

THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Bankruptcy prediction of small and medium enterprises using a flexible generalized extreme value model

Citation for published version:
Calabrese, R, Marra, G & Osmetti, S 2016, 'Bankruptcy prediction of small and medium enterprises using a flexible generalized extreme value model', *Journal of the Operational Research Society*, vol. 67, no. 4, pp. 604-615. https://doi.org/10.1057/jors.2015.64

Digital Object Identifier (DOI):
10.1057/jors.2015.64

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of the Operational Research Society

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Download date: 29. Mar. 2024

Bankruptcy Prediction of Small and Medium Enterprises Using a Flexible Binary Generalized Extreme Value Model

Raffaella Calabrese[1], Giampiero Marra[2] and Silvia Angela Osmetti[3]

[1]Department of Statistics and Quantitative Methods, University of Milano-Bicocca

Via Bicocca degli Arcimboldi 8, 20123 Milano, Italy

raffaella.calabrese1@unimib.it

[2]Department of Statistical Science, University College London,

Gower Street, London WC1E 6BT, U.K.

giampiero.marra@ucl.ac.uk

[3]Department of Statistical Sciences, Università Cattolica del Sacro Cuore di Milano

Largo Gemelli 1, 20123 Milano, Italy

silvia.osmetti@unicatt.it

June 14, 2013

Abstract

We introduce a binary regression accounting-based model for bankruptcy prediction of small and medium enterprises (SMEs). The main advantage of the model lies in its predictive performance in identifying defaulted SMEs. Another advantage, which is especially relevant for banks, is that the relationship between the accounting character-

istics of SMEs and response is not assumed a priori (e.g., linear, quadratic or cubic) and can be determined from the data. The proposed approach uses the quantile function of the generalized extreme value distribution as link function as well as smooth functions of accounting characteristics to flexibly model covariate effects. Therefore, the usual assumptions in scoring models of symmetric link function and linear or pre-specified covariate-response relationships are relaxed. Out-of-sample and out-of-time validation on Italian data shows that our proposal outperforms the commonly used (logistic) scoring model for different default horizons.

Keywords: logistic regression, generalized extreme value distribution, penalized regression spline, scoring model, small and medium enterprises.

1 Introduction

A significant innovation of the Basel II regulatory framework (BCBS, 2005) is the greater use of risk assessments provided by banks' internal systems as inputs to capital calculations. Based on the internal ratings-based approach of the revised framework, banking institutions are allowed to use their own measures as input for their minimum regulatory capital calculations. The main input for these calculations is the probability of default forecasted one year ahead. Therefore, in many credit risk models, such as CreditMetrics (Gupton et al., 1997), CreditRisk+ (Products, 1997) and CreditPortfolioView (Wilson, 1998), default probabilities are essential input parameters. Even if default risk could be estimated for different kinds of loans, i.e. corporate loans, those for small and medium enterprises (SMEs), and the creation of a rational and comprehensive policy for them, play a central role in the European Union (EU) economy (see Small Business Act for Europe, 2008, Brussels: European Commission, http://ec.europa.eu/enterprise/entrepreneurship/docs/sba/SBA_IA). Banks have realized that small and medium sized companies are a distinct kind of clients with needs and peculiarities which require specific risk management tools and methodologies (e.g., Altman

and Sabato, 2006; Dietsch and Petey, 2004; Saurina and Trucharte, 2004).

SMEs are important for the economic system of many countries. Since lending to SMEs is riskier than lending to large corporations (Altman and Sabato, 2006), Basel II (BCBS, 2005) established that banks should develop credit risk models that are specific to SMEs. This generated a lot of interest in the community, which resulted in a number of solutions (e.g., Altman and Sabato, 2006; Berger and Udell, 2002; Ciampi and Gordini, 2008; Fantazzini and Figini, 2009; Lin et al., 2012; Saurina and Trucharte, 2004). We propose a binary generalized extreme value additive (BGEVA) model for predicting SMEs defaults. Specifically, since employing a symmetric link function for a rare event such as default may be problematic (see King and Zeng, 2001, for the use of the logit link), we consider the quantile function of a generalized extreme value (GEV) random variable as link function (Calabrese and Osmetti, 2013; Wang and Dey, 2010). Moreover, because the assumption of a linear or pre-specified (e.g., quadratic or cubic) relationship between the accounting characteristics of SMEs and default is rarely satisfied in practice, we use penalized regression splines to flexibly determine covariate effects from the data (e.g., Ruppert et al., 2003; Wood, 2006). We apply the BGEVA model, and its traditional competitors, to data on 50,160 Italian SMEs for the period $2006-2011$, an important time horizon since it includes the financial crisis of 2008. Our empirical analysis shows that the relationship between accounting characteristics and default is not linear and that our proposal outperforms the commonly used scoring model in terms of predictive performance.

The contribution of this article is twofold. From a methodological point of view, we present a penalized likelihood estimation framework for a binary regression model where the link function is allowed to be asymmetric and the functional dependence of the response on continuous covariates is flexibly determined from the data. Importantly, the methods discussed in this article are implemented in the `R` package `bgeva` (Calabrese et al., 2013), which makes it feasible for practitioners to fit BGEVA models. We elected to follow a frequentist approach because it can particularly appeal to researchers already familiar with traditional frequentist techniques and usually has the advantage of being computationally fast.

From an applied perspective, we improve on the prediction results obtained using classic alternatives. This is crucial for financial institutions and the economic system. If banks can discriminate better between defaulting and non-defaulting SMEs, then the credit system may become more efficient. This is pivotal for small businesses since their financial structure typically depends heavily on financial institutions that provide external funding. This means that shocks to the banking system can have a significant impact on the credit supply to small businesses. As a consequence, SMEs may be subject to funding problems and credit access may be problematic during recession periods in financial markets. As this topic is pivotal in the EU and USA (Berger and Udell, 2006; Pederzini, 2012), we would expect our proposal to be generally useful for analyzing the SMEs of many countries.

The paper is organized as follows. The next section provides a review of the main literature on scoring models for SMEs. In Sections 3 and 4, we discuss the main drawbacks of the commonly used (logistic) regression for credit default applications and introduce the BGEVA scoring model. Section 5 shows the results obtained from applying the traditional and proposed approaches to data on Italian SMEs. The last section is devoted to the conclusions.

2 Literature review

Credit risk models which can separate defaulting and non-defaulting firms as well as predict corporate bankruptcy can be classified into two groups: market-based models and accounting-based models. The majority of them belong to the first group and are based on structural and reduced-form approaches (Merton, 1974; Jarrow and Turnbull, 1995). Since these models make use of capital market information, which is not available for small firms owing to their information opacity and unattainability, we focus on the second group which uses accounting variables of firms from financial statements. More details on credit risk assessment can be found in Hand and Henley (1997).

Altman (1968) was the first to introduce an accounting-based model for estimating the default probabilities of firms. This was achieved calculating his well-known Z-Score using a

standard discriminant model. Since then, different methodologies, such as linear regression, logistic regression, classification trees and neural networks have been employed for credit risk assessment (Thomas et al., 2002). The widely used model is logistic regression (Altman and Sabato, 2006; Lin et al., 2012), which shows, however, some important drawbacks for default prediction. Since default can be viewed as a rare event (because the number of defaults in a sample is typically very small) and the logit link function is symmetric, the default probability could be underestimated (Calabrese and Osmetti, 2013). Moreover, the bias of maximum likelihood parameter estimators for logistic regression is amplified in rare event studies (King and Zeng, 2001).

Accounting-based models can be applied to different kinds of firms. As mentioned in the introduction, because of the Basel II regulatory framework, many scholars have focused on developing tools specifically designed for SMEs. In Italy, SMEs form the 99.9% of the firms (see the SBA fact sheet 2012 for Italy). These employ around the 81% of the individuals in the workforce and contributed for the 68.3% of the Italian Added Value in 2011 (EC, 2012). More generally, SMEs play a pivotal role in many countries. For example, Altman and Sabato (2006) analyzed 2,010 US SMEs for the period $1994 - 2002$. They stressed the importance of using a model designed for assessing the rating of an SME, instead of employing a generic model built for both SMEs and large corporations. According to Altman and Sabato (2006), using different models for small/medium enterprises and large firms may lead banks to lower the required capital as established by Basel II (BCBS, 2005). In a later work, Altman et al. (2009) significantly improved the default prediction power of risk models for UK SMEs by using explanatory variables such as legal action by creditors to recover unpaid debts, company filing histories and comprehensive audit report/opinion. Lin et al. (2012) analyzed UK SMEs adopting different definitions of a failing small business. Specifically, for 429 SMEs, whose performance at the end of 2004 and financial ratios from 2001 were observed, they investigated the impacts that several default definitions have on the choice of predictor variables and model's predictive accuracy.

3 Main drawbacks of the traditional credit scoring model

The most commonly used model for credit scoring applications is logistic regression (e.g., Altman and Sabato, 2006; Becchetti and Sierra, 2002; Lin et al., 2012; Zavgren, 1998). Let y_i be a binary response which describes the event default ($y_i = 1$) and non-default ($y_i = 0$) for the i^{th} SME. The logistic model can be written as

$$logit(PD_i) = \ln\left(\frac{PD_i}{1 - PD_i}\right) = \alpha + \sum_{j=1}^{p} \beta_j x_{ji} = \eta_i, \quad j = 1, 2, ..., p, \quad i = 1, 2, ..., n, \qquad (1)$$

where PD_i is the default probability, α is an intercept, the x_{ji} are p financial and economic continuous variables, the β_j are regression coefficients and n represents the sample size. Details on estimation methods and inferential procedures can be found in McCullagh and Nelder (1989).

There are two main drawbacks associated with model (1). First, PD_i may be underestimated. Since the number of defaults in a sample is usually very small (e.g., Kiefer, 2010; Lin et al., 2012), default can be viewed as a rare event. Hence, the use of the logit link function may not be appropriate because of its symmetry around 0.5, which implies that the response curve, $PD_i = 1/\{1 + \exp(-\eta_i)\}$, approaches zero at the same rate as it approaches one. This is not ideal as the characteristics of defaults are more informative than those of non-defaults and as a consequence PD_i will be underestimated. This suggests using an asymmetric link function as in Calabrese and Osmetti (2013) and Wang and Dey (2010). Second, model (1) assumes a linear relationship between the accounting characteristics of SMEs and response (e.g., Chuang and Lin, 2009; Gestel et al., 2005; Huang et al., 2006; Lee and Chen, 2005; Ong et al., 2005). One could easily include quadratic or cubic terms to relax the assumption of linearity but such a procedure would require choosing a priori the order of the polynomial function for each covariate and it is not in general recommended (e.g., Marra and Radice, 2010; Wood, 2006). This issue can mask possibly interesting non-linear patterns which can help improve our understanding of the underlying covariate-response relationships and per-

haps improve the prediction accuracy of the scoring model as well. This calls for an approach which can flexibly determine covariate effects from the data (see, for instance, Berg (2007) who employed a logistic additive model).

The next section describes a methodology that can overcome the two aforementioned model (1)'s drawbacks by blending the strengths of using an asymmetric GEV link function (Calabrese and Osmetti, 2013; Wang and Dey, 2010) and an additive predictor (Berg, 2007).

4 The BGEVA model

Since the percentage of defaults is typically very low, even for SMEs, the defaulters' characteristics are more informative than those of non-defaulters (Berg, 2007; Kiefer, 2010; Lin et al., 2012). This means that defaulters' features are better represented by the tail of the response curve for values close to one, which can be modeled using a GEV random variable (Kotz and Nadarajah, 2000; Falk et al., 2010). Because our focus is on defaulters, as in Calabrese and Osmetti (2013), we exploit the quantile function of a GEV random variable and specify the link function

$$\frac{[-\ln(PD_i)]^{-\tau} - 1}{\tau}, \tag{2}$$

where $\tau \in \Re$ is the tail parameter. Hence, in (1) we replace $logit(PD_i)$ with (2). Since a GEV link can be asymmetric, underestimation of the default probability may be overcome. As discussed, for instance, in Calabrese and Osmetti (2013), depending on the value of τ, several special cases can be recovered; e.g., when $\tau \to 0$ the GEV random variable follows a Gumbel distribution and its cumulative distribution is the log-log function (Agresti, 2002).

Moving on to the right hand side of equation (1), as explained in the previous section, the assumption of linear or pre-specified covariate-response relationships may be restrictive in applications. Therefore, we replace $\alpha + \sum_{j=1}^{p} \beta_j x_{ji}$ with $\alpha + \sum_{j=1}^{p} f_j(x_{ji})$, where the $f_j(\cdot)$ are unknown one-dimensional smooth functions of the continuous covariates x_{ji}. The smooth terms are represented using the regression spline approach (Eilers and Marx, 1996; Ruppert

et al., 2003; Wood, 2006). Specifically, $f_j(x_{ji})$ is approximated by a linear combination of known (e.g., cubic or thin plate regression) spline bases, $b_k(x_{ji})$, and unknown regression parameters, γ_{jk}. That is, $f_j(x_{ji}) = \sum_{k=1}^{K_j} \gamma_{jk} b_k(x_{ji})$, where K_j is the number of basis functions. In other words, calculating $b_k(x_{ji})$ for each observation point and k yields K_j curves encompassing different degrees of complexity which multiplied by some real valued parameters γ_{jk} and then summed give an estimated curve for the smooth component (see Ruppert et al. (2003) for a more detailed introduction). Smooth terms are typically subject to identifiability centering constraints such as $\sum_{i=1}^{n} f_j(x_{ji}) = 0 \ \ \forall j$ (Wood, 2006).

The use of both a GEV link and smooth components leads to

$$\frac{[-\ln(PD_i)]^{-\tau} - 1}{\tau} = \alpha + \sum_{j=1}^{p} f_j(x_{ji}) = \eta_i, \tag{3}$$

where η_i can be written as $\mathbf{B}_i^{\mathsf{T}}\boldsymbol{\delta}$ with $\mathbf{B}_i^{\mathsf{T}} = [1, b_1(x_{1i}), \ldots, b_{K_1}(x_{1i}), \ldots, b_1(x_{pi}), \ldots, b_{K_p}(x_{pi})]$ and $\boldsymbol{\delta}^{\mathsf{T}} = (\alpha, \gamma_{11}, \ldots, \gamma_{1K_1}, \ldots, \gamma_{p1}, \ldots, \gamma_{pK_p})$. This represents the BGEVA model that will be employed in Section 5. For our case study, a more general additive predictor was not deemed to be required. It is worth mentioning, however, that the `R` package `bgeva` (Calabrese et al., 2013) supports the inclusion, for instance, of parametric terms (i.e., binary and categorical predictors), of terms obtained by multiplying one or more smooth components by some predictor(s), and of smooth functions of two or more continuous covariates (see Wood (2006) for details on these alternative specifications). Of course, $\mathbf{B}_i$ and $\boldsymbol{\delta}$ would have to be modified accordingly but this is a minor change. Finally, note that model (3) is only defined for those observations for which $1 + \tau\eta_i \geq 0$ (e.g., Calabrese and Osmetti, 2013).

4.1 Parameter estimation

In model (3), replacing the smooth terms with their regression spline expressions yields essentially a parametric model whose design matrix includes spline bases. This means that a BGEVA model can be estimated by maximum likelihood (ML). However, classic ML estima-

tion of (3) is likely to result in smooth function estimates that are too rough to be useful for empirical analysis (e.g., Ruppert et al., 2003). This issue can be overcome by penalized ML, where the use of penalty matrices allows us to suppress that part of smooth term complexity which has no support from the data (e.g., Ruppert et al., 2003; Wood, 2006). Specifically, each smooth has an associated penalty, $\boldsymbol{\gamma}_j^{\mathsf{T}}\mathbf{S}_j\boldsymbol{\gamma}_j$, where $\boldsymbol{\gamma}_j^{\mathsf{T}} = (\gamma_{j1}, \ldots, \gamma_{jK_j})$ and $\mathbf{S}_j$ is a positive semi-definite square matrix of known coefficients measuring the roughness of the j^{th} smooth component; for instance, the second-order roughness measure for a univariate spline penalty evaluates $\int\{f''(x_j)\}^2 dx_j$. The formulas for the $b_k(x_{ji})$ and $\mathbf{S}_j$ depend on the type of spline basis employed and we refer the reader to Ruppert et al. (2003) and Wood (2006) for these details.

For a fixed value of τ, the BGEVA model is estimated by maximization of

$$\ell_p(\boldsymbol{\delta}) = \ell(\boldsymbol{\delta}) + \frac{1}{2}\sum_{j=1}^{p}\lambda_j\boldsymbol{\gamma}_j^T\mathbf{S}_j\boldsymbol{\gamma}_j \quad \text{w.r.t.} \quad \boldsymbol{\delta}, \tag{4}$$

where

$$\ell(\boldsymbol{\delta}) = \sum_{i=1}^{n} -y_i(1+\tau\eta_i)^{-1/\tau} + (1-y_i)\ln\{1-\exp[-(1+\tau\eta_i)^{-1/\tau}]\},\ \eta_i = \mathbf{B}_i^{\mathsf{T}}\boldsymbol{\delta}.$$

Penalized log-likelihood (4) is essentially maximized iterating

$$\hat{\boldsymbol{\delta}}^{[a+1]} = \hat{\boldsymbol{\delta}}^{[a]} + (\boldsymbol{\mathcal{J}}^{[a]} - \mathbf{S}_{\boldsymbol{\lambda}})^{-1}(\mathbf{S}_{\boldsymbol{\lambda}}\hat{\boldsymbol{\delta}}^{[a]} - \boldsymbol{\mathcal{U}}^{[a]}) \tag{5}$$

until convergence, where a is the iteration index, $\mathbf{S}_{\boldsymbol{\lambda}} = \text{diag}(0, \lambda_1\mathbf{S}_1, \ldots, \lambda_p\mathbf{S}_p)$ (when the additive predictor is specified as in (3)),

$$\boldsymbol{\mathcal{U}}(\boldsymbol{\delta}) = \frac{\partial\ell(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} = \sum_{i=1}^{n}\left\{y_i(1+\tau\eta_i)^{-1-\frac{1}{\tau}} + (1-y_i)\frac{\exp[-(1+\tau\eta_i)^{-1/\tau}](1+\tau\eta_i)^{-1-\frac{1}{\tau}}}{1-\exp[-(1+\tau\eta_i)^{-1/\tau}]}\right\}\mathbf{B}_i$$

and

$$\begin{aligned}\boldsymbol{\mathcal{J}}(\boldsymbol{\delta}) &= \frac{\partial^2 \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^{\mathsf{T}}} = \sum_{i=1}^{n} \left[\left\{ -\frac{\exp[-2(1+\tau\eta_i)^{-1/\tau}](1+\tau\eta_i)^{2-\frac{2}{\tau}}}{\left(1-\exp[-(1+\tau\eta_i)^{-1/\tau}]\right)^2} - \frac{\exp[-(1+\tau\eta_i)^{-1/\tau}](1+\tau\eta_i)^{2-\frac{2}{\tau}}}{1-\exp[-(1+\tau\eta_i)^{-1/\tau}]} + \right.\right. \\ &+ \left.\left. \frac{\exp[-(1+\tau\eta_i)^{-1/\tau}](1+\tau)(1+\tau\eta_i)^{2-\frac{1}{\tau}}}{1-\exp[-(1+\tau\eta_i)^{-1/\tau}]} \right\} (1-y_i) - (1+\tau)\left(1+\tau\eta_i\right)^{2-1/\tau} y_i \right] \mathbf{B}_i^{\mathsf{T}} \mathbf{B}_i.\end{aligned}$$

In practice, we use the more efficient and safe trust region algorithm, which is based on (5) (Nocedal and Wright, 2006).

The λ_j are smoothing parameters controlling the trade-off between fit and smoothness, and in the above optimization problem they are fixed to some values. This is because joint estimation of $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ and $\boldsymbol{\delta}$ via maximization of (4) would clearly lead to over-fitting since the highest value for $\ell_p(\boldsymbol{\delta})$ would be obtained when $\boldsymbol{\lambda} = \mathbf{0}$. In fact, $\boldsymbol{\lambda}$ should be selected so that the estimated smooth components are as close as possible to the true functions. Automatic multiple smoothing parameter selection can be achieved in several ways; see Ruppert et al. (2003) and Wood (2006) who provide excellent overviews. Here, we elected to use a generalization of the approximate unbiased risk estimator (UBRE, Craven and Wahba, 1979). Specifically, $\hat{\boldsymbol{\lambda}}$ is the solution to the problem

$$\text{minimize} \quad \frac{1}{n}\|\sqrt{\mathbf{W}}(\mathbf{z} - \mathbf{X}\boldsymbol{\delta})\|^2 - 1 + \frac{2}{n}\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}}) \quad \text{w.r.t.} \quad \boldsymbol{\lambda}, \tag{6}$$

where $\sqrt{\mathbf{W}}$ is a weight diagonal matrix square root, $\mathbf{z}_i = \mathbf{B}_i^{\mathsf{T}}\boldsymbol{\delta}^{[a]} + \mathbf{W}_{[ii]}^{-1}\mathbf{d}_i$, $\mathbf{d}_i = \partial \ell(\boldsymbol{\delta})_i / \partial \eta_i$, $\mathbf{W}_{[ii]} = -\partial^2 \ell(\boldsymbol{\delta})_i / \partial \eta_i \partial \eta_i$, $\mathbf{A}_{\boldsymbol{\lambda}} = \mathbf{B}(\mathbf{B}^{\mathsf{T}}\mathbf{W}\mathbf{B} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}\mathbf{B}^{\mathsf{T}}\mathbf{W}$ is the hat matrix, $\mathbf{B} = (\mathbf{B}_1^{\mathsf{T}}, \ldots, \mathbf{B}_n^{\mathsf{T}})^{\mathsf{T}}$ and $\text{tr}(\mathbf{A}_{\boldsymbol{\lambda}})$ the estimated degrees of freedom (*edf*) of the penalized model. (Superscript $[a]$ has been suppressed from $\mathbf{d}_i$, $\mathbf{z}_i$ and $\mathbf{W}_i$ to avoid clutter.) The working linear model quantities are constructed for a given estimate of $\boldsymbol{\delta}$, obtained from a trust region iteration. (6) will then produce an updated estimate for $\boldsymbol{\lambda}$ which will be used to obtain a new parameter vector estimate for $\boldsymbol{\delta}$. The two steps, one for $\boldsymbol{\delta}$ and the other for $\boldsymbol{\lambda}$, are iterated until convergence. Problem (6) is solved employing the approach by Wood (2004) which can evaluate the

approximate UBRE and its derivatives in a way that is both computationally efficient and stable.

As for the tail parameter in the GEV link, it is in principle possible to estimate jointly τ and $\boldsymbol{\delta}$ by maximizing (4). As shown by Smith (1985), however, in a certain range of values of τ the usual regularity conditions for the ML estimator of this parameter do not hold. For this reason, we propose fitting as many BGEVA models as the number of a set of sensibly chosen values of τ and select the model that yields the best empirical predictive performance. In our case study, this proved to be a practical and effective means of handling this parameter.

4.2 Confidence intervals and p-values

Inferential theory for penalized estimators is complicated by the presence of smoothing penalties which undermines the usefulness of classic frequentist results (e.g., Wood, 2006). As shown in Marra and Wood (2012), reliable point-wise confidence intervals for the terms of a model involving penalized regression splines can be constructed using

$$\boldsymbol{\delta}|\mathbf{y} \dot{\backsim} \mathcal{N}(\hat{\boldsymbol{\delta}}, \mathbf{V}_{\boldsymbol{\delta}}), \tag{7}$$

where $\mathbf{y}$ refers to the response vector, $\hat{\boldsymbol{\delta}}$ is an estimate of $\boldsymbol{\delta}$ and $\mathbf{V}_{\boldsymbol{\delta}} = (-\boldsymbol{\mathcal{J}} + \mathbf{S}_{\boldsymbol{\lambda}})^{-1}$. The structure of $\mathbf{V}_{\boldsymbol{\delta}}$ is such that it includes both a bias and variance component in a frequentist sense, a fact that makes such intervals have close to nominal coverage probabilities (Marra and Wood, 2012). Given (7), confidence intervals for linear and non-linear functions of the model parameters can be easily obtained. For instance, for $\hat{f}_j(x_{ji})$ these can be obtained using

$$\hat{f}_j(x_{ji}) \dot{\backsim} \mathcal{N}(f_j(x_{ji}), \mathbf{B}_{ji}(x_{ji})^{\mathsf{T}} \mathbf{V}_{\boldsymbol{\delta}_j} \mathbf{B}_{ji}(x_{ji})),$$

where $\mathbf{V}_{\boldsymbol{\delta}_j}$ is the sub-matrix of $\mathbf{V}_{\boldsymbol{\delta}}$ corresponding to the regression spline parameters associated with j^{th} smooth term, and $\mathbf{B}_{ji}(x_{ji})^{\mathsf{T}} = [b_1(x_{ji}), \ldots, b_{K_j}(x_{ji})]$. For parametric model components, such as binary and categorical predictors, using (7) is equivalent to using classic

results because such terms are not penalized.

For smooth components, point-wise confidence intervals cannot be used for variable selection (e.g., Ruppert et al., 2003, Chapter 6). For this purpose, p-values can be employed. To construct them, we need the distribution of the $\mathbf{f}_j$, where $\mathbf{f}_j = [f_j(x_{j1}), f_j(x_{j2}), \ldots, f_j(x_{jn})]^\mathsf{T}$. As shown by Wood (2013), in the context of extended generalized additive models, $\hat{\mathbf{f}}_j \dot{\sim} \mathcal{N}(\mathbf{f}_j, \mathbf{V}_{\mathbf{f}_j})$ where $\mathbf{V}_{\mathbf{f}_j} = \mathbf{B}_j(\mathbf{x}_j)\mathbf{V}_{\boldsymbol{\delta}_j}\mathbf{B}_j(\mathbf{x}_j)^\mathsf{T}$ and $\mathbf{B}_j(\mathbf{x}_j)$ denotes a full column rank matrix such that $\hat{\mathbf{f}}_j = \mathbf{B}_j(\mathbf{x}_j)\hat{\boldsymbol{\gamma}}_j$. It is then possible to obtain approximate p-values, for testing the hypothesis $\mathbf{f}_j = \mathbf{0}$, based on

$$T_{r_j} = \hat{\mathbf{f}}_j^\mathsf{T} \mathbf{V}_{\mathbf{f}_j}^{r_j -} \hat{\mathbf{f}}_j \dot{\sim} \chi^2_{r_j},$$

where $\mathbf{V}_{\mathbf{f}_j}^{r_j -}$ is the rank r_j Moore-Penrose pseudo-inverse of $\mathbf{V}_{\mathbf{f}_j}$, which can deal with rank deficiencies. Parameter r_j is selected using the established notion of *edf* employed in (6). Because *edf* is not an integer, it can be rounded as follows (Wood, 2013)

$$r_j = \begin{cases} \text{floor}(edf_j) & \text{if } \ edf_j < \text{floor}(edf_j) + 0.05 \\ \text{floor}(edf_j) + 1 & \text{otherwise} \end{cases}.$$

5 Case Study

The Italian SME sector is the largest in the EU, with 3,813 million of SMEs (EC, 2012). The vast majority of Italy's SMEs are micro-firms with less than 10 employees. In fact, Italy's share of micro-firms in all businesses, at 94.6%, exceeds the EU-average (92.2%). Recovery from the financial crisis has been much weaker than that in other EU countries, and SMEs have been the hardest hit companies.

5.1 Data

The empirical analysis is based on annual data for the period $2006 - 2011$ for $50,160$ Italian SMEs. The data are from AIDA-Bureau van Dijk, a large Italian financial and balance sheet information provider. The time horizon considered here is of some interest as it includes the

financial crisis of 2008.

The definition of SME by the European Commission is adopted. That is, a business must have an annual turn-over of less than 50 million of Euros and the number of employees must not exceed 250 (http://ec.europa.eu/enterprise/policies/sme/facts-figures-analysis/sme-definition/index.htm). In this work, default is intended as the end of the SME's activity, that is the status in which the SME needs to liquidate its assets for the benefit of its creditors. In practice, we consider a default to have occurred when a specific SME enters a bankruptcy procedure as defined by the Italian law.

Following Altman and Sabato (2006), we applied choice-based or endogenous stratified sampling, where data are stratified based on the values of the response. Specifically, we drew randomly observations within each stratum defined by the two categories default and non-default, and considered all the defaulted firms. We then selected a random sample of non-defaulted SMEs (over the same years of defaults) to obtain a default percentage in our sample that was as close as possible to that for Italian SMEs, which is around 5% (Cerved-Group, 2011). Choice-based sampling makes observations dependent. Since the sample size is high, according to the super-population theory of Prentice (1986), we assume that ours is a random sample (e.g., Altman and Sabato, 2006). We excluded firms for which information on accounting characteristics is not available.

Financial ratio analysis uses specific formulas to gain insight into a company and its operations. The main measures based on the balance sheet are the financial strength and activity ratios. The former provide information on how well a company is able to meet its obligations and how it is leveraged. This gives investors an idea of the financial stability of the company and its ability to finance itself. Activity ratios focus mainly on current accounts and show how well a company manages its operating cycle. These ratios provide insight into the operational efficiency of the company.

Before fitting the models, we carried out a multicollinearity analysis and discarded variables with a variance inflation factor larger than 5 (Greene, 2012, pp. 257–258). As a result,

we considered the 21 continuous covariates: number of employees, turnover per employee, long and medium term liabilities/total assets, current ratio, leverage, liquidity ratio, solvency ratio, banks/turnover, return on sales, return on equity, added value per employee, turnover/staff costs, return on shareholders' funds, return on capital employed, stock turnover ratio, profit per employee, average remuneration per year, pre-tax profit margin, cost of employees/turnover, total liquid funds and debt/equity.

5.2 Model fitting results

We employed the BGEVA model, logistic additive regression as well as log-log additive regression as it is a well known alternative to the logistic model. Computations were performed in the `R` environment (R Development Core Team, 2013) using the package `bgeva` (Calabrese et al., 2013) which can be used to fit the three models. The predictor specification was consistent with that shown in (3). Smooth components were represented using penalized thin plate regression splines with basis dimensions equal to 20 and penalties based on second-order derivatives (Wood, 2006). Using fewer or more spline bases did not lead to tangible changes in the results reported below and in the next section. Figure 1 shows the PD function of the BGEVA model (obtained from equation (3)) for τ equal to -0.25, 0.001 (corresponding to the log-log response curve) and 0.25. The skewness and approaching rate to 1 and to 0 depend on τ. Compared with the log-log curve, for a negative value of τ the PD of the BGEVA model approaches 0 slowly but 1 more sharply. Viceversa when τ is positive. Since in the context of our case study we need a curve which approaches 1 rapidly, negative values for τ are the obvious choice. For the BGEVA model, values of τ in the set $(-1.00, -0.75, -0.50, -0.25)$ were tried out. The best empirical predictive performance was produced for $\tau = -0.25$ (see next section). Using a finer grid of values and extending their range did not lead to improvements in terms of prediction.

After applying a backward selection procedure at the 5% significance level, using the p-value definition discussed in Section 4.2, only the 12 variables shown in Table 1 were kept. This

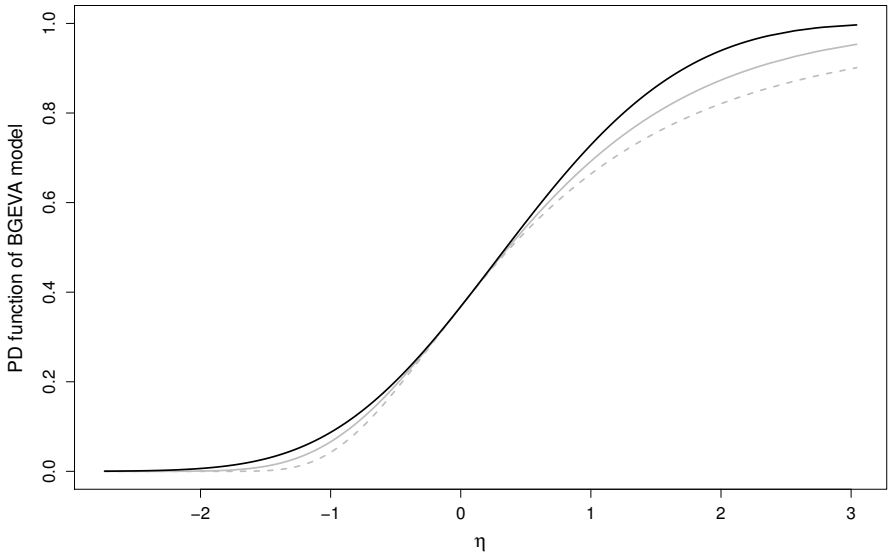


Figure 1: PD function of the BGEVA model for $\tau = -0.25$ (black solid line), $\tau = 0.001$ (grey line) and $\tau = 0.25$ (grey dashed line).

selection result was consistent across all models. Some of the covariate effects are reported in the parametric part of the BGEVA model since their smooth function estimates were linear (i.e. their *edf*s were all equal to 1). For the other explanatory variables, the estimated smooths exhibit *edf*s considerably greater than 1 and are displayed in Figure 2.

As discussed in Calabrese and Osmetti (2013), if the estimated coefficient of a covariate is positive, an increase in its value will result in a decrease in the estimated PD, holding constant all the other variables. Viceversa when the coefficient is negative. The leverage result in Table 1 is consistent with financial theory. Firms with higher levels of debt are more subject to interest rate volatility, hence must produce higher financial returns to balance the risk associated with their financial structure (Majocchi and Zucchella, 2003). Turnover/staff costs shows a negative relationship with PD which is coherent with Ciampi and Gordini (2008) who considered turnover per employee. A striking result could be that profit per employee is positively related to the likelihood of being a defaulted SME. This may be due to government's tax inspection on firms which announce a loss on the accounting year. To avoid inspection, SMEs in Italy, similarly for example to Turkey (Arslan, 2009), usually attempt to

Variable names of parametric model part	*Estimate*	*Std. Error*	*z value*	*p-value*
Intercept	-1.675	2.926e-02	-57.247	$<$ 2e-16
Leverage	0.001	4.642e-04	2.700	0.006943
Turnover/staff costs	0.004	1.061e-03	3.428	0.000607
Profit per employee	-0.009	1.126e-03	-7.697	1.39e-14
Cost of employees/turnover	0.004	8.835e-04	4.902	9.46e-07
Total liquid funds	0.001	1.231e-06	117.892	$<$ 2e-16
Variable names of smooth terms	*Edf*	*Est.rank*	*Chi.sq*	*p-value*
Long-medium term liabilities/total assets	5.42	6	124.70	$<$ 2e-16
Solvency ratio	7.82	8	123.78	$<$ 2e-16
Banks/turnover	3.29	4	37.33	1.54e-07
Added value per employee	4.72	5	21.64	0.000613
Return on shareholders' funds	7.84	8	106.09	$<$ 2e-16
Return on capital employed	7.81	8	71.22	2.81e-12
Stock turnover ratio	6.91	7	92.48	$<$ 2e-16

Table 1: Parametric and smooth component summaries obtained from applying the BGEVA model to a sample of Italian SMEs (2,384 defaulters and 45,296 non-defaulters) for the period $2006 - 2011$. Full details on the calculation of the test statistic for the smooth terms are given in Section 4.2.

avoid reporting a loss through manipulations on their activities out of their main operations. Cost of employees/turnover shows a negative relationship with PD which is not in agreement with the literature as well as expectations. A coherent result is obtained by Ciampi and Gordini (2008) for total personnel costs/added value and by Fantazzini and Figini (2009) for personnel expenses/sales, which show a positive relationship with PD. We suspect that the difference between our result and those obtained in the literature is due to the increasing fiscal pressure on the labor market in Italy for SMEs since the financial crisis (Eurostat, 2012). To explain the negative relationship between total liquid funds and PD, we refer to an interesting European Commission paper by Ferrando et al. (2007). Here, liquid funds are higher if there is a higher probability of a shortage of funds in the future that could be due to financial distress in SMEs.

As for the plots in Figure 2, in line with the interpretation for the parametric components, if the estimated smooth function of a covariate is increasing then the estimated PD decreases, and viceversa. An interesting result is obtained for bank loans/turnover. For low values of this covariate the relationship to PD is negative: for low debt loans, bank loans increase

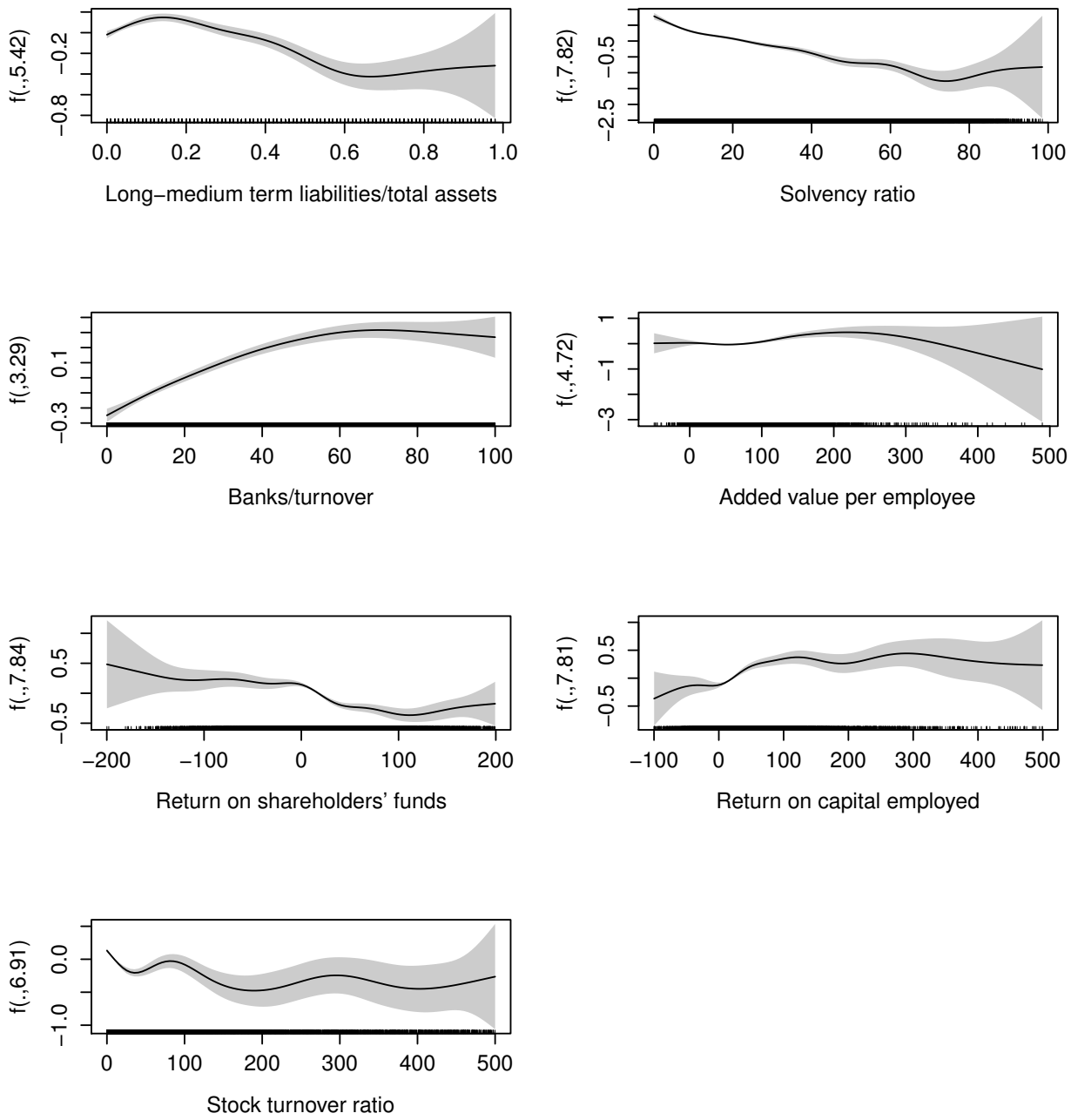


Figure 2: Smooth component estimates of the 7 (out of 12) continuous variables that exhibit a non-linear pattern. These were obtained from applying the BGEVA model on the Italian SME data described in Section 5.1. Results are on the scale of the predictor. Shaded areas represent 95% confidence intervals and the rug plot, at the bottom of each graph, is used to show the covariate values. The numbers in brackets in the y-axis captions are the estimated degrees of freedom (*edf*) of the smooth curves.

SME's solvency. When the debt load is too high, an increase in bank loans to an SME results in a slight (perhaps not significant) increase in the PD. Instead, Ciampi and Gordini (2008) found a positive relationship between bank loans/turnover and PD. Alternatively, Altman and Sabato (2006) considered short term debt/equity book value to model PD and found an inverse relationship. On the contrary, Fantazzini and Figini (2009) showed that short term debt has a positive influence on PD, which is consistent with expectations. These results are not all in agreement. This is probably because these authors modeled covariate effects assuming linearity whereas we employed smooth functions which allowed us to estimate flexibly such effects. Another interesting result is obtained for long-medium term liabilities/total assets: for low values this variable shows an inverse relationship to PD which then becomes positive for high covariate values. When the ratio is higher than 0.7, the estimated curve is slightly increasing although the confidence intervals are wide and hence it is not possible to be conclusive. The fact that the positive impact of long-medium term liabilities/total assets is overall dominant may justify the results obtained by some authors who found a positive impact assuming linearity (e.g., Arslan, 2009).

For values of added value per employee lower than about 200 thousands of Euros, the relationship to PD is negative and coherent with expectations as well as the results obtained by Calabrese and Osmetti (2013) and Ciampi and Gordini (2008). For high values of this covariate, Figure 2 shows a positive relationship, although the confidence intervals are wide and contain the case of an inverse relationship as well. A few analyzes have included return on capital employed as an explanatory variable in scoring models for SMEs (e.g., Toyli et al., 2008). This covariate shows a negative relationship with the likelihood of being a defaulted SME, except for high covariate values where the confidence intervals are so wide that many interpretations are plausible. As for the other variables, a clear interpretation can not be provided and further research is needed. The estimates of the parametric and smooth components obtained by applying the logistic and log-log additive models (not reported here) did not differ significantly from each other and are in agreement with those presented in this section.

5.3 Empirical predictive performance

In this section, we compare the predictive performance of the BGEVA, logistic and log-log additive models. Predictive accuracy can be assessed using the Mean Squared Error (MSE) and Mean Absolute Error (MAE) based on observed default and predicted PD. Scoring models with lower MSE and MAE should forecast defaults and non-defaults more accurately. However, it is the identification of defaulters that is a pivotal aim for banks' internal scoring models. In fact, it is much more costly to classify an SME as a non-defaulter when it is a defaulter than to classify it as a defaulter when it is not. If a defaulted firm is classified as a non-defaulter by the scoring model, then the bank will give a loan. If the borrower becomes defaulted, then the bank may lose the whole or a part of its credit exposure. On the contrary, when a non-defaulter is classified as a defaulter, the bank only loses interest on loans. We, therefore, computed the MAE and MSE only for defaulted SMEs and denote them by MAE^+ and MSE^+. We also calculated the area under the curve (AUC) index (Hand et al., 2001) and H-measure (Hand, 2010), using the `R` package `hmeasure` with a severity ratio of 0.01 for the H-measure. Scoring models with higher AUC and H-measure can forecast defaults and non-defaults more accurately. Note that the H-measure overcomes the drawbacks of AUC when the class sizes and classification error costs are extremely unbalanced (Hand, 2009, 2010), which is especially true for credit scoring applications.

To avoid sample dependency, models were validated on observations that were not included in the sample used to estimate the model. Specifically, we used out-of-sample and out-of-time tests. In the former, the models were estimated using all observations in $2006 - 2011$ but a randomly drawn control sample of size corresponding to the 10% of all observations. In the latter, we analyzed the models' performance for different sample sizes and default horizons. This was achieved in two ways. In the first case, models were fitted using data from the intervals $2006 - 2010$, $2006 - 2009$ and $2006 - 2008$ and their predictive accuracy evaluated on SMEs belonging to 2011, 2010 and 2009, respectively. In the second case, we estimated the models using data for the period $2006 - 2008$ and tested them on SMEs belonging to

Type of control sample	*measure*	*BGEVA*	*log-log*	*logistic*
Out-of-sample	MAE^+	0.754	0.791	0.889
	MSE^+	0.626	0.660	0.795
	H	0.184	0.050	0.018
	AUC	0.682	0.680	0.811
Out-of-time 2009	MAE^+	0.837	0.849	0.883
	MSE^+	0.741	0.754	0.786
	H	0.136	0.052	0.020
	AUC	0.697	0.692	0.809
Out-of-time 2010	MAE^+	0.770	0.798	0.881
	MSE^+	0.620	0.657	0.783
	H	0.081	0.078	0.020
	AUC	0.722	0.722	0.811
Out-of-time 2011	MAE^+	0.843	0.862	0.894
	MSE^+	0.763	0.781	0.804
	H	0.050	0.027	0.030
	AUC	0.619	0.599	0.808

Table 2: Forecasting accuracy measures for out-of-sample and out-of-time exercises obtained from applying the BGEVA, log-log and logistic additive models. Default horizon is of one year.

$2009 - 2010$ and $2009 - 2011$. The estimation period of $2006 - 2008$ was chosen because it was characterized by an economic growth in Italy which was then followed by a decline after 2008.

Table 2 reports the values of MAE^+, MSE^+, AUC index and H-measure for the out-of-sample and out-of-time exercises when the default horizon is of one year. The MAE^+ and MSE^+ for the BGEVA model are lower than those for the log-log and logistic additive models. The AUC of the BGEVA and log-log models are lower than that of logistic regression in all control samples. However, as pointed out earlier, AUC is not reliable when the class sizes and classification error costs are unbalanced (Hand, 2009, 2010), case in which the H-measure is more appropriate (Hand, 2010). The H-measure for BGEVA is higher than that of its competitors in all control samples. Increasing the default horizon does not lead to different conclusions (see Table 3). The forecasting periods considered here are important since the effects of the financial crisis on Italian SMEs have been strong in 2009, where the number of defaulted Italian SMEs in the sample was very high (701), and decreased in 2010 and 2011

Type of control sample	*measure*	*BGEVA*	*log-log*	*logistic*
Two years: $2009-2010$	MAE^+	0.848	0.862	0.882
	MSE^+	0.746	0.760	0.785
	H	0.188	0.066	0.019
	AUC	0.712	0.711	0.811
Three years: $2009-2011$	MAE^+	0.832	0.845	0.881
	MSE^+	0.724	0.740	0.783
	H	0.190	0.067	0.019
	AUC	0.713	0.714	0.811

Table 3: Forecasting accuracy measures for out-of-sample and out-of-time exercises obtained from applying the BGEVA, log-log and logistic additive models. Default horizons are of two and three years.

(429 and 124, respectively).

Our analysis suggests that the accounting-based BGEVA model has a superior predictive performance in identifying defaulted SMEs than that of its traditional alternatives. Since identification of defaulters is crucial for banks, the proposed tool could be employed as their internal scoring model to identify defaulted SMEs.

6 Concluding remarks

We introduced a scoring model to forecast defaulted SMEs. Since lending to SMEs is risky and because they play a crucial role in the economy of many countries, we would expect the BGEVA model to be generally useful for analyzing SMEs.

The proposed approach is based on a penalized likelihood estimation framework where the link function is allowed to be asymmetric (through the use of the quantile function of a GEV random variable) and the functional dependence of the response on continuous covariates is flexibly determined from the data (through the use of penalized regression splines). The developments discussed in this article are implemented in the `R` package `bgeva` (Calabrese et al., 2013) which can be particularly attractive to researchers and practitioners who wish to fit BGEVA models.

Our proposal and its competitors were applied to data on 50,160 Italian SMEs for the

period $2006 - 2011$. The empirical results confirmed that the first main advantage of the BGEVA model lies in its superior performance in forecasting defaulted SMEs for different default horizons. The second is the relaxation of the linearity assumption which was not clearly supported by the data.

Banks and financial institutions could improve their internal assessments and efficiency by using BGEVA models in that they can better identify defaulted SMEs and can shed light on the nature of the relationships between response and SME characteristics. It would be interesting to explore the impact that non-random sample selection (Banasik and Crook, 2007) has on bankruptcy prediction of SMEs and future research will look at the possibility of extending the BGEVA model to account for selection bias.

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, New York.

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, **23**(4), 589–609.

Altman, E. and Sabato, G. (2006). Modeling credit risk for smes: Evidence from the us market. *ABACUS*, **19**(6), 716–723.

Altman, E., Sabato, G., and Wilson, N. (2009). The value of nonfinancial information in sme risk management. In *Credit Scoring & Credit Control XI Conference, Edinburgh*.

Arslan, O. M. B. (2009). Credit risks and internationalization of smes. *Journal of Business Economics and Management*, **10**(4), 361–368.

Banasik, J. and Crook, J. (2007). Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, **183**, 1582–1594.

BCBS (2005). International convergence of capital measurement and capital standards: A revised framework. *Bank for International Settlements*, **25**.

Becchetti, L. and Sierra, J. (2002). Bankruptcy risk and productive efficiency in manufacturing firms. *Journal of Banking and Finance*, **27**, 2099–2120.

Berg, D. (2007). Bankruptcy prediction by generalized additive models. *Applied Stochastic Models in Business and Industry*, **23**, 129–143.

Berger, A. N. and Udell, G. F. (2002). Small business credit availability and relationship lending: The importance of bank organisational structure. *Finance and Economics Discussion Series paper*, **23**, 2004–2012.

Berger, A. N. and Udell, G. F. (2006). A more complete conceptual framework for sme finance. *Journal of Banking and Finance*, **30**, 2945–2966.

Calabrese, R., Marra, G., and Osmetti, S. A. (2013). *bgeva: Binary Generalized Extreme Value Additive Models*. R package version 0.1.

Calabrese, R. and Osmetti, S. A. (2013). Modelling sme loan defaults as rare events: the generalized extreme value regression model. *Journal of Applied Statistics*, **Forthcoming**.

Cerved-Group (2011). Caratteristiche delle imprese, governance e probabilità di insolvenza. In *Report, Milan, IT*.

Chuang, C. L. and Lin, R. H. (2009). Constructing a reassigning credit scoring model. *Expert Systems with Applications*, **36**, 1685–1694.

Ciampi, F. and Gordini, N. (2008). Using economic-financial ratios for small enterprise default prediction modeling: An empirical analysis. In *Proceedings of the Oxford Business & Economics Conference, Oxford, UK*, pages 1–21.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377–403.

Dietsch, M. and Petey, J. (2004). Should sme exposure be treated as retail or as cor- porate exposures? a comparative analysis of default probabilities and asset correlation in french and german smes. *Journal of Banking and Finance*, **28**, 773–788.

EC (2012). Sba fact sheet 2012 for italy. *Enterprise and Industry working paper*.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**(2), 89–121.

Eurostat (2012). *Taxation trends in the European Union*. Statistical Books, Taxation and Custom Union.

Falk, M., Haler, J., and Reiss, R. (2010). *Laws of Small Numbers: Extremes and Rare Events*. Springer.

Fantazzini, D. and Figini, S. (2009). Random survival forests models for sme credit risk measurement. *Methodology and Computing in Applied Probability*, **11**, 29–45.

Ferrando, A., Kohler-Ulbrich, P., and Pal, R. (2007). Is the growth of euro area small and medium-sized enterprises constrained by financing barriers? *Industrial Policy and Economic Reforms, Enterprise and Industry Directorate, General European Commission*, (6).

Gestel, T. V., Baesens, B., Dijcke, P. V., Suykens, J. A. K., Garcia, J., and Alderweireld, T. (2005). Linear and non-linear credit scoring by combining logistic regression and support vector machines. *Journal of Credit Risk*, **1**(4), 31–60.

Greene, W. H. (2012). *Econometric Analysis*. Prentice Hall, New York.

Gupton, G. M., Finger, C. C., and Bhatia, M. (1997). Creditmetrics. In *Technical document*. J. P. Morgan.

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the roc curve. *Machine Learning*, **77**, 103–123.

Hand, D. J. (2010). Evaluating diagnostic tests: the area under the roc curve and the balance of errors. *Statistics in Medicine*, **29**, 1502–1510.

Hand, D. J. and Henley, W. E. (1997). Some developments in statistical credit scoring. In Nakhaeizadeh, N. and Taylor, C., editors, *Machine learning and statistics: the interface*, pages 221–237. Wiley, New York.

Hand, D. J., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. MIT Press.

Huang, J. J., Tzeng, J. H., and Ong, C. S. (2006). Two-stage genetic programming (2sgp) for the credit scoring model. *Applied Mathematics and Computation*, **174**, 1039–1053.

Jarrow, R. A. and Turnbull, S. M. (1995). Pricing derivatives on financial securities subject to credit risk. *Journal of Finance*, **50**, 53–85.

Kiefer, N. M. (2010). Journal of business and economic statistics. *Journal of Business Finance & Accounting*, **28**(2), 320–328.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, **9**, 321–354.

Kotz, S. and Nadarajah, S. (2000). *Extreme Value Distributions. Theory and Applications*. Imperial College Press, London.

Lee, T. S. and Chen, I. F. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, **28**, 743–752.

Lin, S. M., Ansell, J., and Andreeva, G. (2012). Predicting default of a small business using different definitions of financial distress. *Journal of the Operational Research Society*, **63**, 539–548.

Majocchi, A. and Zucchella, A. (2003). Internationalization and performance: Findings from a set of italian smes. *International Small Business Journal*, **21**(3), 249–268.

Marra, G. and Radice, R. (2010). Penalised regression splines: Theory and application to medical research. *Statistical Methods in Medical Research*, **19**, 107–125.

Marra, G. and Wood, S. (2012). Coverage properties of confidence intervals for generalized additive model components. *Scandinavian Journal of Statistics*, **39**, 53–74.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman Hall, New York.

Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, **29**, 449–470.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer-Verlag, New York.

Ong, C. S., Huanga, J. J., and Tzeng, G. H. (2005). Building credit scoring models using genetic programming. expert systems with applications. *Expert Systems with Applications*, **29**, 41–47.

Pederzini, E. (2012). European policies to promote the access to finance of smes. In B. Dallago, C. G., editor, *The Consequences of the International Crisis on European SMEs - Vulnerability and Resilience*, pages 89–106. Routledge.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, **66**, 403–411.

Products, C. S. F. (1997). Creditrisk+: A credit risk management framework. In *Credit Suisse First Boston*.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, London.

Saurina, J. and Trucharte, C. (2004). The impact of basel ii on lending to small- and medium-sized firms: A regulatory policy assessment based on span- ish credit register data. *Journal of Finance Service Research*, **26**(3), 121–144.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika*, **72**, 67–90.

Thomas, L., Edelman, D., and Crook, J. C. (2002). *Credit Scoring and Its Applications*. Society for Industrial and Applied Mathematics, Philadelphia.

Toyli, J., Hakkinen, L., Ojala, L., and Naula, T. (2008). Logistics and financial performance. an analysis of 424 finnish small and medium-sized enterprises. *International Journal of Physical Distribution & Logistics Management*, **38**(1), 57–80.

Wang, X. and Dey, D. K. (2010). Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. *The Annals of Applied Statistics*, **4**(4), 2000–2023.

Wilson, T. C. (1998). The impact of basel ii on lending to small-and medium-sized firms: A regulatory policy assessment based on spanish credit register data. *Economic Policy Review*, **4**, 71–82.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99**, 673–686.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall, Boca Raton.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika*, **100**, 221–228.

Zavgren, C. (1998). The prediction of corporate failure: the state of the art. *Journal of Accounting Literature*, **2**, 1–37.