Y_i, f(X_i)\}\,. \tag{1}$$

One of the most common setups is linear regression with squared loss function, which we use extensively throughout the paper. In this case $f(X_i) = \sum_{j=1}^{p} X_{ij}\beta_j = X_i^T\beta$, where the p-vector β is chosen to minimise:

$$\frac{1}{n}\sum_{i=1}^{n} (Y_i - X_i^T\beta)^2\,. \tag{2}$$

Many other types of setups, including generalised linear models (GLMs) are possible under the general formulation of (1). However, the linear model is instructive and often the most tractable algebraically.

2.2 Introductory example

Suppose we were dealing with setting industry premium relativities (that is, the relative amount that an industry premium should differ from the average premium) for a workers' compensation scheme. There are n insured companies that belong to p different industries. The average relativity for each company is observed and the task is to choose appropriate industry specific relativity based on the data.

Figure 1 Illustration of setting industry premium relativities

	Industry	Observed company Relativity
Company A	Mining	5.0
Company B	Mining	2.4
Company C	Agriculture	0.7
Company D	Manufacturing	1.2
	…	

→

Industry	Adopted Relativity
Mining	3.0
Agriculture	0.7
Manufacturing	1.2
…	

In this setup let:

- There are p industries, with potentially different numbers if companies in each
- The true industry relativities are normally distributed around $\mu = 1$, $R_j \sim N(\mu, \tau^2)$
- The observed company relativities are distributed around their corresponding industry relativity, $Y_i \sim N\left(R_{J(i)}, \sigma^2\right)$

For convenience, we assume each company has equal weight.

It would be possible to simply adopt the average relativity observed within each industry. However, this would typically be suboptimal; industries with fewer companies would be subject to significant variability, so adding some reversion to the mean would likely improve the adopted relativities.

The example presented is a fairly standard actuarial credibility problem. Although nominally a problem of relativity estimation, the idea of assigning partial credit to the observed data is common to many actuarial problems and so the example and its intuition represent a fairly general situation.

We also present some numerical results related to this example throughout. To do this we have created a simple dataset of the above type:

- There are 10 industries, with the number of companies per industry sampled from a Negative binomial distribution with mean 10 and $= 1$. The smallest industry has 2 companies and the largest has 39.
- We have set $\tau^2 = 0.2^2$ and $\sigma^2 = 0.3^2$

The full dataset is provided in the Appendix.

2.3 A Credibility approach

2.3.1 A Bayesian solution

A standard actuarial credibility approach to this example problem would be to apply Bayes' Theorem to obtain a posterior distribution. For a given industry J, we have a prior distribution on the industry relativity:

$$f(r) = (2\pi)^{-1/2} \exp\{-(r-1)^2/2\tau^2\} \tag{3}$$

And a likelihood conditional on a company's industry relativity:

$$f\left(y|R_J\right) = \prod_{i:J(i)=J} (2\pi)^{-1/2} \exp\left\{-\left(y_i - R_J\right)^2/2\sigma^2\right\} \tag{4}$$

Applying Bayes' theorem gives a new normal distribution with mean estimate (if n_j is the number of companies in industry j and $\bar{Y}_j$ is the average observed relativity for that industry.

$$\hat{R}_j = 1 + \frac{n_j}{n_j + \sigma^2/\tau^2}(\bar{Y}_j - 1) \tag{5}$$

We recognise this as applying a proportion, $\frac{n_j}{n_j+\sigma^2/\tau^2}$, of the observed industry effect to the final relativity. This proportion grows as the number of companies within an industry increases.

In practice, the variance terms τ^2 and σ^2 would have to be estimated as well. There are standard formulae for doing this. For instance (Ohlsson, 2008 or Bühlmann, H., & Gisler, A., 2006) we could set:

$$\hat{\sigma}^2 = \frac{\sum_i \left(Y_i - \bar{Y}_{J(i)}\right)^2}{n-p}, \qquad \hat{\tau}^2 = \frac{\sum_j \left(\bar{Y}_j - \bar{Y}\right)^2 - (p-1)\,\hat{\sigma}^2}{n - \sum_j n_j^2/n} \tag{6}$$

We refer to estimates such as these as 'plug-in' estimates, as they can be calculated directly from the data and incorporated into the resuts.

The Bayesian solution will provide optimal estimates of industry relativities against the true relativities, where error is measured by average squared error:

$$\frac{1}{p}\sum_j \left(\bar{Y}_j - R_j\right)^2 \tag{7}$$

We can measure the performance on our example dataset by generating a large test dataset on the same basis – here with 5,000 companies per industry. The credibility approach performs significantly better compared to alternatives such as setting the relativity to 1 for all industries (the 'constant model'), or using the observed industry average as the final relativity (which we call the ordinary least squares or 'OLS model', as it corresponds to the solution of solving the standard linear model to obtain industry averages).

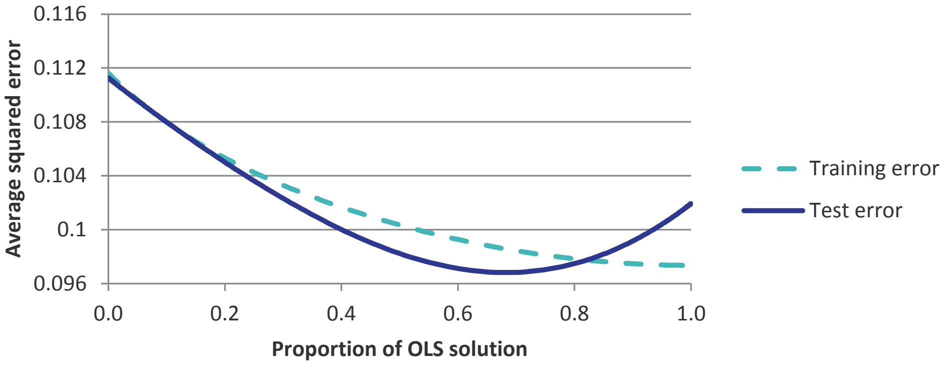
Table 1 Average squared error as given in equation (7) on example dataset

Approach	Avg error on test data
Global average relativity ('constant model')	0.107
Bayesian estimates	0.097
Industry average ('OLS model')	0.102

2.3.2 A further thought experiment on the credibility approach

Suppose that we did not have a plug-in estimate for τ^2 and instead had to find it using some form of experimentation. Such situations do occur in practice; variance parameters such as τ^2 can sometimes be difficult to estimate while maintaining plausible credibility estimates. If we set τ^2 close to zero, then this corresponds to a belief that relativities should be very close to 1 for all industries and very little credibility is given to the observed data. Conversely, if we set τ^2 to be very large then we have little pre-existing belief around the relativities, and the corresponding estimates will rely almost exclusively on the observed data. Thus as τ^2 varies from 0 to infinity, it indexes a series of models that vary from the constant model through to the OLS model – the two extremes shown in Table 1. For our example dataset, we vary τ^2 in this way and for each value measure the average training error (the average error on the original dataset) and the average test error (the error on the large test dataset). The results are summarised in the figure below.

Figure 2 Performance of industry relativity estimation as τ^2 is varied



Rather than using τ^2 directly, the x-axis in the figure is the proportion of progress towards the OLS solution as measured by the degree to which the estimated industry relativities differ from the overall mean:

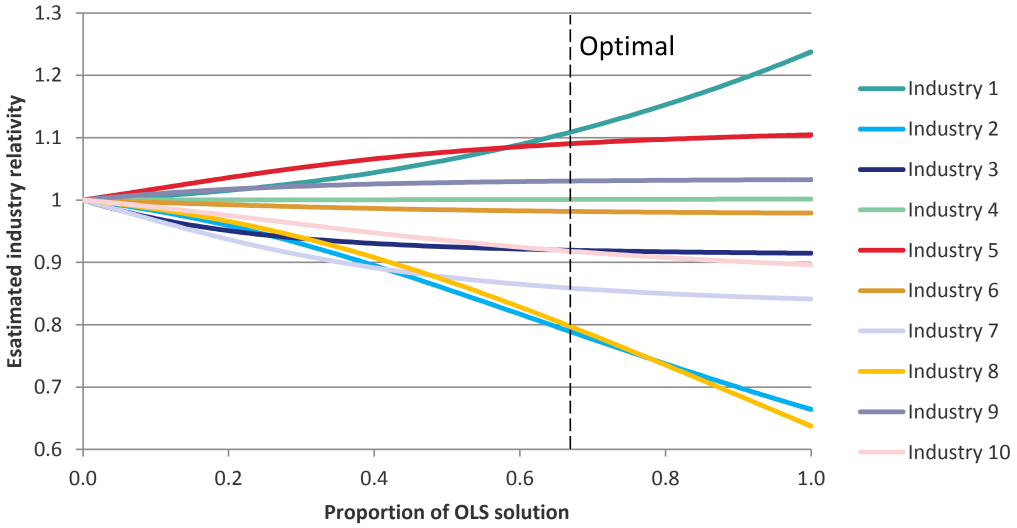
$$\text{Proportion of OLS solution} = \frac{\sum_j \left|\hat{Y}_j - \bar{Y}\right|}{\sum_j \left|\bar{Y}_j - \bar{Y}\right|} \tag{8}$$

This proportion corresponds to the average credibility factor across industries. While the training error unsurprisingly favours the full OLS, the optimal test error favours an intermediate value where the average credibility factor is about 0.68. This corresponds to a τ^2 value of 0.202^2, very close to the true value.

Thus by allowing τ^2 to vary, we have effectively turned a Bayesian credibility problem in a frequentist tuning one; the 'best' model is the one across the range of choices that minimises test dataset error.

It is also instructive to examine how the individual industry relativities are changing as τ^2 grows; this is shown in the figure below. When $\tau^2 = 0$ all industry relativities equal 1 and they then grow with τ^2. The speed of growth is not uniform amongst parameters. For instance, the parameter for Industry 5 (the second from top line) initially grows faster than Industry 1 (the topmost line), but is then overtaken. This reflects the relative size of each industry; industry 5 has fifteen companies in the dataset so is given greater credibility early, whereas Industry 1 has only two companies and so is given little credibility until τ^2 grows relatively large.

Figure 3 Change in estimated industry relativities as τ^2 is varied



2.4 A penalised regression approach

2.4.1 Ridge regression

To apply penalised regression, we first restate the example in the notation of Section 2.1. Let X be an $n \times p$ matrix, with $X_{ij} = 1$ if the i^{th} company belongs to the j^{th} industry, and $X_{ij} = 0$ otherwise. The OLS solution is then recovered by solving equation (2) for $\beta = (\beta_1, \dots, \beta_p)$ with each β_j corresponding to the j^{th} industry relativity. The penalised regression solution differs from equation (2) in that adds a penalty term that is a function of the β.

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + P(\beta) \tag{9}$$

The penalty is chosen to encourage 'good' behaviour amongst the parameters, using penalising them for being too high. In our current context, we may believe it reasonable to penalise relativities that are significantly different from 1:

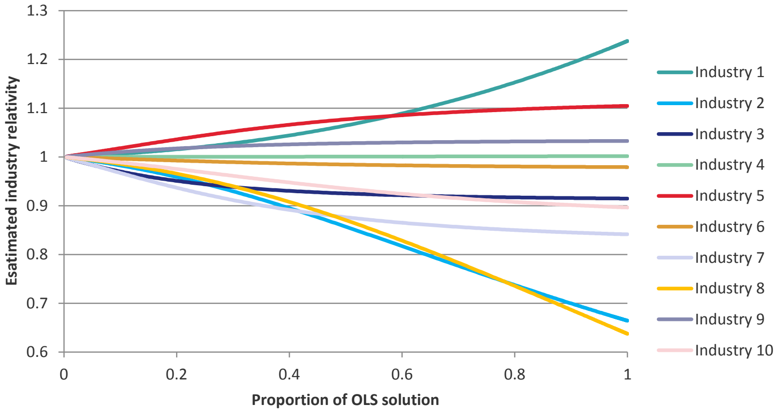
$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\sum_{j}\left(\beta_j - 1\right)^2 \tag{10}$$

An equivalent, but slightly more common presentation would be to take the mean ($\mu = 1$ here) out of the β_j, making the required problem choosing β to minimise

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - (\mu + X_i^T\beta)\}^2 + \lambda\sum_{j}{\beta_j}^2\,, \tag{11}$$

The final relativities for each industry are then $\mu + \beta_j$. Equation (11), linear regression with a squared penalty term, is known as ridge regression (Hoerl & Kennard, 1970). When λ is large, the penalty forces the β_j to be close to zero, yielding the constant model. When $\lambda = 0$ the penalty has no impact and the OLS solution is recovered. So varying λ from infinity through to 0 creates an index of models ranging from the constant to the OLS mode. The figure below shows the parameter evolution for this index of models, utilised in the same x-axis as previous figures.

Figure 4 Change in estimated industry relativities using ridge regression as λ is varied



Inspection of the results shows that the ridge regression parameter evolutions are **exactly the same** as those arising from the credibility approach. Further, the choice of λ is typically chosen with reference to performance on a test dataset, giving the same 'optimal' choice as seen in Figure 3.

The equivalence between the credibility result with normal priors and ridge regression is shown in Corollary 1 of Section 4. This observation stems from a much deeper result that virtually all credibility problems can be expressed as a penalised regression problem; we formalise this idea in Theorem 1 in Section 4.

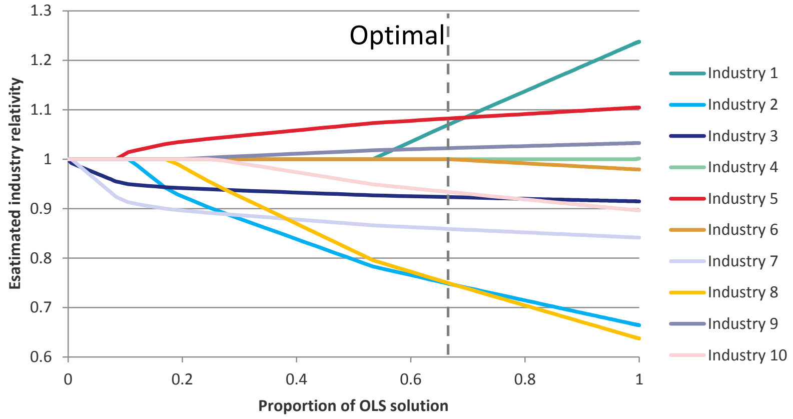
2.4.2 Other types of penalised regression

One of the most popular (and heavily researched) penalties is the LASSO, which uses an absolute penalty on the coefficients:

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - (\mu + X_i^T\beta)\}^2 + \lambda\sum_{j}|\beta_j| \tag{12}$$

As with ridge regression, varying λ will index a series of models from the constant through to OLS. In contrast to the ridge regression, the LASSO has the property of producing **sparse** solutions where some of the coefficients are held at zero while others grow. This is illustrated in the figure below. The optimal solution (found via test dataset error) is at proportion 0.65 towards the OLS solution, and at this point the relativity for Industries 4 and 6 are still exactly equal to 1. The implication is that it is unnecessary to vary the estimate for these industries away from the 'default' relativity, which has an attractive intuition to it. We also see in this example that industries that contain more companies tend to be added first, with Industry 1 (which has few companies but large observed effect) added relatively late in the parameter evolution.

Figure 5 Change in estimated industry relativities using LASSO regression as λ is varied



More generally, there are a large number of different possible penalties, which will tend to have different parameter evolution paths. The power index, $P(\beta) = \sum_j |\beta_j|^\gamma$, is a common class of penalties that includes the ridge and LASSO. More complex functions are also possible and we give some examples of possible choices in Section 3. A property often required of the penalty function is convexity; imposing this and combining with a convex loss function will generally result in a numerically tractable computation.

2.5 Further discussion

The equivalence between actuarial credibility approaches and penalised regression has some significant implications. We discuss some of these, plus some additional considerations:

- Bayesian credibility analysis is usually confined to distributions (of priors and observations) that are algebraically tractable. Re-interpreting a given setup as a penalised regression optimisation potentially allows a broader class of distributions to be used.
- The sparse nature of LASSO allows it to produce estimates in situations where there are **more** predictors than observations. This is because the LASSO performs a dual function of parameter estimation and variable selection, powerful in such situations.
- The predictors in our example were all indicator functions, but there is no reason why equation (8) need be limited to such predictors. Penalising a continuous variable has the same intuition; it adds a portion of the measured continuous effect, rather than the full amount, with a view towards improving performance on future unseen data. If there are multiple continuous variables it is usual to first scale them to unit variance, so that the penalty applies equally to each.
- In Bayesian credibility estimating the variance parameters can be difficult in practice, especially when the true distributions of the priors are unknown, or if the dataset is small. For instance, we have experienced cases where the formulae in equation (6) have produced negative variance estimates! Treating such parameter selection as a frequentist error minimisation problem can offer practical advantages in such situations.
- We have restricted the examples to squared error, which is naturally tied to ordinary least squares regression problems. However the approach is general, and penalised regressions for generalised linear models have been implemented in a number of popular statistical packages.
- The original motivation for ridge regression was to improve stability when predictor variables were highly correlated. In such situations unpenalised fits often produce wild

parameters that largely offset each other. Penalised regression still helps address this issue, even if the original motivation for the approach has shifted.

3 Actuarial applications

3.1 Introduction

We have already provided one explicit example where an actuarial credibility problem can be reinterpreted as a penalised regression problem. This may have some practical advantages (particularly if the analyst wants to explore alternative penalty functions or priors), but is essentially re-solving an existing problem. However, there are other areas where penalised regression may add genuine value to other actuarial problems. In this section we give a brief discussion of some possibilities.

3.2 Reserving

Suppose we were estimating continuance rates for a workers compensation scheme. The value of each cell in the triangle, N_{ij} represents the number of active claimants at the end of the year for accident year i and development year j. A pseudo-dataset is presented below to illustrate. Suppose further that:

- The continuance rate (an estimate of $C_j = N_{i,j}/N_{i,j-1}$) for each development period had been estimated in the previous valuation with a weight attached to each estimate. Call this $\hat{C}_{j,old}$, with weight $W_{j,old}$
- The latest year continuance rates are now observable $\hat{C}_{j,new}$, with weight $W_{j,new}$

Table 2 Example actives triangle for estimation of continuance rates

Accident	Development year														
Year	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	698	603	245	188	147	108	99	89	80	71	60	60	51	42	45
2	713	668	320	227	180	167	131	105	87	81	77	68	62	60	
3	764	719	291	191	143	93	95	74	69	68	74	65	51		
4	702	657	273	171	132	111	96	72	75	71	60	60			
5	636	593	203	116	87	66	59	56	48	42	36				
6	726	647	314	207	152	138	116	102	87	87					
7	749	665	296	206	153	144	119	102	99						
8	725	650	264	152	117	107	92	80							
9	672	645	275	176	132	108	98								
10	716	705	288	188	147	122									
11	848	750	321	216	150										
12	836	752	327	209											
13	845	728	290												
14	854	773													
15	945														
Avg prev 3 (excl last)		0.882	0.424	0.656	0.767	0.893	0.841	0.884	0.913	etc					
Latest		0.905	0.398	0.639	0.694	0.830	0.907	0.870	0.971						

A common approach selecting continuance rates is to set the new selection equal to the previous year's selection unless the latest period is 'sufficiently different' to motivate a change. A penalised regression formulation of this idea could be attempting to mimimise:

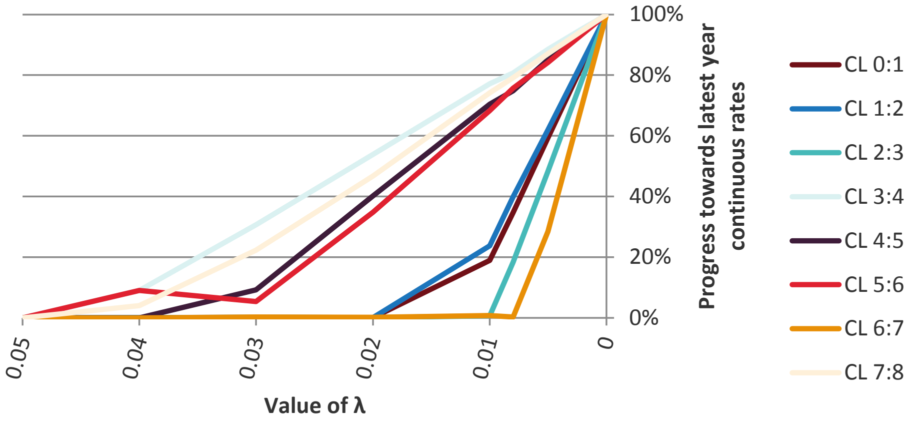
$$\sum_{j=1}^{J} W_{j,new}\left(C_{j,selection} - \hat{C}_{j,new}\right)^2 + \lambda \sum_{j=1}^{J} W_{j,old}\left|\hat{C}_{j,new} - \hat{C}_{j,old}\right| \qquad (13)$$

Using (13), we see that a large value of λ will favour the old estimates, and $\lambda = 0$ will entirely rely on the latest data. For the example in Table 2 we have:

- Set the 'old' continuance rates based on the weighted average of the latest three years (excluding the very last). The weight is the sum of actives for those three accident years in the prior development period
- Set the new continuance rates based on the latest development year, with weight the number of actives in the immediately prior period.

The figure below shows the results as λ is varied from 0.05 (where the old rates are entirely adopted) to 0.00 (where the new rates only are selected. The LASSO-like penalty means that for intermediate values a subset of parameters are held at their 'old' values, consistent with the original intuition. In this case choosing an intermediate value of $\lambda = 0.02$ would lead to adjustment of 4th, 5th, 6th and 8th continuance rate factors, but leave the others unchanged. The adjusted factors can be thought of as a weighted average between the two extremes. This behaviour appears consistent with the actual experience in the triangle; these middle years have reasonable weight and the continuance rate behaviour has changed markedly.

Figure 6 Progress of continuance rate selection towards latest data as λ is decreased to zero



The penalisation behaviour of (13) is effective in stabilising changes in factors from year to year. Other additional penalties could also be added to encourage other desirable behaviour, such as monotonicity of differences in selected continuance rates across development periods.

Interestingly, if actuarial selections can be encoded as a penalised regression, this can then be interpreted in the equivalent credibility setting and it would be possible to then obtain a 'revealed' estimate of the priors used in setting assumptions.

3.3 Pricing

Pricing models for an insurer attempt to estimate the risk or claims cost associated with a customer, given their characteristics. A common feature of pricing models is the 'winner's curse' where pockets of an insurers' portfolio might be priced lower than would otherwise be the case due to statistical noise. If they chase such pockets with lower prices, they may then discover worse than expected performance as a reversion to the mean is observed in subsequent periods. There are two ways in which a penalised regression (or Bayesian credibility) approach can ameliorate this situation:

- **Penalising model updates:** Suppose that we have an existing least squares linear model $\hat{Y}_{i,old} = X_i^T \beta_{old}$ which we desire to update with new parameters (here least squares is again used for simplicity in presentation – a generalised linear model would be more standard). Letting $\hat{Y}_{i,new} = X_i^T \beta_{new}$, we could choose $\beta_{new}$to minimise

  $$\sum_{i=1}^{n}(Y_i - X_i^T \beta_{new})^2 + \lambda \sum_{i=1}^{n}\left(\hat{Y}_{i,new} - \hat{Y}_{i,old}\right)^2 \tag{14}$$

  Thus we can explicitly control change the degree of change in the model by penalising the departure from the previous predictions. This is consistent with standard credibility premium updating formulae, but the equation above will generally be more tractable for more complex pricing models. Such penalisation has the additional desirable property that it resists model changes that significantly redistribute risks. It is not uncommon to observe pricing models with similar aggregate performance (in terms of predictive power) but having very different predictions on portions of the customer base – the above approach helps to mitigate this problem.
- **Penalising model terms directly:** Usually some variables are harder to model accurately than others. For instance in car insurance, car make is typically difficult as some makes have large exposure while some other makes might have small exposure (but potentially large observed effects). Similarly a detected interaction might appear extreme, but the analyst usually has to make an 'all-or-nothing' decision on its inclusion. A ridge regression such as (11) can be an effective way to balance these risks. In practice it is not uncommon to select only a subset of variables to include in the penalty. These might be interactions or parameters related to categorical variables with many levels. Thus it is possible to build some reversion to the mean into the model, validated on test datasets. In situations where many potential variables are considered, the LASSO might also be an effective way to perform variable selection or interaction detection.

As an interesting aside, Lovick and Lee (2012) use the 'case deleted deviance' to stabilise model coefficients in a GLM. This approach can also be restated in a penalised regression framework of the second bullet – a GLM with a ridge regression coefficient applied to the coefficients.

3.4 Automatic curve fitting

Many actuarial tasks can be thought of as curve fitting problems. The challenge is often different to classical parametric curve fitting, since there is often no expectation that a parametric shape (such as a line or parabola) is appropriate to the problem. However other intuitions remain, such as a preference for smoothly varying estimates unless there is clear evidence of a step change. Such intuitions can be encoded in a penalised regression.

We illustrate with an example – suppose we observe the number of claims $Y_1, \dots, Y_n$, distributed as a Poisson variable, across 30 consecutive time intervals $t = 1, \dots, n$. Let $\beta_i = \log(\mu_i)$ be the underlying mean that we wish to estimate in each period. An unpenalised log-likelihood is (up to a constant):

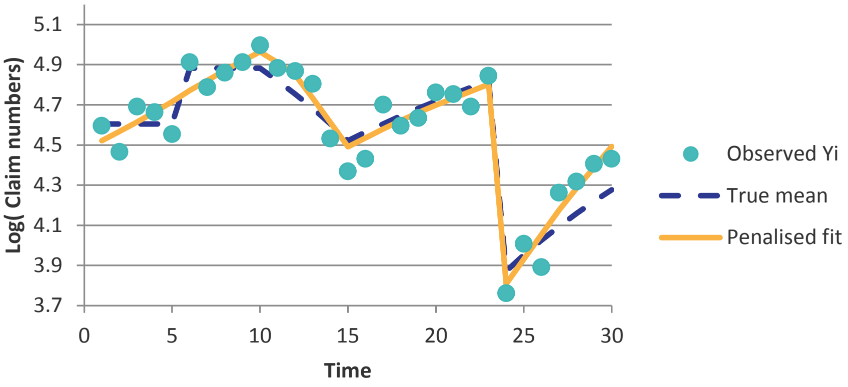
$$-\sum_{i=1}^{n} Y_i\beta_i - e^{\beta_i} \tag{15}$$

Solving this for maximum likelihood would result in $\widehat{\beta_\iota} = \log Y_i$ for each i, applying no smoothing to the fit. However the expression:

$$-\sum_{i=1}^{n}\left(Y_i\beta_i - e^{\beta_i}\right) + \lambda_1\sum_{i=1}^{n-1}|\beta_{i+1}-\beta_i| + \lambda_2\sum_{i=1}^{n-2}|(\beta_{i+2}-\beta_{i+1})-(\beta_{i+1}-\beta_i)| \tag{16}$$

introduces a penalty on a prediction being different than the previous (the λ_1 term), plus an additional penalty on the slope between successive predictions changing (the λ_2 term). This is enough structure to produce many plausible types of fits. We show an example in Figure 7 below where we have specified a true mean function, sample using the Poisson assumption and then applied a penalised fit with $\lambda_1 = 0$ and $\lambda_2 = 8$ (that is, we impose a change in gradient penalty but no change in level penalty). In this case the fit reproduces the underlying shape well.

Figure 7 Penalised regression fit to Poisson distributed data.



3.5 Hierarchical credibility

Penalised regression can also be applied to hierarchal credibility. In the example of 2.2, suppose that each company had a number of subsidiaries Z_i that we observed instead of just a company relativity. The subsidiary relativities are normally distributed around the company relativity. This is a standard actuarial problem (see Taylor, 1979). If we create a subsidiary indicator matrix Z where each $Z_{ik} = 1$ whenever the i^{th} observation belongs to the k^{th} subsidiary and $Z_{ik} = 0$ otherwise, then the extended ridge regression:

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - (\mu + X_i^T\beta + Z_i^T\gamma)\}^2 + \lambda_1\sum_j {\beta_j}^2 + \lambda_2\sum_k {\gamma_k}^2\,, \tag{17}$$

will reproduce the standard hierarchical credibility estimates for appropriate choices of λ_1 and λ_2.

4 Theoretical results

The theoretical results in this section are not difficult; indeed the main theorem below is largely tautological. However, they are still instructive in explaining the connection between credibility and penalised regression.

Theorem 1: Finding the mode of a Bayesian posterior can be expressed as finding the solution to a penalised regression problem.

Proof: Suppose we have Bayesian parameters θ with prior $f_\Theta(\theta)$, and that we observe data that depends on the paramters with likelihood $f_{X|\Theta=\theta}(x)$. The by Bayes theorem the posterior is

$$\begin{aligned} f_{\Theta|\mathrm{X=x}}(\theta) &\propto f_{X|\Theta=\theta}(x) f_\Theta(\theta) \\ &= \exp\left[\log\{f_{X|\Theta=\theta}(x)\} + \log\{f_\Theta(\theta)\}\ \right] \end{aligned} \tag{18}$$

So finding the mode of θ in the posterior given data x is equivalent to choosing θ to minimise

$$-\log\{f_{X|\Theta=\theta}(x)\} - \log\{f_\Theta(\theta)\} \tag{19}$$

Interpreting the $-\log\{f_{X|\Theta=\theta}(x)\}$ term as the loss function on the data and $-\log\{f_\Theta(\theta)\}$ as the penalty term on the parameter, this can be viewed as a penalised optimisation. ■

Corollary 1: Bayesian credibility with normal priors on the coefficients and normal errors is equivalent to the ridge regression model.

Proof: Identifying θ with the R_j in Section 2.2, and using the normal distribution pdfs for industries and companies, we can write down Equation (19) up to a constant as:

$$\sum_i \frac{\left(Y_i - R_{J(i)}\right)^2}{2\sigma^2} + \sum_j \frac{\left(R_j - \mu\right)^2}{2\tau^2} \tag{20}$$

Minimising this is equivalent to finding the mode of the posterior. However, we know that the posterior is also normally distributed in the parameters, so the mode equals the mean. Applying the change in notation used in 2.4 and identifying β_j with $R_j - \mu$, we see that minimising (20) is equivalent to minimising

$$\frac{1}{n}\sum_i (Y_i - (\mu + X_i^T\beta))^2 + \frac{\sigma^2}{n\tau^2}\sum_j \beta_j^2\ , \tag{21}$$

recovering the ridge regression (11) with $\lambda = \sigma^2/(n\tau^2)$. ■

Corollary 2: The LASSO is equivalent to finding the mode of the Bayesian posterior with normal errors and a double exponential prior on coefficients.

Proof: Assume that the $R_j \sim \mathrm{dbl.exp}(\mu, \phi)$. That is, $f_R(r) = (2\phi)^{-1}\exp(-|r - \mu|/\phi)$.

Applying a similar argument to the previous corollary gives a log-posterior of the form

$$\sum_i \frac{\left(Y_i - R_{J(i)}\right)^2}{2\sigma^2} + \sum_j \frac{|R_j - \mu|}{2\phi}\ , \tag{22}$$

This can be similarly viewed as a LASSO in (12) with $\lambda = \sigma^2/n\phi$. ■

One feature visible in the corollaries is that the λ in the penalised regression has a direct connection to the variance parameters in the credibility model. This is a reasonable general result, and of practical value; if an optimal penalised regression is found, the λ can then be used to determine implied consequences for the variances of the priors.

5 Boosting models and mixed models

5.1 Boosting

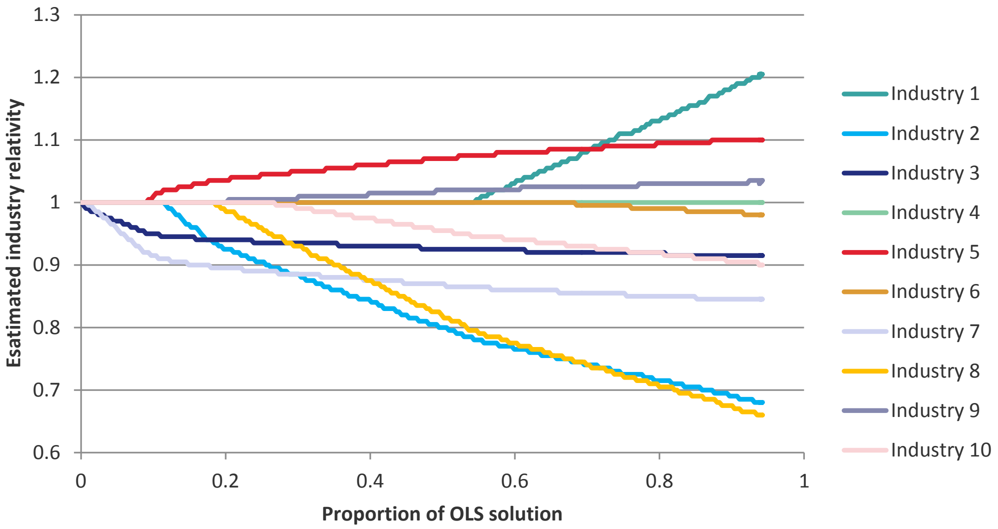
The key idea of a boosted model is that a series of simple models are iteratively added together to produce a superior model. Just as there are important links between credibility and penalised regression, there are similarly important links between boosted models and penalised regression approaches. We give one example of this relationship here. Consider the setup presented in 2.4, and the following algorithm:

L_0 Boosting algorithm:

```
1. Start with $g_0(x) = \mu$, the constant model
2. For $m = 1,2, \dots, M$
    a. Find $j$ that maxmimises $|\sum_i X_{ij}\{Y_i - g_{m-1}(X_i)\}|$
    b. Set $g_m(x) = g_{m-1}(x) + \delta x_j \text{sign}(\sum_i X_{ij}\{Y_i - g_{m-1}(X_i)\})$, for $\delta$ a small constant
3. Choose $g_m(x)$ that performs best on test data as the final model
```

The algorithm adds a small linear effect at each stage, so each of the $g_m(x)$ is a linear model. Further, it can be shown that $\sum_j |\sum_i X_{ij}\{Y_i - g_{m-1}(X_i)\}|$ decreases as the algorithm runs and the ultimate end point when each $\sum_i X_{ij}\{Y_i - g_{m-1}(X_i)\} = 0$ corresponds to the OLS solution; this last statement can be seen directly by taking the derivative of the linear model at (2) with respect to β_j. So once again we have an index of models that moves from the constant model to the OLS solution. We present the results for our relativity example, using $\delta = 0.005$ in the figure below.

Figure 8 Change in estimated industry relativities using LASSO regression as λ is varied



The results are nearly identical to the LASSO evolution in Figure 5. The only difference is that the curves are ‘blocky’ because of the discrete steps; this difference can be driven to zero by reducing the value of δ and using more steps.

This link between boosting and penalised regression gives some insight into what the LASSO is actually doing as parameters evolve; it is essentially increasing parameters that have the highest correlation between predictors and the running residuals. This interpretation was first identified by Efron et al. (2004). More generally, it has been found that a large class of penalised regression problems can be reinterpreted as boosting (or coordinate descent) models. Interested readers are referred to other papers that explore these links, such as Friedman (2012).

More practically, boosting type interpretations further enhance the computability of Bayesian and penalised regression problems. Boosting procedures tend to rely on differentiability of the loss function and some related constraints on the penalty function – given these, numerically generating a solution path of parameters via a boosting type algorithm is generally straightforward.

5.2 Mixed models

One other related approach is the mixed model. Such models are typically of the form:

$$Y_i = X_i^T \beta + Z_i^T u + \epsilon_i \tag{23}$$

The $X_i^T \beta$ term are fixed effects, similar to standard regression. The $Z_i^T u$ are random effects, with a prior distribution and related correlation structure assumed amongst the u components. The random effects part in particular can be directly interpreted in a Bayesian sense and so this formulation is often a convenient way to encode credibility problems. Additionally, many statistical packages have mixed model fitting built in, making it easier to obtain credibility estimates.

Interested readers are referred to the textbooks by Ruppert et al. (2003) and McCulloch & Neuhaus (2001).

6 Conclusion

This paper demonstrates and proves deep links between credibility theory, popular in actuarial science, and penalised regression, popular in modern statistics research. Other links to boosted and mixed models are also introduced. The ability to reframe problems in this way is potentially powerful, and creates significant opportunities to extend the use of such approaches to many actuarial problems.

Given the similarities, it is natural to ask whether one approach is better than the other. We believe that both credibility and penalised regression offer significant value. Penalised regression tends not to have ‘plug-in’ estimates for λ in the same way that credibility models do. Conversely, using penalised regression broadens the range of possible models via bespoke penalties, and more of the computational challenges have been addressed. Importantly, the ability to move between the two allows further insight into how assumptions are being set.

We believe that significant work remains in:

- Ensuring the theoretical links between approaches are developed in specific actuarial problems. This will allow solutions that are both computationally feasible and interpretable in a formal Bayesian model setting.

- Ensuring the practical tools are sophisticated enough to allow users to apply such models easily. For example, for a reserving problem a penalised fit needs to be easily implementable and adaptable so that the burden of applying it is not unduly burdensome compared to a traditional judgemental selection in a spreadsheet.

New advances in both theory and computation mean that the future of actuarial assumption setting appears bright indeed.

References

De Alba, E. (2002). Bayesian estimation of outstanding claim reserves. *North American Actuarial Journal*, 6(4), 1-20.

Bühlmann, H. (1967). Experience rating and credibility. *Astin Bulletin*, 4(03), 199-207.

Bühlmann, H., & Gisler, A. (2006). *A course in credibility theory and its applications*. Springer Science & Business Media.

Bühlmann, H., & Straub, E. (1970). Glaubwürdigkeit für schadensätze. *Bulletin of the Swiss Association of Actuaries*, 70(1), 111-133.

Candes, E., & Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313-2351.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407-499.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348-1360.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119-139.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The Annals of Statistics, 28(2), 337-407.

Friedman, J. H. (2012). Fast sparse regression and classification. International Journal of Forecasting, 28(3), 722-738.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1), 1.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The elements of statistical learning* (Vol. 2, No. 1). New York: Springer.

Heilmann, W. R. (1989). Decision theoretic foundations of credibility theory. *Insurance: Mathematics and Economics*, *8*(1), 77-95.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Jewell, W. S. (1974). Credible means are exact Bayesian for exponential families. Astin Bulletin, 8(01), 77-90.

Klugman, S. A. (1992). *Bayesian statistics in actuarial science: with emphasis on credibility* (Vol. 15). Springer Science & Business Media.

Lovick, A. C., & Lee, P. K. W. (2012). Redefining the deviance objective for generalised linear models. *British Actuarial Journal*, 17(03), 491-509.

Makov, U. E. (2001). Principal applications of Bayesian methods in actuarial science: a perspective. *North American Actuarial Journal*, 5(4), 53-57.

McCulloch, C. E., & Neuhaus, J. M. (2001). Generalized linear mixed models. John Wiley & Sons, Ltd.

Nelder, J. A., & Verrall, R. J. (1997). Credibility theory and generalized linear models. *Astin Bulletin*, 27(01), 71-82.

Ohlsson, E. (2008). Combining generalized linear models and credibility models in practice. *Scandinavian Actuarial Journal*, 2008(4), 301-314.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). Semiparametric regression (No. 12). Cambridge university press.

Schapire, R. E. (1990). The strength of weak learnability. *Machine learning*, 5(2), 197-227.

Schapire, R. E., Freund, Y., Bartlett, P., & Lee, W. S. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 1651-1686.

Semenovich, D. Applications of Convex Optimization in Premium Rating (2013). *Casualty Actuarial Society E-Forum, Spring 2013*.

Taylor, G. C. (1979). Credibility analysis of a general hierarchical model. *Scandinavian Actuarial Journal*, 1979(1), 1-12.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tibshirani, R., Hastie, T., Narasimhan, B., & Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10), 6567-6572.

Verrall, R. J. (1990). Bayes and empirical Bayes estimation for the chain ladder model. *Astin Bulletin*, 20(02), 217-243.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1), 19-35.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2), 301-320.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.

Appendix A – Dataset used in relativity example of Section 2.2

Industry	Industry relativity	# Companies	Observed company relativities
1	1.1242	2	1.1223, 1.3529
2	0.91118	4	0.4562, 0.2696, 0.7143, 1.2166
3	0.89955	39	0.6923, 1.6192, 1.2483, 0.7732, 1.1582, 0.8665, 0.7123, 1.0053, 1.1924, 1.3101, 0.3289, 1.1636, 0.6345, 1.1169, 1.1776, 0.4681, 0.5965, 0.8537, 0.4481, 1.0470, 1.0121, 0.5398, 0.6954, 0.7111, 1.3638, 0.8028, 0.5014, 0.6354, 1.2545, 1.1902, 0.8959, 0.6497, 1.2869, 1.7166, 0.6418, 0.8090, 0.5729, 0.8441, 1.1335
4	1.04617	3	1.2335, 1.2777, 0.4936
5	0.99527	15	1.327, 2.0093, 1.0743, 0.7538, 1.7095, 1.5055, 0.8248, 0.5009, 1.2022, 1.1815, 0.9912, 0.9468, 0.8021, 1.2561, 0.4837,
6	1.10071	16	1.1371, 1.0042, 1.2048, 1.357, 0.575, 1.3438, 0.8362, 0.8979, 0.6881, 1.388, 0.9544, 1.2752, 0.8618, 0.6445, 0.7294, 0.7669
7	0.77085	19	0.6701, 1.2943, 0.958, 0.7913, 1.0421, 0.9308, 1.0927, 0.5598, 0.5589, 0.9567, 1.0099, 1.1996, 0.524, 0.5067, 0.8383, 0.2109, 1.3797, 0.4358, 1.0238
8	0.69217	3	0.2256, 0.6103, 1.0762
9	0.94736	32	1.2685, 1.4808, 0.8494, 1.2255, 0.847, 0.9476, 1.1787, 1.3984, 0.6495, 0.8763, 1.3731, 0.9947, 0.7941, 0.797, 1.1971, 0.7174, 0.6609, 1.4600, 1.1357, 1.3292, 1.0479, 0.7416, 1.2657, 0.8568, 1.1674, 0.8403, 1.0375, 0.4283, 1.5123, 0.6521, 1.2120, 1.1031
10	0.84724	9	0.9277, 0.9985, 0.5335, 1.1652, 1.0001, 0.9947, 0.4392, 1.1106, 0.8965