Forecasting portfolio credit default rates

Dirk Tasche

Bank of England – Prudential Regulation Authority[1]
dirk.tasche@gmx.net

Cass Business School
February 18, 2015

[1] The opinions expressed in this presentation are those of the author and do not necessarily reflect views of the Bank of England.

Outline

Data

Moody's corporate issuer and default counts in 2008 and 2009[2].

Grade	2008		2009	
	Issuers	Defaults	Issuers	Defaults
Caa-C	421	63	528	182
B	1158	25	962	72
Ba	527	6	511	12
Baa	1025	5	1011	9
A	981	5	964	2
Aa	595	4	527	0
Aaa	145	0	136	0
All	4852	108	4639	277

[2]Source: Moody's (2013)

Forecast problem

Moody's corporate issuer proportions and default rates in 2008 and issuer proportions in 2009[3]. All numbers in %.

Grade	2008		2009	
	Issuers	Default rate	Issuers	Default rate
Caa-C	8.7	15.0	11.4	?
B	23.9	2.2	20.7	?
Ba	10.9	1.1	11.0	?
Baa	21.1	0.5	21.8	?
A	20.2	0.5	20.8	?
Aa	12.3	0.7	11.4	?
Aaa	3.0	0.0	2.9	?
All	100.0	2.2	100.0	?

[3]Source: Moody's (2013)

Comparison of two approaches

Observed default rates (DR) for 2008[4] and 2009 and Total Probability (TP) and Kullback-Leibler (KL) forecasts for 2009. All numbers in %.

Grade	DR 2008	TP 2009	KL 2009	DR 2009
Caa-C	14.96	12.09	30.22	34.47
B	2.16	3.24	9.53	7.48
Ba	1.14	1.46	4.47	2.35
Baa	0.49	0.78	2.42	0.89
A	0.51	0.33	1.02	0.21
Aa	0.67	0.12	0.36	0.00
Aaa	0.00	0.03	0.10	0.00
All	2.23	2.46	6.69	5.97

[4] Default rates for 2008 were smoothed by quasi-moment matching (Tasche, 2013) before being used for the TP and KL forecasts.

Objective

- Default rate forecasts are often based on
  - regression on macroeconomic variables or
  - assumptions on shared portfolio characteristics (e.g. with credit bureau data collections).
- Drawbacks:
  - Long time series of observations are required.
  - Firm specific underwriting policies are not taken into account.
- We **investigate methods** that
  - allow for period-to-period forecasts and
  - only rely on internal data.

Setting

- Formalise setting of slide 4. We only consider **binary classification problem**.
- **Known:**
  - Probability space $(\Omega, \mathcal{A}, \mathrm{P}_0)$ (training set).
  - σ-field $\mathcal{C} \subset \mathcal{A}$ (covariates).
  - Event $A \in \mathcal{A}$, $A \notin \mathcal{C}$ (class of example).
  - Probability measure P_1 on $(\Omega, \mathcal{C})$ (test set without class labels).
- In rating example (slide 4):
  - $\mathcal{C}$ is information provided by rating grade.
  - A means issuer's default. Issuer's default status is not known at the beginning of the year.
  - P_0 is known joint distribution of rating grades at beginning of 2008 and default status at end of 2008.
  - P_1 is known distribution of rating grades at beginning of 2009.

Problem

- Find **extension** P_1^* of P_1 to $\sigma(\{A\} \cup \mathcal{C})$ such that we can compute $\mathrm{P}_1^*[A]$ and $\mathrm{P}_1^*[A \,|\, \mathcal{C}]$.
- The extension should meaningfully incorporate features of P_0.
- In the rating example $\mathrm{P}_1^*[A]$ is a forecast of the portfolio-wide 2009 default rate and $\mathrm{P}_1^*[A \,|\, \mathcal{C}]$ is a forecast of the grade-level default rates.
- **Assumptions:**
  - $\mathrm{P}_0\big|_{\mathcal{C}}$ has a density f with respect to some measure μ on $(\Omega, \mathcal{C})$.
  - Suppose $p_0 = \mathrm{P}_0[A] \in (0, 1)$.
- In the example:
  - μ is the Laplace distribution on $\{\text{Caa-C}, \ldots, \text{Aaa}\}$ and f is given by the rating frequencies.
  - $p_0 = 2.2\%$.

The Law of Total Probability approach

- Classical case: For $\mathcal{C}$-measurable partition $C_1, C_2, \ldots$ of Ω

$$\mathrm{P}[A] \;=\; \sum_{k=1}^{\infty} \mathrm{P}[A\,|\,C_k]\,\mathrm{P}[C_k].$$

- **For classification problem:**
  - Replace $\mathrm{P}[C_k]$ by $\mathrm{P}_1[C_k]$ and $\mathrm{P}[A\,|\,C_k]$ by $\mathrm{P}_0[A\,|\,C_k]$.
  - In general, $\mathrm{P}_1^*[B] = \mathrm{E}_1\big[\mathrm{P}_0[B\,|\,\mathcal{C}]\big]$, $B \in \mathcal{A}$ defines a probability measure on $(\Omega, \mathcal{A})$ if $\mathrm{P}_1 \ll \mathrm{P}_0\big|_{\mathcal{C}}$.
- This gives column "TP 2009" on slide 5 (assuming that $\mathrm{P}_0[A\,|\,\mathcal{C}] = \mathrm{P}_1^*[A\,|\,\mathcal{C}]$).
- In the machine learning literature, this solution is called **covariate shift** approach (Moreno-Torres et al., 2012).

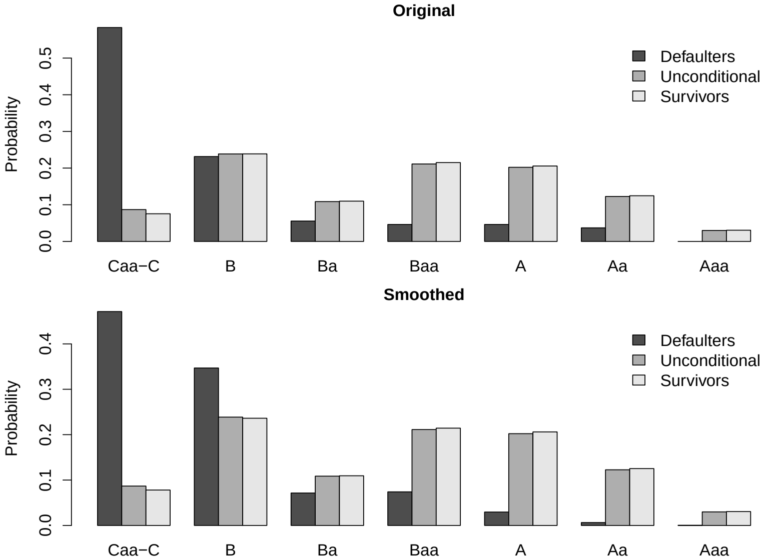
Class Density Ratios

- Since $\mathrm{P}_0\big|_{\mathcal{C}} \ll \mu$ we have μ-densities f_A and f_{A^c} of $\mathrm{P}_0[\cdot\,|\,A]\big|_{\mathcal{C}}$ and $\mathrm{P}_0[\cdot\,|\,A^c]\big|_{\mathcal{C}}$.
- **Assumption:** $f_{A^c} > 0$.
- Define the **density ratio** $\lambda_0 = f_A\,/\,f_{A^c}$.
- On $(\Omega, \mathcal{C}, \mathrm{P}_0)$ then we have

$$\begin{aligned} f &= p_0\,f_A + (1-p_0)\,f_{A^c} \\ \mathrm{P}_0[A\,|\,\mathcal{C}] &= \tfrac{p_0\,\lambda_0}{1-p_0+p_0\,\lambda_0}. \end{aligned} \tag{1}$$

- Hence $\mathrm{P}_0\big|_{\mathcal{C}}$ is a mixture distribution. This suggests estimation of $\mathrm{P}_1^*[A]$ by a **mixture model** approach.

Example: Moody's 2008 rating distributions



The Kullback-Leibler estimator

- Assume that P_1 has density $g > 0$ with respect to μ. Minimise the Kullback-Leibler (KL) distance between g and $p\,f_A + (1-p)\,f_{A^c}$:

$$\begin{aligned} \mathrm{KL}(p) &= \int g \log\left(\tfrac{g}{p\,f_A + (1-p)\,f_{A^c}}\right) d\mu \\ &= \mathrm{E}_1\big[\log(g/f_{A^c})\big] - \mathrm{E}_1\big[\log(p\,\lambda_0 + 1 - p)\big]. \end{aligned} \tag{2}$$

- If E_1 is an empirical measure, minimising the KL distance gives a maximum likelihood estimator of $\mathrm{P}_1^*[A]$.
- **First order condition** for minimum:

$$\mathrm{KL}'(p) = 0 \iff \mathrm{E}_1\left[\frac{\lambda_0 - 1}{1 - p + p\,\lambda_0}\right] = 0. \tag{3}$$

- A solution of (3) is called **KL estimator** of $\mathrm{P}_1^*[A]$.

Exact fit for the KL estimator

- Suppose that $\mathrm{P}_1[\lambda_0 = 1] < 1$. Then there is a unique solution $0 < p_1 < 1$ to (3) if and only if

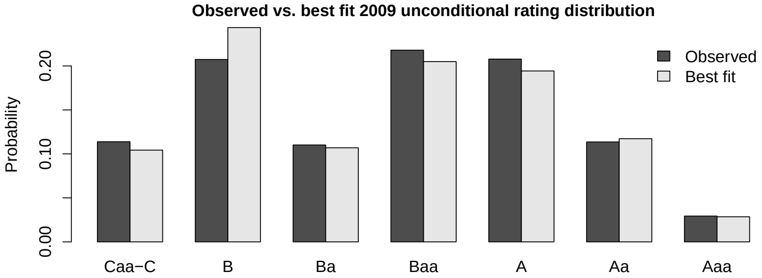
$$\mathrm{E}_1[\lambda_0] > 1 \quad \text{and} \quad \mathrm{E}_1[1/\lambda_0] > 1. \tag{4}$$

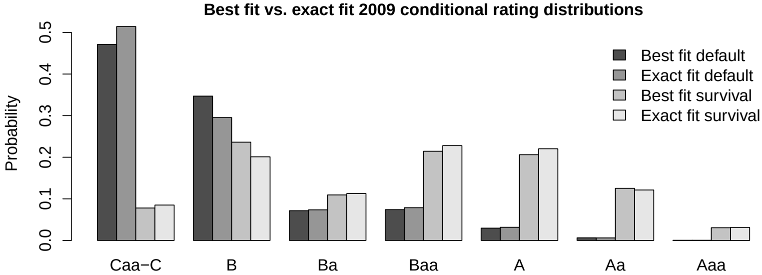
- If there is a solution $0 < p_1 < 1$ to (3) then there is a probability measure P_1^* on $\sigma(\{A\} \cup \mathcal{C})$ such that
  1) $\mathrm{P}_1^*\big|_{\mathcal{C}} = \mathrm{P}_1$,
  2) $\mathrm{P}_1^*[A] = p_1$, and
  3) $\mathrm{P}_1^*[C \,|\, A] = \int_C \frac{g\,\lambda_0}{1-p_1+p_1\,\lambda_0}\,d\mu$ and $\mathrm{P}_1^*[C \,|\, A^c] = \int_C \frac{g}{1-p_1+p_1\,\lambda_0}\,d\mu$ for $C \in \mathcal{C}$.
- Property 1) is called **exact fit.**
- P_1^* is the only probability measure on $\sigma(\{A\} \cup \mathcal{C})$ with 1) and density ratio λ_0. P_1^* is called **KL extension** of P_1.
- The measure extension result still holds if g is a density of P_1 with respect to some measure $\nu \neq \mu$.

Comments

- In the multi-class case, there is no similarly simple condition like (4) for the existence of a solution to the first order equations for the KL minimisation.
- The criterion (4) seems to be satisfied most of the time.
- Is there another way to assess ex ante (before column “DR 2009” on slide 5 is observed) whether Total Probability or KL approach (or none of the two) is better?
- A partial response comes from studying **prior probability shift** (Moreno-Torres et al., 2012):
  - In general, it holds that $g_A = \frac{g\,\lambda_0}{1-p_1+p_1\,\lambda_0} \neq f_A$ and $g_{A^c} = \frac{g}{1-p_1+p_1\,\lambda_0} \neq f_{A^c}$.
  - Prior probability shift denotes special case $g = q\,f_A + (1-q)\,f_{A^c}$. Then it follows that $f_A = g_A$ and $f_{A^c} = g_{A^c}$.

Example: Best vs. exact fit for Moody's 2009 data





Prior probability shift

- Let $q \in (0,1)$ and assume that P_1 is given by

$$\frac{d\,\mathrm{P}_1}{d\,\mu} \;=\; g \;=\; q\,f_A + (1-q)\,f_{A^c}. \tag{5}$$

- Then $p_1 = q$ is the unique solution of (3) in $(0,1)$.
- Moreover, it holds that

$$\mathrm{E}_1\big[\mathrm{P}_0[A\,|\,\mathcal{C}]\big] - q \;=\; (p_0 - q)\,\frac{\mathrm{E}_0[\mathrm{P}_0[A\,|\,\mathcal{C}]\,(1-\mathrm{P}_0[A\,|\,\mathcal{C}])]}{p_0\,(1-p_0)}.$$

- For the regression of $\mathbf{1}_A$ on $\mathcal{C}$ under P_0 we have that

$$1 - R^2 \;=\; \frac{\mathrm{E}_0[\mathrm{P}_0[A\,|\,\mathcal{C}]\,(1-\mathrm{P}_0[A\,|\,\mathcal{C}])]}{p_0\,(1-p_0)}.$$

- Hence the Total Probability and KL estimates of q are the less different the better the forecast of A by $\mathrm{P}_0[A\,|\,\mathcal{C}]$ is on the training set.

Some thoughts

- Under assumption (5), the KL estimator is an unbiased estimator of the class probability.
- Hofer and Krempl (2013) analyse a credit dataset that seems to fulfil (5).
- On Moody's data (Moody's, 2013), KL performs worse than Total Probability on average.
- Clearly, if historical records are available a decision between KL and Total Probability should be based on time series analysis.
- For non-credit applications, sometimes a rationale based on causality can be helpful.

Causality in classification problems

- Classification problem: Infer class Y of an observation based on covariates X.
- Fawcett and Flach (2005) distinguish two types of 'classification domains':
  (i) $X \rightarrow Y$ where the class is causally dependent on the covariates X.
  (ii) $Y \rightarrow X$ where different classes cause different outcomes of X.
- Fawcett and Flach (2005) describe two examples of (ii):
  - Infection status with regard to a disease and illness symptoms.
  - Manufacturing fault status and properties of the produced goods.
- (ii) is considered a justification of assumption (5).
- There is no clear causality in credit classification problems.

A prudent approach to probability of default quantification I

- Let $g > 0$ be a μ-density of P_1 on $(\Omega, \mathcal{C})$. If (4) holds, g has the following decomposition:

$$g \;=\; p_1\, g_A + (1 - p_1)\, g_{A^c},$$

with g_A and g_{A^c} as on Slide 14.

- Define P_0^* on $(\Omega, \sigma(\{A\} \cup \mathcal{C}))$ by

$$\frac{d\,\mathrm{P}_0^*\big|_{\mathcal{C}}}{d\,\mu} \;=\; p_0\, g_A + (1 - p_0)\, g_{A^c}$$

and its KL extension. Then $\mathrm{P}_0^*[A\,|\,\mathcal{C}] = \mathrm{P}_0[A\,|\,\mathcal{C}]$.

- Moreover, with $R_*^2 = 1 - \frac{\mathrm{E}_0^*[\mathrm{P}_0[A\,|\,\mathcal{C}]\,(1-\mathrm{P}_0[A\,|\,\mathcal{C}])]}{p_0\,(1-p_0)}$ we obtain

$$p_1\, R_*^2 + (1 - R_*^2)\, p_0 \;=\; \mathrm{E}_1\big[\mathrm{P}_0[A\,|\,\mathcal{C}]\big].$$

A prudent approach to probability of default quantification II

- Hence, it holds that

$$p_0 \le p_1 \;\Rightarrow\; p_0 \le \mathrm{E}_1\big[\mathrm{P}_0[A\,|\,\mathcal{C}]\big] \le p_1,$$
$$p_0 \ge p_1 \;\Rightarrow\; p_0 \ge \mathrm{E}_1\big[\mathrm{P}_0[A\,|\,\mathcal{C}]\big] \ge p_1.$$

- This observation suggests the following prudent estimation method for $\mathrm{P}_1^*[A]$:
  - Determine p_1 according to (3).
  - If $p_0 \le p_1$ choose $\mathrm{P}_1^*[A] = p_1$.
  - If $p_0 \ge p_1$ choose $\mathrm{P}_1^*[A] = \mathrm{E}_1\big[\mathrm{P}_0[A\,|\,\mathcal{C}]\big]$.
- With this approach, there is an incentive to optimise the accuracy of the conditional probabilities of default $\mathrm{P}_0[A\,|\,\mathcal{C}]$ (see slide 16).

The problem

- For sake of illustration, suppose that on slide 4
  - ’issuers’ is replaced by ’% of exposure’ and
  - ’default rate’ is replaced by ’loss rate’.
- Are then the Total Probability and KL forecast methods applicable?
  - Clearly, ’yes’ for Total Probability because then it is simply assumed that the grade-level loss rates in 2009 are the same as the ones observed in 2008.
  - Less clear for KL because its derivation is heavily based on probability calculus.
- Two interpretations of model (slide 7):
  - Individual: p_0 is one issuer’s probability of default.
  - Collective: p_0 is the proportion of all issuers that default.

The finite measure approach

- With the collective interpretation of the model (slide 7), it is applicable to the 'exposure – loss rate' problem:
    - Probabilities are understood as proportions.
    - Probability calculus is calculus of proportions in terms of finite measures.
    - Conditional probabilities are relative proportions.
    - Bayes' formula is a re-engineering tool without interpretation of causality.
- Limited practical application to a retail credit loss estimation problem was inconclusive with regard to suitability of approach.
- Suggestion to use the prudent approach described before.

- We have studied the problem of **forecasting prior class probabilities** in the presence of a changed covariates distribution.
- Straight-forward forecasts based on Law of Total Probability (TP) may underestimate the amount of change of the prior probabilities.
- Alternative simple finite mixture model approach is promising:
  - Deploying the Kullback-Leibler (KL) estimator provides exact fit of the changed covariates distribution.
  - In the binary classification case, the KL estimator always forecasts more change of the prior probabilities than the TP.
  - In credit risk, this can be used to obtain conservative estimates of probability of default and expected loss.
- This approach may reduce dependence on macroeconomic data and assumptions of similarities of portfolios.
- Loss provisioning and stress testing are **potential applications.**

T. Fawcett and P.A. Flach. A response to Webb and Ting's On the Application of ROC Analysis to Predict classification Performance under Varying Class Distributions. *Machine Learning*, 58(1):33–38, 2005.

V. Hofer and G. Krempl. Drift mining in data: A framework for addressing drift in classification. *Computational Statistics & Data Analysis*, 57(1):377–391, 2013.

Moody's. Annual Default Study: Corporate Default and Recovery Rates, 1920-2012. Special comment, Moody's Investors Service, February 2013.

J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodriguez, N.V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

D. Tasche. The art of probability-of-default curve calibration. *Journal of Credit Risk*, 9(4):63–103, 2013.