Scorecard Modelling Best Practice – Does It Work in Theory?

Alan Forrest

Practitioner in Residence, Credit Research Centre at the Business School, University of Edinburgh

Head of Model Risk Oversight, Virgin Money UK

15th April 2021

Summary

- Introduction
  - Scorecards, best-practice and validation challenges
  - Scorecard model shifts – the key to their model risk quantification
- Information geometry of scorecard model shifts
  - Scorecard shifts have two descriptions – covariant and contravariant
  - Calculating scorecard model shifts in practice
- Geometric insights and conclusion
  - How scorecard best-practices look from this geometric point of view

Background to Information Geometry

Origins

- Rao, C.R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. 37, 81–89.

Contingency Tables

- S.Kullback, (1978) , Information Theory and Statistics
- Chentsov, N. N. (1982). Statistical Decision Rules and Optimal Inferences. Transactions of Mathematics Monograph, # 53 (Published in Russian in 1972).

Modern Theory

- Amari, S. and Nagaoka, H. (2000). Methods of Information Geometry. Oxford University Press.

Recommended Modern Introduction and Survey

- Frank Nielsen (2020) An Elementary Introduction to Information Geometry, [1808.08271v1] An elementary introduction to information geometry (arxiv.org)

Scorecards – the Workhorses of Banking

Factor	Code	Description	Score
intercept			310
Time in address	2	13-36 months	26
	3	37-72 months	37
	4	> 72 months	50
Income	2	£10,000 - < £30,000	16
	0	no record	35
	3	£30,000 - < £60,000	57
	4	£60,000 - < £100,000	70
	5	>=£100,000	85
Existing Customer?	0	Existing Customer	8

Application for a Personal Loan:

Score > 440	Offer 7.2% for up to £10000
390 < Score <=440	8.2% up to £10000
370 < Score <=390	Refer to specialist
Score <= 370	Reject

Built by data transformations and regression.

Used for product pricing, customer management, provisioning/IFRS9, inputs to regulatory capital models etc.

Following established best-practices.

Scorecarding best practices

Logistic Regression (or even linear regression!)

Classified variables

Information value (IV) and Population Stability (PSI)

Bootstrapping, holdout test data and resampling

Sample selection and weighting

Selection and refinement of factors, Akaike Information Criterion

Model sensitivity, Delta Method, Model shifts

Combinations, Ensembles and Dynamic updates of models

Model Risk and Model Shifts

Model Risk Questions:

- How robust is this scorecard? Is it 57 or 58? Is its outcome sensitive to the classes?
- Should we worry about income inflation, or time-dependence of TiA?
- Is the “existing customer” effect just a data artefact?
- Is it appropriate to treat “no record” as an average income?
- What if context changes? – Covid-19, Brexit, Climate change – how should the scores and decision points be adjusted?
- With a new use for this score – should it be adjusted further?

How differently would I build my scorecard if my assumptions changed, or if the data changed, or if the market changed, or fresh performance information came in?

Model Shifts are the key to Quantitative Model Risk Management

Scorecard Model Shifts have two descriptions

Score shifts

- How we implement model changes and recalibrate in practice
- Tangent vector displacements in a model space
- Contravariant

Weights of Evidence shifts

- How we measure model performance and estimate recalibrations
- Maximum likelihood projections onto model subspaces
- Covariant

By the end of this talk, I hope you will understand (or at least appreciate) all the bullets on this page (if you don’t already)!

Interactive model – balanced population

Characteristic		Good	Bad	Default Rate
x2	x1			
1	1	4910	90	1.80%
1	-1	4990	10	0.20%
-1	1	4990	10	0.20%
-1	-1	4910	90	1.80%

Logistic Model: PD/(1-PD) = EXP(Int + beta1*X1 + beta2*X2 [+ beta12*X1*X2])

Model	Int	beta1	beta2	beta12
x1, x2	-4.5951	0.0000	0.0000	-
Full	-5.1057	0.0000	0.0000	1.1065

Shifted data

Characteristic		Good	Bad	Default Rate
x2	x1			
1	1	4955	45	0.90%
1	-1	4995	5	0.10%
-1	1	4980	20	0.40%
-1	-1	4820	180	3.60%

Delta Approach, uses Weights of Evidence (WOE) = LN (BAD / GOOD)

x1	base: good | bad		shift: good | bad		Delta WOE (bads)	Coding
1	9900	100	9935	65	-0.4273	-0.5169
-1	9900	100	9815	185	0.6066	0.5169

Shifted Model	Int	beta1	beta2	beta12
x1, x2	-4.7245	-0.5313	-0.7025	-
Delta		-0.5169	-0.6856	-
Full	-5.1032	-0.0063	-0.7007	1.1087

Imbalanced population

Characteristic		Good	Bad	Default Rate
x2	x1			
1	1	982	18	1.80%
1	-1	4990	10	0.20%
-1	1	4990	10	0.20%
-1	-1	982	18	1.80%

Model	Int	beta1	beta2	beta12
x1, x2	-5.3626	0.0000	0.0000	-
Full	-5.1059	0.0000	0.0000	1.1067

Characteristic		Good	Bad	Default Rate
x2	x1			
1	1	991	9	0.90%
1	-1	4995	5	0.10%
-1	1	4980	20	0.40%
-1	-1	964	36	3.60%

Shifted Model	Int	beta1	beta2	beta12
x1, x2	-5.5391	-0.7997	-1.1756	-
Delta		-0.1721	-0.6896	-
Full	-5.1033	-0.0062	-0.7008	1.1088

Data points and model points

Characteristic		Good	Bad	Default Rate	Modelled Default Rate
x2	x1				
1	1	982	18	1.80%	0.47%
1	-1	4990	10	0.20%	0.47%
-1	1	4990	10	0.20%	0.47%
-1	-1	982	18	1.80%	0.47%

Characteristic		Good	Bad	Default Rate	Modelled Default Rate
x2	x1				
1	1	991	9	0.90%	0.05%
1	-1	4995	5	0.10%	0.27%
-1	1	4980	20	0.40%	0.57%
-1	-1	964	36	3.60%	2.75%

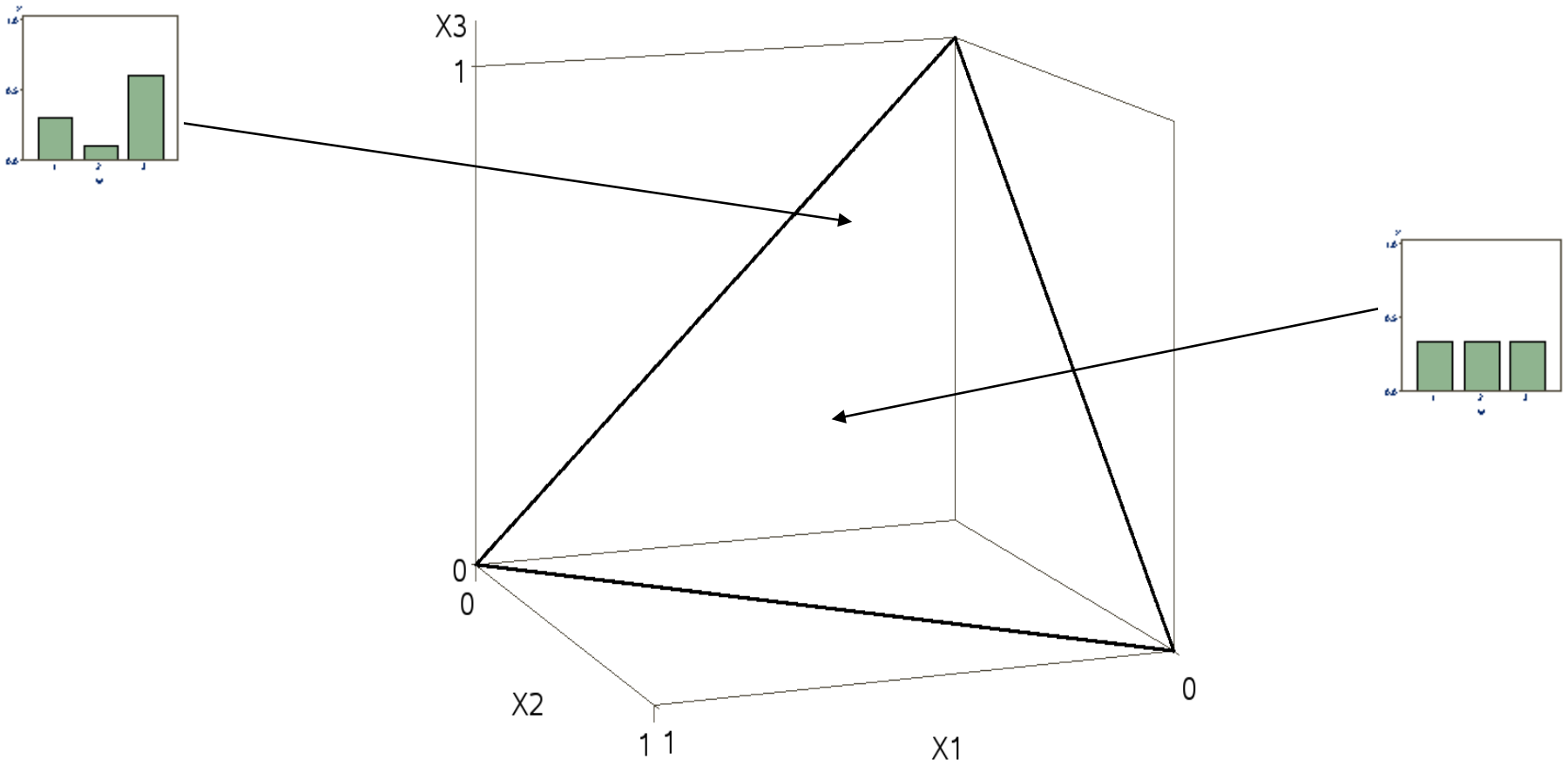
Characteristic			Data (d)	d	m	d'	m'
x1	x2	y					
1	1	1	982	0.0818	0.0829	0.0826	0.0833
-1	1	1	4990	0.4158	0.4147	0.4163	0.4155
1	-1	1	4990	0.4158	0.4147	0.4150	0.4143
-1	-1	1	982	0.0818	0.0829	0.0803	0.0810
1	1	-1	18	0.0015	0.0004	0.0008	0.0000
-1	1	-1	10	0.0008	0.0019	0.0004	0.0011
1	-1	-1	10	0.0008	0.0019	0.0017	0.0024
-1	-1	-1	18	0.0015	0.0004	0.0030	0.0023

Data and Models both live in the same space

A 7-dimensional simplex within 8-dimensional space

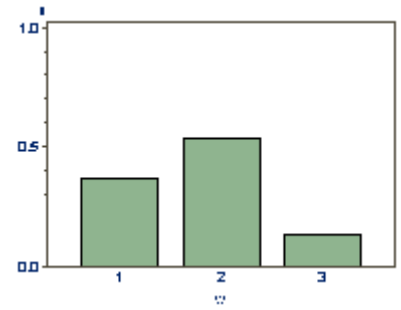
A picture in low dimensions:

3-cell model/data space - 2 dimensional simplex, triangle



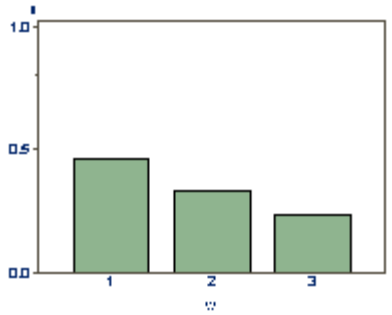
Maximum Likelihood Estimation of a model:

A log-linear example $m_w = ce^{aw}$ w=1,2,3

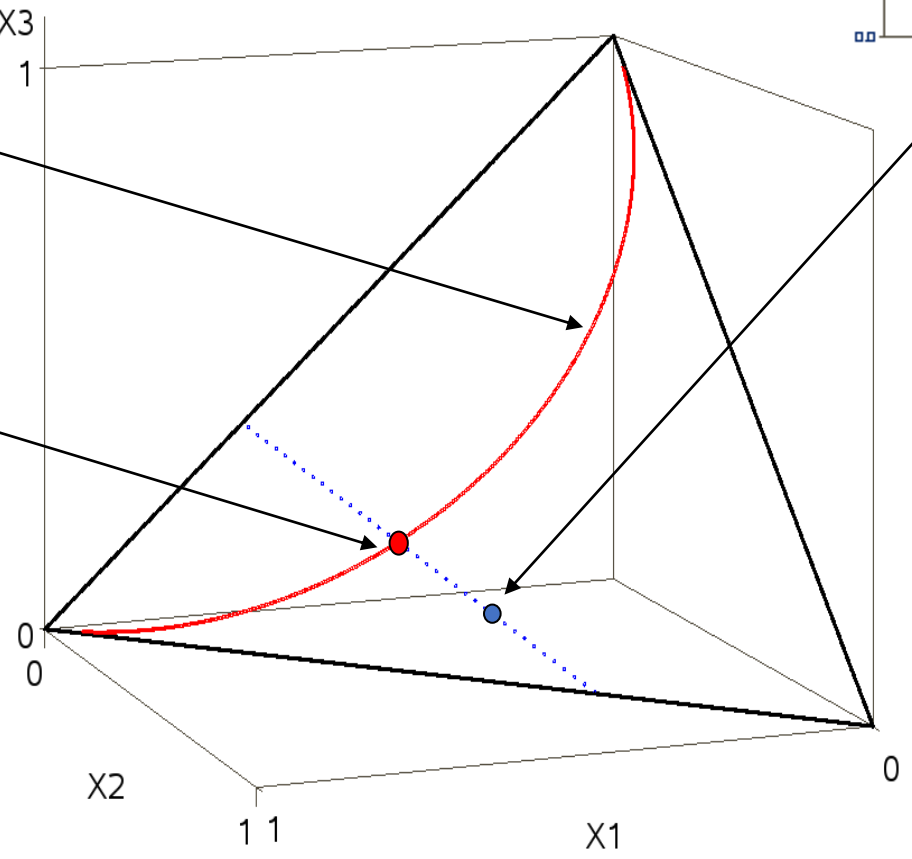


Data

Model Space



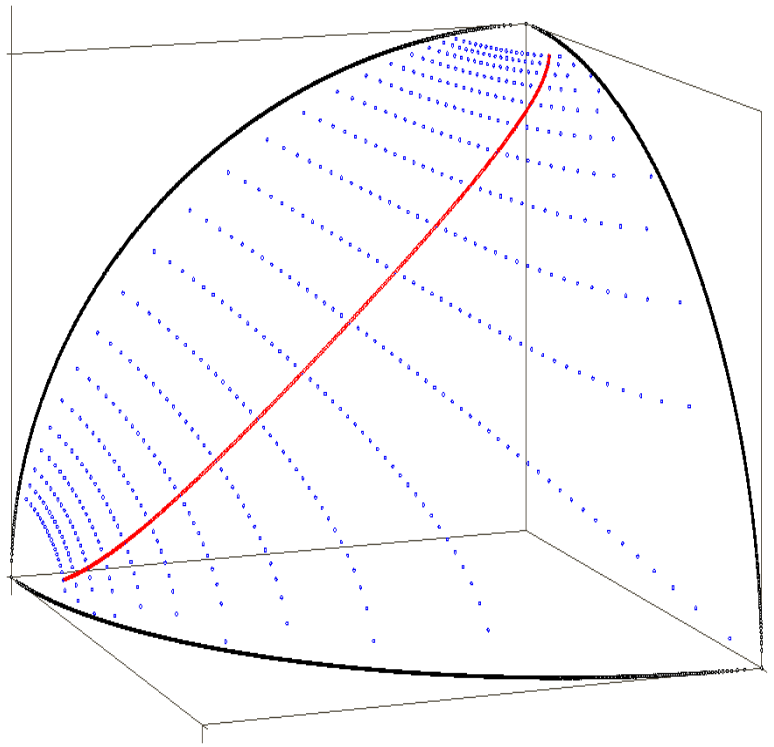
MLE fit



MLE fit minimises

$KL(d,m) = \sum_w d_w \ln(d_w/m_w)$

the Kullback-Leibler divergence from d to m



The most natural metric is Riemannian

$$\delta s^2 = \sum_w \delta x_w^{\,2} / x_w$$

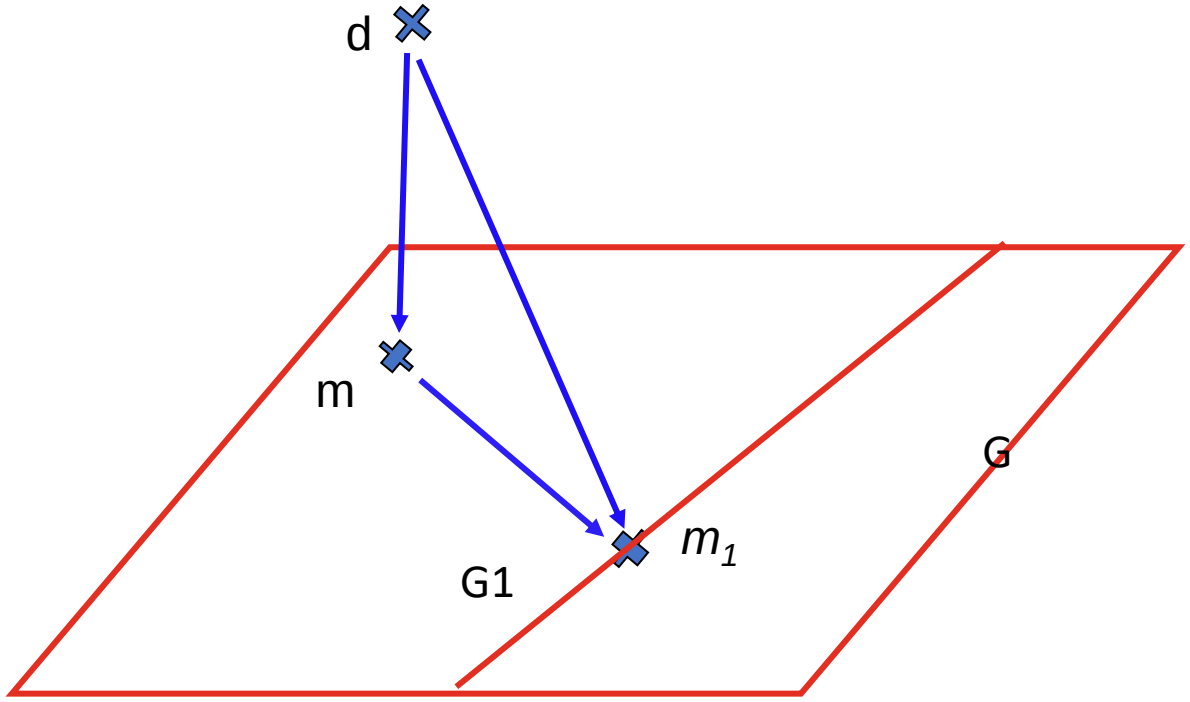
- Locally equivalent to KL divergence:
  $\delta s^2 = 2\, KL(x, x+\delta x) = 2\, KL(x+\delta x, x)$
  up to third order terms.
- Local Chi-squared
- Boot-strapping metric – Dirichlet distribution

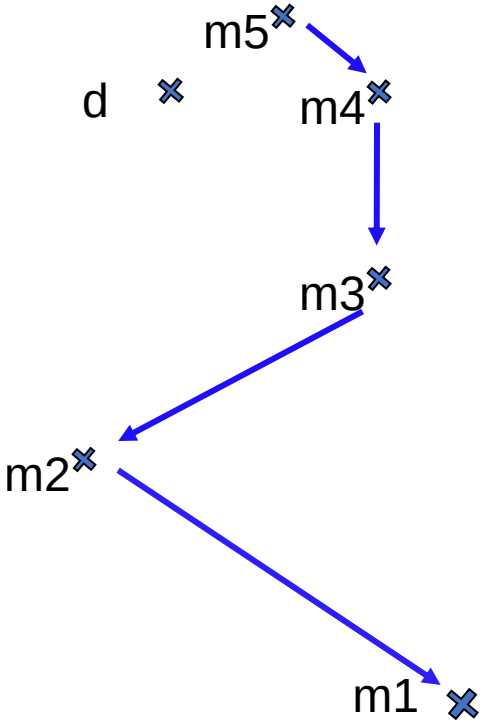
The model fitting foliations (blue) are orthogonal to the model spaces (red) in this metric.

Isometric to a portion of a sphere.

- $2x = u^2$ embeds this space isometrically in Euclidean space.
- Derives the Hellinger Distance

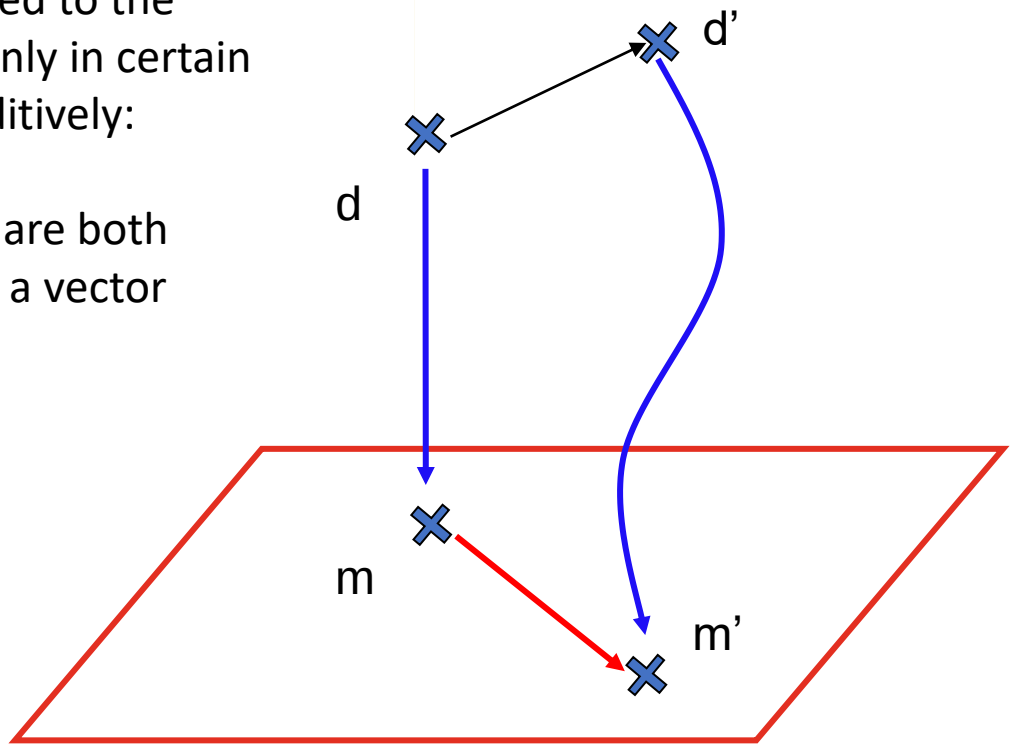
- Global “Pythagorean Theorems” for nested MLE model fits: $KL(d,m_1) = KL(d,m) + KL(m,m_1)$ – allows ANOVA-like analysis of contingency tables
- $KL(d,m_1) = \sum_n KL(m_{n+1}, m_n)$ - an orthogonal decomposition of forward model selection – a natural framework for selection strategies and the Akaike Information Criterion.





Data is projected to the model space only in certain directions, additively:

m-d and *m'-d'* are both constrained to a vector subspace G^o .



Model shifts follow only certain directions, multiplicatively:

ln(m'/m) is constrained to a vector subspace *G* .

In the space T_mM , tangent to the model space, *M*, at *m* , this multiplicative constraint is linearised as $\delta m = mg$ for some *g* in *G* (coordinate-wise multiplication).

Therefore T_mM is naturally embedded in *mG*

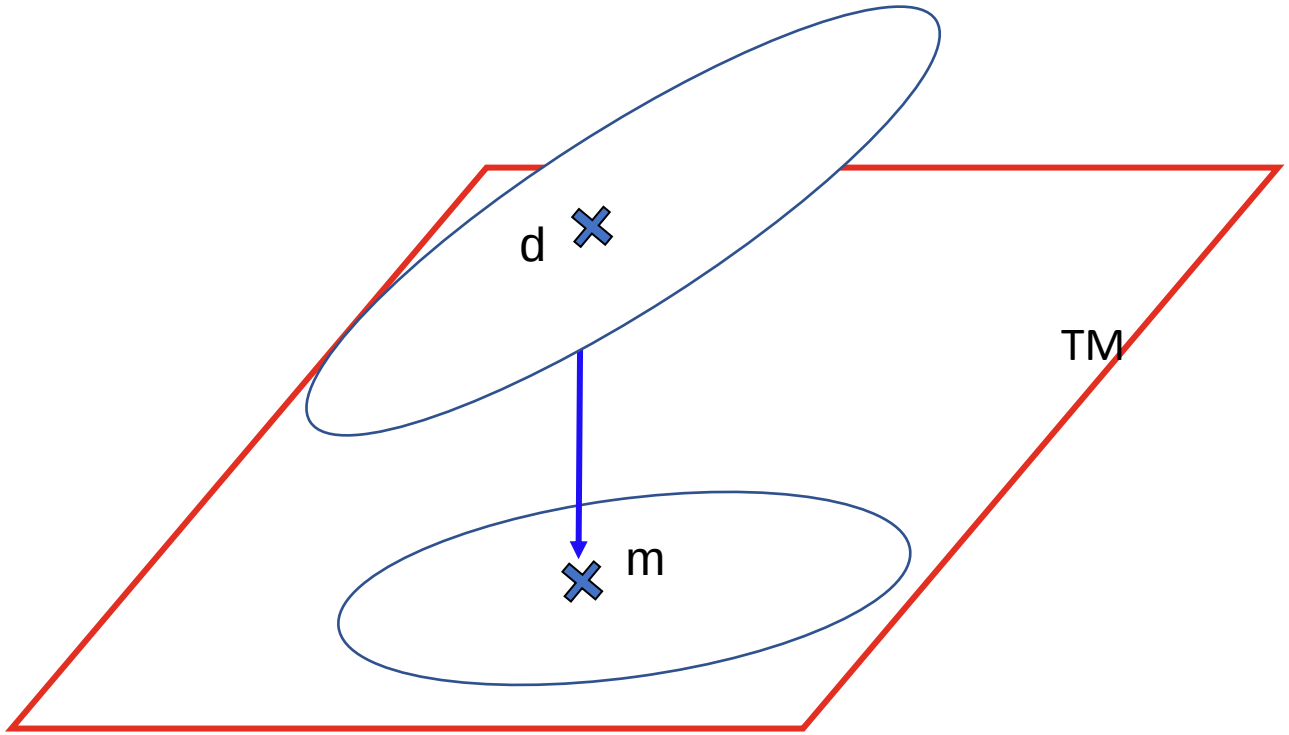
For a model of exponential class, *G* is a fixed subspace, independent of *m* . *G* acts as group of connections on the model space.

G and G^o are ortho-complements in the usual Euclidean inner product (T_mM is orthogonal to G^o in the $\delta m^2/m$ metric).

- Arithmetic averages of data : $MLE(ave(d_n)) = MLE(ave(MLE(d_n)))$ – MLE models are sufficient statistics for their own dynamic updating: we can “forget” some of the old data.

  $KL(ave(d_n) , m) + ave\ KL(d_n , ave(d_n)) = ave\ KL(d_n , m)$ – a barycentric decomposition for KL

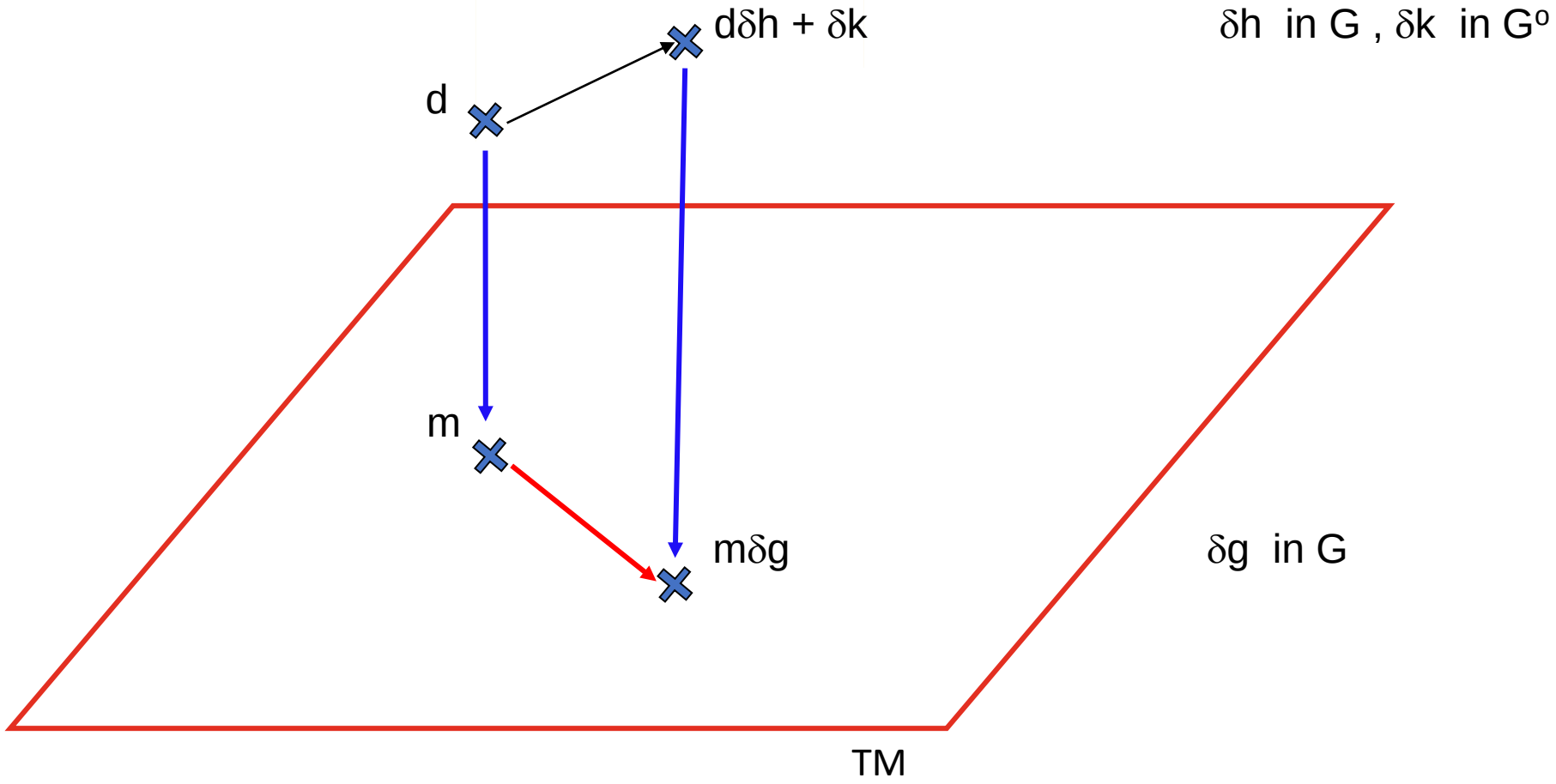
- *M* is closed under (normalised) multiplicative averages – which correspond to combinations in log-odds space.
- Dirichlet distribution centred on the observed data is proportional to $EXP(-N\ KL(d,x))$ – the Pythagorean theorem allows this distribution to restrict to *M* either as a fibre or as a cofibre, equivalently.



KL balls give bootstrap variation around the data; projected to standard error variation around the model.

Test / holdout samples explore the same region

Population Stability Index is closely related to the KL divergence of the data. The geometric relation between data variance and KL, suggests that PSI is a good measure of pressure on model shift.



δg is independent of δk .

To first order, δg is a bilinear function of δh and $m - d$

Basis for 2^2 x 2 Logistic regression

(after Hadamard, Rademacher, Walsh…)

x1	x2	y	Constant	Population			Default Level	Factor X1	Factor X2	Factor X1*X2
1	1	1	1	1	1	1	1	1	1	1
-1	1	1	1	-1	1	-1	1	-1	1	-1
1	-1	1	1	1	-1	-1	1	1	-1	-1
-1	-1	1	1	-1	-1	1	1	-1	-1	1
1	1	-1	1	1	1	1	-1	-1	-1	-1
-1	1	-1	1	-1	1	-1	-1	1	-1	1
1	-1	-1	1	1	-1	-1	-1	-1	1	1
-1	-1	-1	1	-1	-1	1	-1	1	1	-1
			G							G^o

G = < constant , pop , default , X1 , X2 > G^o = < X1*X2 >

Local shifts in new basis

d	m	d'	m'	δk*10^6	δh	δg
0.0818	0.0829	0.0826	0.0833	2.170	0.0092	0.0041
0.4158	0.4147	0.4163	0.4155	-2.170	0.0010	0.0020
0.4158	0.4147	0.4150	0.4143	-2.170	-0.0020	-0.0010
0.0818	0.0829	0.0803	0.0810	2.170	-0.0183	-0.0233
0.0015	0.0004	0.0008	0.0000	-2.170	-0.5014	-2.1477
0.0008	0.0019	0.0004	0.0011	2.170	-0.4974	-0.5505
0.0008	0.0019	0.0017	0.0024	2.170	1.0026	0.1984
0.0015	0.0004	0.0030	0.0023	-2.170	0.9986	1.7755

Convert to Basis coordinates

δh @ d	**δg @ m**
0.1240	0.1245
0.0031	-0.6400
-0.3712	-0.7999
-0.0020	0.0000
-0.1266	-0.1256
0.0031	0.6460
0.3788	0.8075
0.0000	0.0000

G

$$d' - d = d\delta h + \delta k$$
$$m' - m = m\delta g$$

$$\delta g = A\ \delta h$$

1	0	0	0.4740	-0.0107	0	0
0	1.5687	0.8531	0	0	-0.5741	-0.8612
0	0.8531	1.5687	0	0	-0.8612	-0.5741
0	0	0	1	-0.0160	0	0
0	0	0	-0.4784	1	0	0
0	-0.5741	-0.8612	0	0	1.5687	0.8531
0	-0.8612	-0.5741	0	0	0.8531	1.5687

Nested models for 2^2 x 2 Logistic regression

Constant	Population			Default Level	Factor X1	Factor X2	Factor X1*X2
1	1	1	1	1	1	1	1
1	-1	1	-1	1	-1	1	-1
1	1	-1	-1	1	1	-1	-1
1	-1	-1	1	1	-1	-1	1
1	1	1	1	-1	-1	-1	-1
1	-1	1	-1	-1	1	-1	1
1	1	-1	-1	-1	-1	1	1
1	-1	-1	1	-1	1	1	-1
G							G'

G = < constant , pop , default , X1 , X2 >

G1 = < constant , pop , default , X1 >

G2 = < constant , pop , default , X2 >

G^o = < X1*X2 >

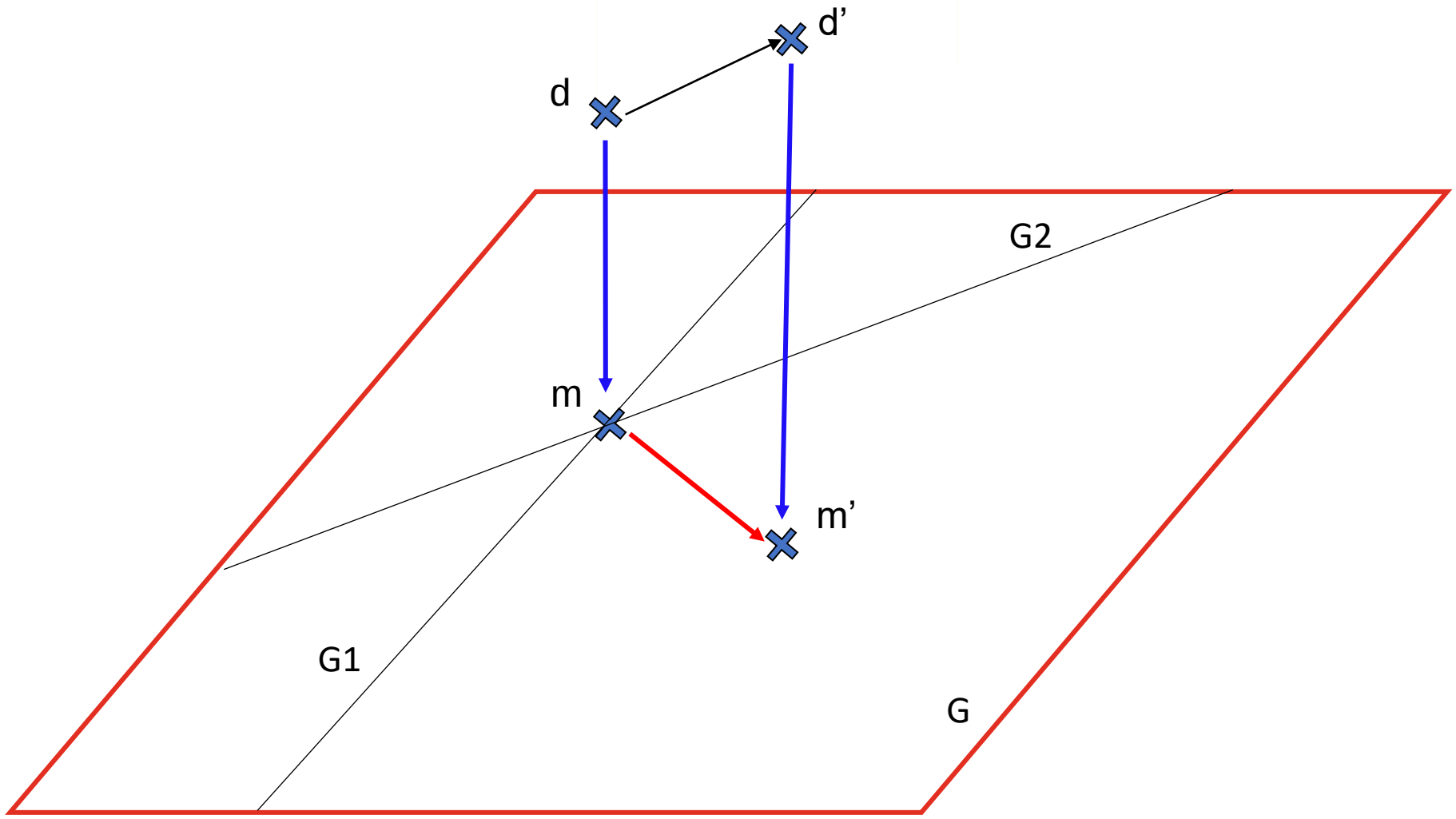
$G1^o$ = < X2 , X1*X2 >

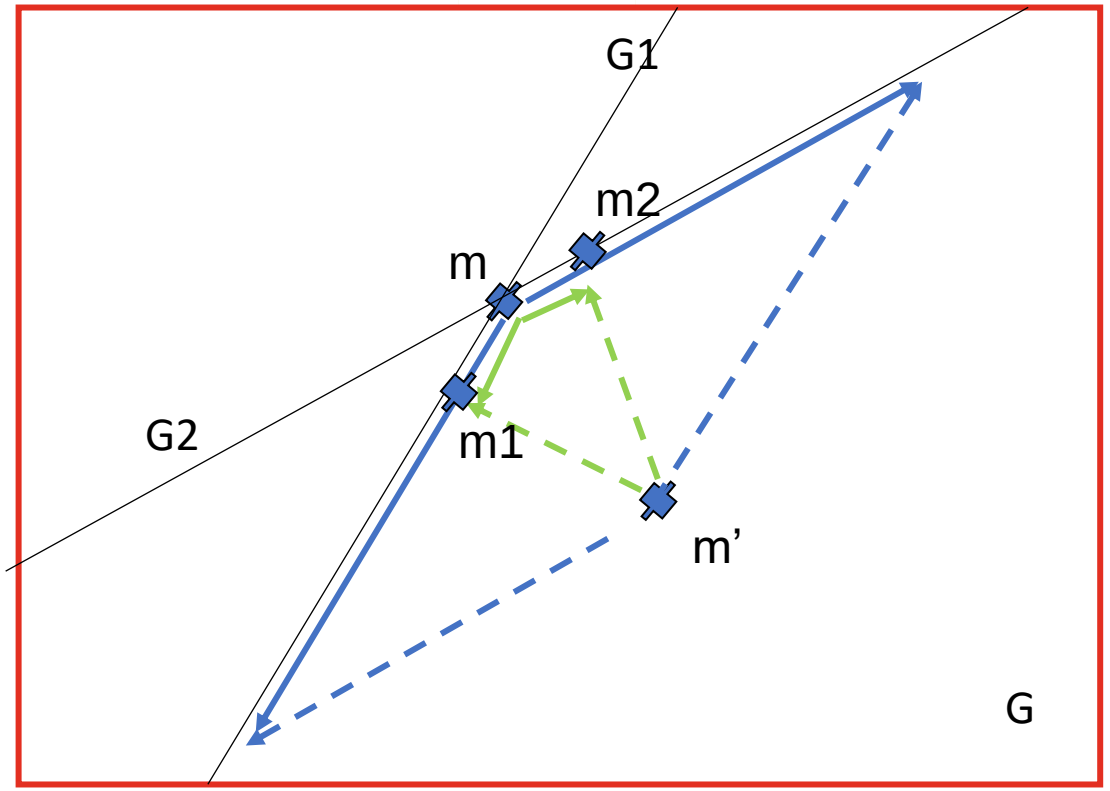
$G2^o$ = < X1 , X1*X2 >

X1,X2 model

X1 model

X2 model


d'
d
G2
m
m'
G1
G



Delta models in contravariant coordinates

δg1	δg2	δg'
0.1036	-0.0006	-0.0928
-0.0861	0	-0.3937
0	-0.3448	-0.5802
0	0	-0.0050
-0.1048	-0.0006	0.0883
0.0871	0	0.3998
0	0.3483	0.5878

Contravariant description — Scores in scorecard

Covariant description — Weights of evidence on single factors

To retrieve the correct scores – apply a local geometric transformation of covariant to contravariant description and then translate to score shifts.

Model shift in covariant coordinates

Data shift $d' - d = d\delta h + \delta k$

Model shift $m' - m = m\delta g$

δh	δg
0.1240	0.1245
0.0031	-0.6400
-0.3712	-0.7999
-0.0020	0.0000
-0.1266	-0.1256
0.0031	0.6460
0.3788	0.8075

Convert from contravariant to covariant coordinates

δh @ d *10^6	δg @m *10^6
0.00	0.00
0.00	0.00
0.00	0.00
0.00	0.00
-29.20	-29.20
25.00	25.01
87.50	87.51

In covariant coordinates: Model shift = Data shift (with G^o component removed)

Heuristic Proof: locally the model shift is related to its MLE submodels by orthogonal projection onto submodel subspaces. This matches corresponding data projections onto its local submodel spaces – the argument is simple for co-dimension 1 and we build this up inductively to creates an identical covariant description of data and model overall.

Scorecarding best practices

Logistic Regression	The natural outcome of KL minimisation in model/data space, with the model space multiplicatively convex, of exponential type – simplest, most natural, max-entropy solution.
Variable classing	Makes possible a finite dimensional data space and the convenience of model space as a subspace of data space. Removes (most) worries about infinities and continuity.
information value (IV) and Population Stability (PSI)	Corresponds closely with KL divergence, the natural metric of model shift and MLE estimation in the data space. The metrics have the best chance to detect possible pressure on model shift and model selection.
Bootstrapping, test data and resampling	Are good ways of exploring the KL metric variation / Dirichlet distribution around the data point.
Sample selection and weighting	We can use the experimental design principle "follow the greatest information gradient". This supports and modulates generally the recognised strategy of evening up bad counts by reweighting; and of monitoring “soft spots”.
Forward selection of factors	Geometrically, we see a progress towards the data point by mutually orthogonal steps: KL divergences add and AIC is natural.
Model sensitivity, Model shifts	Delta is a good approach, once it is corrected for its contravariance. In a covariant coordinate system, Model shift = data shift - let's get exploring!
Combinations, Ensembles and Dynamic updates of models	Geometrically, it’s most natural to take additive combinations or barycentres of data; but to take multiplicative combinations of models (i.e. additive in log-odds space).

Questions and discussion

Sampling, monitoring and weighting samples

Assuming a model is already in place and we're looking to challenge it optimally: What customers should we sample / acquire / monitor to improve the standard error of the model most?

Experimental design principles say "follow the greatest expected information gradient"

- Geometrically, we shift the data point in data space by adding new customers, to move the model point by as large a local metric distance (KL divergence) as possible
- But our shifts are restricted to purely population movements (we don't know the default outcome in advance)

1. For each kind of customer, find the model shift that happens when one is added
2. Select the customer that maximises the KL divergence of that shift

Note that because of MLE sufficiency, we can work locally at the incumbent model and remember the old data only so far as it generates expected model outcomes.

This process naturally weights up the parts of the sample with the poorest expected model accuracy (IV measure) and the lowest counts. This is likely to be where the bad rates are highest