

## Deriving Consensus Ratings of the Big Three Rating Agencies

Grün, Bettina; Hofmarcher, Paul; Hornik, Kurt; Leitner, Christoph; Pichler, Stefan

Published: 01/12/2010

### *Document Version*

Publisher's PDF, also known as Version of record

[Link to publication](#)

### *Citation for published version (APA):*

Grün, B., Hofmarcher, P., Hornik, K., Leitner, C., & Pichler, S. (2010). *Deriving Consensus Ratings of the Big Three Rating Agencies*. (Research Report Series / Department of Statistics and Mathematics; No. 99).

# DERIVING CONSENSUS RATINGS OF THE BIG THREE RATING AGENCIES



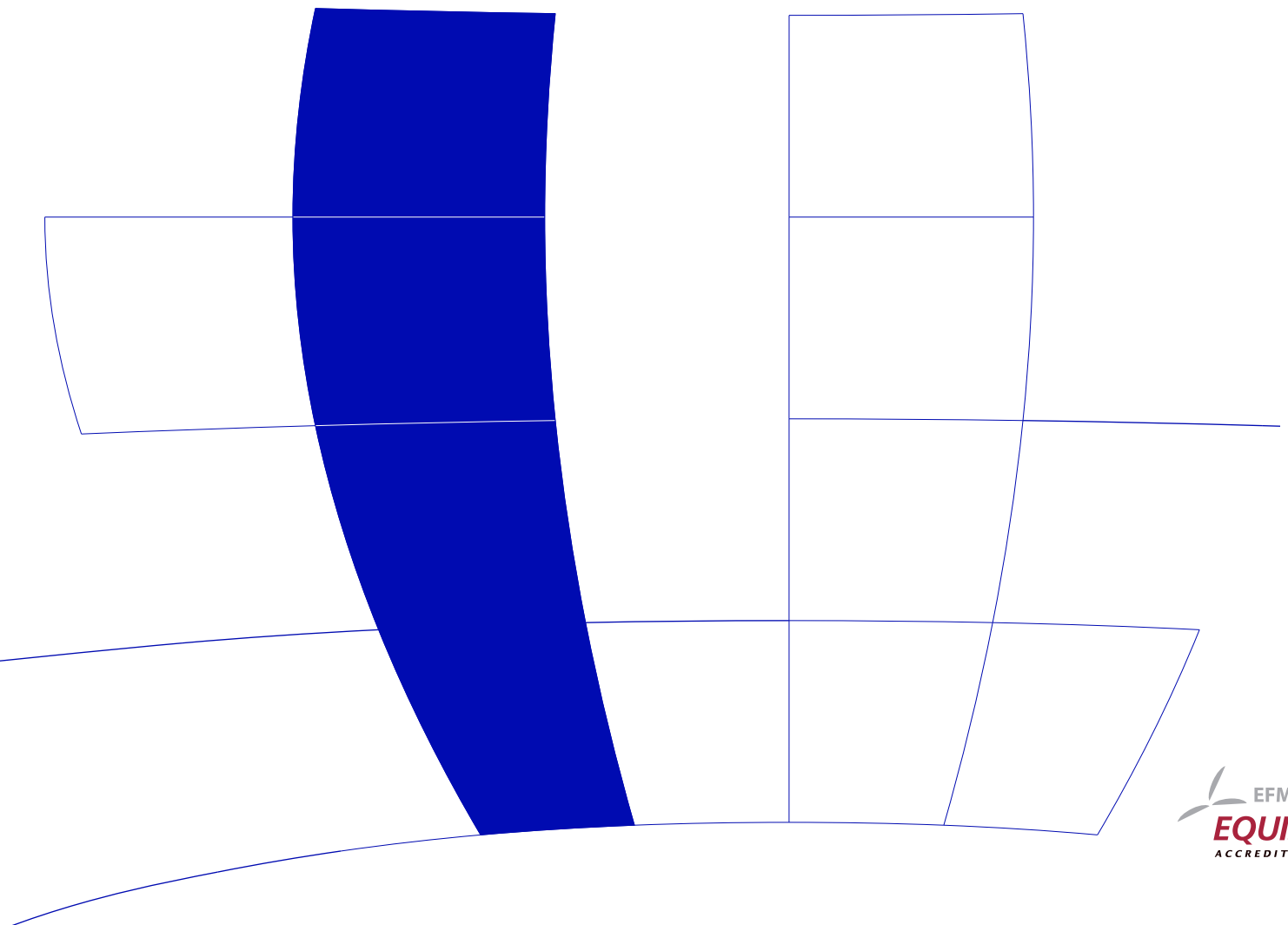
Bettina Grün, Paul Hofmarcher, Kurt Hornik,  
Christoph Leitner, Stefan Pichler

Institute for Statistics and Mathematics  
WU Wirtschaftsuniversität Wien

## Research Report Series

Report 99  
March 2010

<http://statmath.wu.ac.at/>



# Deriving Consensus Ratings of the Big Three Rating Agencies

Bettina Grün<sup>a</sup>, Paul Hofmarcher<sup>a</sup>, Kurt Hornik<sup>a</sup>,  
Christoph Leitner<sup>a</sup>, Stefan Pichler<sup>b</sup>

March 15, 2010

<sup>a</sup> Institute for Statistics and Mathematics, WU Wien,  
Augasse 2-6, A-1090 Vienna, Austria

<sup>b</sup> Institute for Finance, Banking and Insurance, WU Wien,  
Heiligenstädter Straße 46-48, A-1190 Vienna, Austria

---

## Abstract

This paper introduces a model framework for dynamic credit rating processes. Our framework aggregates ordinal rating information stemming from a variety of rating sources. The dynamic of the consensus rating captures systematic as well as idiosyncratic changes. In addition, our framework allows to validate the different rating sources by analyzing the mean/variance structure of the rating errors.

In an empirical study for the iTraxx Europe companies rated by the big three external rating agencies we use Bayesian techniques to estimate the consensus ratings for these companies. The advantages are illustrated by comparing our dynamic rating model to a benchmark model.

**Keywords:** Bayesian estimation, consensus information, credit ratings, external rating agencies, rating validation.

# 1 Introduction

The role of credit ratings provided by the big three external rating agencies Standard&Poor’s, Moody’s and Fitch has increased because modern credit risk pricing requires individual risk parameters, like rating implied default probabilities (PDs) (Kliger and Sarig, 2000). Despite the fact that all three raters express forward-looking opinions about the creditworthiness of firms on an ordinal scale, all three use different rating systems with different granularity as well as different labels (typically, a combination of letters, numbers and/or modifiers). Nevertheless, the agencies consider the likelihood of default to be a centerpiece of creditworthiness and therefore consistent with the goal of an ordinal rating scale, where firms with a lower rating should have a higher PD than firms with a higher rating (e.g., Cantor and Packer, 1997). Obviously, the raters do not always agree on the creditworthiness of the firms (e.g., Cantor and Packer, 1995; Jewell and Livingston, 2002). This resulting rating heterogeneity raises questions regarding the (1) nature, (2) quality and (3) interpretation of the ratings and the corresponding PDs. Are there consistent differences in their rating behavior? Does one agency have somewhat better information than the others regarding the creditworthiness? Or, does the rating heterogeneity just evince the very subjective and probabilistic nature of ratings (Ederington, 1986)? Along with different definitions of ratings, do they measure different variables representing the creditworthiness? Hence, rating heterogeneity nourishes the hypothesis that the rating processes of the agencies are not absolute and the differences in the published ratings may be a result of different sources of information, of different opinions about the obligors or of discriminative focuses in the rating process, e.g., one agency might give more weight to the balance sheet leverage than the other. In addition, unsystematic or random errors may occur in a rating process. Cantor and Packer (1997) assess the problem whether observed rating heterogeneity reflects different rating scales or is simply the result of selection bias. There is a growing literature on the analysis of credit ratings as well as their providers in the context of validation, regulation and information of the credit market (Moon and Stotsky, 1993; Cantor and Packer, 1995; Krahn and Weber, 2001; Jewell and Livingston, 2002; Altman and Rijken, 2004; Stolper, 2009), but to the best of our knowledge there is no literature discussing the topic how to combine different (heterogeneous) ratings of a company into a common rating. Especially in the area of financial modeling, where ratings play an inevitable role, it is essential to be able to deal with rating heterogeneity. To give examples, ratings serve as input parameters in industry models, e.g. CreditMetrics; they are used for regulatory issues, like in the Basel II framework. For the validation and extension of existing rating systems it is indispensable to cope with rating heterogeneity. Our framework addresses these issues.

In order to aggregate information of different raters a measure of “consensus” might be helpful. Zarnowitz and Lamnros (1987) define “consensus” as the degree of agreement among point predictions aimed at the same target by different individuals. It can be computed as the median (Su and Su, 1975) or the mean of all the predictions in the sample (Zarnowitz and Lamnros, 1987). Alternative strategies for the aggregation of predictions are discussed by Cook and Seiford (1982); Schnader and Stekler (1991); Kolb and Stekler (1996). In the context of forecasting the PD of some firms, Hornik et al. (2010) use a static mixed-effects model (Pinheiro and Bates, 2000) to model the consensus PD with rater-specific fixed effects and a random effect for firms.

The aim of this paper is to solve the information problem of combining different rating information stemming from different rating sources by deriving appropriate *consensus* information, i.e., consensus ratings which incorporate the information of several rating sources. Such a consensus measure can be interpreted as a *more informative* rating since it incorporates the whole disposable information about one firm. In addition, based on the consensus ratings and the rating errors, we assess the *precision* and the *agreement* of the different rating sources which may serve as the basis of validating different rating systems. Finally, to justify our framework, we compare it to an intuitive *benchmark* approach, which states that the consensus rating of a company is simply the “mean” of the company’s ratings at any considered time. In contrast to the benchmark approach our model is more appropriate especially

if someone is interested in forecasting rating movements of companies or if the considered data include missingness, e.g., if for a company only two of three ratings are available.

The model framework presented in this paper is related to other studies on credit rating systems (e.g., McNeil and Wendin, 2007; Hornik et al., 2010; Stefanescu et al., 2009). In contrast to Hornik et al. (2010) our model framework estimates the consensus rating on an ordinal scale and in a *dynamic* way. In addition, we make use of a *latent* market variable, describing the overall level of “creditworthiness”, and induces a correlation structure between the estimated consensus ratings. This is a well accepted strategy in the credit risk literature (e.g., Nickell et al., 2000; McNeil and Wendin, 2006, 2007; Stefanescu et al., 2009). Therefore we refer to our model setup as the *dynamic latent trait model*.

In order to illustrate the potential of our dynamic model framework, we apply it to the iTraxx Europe (Series 10) companies rated by the big three external rating agencies. In particular, we use all available ordinal rating information of these companies by the three raters over a time period from 2007-01-01 to 2008-12-31. Using these data, we estimate the consensus ratings and analyze the three raters according to their rating errors and their agreement with the consensus ratings.

The remainder of this paper is organized as follows: Section 2 describes the estimation of the consensus ratings. In Section 2.1 we discuss our dynamic latent trait model and Section 2.2 explains the benchmark approach which is used to validate our dynamic model. Section 3 provides a data description of the iTraxx Europe (Series 10) index and the agency ratings of the firms within this index. Section 4 applies the models described in Section 2 to the data. Bayesian estimation techniques, like Gibbs sampling, are used to estimate the parameters of interest. The benchmark as well as the dynamic model are fitted to the data. The appropriateness of the dynamic model is confirmed by the deviance information criterion (DIC) (Spiegelhalter et al., 2002). The DIC indicates that the dynamic model dominates than the benchmark approach. Section 5 concludes and summarizes the main results and the implications of our framework.

## 2 Consensus modeling

In this section we develop a model framework to derive a consensus rating for raters providing ordinal rating information, e.g., external agency ratings. Our model is designed for a dynamic framework capturing a time dependent rating process. Despite the fact that the raters publish ordinal ratings, we assume that they estimate a numerical variable—representing the creditworthiness of the firm—in an internal rating process. Each firm is then assigned to a particular rating class if this variable lies within a certain interval (e.g., McNeil and Wendin, 2007; Stefanescu et al., 2009). In general, the specific rating process including both the estimation as well as the scale of the variable (representing the creditworthiness) is unknown. In the literature, modeling the creditworthiness, was first discussed by Altman (1968) who introduces the Z-score. Z-scores are used to predict corporate defaults and are an easy-to-calculate control measure for the financial distress status of companies. The Z-score uses multiple corporate income and balance sheet values to measure the financial health of a company. Furthermore, Merton (1974) assumes that the creditworthiness can be reflected by the distance-to-default (DD) capturing the distance of the firm’s asset value to its default threshold on the real line. Alternatively, the creditworthiness variable can also be the result of a ordered probit or logit regression model (e.g., Altman and Rijken, 2004). To obtain ordinal ratings, the estimated DD, the Z-score, or any other numerical variable representing the creditworthiness—which is in the following referred to as “rating score”—is mapped onto an ordinal rating scale by the raters.

Let  $\{1, \dots, K_j\}$  be the set of possible non-default rating classes of rater  $j$  in descending creditworthiness. That is, 1 denotes the best credit quality and  $K_j$  the worst non-default rating class of rater  $j$ . Further,  $S_{ij}(t)$  denotes the estimated rating score (e.g., negative DD, Z-score) and  $r_{ij}(t)$  the associated observed ordinal rating of firm  $i$  by rater  $j$  at time  $t$ . The relationship between  $r_{ij}(t)$  and  $S_{ij}(t)$  is given

by

$$r_{ij}(t) = k \Leftrightarrow S_{ij}(t) \in [\lambda_{k-1,j}, \lambda_{k,j}], \quad (1)$$

for a monotonically increasing sequence  $\lambda_{k,j}$  with  $k = 1, \dots, K_j$ . The class boundaries are assumed to be constant over time. The data consists of observations for  $J$  raters and  $T$  time points. Observing rating  $k$  for a firm by rater  $j$  means that its rating score lies somewhere in the interval  $[\lambda_{k-1,j}, \lambda_{k,j}]$ .

In general, the thresholds  $\lambda_{k,j}$  are not provided by the raters. One possibility to obtain  $\lambda_{k,j}$  is to relate the ratings to the observable empirical default rates. In particular, the thresholds can be computed by using the empirical default rates on an appropriate scale<sup>1</sup>. Assuming that the scores of empirical default rates,  $S_{ij}(t)$ , are defined on the real line we have to fix the lower as well as the upper threshold ( $\lambda_{0,j} = -\infty$  and  $\lambda_{K_j,j} = +\infty$ , respectively). The length of the intervals need not be equal and may differ from rater to rater. Nevertheless, it is expected that firms within the same interval will exhibit roughly the same creditworthiness (Stefanescu et al., 2009).

Due to general informational asymmetry between firm owners and raters<sup>2</sup> which can be due to limited access to the existing information, such as incomplete accounting information (Duffie and Lando, 2001), or delayed observations of the driving risk factors (Guo et al., 2008) the raters cannot estimate the “true” score (reflecting the creditworthiness) of a firm. Assuming that the yielding rating errors can be modeled additively<sup>3</sup> and following Equation (1) the relationship between the estimated rating score  $S_{ij}(t)$  and the latent score  $S_i(t)$  on the score scale is given by

$$S_{ij}(t) = S_i(t) + \epsilon_{ij}(t), \quad (2)$$

where  $\epsilon_{ij}(t)$  denotes the rating error for firm  $i$  by rater  $j$  at time  $t$ . In the following, the latent score  $S_i(t)$  is also referred to as the *consensus* score.

On the right hand side of Equation (2) we find two terms, which have to be specified: (1) The latent score  $S_i(t)$  which describes the consensus creditworthiness and (2) the error term  $\epsilon_{ij}(t)$  which captures the accuracy of the rating system of a specific rater. In the following those terms are specified for both the dynamic latent trait model and the benchmark approach.

Despite the fact that the scores  $S_{ij}(t)$  are unknown, the latent scores  $S_i(t)$  and the bias/variance structure of the rating errors can be estimated in our framework by specifying the distribution of the rating errors and using the interval thresholds  $\lambda_{\cdot,j}$  along with the relationship of Equation (1). The estimated consensus scores  $S_i(t)$  can then be mapped on the rater-specific ordinal scale to derive the consensus ratings  $r_{ij}^*(t)$  which obviously depend on the used rating system (of rater  $j$ ). Since  $r_{ij}(t)$  and  $r_{ij}^*(t)$  for all  $i$  and  $j$  are on the same rating scale one can easily compare these ratings and derive inference about the quality of the ratings  $r_{ij}(t)$ .

## 2.1 Dynamic latent trait model

**Latent consensus score.** In order to specify the latent scores  $S_i(t)$ , we follow the lines of McNeil and Wendin (2007); Stefanescu et al. (2009) and assume that the scores are driven by market- (systematic risk) as well as firm-specific effects (idiosyncratic risk). We define a time-dependent process  $m_i(t)$  capturing the idiosyncratic changes and a latent market factor  $f(t)$  capturing the systematic development of the latent scores  $S_i(t)$ . The idiosyncratic changes  $m_i(t)$  capture the firm-specific risk and can be modeled as

<sup>1</sup>Beside this, we assume that raters do not change their rating technology during the desired time period, i.e., they are always measuring creditworthiness on the same scale. This assumption justifies time independent  $\lambda_{k,j}$ .

<sup>2</sup>The general informational asymmetry between firm owners and raters constitutes the cornerstone of modern corporate finance (e.g., Leland and Pyle, 1977; Berk and DeMarzo, 2007).

<sup>3</sup>This is in line with Duffie and Lando (2001) who build their model on a Merton-type log normal firm value process and assume that the error in the observation of the firm value is normal and additive to the log of the firm value.

an adequate time series process to cope with repeated observations. The latent market  $f(t)$ , capturing the development of the market, implies a correlation structure between the different firms and can also be modeled by an adequate time-dependent process, e.g., a stationary auto-regressive process or a random walk. Let  $\nu_i$  be the firm specific long-term mean of firm  $i$  which can be interpreted as the historical average creditworthiness of the firm. The development of the latent scores  $S_i(t)$  on the score scale is given by

$$S_i(t) = \nu_i + m_i(t) + \alpha f(t), \quad (3)$$

where the factor loading  $\alpha$  captures the dependence of  $S_i(t)$  on  $f(t)$ .

In order to estimate the consensus scores  $S_i(t)$  we have to specify the underlying processes and distributions of this framework. We specify the time-dependent processes, describing the development of  $S_i(t)$  (Equation (3)), the firm-specific changes  $m_i(t)$  and the latent market factor  $f(t)$  as AR(1) processes

$$m_i(t) = \beta_i m_i(t-1) + \omega_i(t), \quad (4)$$

$$f(t) = \gamma f(t-1) + \xi(t). \quad (5)$$

$m_i(t)$  and  $f(t)$  are assumed to start with zero at  $t = 0$ .  $\omega_i(t)$  is a normal distributed error term with mean zero and a constant variance across time and firms, and  $\xi(t)$  is a standard normal distributed error term.  $\beta_i$  ( $|\beta_i| < 1$ ) and  $\gamma$  ( $|\gamma| < 1$ ) reflect the dependence on period  $t - 1$  (inter-temporal correlation).

**Rating error.** In order to specify the rating errors  $\epsilon_{ij}(t)$ , we assume that they are independent of the firms and their characteristics (in particular, their creditworthiness itself) and the general rating process does not change over time  $t$  (see Hornik et al., 2010). Assuming that  $\mu_j$  and  $\sigma_j$  denote the mean and standard deviation of the rating errors  $\epsilon_{ij}(t)$ , respectively, the rating errors  $\epsilon_{ij}(t)$  are given by

$$\epsilon_{ij}(t) = \mu_j + \sigma_j Z_{ij}(t) \quad (6)$$

where  $Z_{ij}(t)$  is assumed to be independent standard normal distributed over  $i$ ,  $j$  and  $t$ .

## 2.2 Benchmark Model

In addition to the dynamic latent trait model, we define an intuitive benchmark approach and compare it with our dynamic latent trait model. Being conservative, one could consider to take the companies' worst rating as the benchmark. This is inappropriate for two reasons. Firstly, such an approach disregards the information contained in the other available rating sources. Secondly, from an economic point of view a rated company must be convinced that its creditquality lies somewhere *between* its ratings and is not represented by the worst rating. Otherwise there would be little reason to obtain several ratings (Hsueh and Kidwell, 1988). Hence, without any rater specific characteristics, the "mean" of the observed ratings could serve as a consensus benchmark.

**Latent consensus score.** Our benchmark model follows the idea that for any time  $t$ , the consensus score  $S_i(t)$  of a company is simply the *mean* over rating scores  $S_{ij}(t)$ . In doing so, we do not assume any time-dependent process driving the development of  $S_i(t)$ , i.e., for any time  $t$ ,  $S_i(t)$  is independent of  $S_i(t-1)$ .

**Rating errors.** For the rating errors, we assume that there are no rater specific error terms  $\mu_j$  and  $\sigma_j$ , but a constant standard deviation  $\sigma$  of the rating errors between the raters. This implies that all raters are weighted equally in the estimation process. Within our model framework the relationship between

consensus score  $S_i(t)$  and the estimated scores  $S_{ij}(t)$  for the benchmark model is given by

$$S_{ij}(t) = S_i(t) + \sigma Z_{ij}(t), \tag{7}$$

with  $Z_{ij}(t)$  distributed as in the dynamic case.

One drawback of this model specification is that these assumptions for the rating errors and the latent scores may lead to distorted results for rating data including missings, i.e., some companies are not rated by all agencies (see Figure 1).

### 3 Data

**Ordinal ratings of the iTraxx Europe companies.** We use historical long-term issuer ratings of the constituents of the iTraxx Europe index (Series 10) from February 2007 to January 2009 provided by the big three external rating agencies Standard&Poor’s, Fitch and Moody’s. The iTraxx Europe index series consists of the 125 most-liquid CDS referencing European investment-grade entities and a new series is determined by dealer liquidity poll every six months. Most of the 125 names in the indexes are large multinationals and have traded equity. We choose the iTraxx Europe index, because it forms a representative contingent of the overall European credit derivative market and its constituents have a high number of co-ratings (occurrences of ratings of a single firm by two different raters) from the big three rating agencies. The time series was constructed using historical ordinal rating announcements taken from Reuters Credit Views. We exclude all companies for which we do not have rating information of at least two agencies for the complete time period, i.e., those with withdrawn ratings and entities which acquire a rating for the first time within the selected time frame. This process yields a sample of 5616 monthly ratings for 95 companies over 24 months (February 2007 to January 2009). Table 1 shows the co-ratings structure of the three raters. The average number of ratings for each month is 2.46.

	Fitch	Moody’s	S&P
Fitch	88	44	88
Moody’s	44	51	51
S&P	88	51	95

Table 1: Co-ratings structure for 95 out of the 125 iTraxx Europe (Series 10) companies of the big three external rating agencies Fitch, Moody’s and Standard&Poor’s (S&P).

As described in Section 1, the three rating agencies use different rating systems. Moody’s rating system for global corporates contains 20 non-default rating categories, ranging from *Aaa* to *C* and is so in the near default ratings more granular than the rating systems of Fitch and Standard&Poor’s (Emery and Ou, 2009). These two agencies assign 17 non-default rating categories (*AAA* to *CCC/C*) to global corporates (Needham and Verde, 2009; Vazza et al., 2009). Table 2 shows the number of ratings (per rating category and rater) of the monthly ratings from February 2007 to January 2009 for the rating agencies Fitch, Moody’s and Standard&Poor’s.

According to the three rating distributions of this rating data, only one firm is rated as a non-investment firm (ContinentalAG) and this only by Standard&Poor’s (see Crouhy et al., 2001, for a description of investment grades and speculative grades). The distributions show also that the granularity of the three rating systems is equal in the relevant segment of this rating data.

The rating history of 57 firms (60%) changed over the considered time period. Fitch changed the ratings of 35 firms, where 29 firms were downgraded and 4 firms were upgraded. The remaining two companies experienced a downgrade as well as an upgrade. Moody’s changed the ratings of 17 firms, where 8 firms were downgraded and 8 firms were upgraded (the remaining company experienced two



	Fitch		Moody's		S&P	
	label	no.	label	no.	label	no.
1	AAA	6	Aaa	18	AAA	0
2	AA+	85	Aa1	176	AA+	45
3	AA	148	Aa2	41	AA	167
4	AA-	193	Aa3	54	AA-	233
5	A+	226	A1	79	A+	170
6	A	243	A2	153	A	251
7	A-	410	A3	225	A-	473
8	BBB+	454	Baa1	231	BBB+	576
9	BBB	315	Baa2	183	BBB	292
10	BBB-	30	Baa3	64	BBB-	72
11	BB+	2	Ba1	0	BB+	0
12	BB	0	Ba2	0	BB	1
13	worse	0	worse	0	worse	0

Table 2: Number of ratings (per rating category and rater) of the 95 out of the 125 iTraxx Europe companies.

upgrades as well as two downgrades). Standard&Poor's changed the ratings of 45 firms, where 29 firms were downgraded and 12 firms were upgraded (the remaining four company experienced upgrade(s) as well as downgrade(s)). Hence, a clear tendency of downgrading is observable in this period.

In order to model the consensus ratings (Equation 2), each ordinal rating is identified with a numerical interval reflecting the upper and lower bound of the creditworthiness on the real line (see Equation 1). Here, we estimate the thresholds for the ordinal ratings using the empirical default rates (1990–2006) provided by the external raters (Needham and Verde, 2009; Emery and Ou, 2009; Vazza et al., 2009). A detailed description of this estimation is given in Appendix A.

**Dow Jones EURO STOXX 50.** For comparison reason we use the Dow Jones EURO STOXX 50 as a representative market development of the iTraxx Europe portfolio from February 2007 to January 2009 (see Figure 2). The Dow Jones EURO STOXX 50 is the leading stock (price) index for the Eurozone and covers 50 stocks from 12 Eurozone countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, and Spain. At January 2009, stocks of 30 out of the 95 companies are contained in the EURO STOXX 50.

## 4 Analysis of the big three rating agencies using their ratings for the iTraxx Europe companies

### 4.1 Model estimation

Using the available ordinal ratings  $r_{ij}(t)$  for each company  $i = 1, \dots, 95$  (out of the 125 iTraxx Europe companies) and external rating agency  $j = \{F, M, SP\}$  from  $t = 1, \dots, 24$  (February 2007 to January 2009) and the associated thresholds  $\lambda_{j,k}$  for  $k = 1, \dots, K_j$  with  $K_F = 17$ ,  $K_M = 20$ , and  $K_{SP} = 17$  we estimate the model parameters of our dynamic latent trait model as well as the parameters of our benchmark model. For the estimation frequentist as well as Bayesian techniques can be used. E.g., Hornik et al. (2010) estimated their model by standard maximum likelihood estimation. Here, we follow McNeil and Wendin (2007) and Stefanescu et al. (2009) and choose a Bayesian estimation approach using Markov chain Monte Carlo methods (MCMC) and Gibbs sampling (Carlin and Louis, 2009). Such an approach

requires prior distributions to be chosen for the parameter set. In order to minimize the influence of the prior distributions on the posterior distribution we have specified non-informative priors for all our parameters.

In particular, we run four parallel Markov chains, each initialized with a different seed and a different random number generator. The Gibbs sampler ran for 50,000 iterations, using a thinning of 10 whereby the first 5,000 were discarded as burn-in period. This yields 4,500 draws from the posterior for each parameter for each chain. Trace plots as well as the Geweke diagnostic and the Gelman Rubin’s convergence diagnostic indicated satisfactory convergence of all chains (e.g., Gelman and Rubin, 1992; Plummer et al., 2008).

**Model selection.** In order to compare our dynamic latent trait model with the benchmark model we use the *deviance information criterion* (DIC; according to Spiegelhalter et al., 2002). The DIC is a generalization of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for hierarchical models. In contrast to the AIC and BIC, DIC allows to compare Bayesian hierarchical models where the effective number of parameters is not clearly defined. Similar to the other information criteria a trade-off between model fit and model complexity is evaluated. The DIC contains one penalty term for the effective number of parameters used measuring model complexity and one term equal to the deviance of the likelihood measuring model fit. A lower DIC value indicates a better model fit. According to Spiegelhalter et al. (2002), if the difference in DIC is greater than 10, then the model with the larger DIC value has considerably less support than the model with the lower DIC value.

For our models, the lower DIC value of our dynamic latent trait model (DIC = 9581.55) indicates that this model dominates in the terms of model fit as well as model complexity the obvious benchmark model (DIC = 16242.89).

## 4.2 Results for the dynamic latent trait model

**Rating errors.** We begin our analysis of the estimation results with the rating errors. Our dynamic latent trait model captures estimates for the rating bias  $\mu_j$  and the standard deviation  $\sigma_j$  of the rating error of the big three external rating agencies on the score scale. Table 3 shows the results for the estimated posterior distribution of the parameters for the three raters  $\mu_j$  and  $\sigma_j$ , respectively. The posterior distributions of the parameters are characterized by the mean values (mean) and the standard deviations (SD) of the 18,000 ( $4 \times 4,500$ ) posterior draws.

	$\mu_j$		$\sigma_j$	
	mean	SD	mean	SD
Fitch	0.0156	0.0018	0.0751	0.0021
Moody’s	-0.0887	0.0024	0.1013	0.0028
S&P	0.0732	0.0017	0.0642	0.0017

Table 3: Estimated rating bias  $\mu_j$  and standard deviations  $\sigma_j$  for the rating errors (on the score scale) of the big three external rating agencies Fitch, Moody’s and Standard&Poor’s. The posterior distributions of the parameters are characterized by the mean values (mean) and the standard deviations (SD) of the 18,000 ( $4 \times 4,500$ ) posterior draws.

We infer from Table 3 that Fitch has the smallest absolute rating bias from the consensus on the score scale with respect to the posterior mean (0.0156). Moody’s clearly seems to be too optimistic in its credit assessment yielding a posterior mean for the rating bias  $\mu$  of  $-0.089$  on the score scale. Note, that our model is based on the thresholds  $\lambda_{j,k}$  (and therefore PD equivalents) which are clearly lower for Moody’s than the other two raters. Despite the high difference (on the score scale: 0.139) in the PD

equivalents of Moody’s and Standard&Poor’s indicated in the Appendix (see Figure 3), Moody’s is still more optimistic by rating investment-grade firms than Standard&Poor’s. In this study, Standard&Poor’s is with a posterior mean of the rating bias of 0.073 the most conservative rater out of the three considered rating agencies.

In addition to the rating biases, our model captures the standard deviation (precision) of the rating errors of the three raters (Table 3). Whereas the posterior mean of the standard deviation  $\sigma$  of the rating errors is rather similar for Fitch and Standard&Poor’s (0.075, 0.064), Moody’s has a higher posterior mean of the standard deviation (0.101), indicating that its ratings deviate more strongly from the consensus ratings.

**Consensus score.** In addition to the analysis of the bias/variance structure of the rating errors, we analyze the estimated consensus scores of our dynamic latent trait model. Instead of showing the consensus scores of all iTraxx Europe companies, Figure 1 shows the estimated consensus rating scores of four sample companies (ENELSPA, NESTLE, GLENCORE INT. AG, ROYAL BANK OF SCOTLAND) and compares them with the original ratings (mapped onto the score scale) of the three raters Fitch, Moody’s and Standard&Poor’s as well as with the mean rating score of the three raters.

Due to the fact that the companies ENELSPA and NESTLE are rated by all three raters, the consensus score (solid line) is very similar to the mean score (dashed line). In the case of the two other companies GLENCORE INT. AG and ROYAL BANK OF SCOTLAND where for each company ratings of only two raters are available, Figure 1 shows remarkable differences between the consensus and the mean score. Due to rater specific error terms, our latent consensus score is able to incorporate such a missingness structure.

Furthermore, we can confirm the need of a latent market factor in our dynamic latent trait model by showing the strong relationship between our latent market  $f(t)$  and a reference market, the Dow Jones EURO STOXX 50 index (correlation:  $-0.947$ )<sup>4</sup>.

**Consensus rating.** In addition to the analysis of the consensus scores, we can use the consensus ratings derived by mapping the scores onto the rater’s rating scales to analyze the rating agreement of the raters.

An intuitive way for this is the Hit-Miss-Match (HMM) Matrix which counts how many consensus ratings exactly match the ratings provided by a rater. Table 4, 5 and 6 show the HMM matrix for each rater.

In Table 4 most ratings are on the main diagonal or one rating notch below or above indicating a high agreement between Fitch’s ratings and the consensus ratings. Table 5 shows that Moody’s ratings are rather one or more rating notches below the consensus ratings, confirming the negative rating bias shown in Table 3. In contrast to Moody’s ratings, Standard&Poor’s ratings are rather one or more rating notches above the consensus ratings, confirming the positive rating bias shown in Table 3.

Furthermore, we can compute the proportion of ratings for each rating deviation (measured in rating notches) between the consensus ratings and the ratings provided by the raters (shown in Table 7).

Table 7 shows that Fitch’s ratings have a very high accordance (72.4%) with the estimated consensus ratings. According to the estimated rating biases (see Table 3) Moody’s is rather too “optimistic” than the other raters. These effect is also seen in Table 7. Only 27.6% of Moody’s ratings exactly hit the consensus rating. 84.7% are within one rating notch and 67.9% are more optimistic, i.e., are at least one rating category better than our estimated consensus rating. For Standard&Poor’s we obtain that 54.1% are within one rating category in comparison to the consensus rating. In contrast to Fitch, Standard&Poor’s have few ratings which are even 4 rating classes below the estimated consensus rating.

---

<sup>4</sup>Note, that the negative correlation is due to the fact that an increase in  $f(t)$  on the score scale indicates a decrease in the creditworthiness.

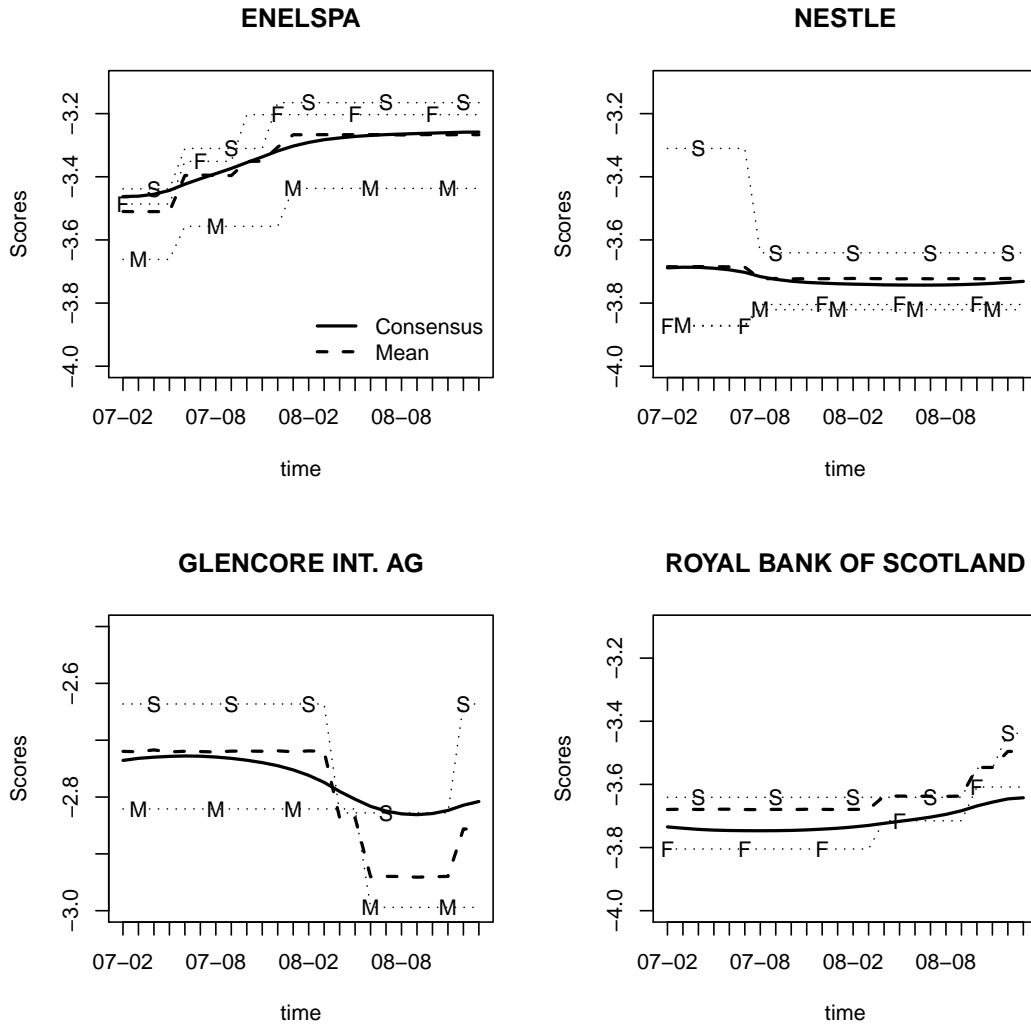


Figure 1: Estimated consensus score, the mean score, and the original ratings mapped onto the score scale of the big three external rating agencies Fitch (F), Moody’s (M) and Standard&Poor’s (S).

## 5 Discussion

In this paper we investigate a new dynamic framework for aggregating credit-rating information in a multi-rater set-up, i.e., in situations where ordinal ratings from different sources for the same firm are available. In our model we assume that the raters do not directly estimate the ordinal ratings, but they estimate a numerical variable—representing the creditworthiness of the firm—in an internal rating process. We treat the true unobservable numerical variable of a firm as a latent variable and model its dynamic by using systematic as well as idiosyncratic changes. In contrast to other methods, our model class allows missingness in the data and captures the panel structure of the data.

In addition to the solution for the aggregation problem, our model is useful in the validation of the different sources. The analysis of the mean/variance structure of the rating errors yields to rater-specific rating biases as well as the precision of the different rating systems.

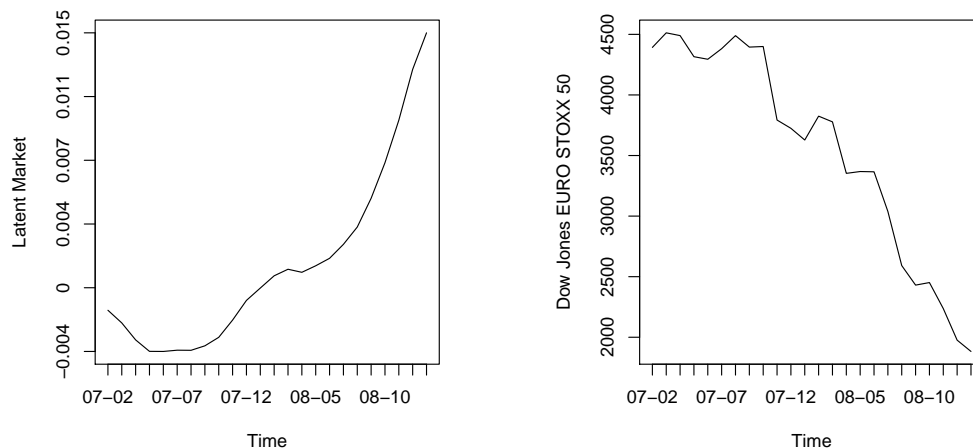


Figure 2: Estimated latent market factor  $f(t)$  and the Dow Jones EURO STOXX 50 index over the full time period (2007-02 to 2009-01).

Consensus rating	Fitch rating										
	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+
AAA	0	0	0	0	0	0	0	0	0	0	0
AA+	0	33	0	0	0	0	0	0	0	0	0
AA	6	52	124	17	0	0	0	0	0	0	0
AA-	0	0	21	157	44	14	0	0	0	0	0
A+	0	0	3	19	148	50	0	0	0	0	0
A	0	0	0	0	34	166	33	0	0	0	0
A-	0	0	0	0	0	13	308	82	3	0	0
BBB+	0	0	0	0	0	0	69	350	93	0	0
BBB	0	0	0	0	0	0	0	22	218	4	0
BBB-	0	0	0	0	0	0	0	0	1	26	2
BB+	0	0	0	0	0	0	0	0	0	0	0

Table 4: Hit-Miss-Match Matrix between the estimated consensus ratings and the ratings provided by Fitch, measured on the Fitch rating scale.

The suggested framework for modeling consensus of a multi-rater panel is very general and allows for a variety of possible enhancements. We could aim at employing more flexible models for the distributions of the rating scores and rating errors, e.g., via suitable mixtures of normals. We could also allow more flexibility in the specification of the factor loading  $\alpha$  capturing the dependence between the latent scores and the latent market (see Equation 3) by using a firm- or industry-specific factor loading. In addition, it would be interesting to allow for industry-specific parameters for the rating bias, the standard deviation of the rating error and the long-term mean (see Hornik et al., 2010). We could also try to use an external market factor (e.g., the Dow Jones EURO STOXX 50) instead of a latent market factor to describe the systematic changes of the latent scores. The use of Bayesian estimation techniques allows very flexible specification of models, so that we intend to explore these possible enhancements in our future research.

By using the ratings for the iTraxx Europe companies (Series 10) provided by the big three rating

Consensus rating	Moody's rating										
	Aaa	Aa1	Aa2	Aa3	A1	A2	A3	Baa1	Baa2	Baa3	Ba1
Aaa	0	0	0	0	0	0	0	0	0	0	0
Aa1	1	0	0	0	0	0	0	0	0	0	0
Aa2	10	80	7	2	0	0	0	0	0	0	0
Aa3	7	96	31	3	0	0	0	0	0	0	0
A1	0	0	3	16	34	24	0	0	0	0	0
A2	0	0	0	33	18	3	0	0	0	0	0
A3	0	0	0	0	3	126	73	0	0	0	0
Baa1	0	0	0	0	24	0	150	75	4	21	0
Baa2	0	0	0	0	0	0	2	156	101	3	0
Baa3	0	0	0	0	0	0	0	0	76	37	0
Ba1	0	0	0	0	0	0	0	0	2	3	0

Table 5: Hit-Miss-Match Matrix between the estimated consensus ratings and the ratings provided by Moody's, measured on the Moody's rating scale.

Consensus rating	Standard&Poor's rating											
	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB
AAA	0	43	46	0	0	0	0	0	0	0	0	0
AA+	0	2	93	7	0	6	0	0	0	0	0	0
AA	0	0	28	136	3	0	0	0	0	0	0	0
AA-	0	0	0	89	107	9	0	0	0	0	0	0
A+	0	0	0	1	60	121	2	0	0	0	0	0
A	0	0	0	0	0	99	158	0	0	0	0	0
A-	0	0	0	0	0	16	312	185	0	0	0	0
BBB+	0	0	0	0	0	0	1	391	89	0	0	0
BBB	0	0	0	0	0	0	0	0	203	40	0	0
BBB-	0	0	0	0	0	0	0	0	0	32	0	1
BB+	0	0	0	0	0	0	0	0	0	0	0	0
BB	0	0	0	0	0	0	0	0	0	0	0	0

Table 6: Hit-Miss-Match Matrix between the estimated consensus ratings and the ratings provided by Standard&Poor's, measured on the Standard&Poor's rating scale.

	-4	-3	-2	-1	0	1	2	3
Fitch	0.000	0.000	0.008	0.154	0.724	0.109	0.004	0.000
Moody's	0.000	0.000	0.017	0.027	0.276	0.544	0.115	0.020
S&P	0.003	0.030	0.426	0.533	0.008	0.000	0.000	0.003

Table 7: Proportion of ratings per rating class deviation between the consensus ratings and the origin ratings provided by the big three rating agencies Fitch, Moody's and Standard&Poor's.

agencies Fitch, Moody's and Standard&Poor's we compute a more informative rating, the consensus rating for each company and show that there are remarkable differences in the rating behavior and rating systems of the three raters. In particular, we infer from our results, that Moody's is the most favorable and Standard&Poor's the most pessimistic rater.

## Computational details

All computations were carried out in the R system (version 2.10.1) for statistical computing (R Development Core Team, 2009). In particular, the R package `rjags` (Plummer, 2009) was used for Gibbs sampling and model selection, and the R package `coda` (Plummer et al., 2008) was used for the output diagnostic.

### A Estimation of the rating thresholds

In order to map the ordinal ratings provided by the three external rating agencies to PD ratings (PD equivalents) we follow the approach proposed by Neagu et al. (2009). They relate empirical PDs to ratings on an appropriate score scale. The score variable represents a rank ordering of risk of default over some future time horizon (we use a one year future time period). The task is to find a transformation of the score variable into an empirical PD. In other words, this method aims at finding a function  $F$  such that:

$$\text{PD} = F(\text{score}),$$

which can be written by using a default indicator as:

$$\text{Prob}(\text{default indicator} = 1) = F(\text{score})$$

and gives the base formulation for the binary response class of models. Different types of models, utilizing different forms for the function  $F$ , can be fit. Neagu et al. (2009) suggest to try the three most commonly used binary response models: logit, probit, and complementary log-log (CLL) models. These models can be applied directly to the score data, but in real-world applications the score data tends to exhibit a high degree of skewness. In this case it is recommended that a transformation of the score variable is made: a Box-Cox power transformation (Fox, 1997) or a Box-Tidwell transformation (Granger and Newbold, 1977).

In particular, we use the published historical empirical global corporate default rates of the three external rating agencies from 1990 to 2006 (Needham and Verde, 2009; Emery and Ou, 2009; Vazza et al., 2009). In order to yield one-year empirical default rates we compute the averages over the time period. We then fit all combinations of binary response class models (probit, logit, and CLL) and transformations (Box-Cox power and Box-Tidwell) to the average default rates. A probit score model (as described in Section 2) with Box-Tidwell transformation is selected as the best method according to the Hosmer-Lemeshow statistic (Hosmer and Lemeshow, 2000). Figure 3 shows the estimated “mapping” lines using a probit score model with Box-Tidwell transformation for the three different rating systems of Fitch, Moody’s, and Standard & Poor’s using the empirical default rates from 1990 to 2006. Note, that the rating system of Moody’s is finer on the upper side, i.e., assigning four more rating grades to the high PD segment than the other two raters.

Whereas the empirical default rates and the PD mapping of Fitch and Standard&Poor’s seem to be rather similar, Moody’s empirical default rates and mapping line is clearly below the other two. E.g., in average the difference on the probit scale between the investment grades of Standard&Poor’s and Moody’s is 0.139.

In order to cleave to the ordinal structure of ratings, thresholds for the mapping PDs derived from the empirical default rates have to be computed. We compute the thresholds by the means of two adjacent mapping PDs on the logit scale for each rater  $j$ . I.e., the upper threshold  $\lambda_k$  of rating class  $k = 1, \dots, K_j - 1$  of rater  $j$  is given by  $\lambda_k = 1/2(\text{logit}(\text{PD}_{k+1}) + \text{logit}(\text{PD}_k))$  and the “lower” threshold of the best rating class is  $-\infty$  and the “upper” threshold of the worst rating class is  $+\infty$  (Altman and Rijken, 2004).

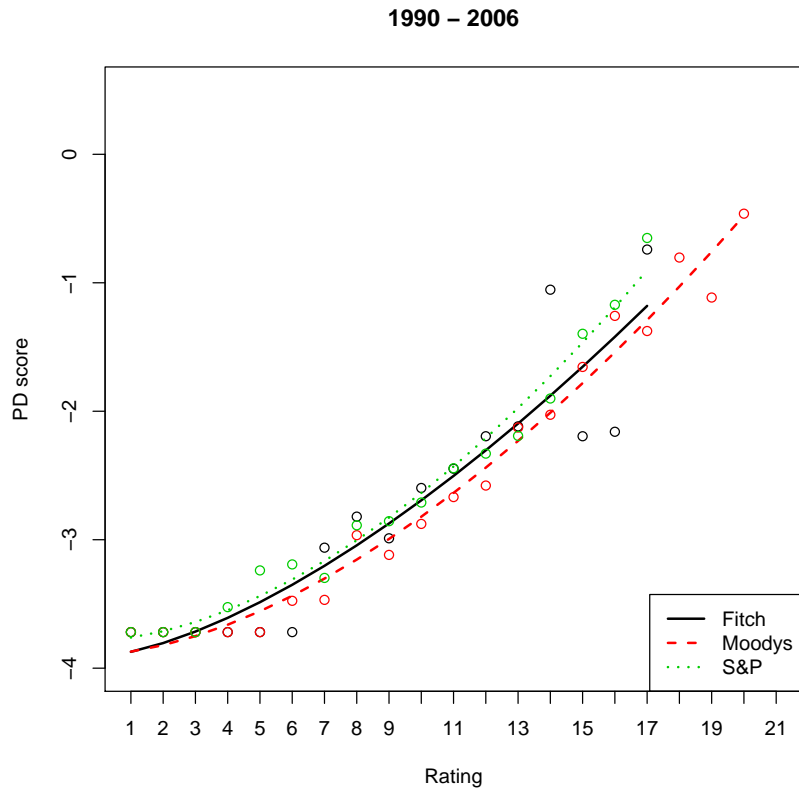


Figure 3: Mapping of the empirical default rates stemming from the three raters on the score scale based on a probit score model with Box-Tidwell transformation using the empirical default rates from 1990 to 2006.



## References

- E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23:189–209, 1968.
- E. I. Altman and H. A. Rijken. How rating agencies achieve rating stability. *Journal of Banking & Finance*, 28: 2679–2714, 2004.
- J. Berk and P. DeMarzo. *Corporate Finance*. Statistics and Computing. Pearson International Edition, Boston, USA, 2007. ISBN 0-321-41680-5.
- R. Cantor and F. Packer. Sovereign credit ratings. *Current Issues in Economics and Finance*, 1(3), 1995.
- R. Cantor and F. Packer. Differences of opinion and selection bias in the credit rating industry. *Journal of Banking & Finance*, 21:1395–1417, 1997.
- B. P. Carlin and T. A. Louis. *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC, 2009.
- W. D. Cook and L. M. Seiford. On the Borda-Kendall consensus method for priority ranking problems. *Management Science*, 28(6):621–637, 1982.
- M. Crouhy, D. Galai, and R. Mark. Prototype risk-rating system. *Journal of Banking & Finance*, 25:47–95, 2001.
- D. Duffie and D. Lando. Term structures of credit spreads with incomplete information. *Econometrica*, 69: 633–664, 2001.
- L. Ederington. Why split ratings occur. *Financial Management*, 15(1):37–47, 1986.
- K. Emery and S. Ou. Corporate default and recovery rates, 1920–2008. Moody’s global credit policy, Moody’s, New York, USA, 2009. URL <http://www.moody.com>.
- J. Fox. *Applied Regression Analysis, Linear Models, and Related Methods*. Sage, London, 1997.
- A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7: 457–511, 1992.
- C. Granger and P. Newbold. *Forecasting Economic Time Series*. Academic, New York, 1977.
- X. Guo, R. A. Jarrow, and Y. Zeng. Credit risk models with incomplete information. *Mathematics of Operations Research*, 34(2):320–332, May 2008.
- K. Hornik, R. Jankowitsch, C. Leitner, S. Pichler, M. Lingo, and G. Winkler. A latent variable approach to validate credit rating systems. In D. Rösch and H. Scheule, editors, *Model Risk in Financial Crises*, pages 277–296. Risk Books, London, 2010.
- D. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 2nd edition, 2000.
- P. Hsueh and D. Kidwell. Are two better than one? *Financial Management*, 17:46–53, 1988.
- J. Jewell and M. Livingston. A comparison of bond ratings from Moody’s, S&P and Fitch IBCA. *Financial Markets, Institutions & Instruments*, 8:1–45, 2002.
- D. Kliger and O. Sarig. The information value of bond ratings. *The Journal of Finance*, 6:2879–2902, 2000.
- R. Kolb and H. O. Stekler. Is there a consensus among financial forecasters? *International Journal of Forecasting*, 12:455–464, 1996.
- J. P. Krahn and M. Weber. Generally accepted rating principles: A primer. *Journal of Banking & Finance*, 25:3–23, 2001.

- H. E. Leland and D. H. Pyle. Informational asymmetries, financial structure, and financial intermediation. *Journal of Finance*, 32:371–387, 1977.
- A. J. McNeil and J. P. Wendin. Dependent credit migrations. *Journal of Credit Risk*, 2:87–114, 2006.
- A. J. McNeil and J. P. Wendin. Bayesian inference for generalized linear mixed model of portfolio credit risk. *Journal of Empirical Finance*, 14:131–149, 2007.
- R. C. Merton. On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29:449–470, 1974.
- C. G. Moon and J. G. Stotsky. Testing the differences between the determinants of Moody’s and Standard & Poor’s ratings. *Journal of Applied Econometrics*, 8:51–69, 1993.
- R. Neagu, S. Keenan, and K. Chalermkraivuth. Internal credit rating systems: Methodology and economic value. *The Journal of Risk Model Validation*, 3(2):11–34, 2009.
- C. L. Needham and M. Verde. Fitch ratings global corporate finance 2008 transition and default study. Credit market research, Fitch Ratings, 2009. URL <http://www.fitchratings.com>.
- P. Nickell, W. Perraudin, and S. Varotto. Stability of rating transitions. *Journal of Banking & Finance*, 24:203–227, 2000.
- J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer-Verlag, New York, USA, 2000. ISBN 0-387-98957-9.
- M. Plummer. *rjags: Bayesian Graphical Models Using MCMC*, 2009. URL <http://mcmc-jags.sourceforge.net>. R package version 1.0.3-5.
- M. Plummer, N. Best, K. Cowles, and K. Vines. *coda: Output Analysis and Diagnostics for MCMC*, 2008. URL <http://CRAN.R-project.org/package=coda>. R package version 0.13-3.
- R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL <http://www.R-project.org/>. ISBN 3-900051-07-0.
- M. Schnader and H. O. Stekler. Do consensus forecasts exist? *International Journal of Forecasting*, 7:165–170, 1991.
- D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 4:583–639, 2002.
- C. Stefanescu, R. Tunaru, and S. Turnbull. The credit rating process and estimation of transition probabilities: A Bayesian approach. *Journal of Empirical Finance*, 16:216–234, 2009.
- A. Stolper. Regulation of credit rating agencies. *Journal of Banking & Finance*, 33:1266–1273, 2009.
- V. Su and J. Su. An evaluation of ASA/NBER business outlook survey forecasts. *Explorations in Economic Research*, 2:588–618, 1975.
- D. Vazza, D. Aurora, and N. Kraemer. 2008 annual global corporate default study and rating transition. Ratings direct, Standard&Poor’s, 2009. URL <http://www.standardandpoors.com/ratingsdirect>.
- V. Zarnowitz and L. A. Lamnros. Consensus and uncertainty in economic prediction. *The Journal of Political Economy*, 95(3):591–621, June 1987.