

Analysis of Default Data Using Hidden Markov Models

GIACOMO GIAMPIERI, MARK DAVIS AND MARTIN CROWDER

Department of Mathematics, Imperial College, London SW7 2AZ

Abstract. The occurrence of defaults within a bond portfolio is modeled as a simple hidden Markov process. The hidden variable represents the risk state, which is assumed to be common to all bonds within one particular sector and region. After describing the model and recalling the basic properties of hidden Markov chains, we show how to apply the model to a simulated sequence of default events. Then, we consider a real scenario, with default events taken from a large database provided by Standard & Poor's. We are able to obtain estimates for the model parameters, and also to reconstruct the most likely sequence of the risk state. Finally, we address the issue of global vs. industry-specific risk factors. By extending our model to include independent hidden risk sequences, we can disentangle the risk associated with the business cycle from that specific to the individual sector.

1. Introduction

Interaction effects are a key component of portfolio credit risk, but how to quantify these effects in a credible way has been the subject of some controversy. For large portfolios of, say, $n = 50$ bonds, it is generally unfeasible to model the default risk of each individual issuer and the 'correlation' (however defined) with other issuers, since this leads to a high-dimensional model with an enormous number of parameters, which cannot be reliably estimated. Instead, one is looking for a simple description of the interaction process, justifiable on economic and/or empirical grounds, that is characterized by a small number of parameters. A number of such models has been proposed and some of them are in widespread industrial use. A good example is Moody's 'binomial expansion techniques' (BET) (see [1] or §11.3 of [2]), or Vasicek's 'large homogenous portfolio' approximation [3]. In the BET, the original portfolio of size n is replaced by a smaller portfolio of size $n' < n$, the members of which are supposed to default independently, leading to a binomial distribution for the number of defaults over a fixed time horizon. The number n' is determined by a 'diversity score' analysis in which issuers in different industry sectors are deemed independent while those in the same sector are coupled in a quantified way. The bottom line is that portfolios with low diversity have greater tail risk. A similar effect was obtained by Davis and Lo [4] in an infection model which assumes that a defaulting bond may trigger off defaults in other bonds. The model only has two parameters, an individual default probability p and an infection parameter q .

As the latter is increased, default distributions very similar to the Moody's model are produced.

These models are *static* in that they only concern the total number of defaults in a specified period. For applications such as CDOs (collateralized debt obligations) the *timing* of defaults is as important as the total number, and one needs a dynamic – i.e. stochastic process – model. In [5], Davis and Lo define the so-called *enhanced risk* model as a dynamic version of infectious defaults. The portfolio is assumed to be in one of two states: normal risk and enhanced risk. It starts in normal risk, but as soon as a default occurs it moves to enhanced risk, where the hazard rates for all remaining issuers are multiplied by an enhancement factor $\kappa > 1$. The portfolio stays in the enhanced risk state for an exponentially-distributed random time before dropping back to normal risk. The two states can be thought of as a general ‘good times/bad times’ economic variable. This interpretation is the one examined below in this paper. A somewhat similar approach, in which default of ‘primary issuers’ affects the hazard rate of ‘secondary issuers’, has been taken by Jarrow and Yu [6].

A very common idea, which one sees in, for example, the Credit Metrics model [7] or the CDO model in §11.3 of [2], is to suppose that each issuer is exposed to three default risk factors: a business cycle factor affecting all issuers, an industry-specific factor affecting only firms in the same industry sector, and an idiosyncratic factor specific to the issuer itself. One of the purposes of this research is to determine whether these factors are supported by the data.

Turning to the work described below, we consider a simplified enhanced risk model along the lines of Davis and Lo [5] with two, not directly observed, states corresponding to normal and enhanced risk. There is no ‘infection’ effect: we suppose that the hidden variable is a two-state Markov process in discrete time (time is quantised into quarterly intervals), not depending on the default events. Within each time period defaults are supposed to be binomially distributed, with higher mean in the enhanced risk state. This is a ‘hidden Markov model’, for which theory and estimation algorithms (mainly developed for signal processing applications) are available [8],[9]. We estimate model parameters and most likely paths for the hidden state using a large database of default histories made available to us by Standard & Poor's. Section 2 below describes the model, while Section 3 describes the estimation techniques and gives parameter estimates and most likely hidden Markov process sample paths for data drawn from four different industry sectors. Error estimates for the parameters can be obtained by a bootstrapping technique described in Section 5. As a further diagnostic test, we investigate in Section 6 whether the model prediction of binomial default distribution within each risk category is supported by the data. Section 7 is a preliminary examination of the influence of global versus industry-specific economic factors. We now aggregate the data from all the sectors previously considered and re-estimate to find the most likely path for a global economic factor. It seems that the global factor probably accounts for most of the interaction between issuers, but that one could possibly distinguish secondary effects related to industry sectors. Concluding

remarks and suggestions for further work are given in the final section, Section 8.

2. Description of the model

A discrete state, discrete time hidden Markov model (HMM) consists of a set of n nodes, each of which is associated with an observed quantity. The current state is not observable, but produces an output with a certain probability distribution. The parameters of the model are an initial state π which describes the distribution over the initial node, a transition matrix a_{ij} for the transition probability from node i to node j conditional on node i , and an observation matrix $b_i(m)$ for the probability of observing m conditional on node i .

More specifically, in our case the hidden state is associated with the risk state, which can take two values: 0 (normal risk), and 1 (enhanced risk). In the normal risk state, the number m of observed defaults in each time step is binomially distributed, with parameter λ :

$$p_0(m) = \binom{N_s}{m} \lambda^m (1 - \lambda)^{(N_s - m)} \quad (1)$$

where N_s is the number of surviving bonds at time s . In the enhanced risk state, the p.d.f. $p_1(m)$ is still given by eq. (1), after multiplying λ by a factor $\kappa \geq 1$. In the limiting case $\kappa = 1$, the two hidden states become equivalent, i.e., $p_0(m) = p_1(m)$. The transition matrix a_{ij} is assumed to be constant, and is parametrized as

$$a_{ij} = \begin{pmatrix} q & 1 - q \\ 1 - p & p \end{pmatrix} \quad (2)$$

where q is the probability of remaining in the normal risk state, and p is the probability of remaining in the enhanced risk state. Thus, our model is fully described by four parameters: λ, κ, q, p . Note that, although a_{ij} is time independent, the model for the observed data turns out to be time dependent, since the number of defaults in each time step depends on the number of surviving bonds at the beginning of the period (see eq. (1)).

3. Estimation of parameters

Given the model and the observation sequence, the model parameters can be estimated with standard Maximum Likelihood techniques. In particular, an algorithm developed by Baum and Welch for signal processing applications (see, e.g., [9]) can be applied. Implementation of the Baum-Welch algorithm involves computation of two different probability terms. First, the forward path probability $\alpha_t(i) = P(m_1 m_2 \dots m_t, i)$ is defined as the joint probability of having generated a partial observation sequence in the forward direction (i.e., from the start of the sequence) and having arrived at a certain hidden state i at time t . Next, the backward path probability $\beta_t(i) = P(m_{t+1} \dots m_T | i)$ denotes the probability of generating a partial observation sequence in the reverse

direction (from the final time T), given that the state sequence starts from a certain hidden state i at time t . In addition, the probability $\gamma_t(i) = P(i|\mathcal{M})$ of being in a given hidden state i at time t , given the whole observation sequence $\mathcal{M} = m_1, \dots, m_T$, can be expressed in terms of the forward-backward variables as

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=0}^1 \alpha_t(j)\beta_t(j)} \quad (3)$$

Given an observed time series of default events, the parameters of the HMMs, as well as the most likely hidden risk sequence, can be estimated by using these probabilities. Full details of the Baum-Welch procedure for parameter estimation, as well as the various implementation issues, are described in [9].

First, we apply the Baum-Welch algorithm for estimating the parameters of the hidden Markov model, using a simulated data set. We have started with $N=1000$ bonds, and simulated a sequence of defaults events over a time period of 20 years, using a time step of 90 days. The parameters chosen in this simulation are $\lambda = 0.004, \kappa = 5, q = p = 0.9$, meaning that in each of the 90-day periods, 0.4% of surviving bonds will, on average, default, and five times that many if we are in the enhanced risk state. The probability of jumping from one risk state to the other is 10% during each time step. We stress that the Baum-Welch algorithm only finds local maxima of the probability function, so that the starting model has to be chosen with care. In our case, we start from the initial guess $\lambda = 0.001, \kappa = 2, q = p = 0.5$, and after only five iterations we find $\lambda = 0.0038, \kappa = 4.6727, q = 0.9075, p = 0.9043$, in excellent agreement with the true values. The simulated time sequence, along with the true and estimated hidden sequences, is shown in Fig. 1.

As one may expect, the number of iterations increases, and the final accuracy decreases, as κ gets smaller. For example, decreasing κ from 5 to 3 (keeping all other parameters the same) produces the following output: $\lambda = 0.0046, \kappa = 2.5733, q = 0.8353, p = 0.8740$. We have also verified that in the degenerate case $\kappa = 1$, the transition probabilities q and p become meaningless.

4. Results from S&P's database

The Standard & Poor's CreditPro 6.2 database provides the history of 9928 bonds belonging to 13 industry sectors. Subsector and country are also specified, along with the date of first rating, and all rating transitions from 1/1/81 to 31/12/02. We have only considered US issues in four sectors: Consumer, Energy, Media, and Transport. In particular, we have extracted the default times of bonds in each of the considered US sector. Since the S&P's database provides also the time when each bond was first rated, the HMM model fitted is conditional on these observed times. We have grouped the default events in quarterly periods, and applied the estimation algorithm to the resulting sequence.

Table 1 shows, for each considered sector, the total number of bonds issued over the whole period, the number of defaulted bonds, and the parameters obtained from the

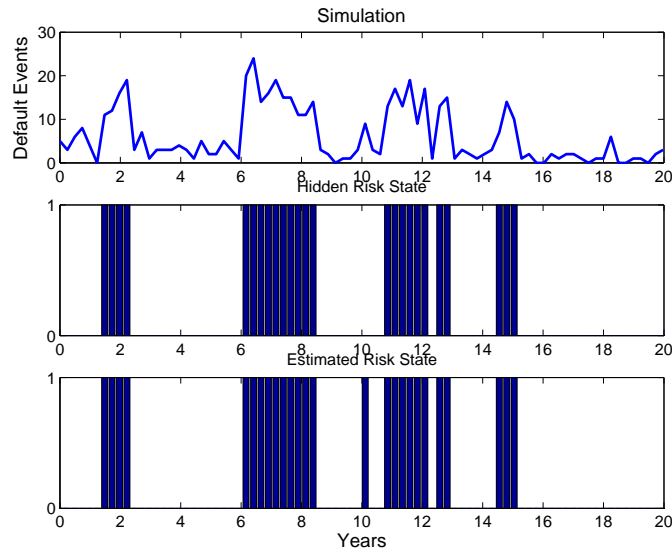


Figure 1. Simulated data. The top graph gives the number of defaults as function of time. The solid bars in the middle plot indicate the enhanced risk periods. The bottom plot shows the reconstruction of the risk state, which agrees very well with the true distribution.

forward-backward procedure. Note that the intensity λ varies by as much as a factor two between different sectors, whereas κ , q , and p remain rather consistent. Also, p is always smaller than q , meaning that the hidden risk is more likely to move from high to low than viceversa, at each time step.

Figures 2-5 show the time sequence of default events, along with the estimated risk-state sequence, for each of the four sectors considered. The optimal (i.e., most likely) hidden sequence is computed via the Viterbi algorithm [9]. Note how the risk state is correlated among different sectors, especially in the few final years, which raises the possibility of investigating a cross-sector infection effect. This issue will be addressed in the final section.

Table 1. Results from S&P’s database. For each industry, the table provides: total number of bonds issued in the observed period (N_{tot}), number of defaults (N_{def}), and the estimated parameters (λ, κ, q, p).

Sector	N_{tot}	N_{def}	λ	κ	q	p
consumer	1041	251	0.0026	6.1	0.95	0.81
energy	420	71	0.0014	7.1	0.95	0.88
media	650	133	0.0027	7.2	0.96	0.83
transport	281	59	0.0025	8.9	0.97	0.78

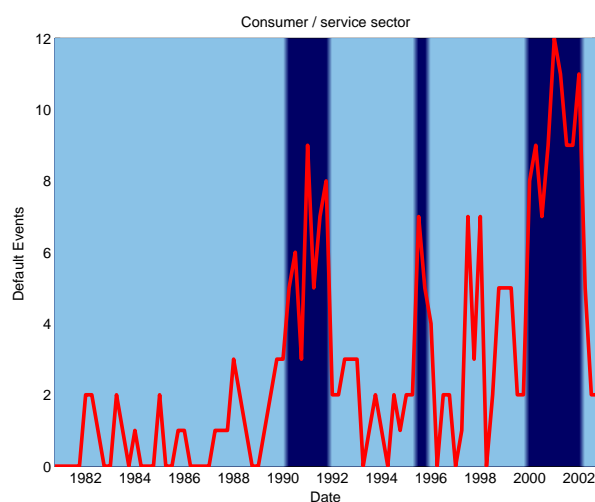


Figure 2. Hidden risk state and default history for the Consumer sector. The red line gives the number of defaults in each quarter, from 1/1981 to 12/2002. The background color gives the estimated risk state, varying from light blue (low risk) to dark blue (high risk).

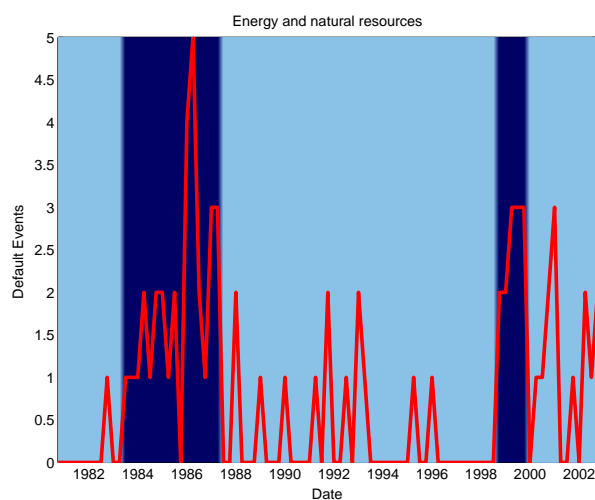


Figure 3. Same as figure 2, for the Energy sector.

5. Parametric bootstrap

The bootstrap procedure consists in choosing random samples generated using a particular set of parameters. The range of sample estimates thus obtained allows determining the uncertainty of the quantity we are estimating. In particular, using this technique we can estimate the covariance matrix for the four parameters found with the Maximum Likelihood estimator. For simplicity, we focus our attention to the one of the four cases considered in Table 1, namely to the US Consumer industry. The bootstrap technique requires simulating a large (e.g., $N = 100$) number of realization

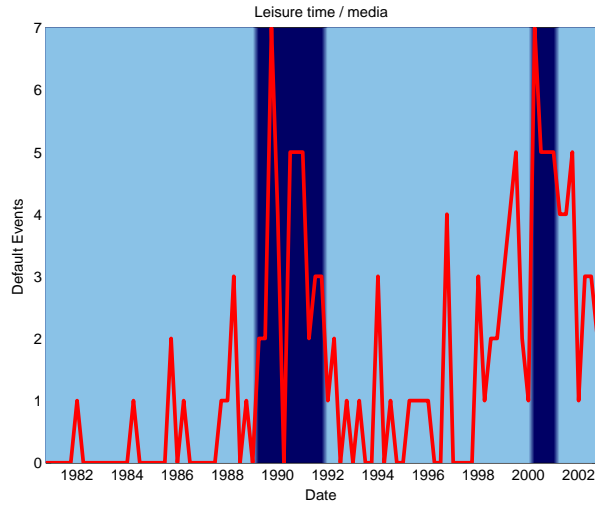


Figure 4. Same as figure 2, for the Media sector.

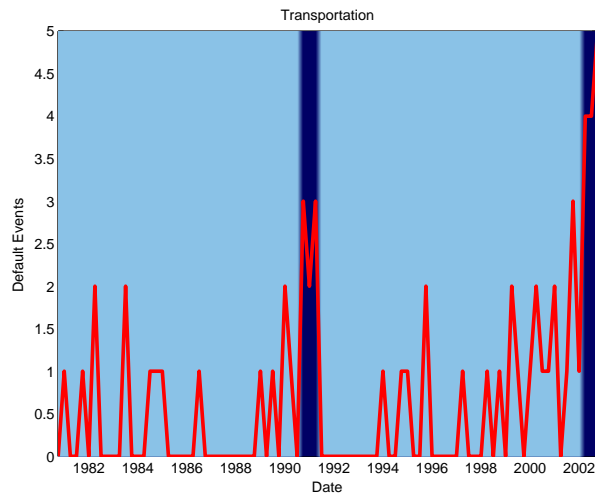


Figure 5. Same as figure 2, for the Transport sector.

of the fitted HMM, that is, the HMM with the parameters shown in the second line of Table 1. For each generated realization, we estimate the four parameters with the same Baum-Welch algorithm used above. Each result is stored in a vector θ_i , and the covariance matrix estimator is then given by

$$\mathbf{C} = \frac{1}{N-1} \sum_{i=1}^N (\theta_i - \hat{\theta})' \cdot (\theta_i - \hat{\theta}) \quad (4)$$

where

$$\hat{\theta} \equiv \frac{1}{N} \sum_{i=1}^N \theta_i \quad (5)$$

For the US consumer case we find that the mean $\hat{\theta} = (0.0027, 6.8, 0.95, 0.77)$ coincides with good approximation with the reference value, while the covariance matrix is given

by

$$\mathbf{C} = \begin{pmatrix} 8.8\text{E-}8 & -2.0\text{E-}4 & 1.9\text{E-}6 & 2.4\text{E-}6 \\ -2.0\text{E-}4 & 2.4\text{E}0 & 9.4\text{E-}3 & 3.5\text{E-}2 \\ 1.9\text{E-}6 & 9.4\text{E-}3 & 1.8\text{E-}3 & -6.2\text{E-}4 \\ 2.4\text{E-}6 & 3.5\text{E-}2 & -6.2\text{E-}4 & 3.7\text{E-}2 \end{pmatrix} \quad (6)$$

Assuming a Gaussian distribution around the mean, the square root of each diagonal component gives the standard deviation of the parameter, thus

$$\lambda = 0.0019 \pm 0.0003$$

$$\kappa = 6.2 \pm 1.5$$

$$q = 0.93 \pm 0.04$$

$$p = 0.80 \pm 0.19$$

As seen in §2, in the degenerate case $\kappa = 1$, the hidden risk state has no effect on the default probability. However, our analysis implies that κ is significantly greater than 1. We can therefore conclude that there is indeed a hidden variable which determines the level of risk for each bond in the sector, in agreement with our model's hypothesis.

In addition, the bootstrap procedure provides the distribution of the parameters θ_i around the mean $\hat{\theta}$. Fig. 6 shows the four histograms corresponding to each of the four estimated parameters. The distribution of λ and κ is reasonably close to being normal, although that of λ has fatter tails. Note that q and p , being bounded between 0 and 1, cannot be, strictly speaking, normally distributed. In the Consumer sector under consideration, p has a larger uncertainty than q , as it results from the covariance matrix 6. Thus, in a few cases, the model estimates a value of p which is rather small, as can be seen in the relative frequency histogram shown in Fig. 6.

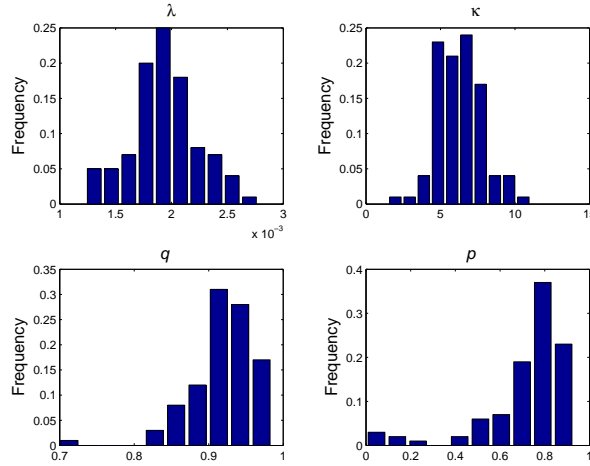


Figure 6. Frequency distribution of estimated parameters produced by the parametric bootstrap technique, for the Consumer sector. The reference values are from Table 1.

6. Goodness-of-fit of the binomial distribution

As explained in §2, our model assumes that, at each time step, the number m of defaults is distributed according to the binomial p.d.f. $p_0(m)$ or $p_1(m)$, depending on the risk state. We can now check whether this is verified in reality. In doing so, we must remember that the parameters of the binomial distribution (for each of the two risk states) are not constant, because of the fact that N_s varies with time. Thus, when we look at the distribution of the number of defaults, conditioned on being in risk state 0 and 1 respectively, we do not expect to find a pure binomial curve. We have considered the enhanced risk periods for four of the US sectors considered in §4 (solid blue regions in figs. 2-5). Since N_s is far from being constant, one should expect a deviation from a purely binomial curve. The distributions found using the actual data (fig. 7) are not in contradiction with the expected ones, indicating that the binomial distribution is indeed a viable assumption.

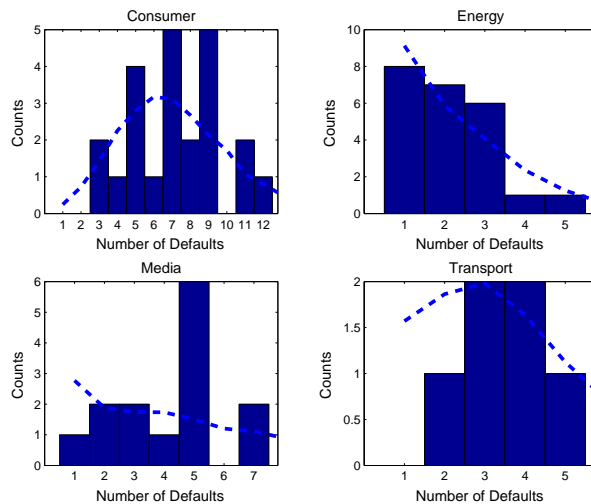


Figure 7. Histogram of the number of defaults during the enhanced risk quarters in the four sectors: Consumer, Energy, Media, and Transport. The dotted line in each plot is the expected curve for a HMM, with the estimated parameters obtained from Table 1.

7. Global vs. industry specific effects

The main motivation behind our work is to model the risk factor affecting all individual firms within a specific sector. However, in some cases, the enhanced risk could be related to global economic factors, affecting all sectors at the same time, as noted at the end of §4. We have thus applied our model to the whole database (US issuers only), without distinction for industry type. The total number of bonds considered is 6775, of which 1013 defaulted during the considered period. The Baum-Welch procedure, applied to the whole set of data, produced the following parameters:

$\lambda = 0.0017, \kappa = 4.9, q = 0.94, p = 0.85$. The default sequence, along with the estimated hidden state, is shown in fig. 8. Also shown are the historical recession periods of the business cycle [10]. Note how the high risk state overlaps with the recession periods during the two most serious recessions in the last 20 years, namely the 7/1990-3/1991 and the 3/2001-11/2001 contractions of the business cycle. Interestingly, in these two notable cases the high risk state anticipates the onset of recession, and hang on for a few more months after the end of the recession.

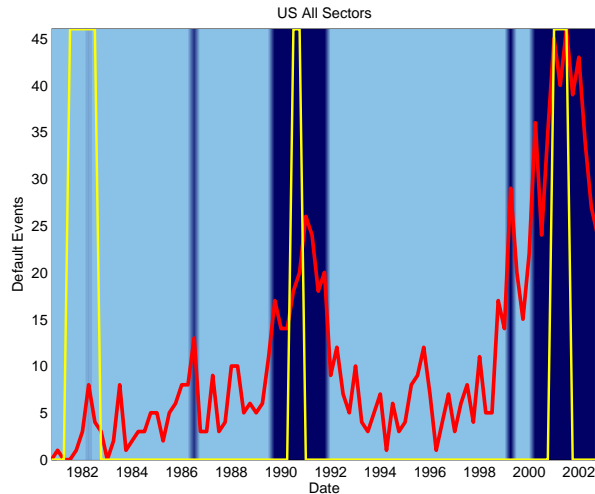


Figure 8. Same as figs. 2-5, with all available data. Recession periods (indicated in yellow) are from NBER [10].

Note also that the presence of global enhanced risk periods, especially the one at the end of the observing time, could have easily been guessed by overlapping the analogous plots for the individual sectors (shown in figs. 2-5). However, some of the enhanced risk periods shown in figs. 2-5 do not have a counterpart in fig. 8. This suggests that one may be able to disentangle, a posteriori, the default risks associated with the global economy from those specific to a particular sector. In order to do so, one can modify the HMM in the following way. We first take as given the globally-estimated sequence of hidden states describing the global risk, resulting from the above analysis. Then we postulate that the intensity for a specific sector j is increased by a factor κ'_j (different from κ_j) during the enhanced global risk periods. Thus, the intensity for sector j can now take four different values, instead of just two (we omit the subscript j henceforth): λ during normal periods, $\kappa\lambda$ during periods of high-risk specific to the sector under consideration, $\kappa'\lambda$ during the periods of enhanced global risk, and finally $\kappa\kappa'\lambda$ when both the sector-specific and global risks are enhanced. In this way, we hope to be able to determine a hidden sequence for the industry, which has an additional effect to that of the global economy. The parameters κ and κ' determine the enhancement factor for the intensity associated with each of the two risk factors, respectively.

As an example, we have considered the US Consumer sector. The Baum-Welch

algorithm applied to the modified model gives the following values for the parameters: $\lambda = 0.0025$, $\kappa = 5.2$, $q = 0.95$, $p = 0.22$, $\kappa' = 4.5$. As shown in Fig. 9, the hidden state associated with the industry-specific risk factor turns out to be on state 0 (low risk) during the the global high-risk periods (indicated by the green line). In other words, the increased number of defaults during the 1990 and 2001 recessions can be ascribed to global risk factors, without the need of adding idiosyncratic effects. However, there are short high-risk periods between 1996 and 2000 which are specific to this sector, and which are not related to the business cycle. Note that the short global high-risk period in 1986 (which is not associated with a recession, see fig. 8) does not cause an increase in default risk for the consumer sector. The fact that κ and κ' are almost identical can be interpreted as an indication that the increase in default risk is similar for the two underlying causes.

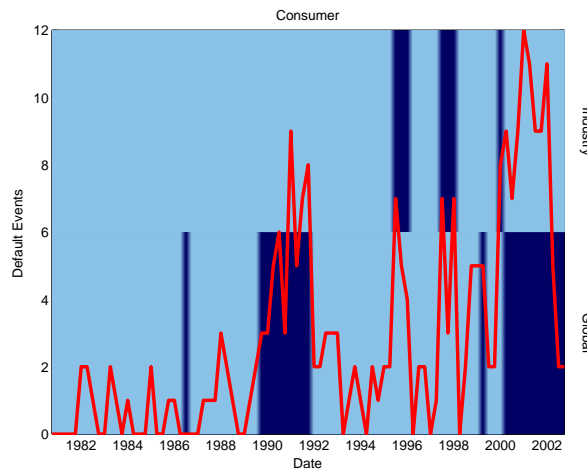


Figure 9. Hidden risk state for the Consumer sector, after subtracting global risk factors. The solid blue regions in the bottom half indicate the enhanced risk periods for the global economy (from fig. 8). The remaining industry specific risk is shown in the top half. The red line gives the number of defaults, as in fig. 2.

A very different scenario occurs in the Energy sector. As can be seen by comparing figs. 3 and 8, this sector seems to be highly uncorrelated with the global cycle, at least during the most serious recessions. Indeed, our new model, when applied to this sector, gives: $\lambda = 0.0014$, $\kappa = 7.2$, $q = 0.95$, $p = 0.88$, $\kappa' = 1.0$. The first four parameters are essentially the same as in Table 1, whereas κ' being almost exactly 1 indicates that global risk periods do not increase the chances of default for energy firms. Since global risk is found not to affect the default intensity for this particular sector, the peaks in the number of defaults that occurred in 1986 and 1999 are associated by our model to an industry-specific risk, despite the fact that the global risk turns out to be in an enhanced state as well. Note that the bottom sequence in fig. 10 is irrelevant (since $k' = 1$), whereas the top one is identical to that shown in fig. 3.

One final remark concerns the effective independence of the two hidden risk states. Since we have obtained the risk for the global economy from the whole database, it is

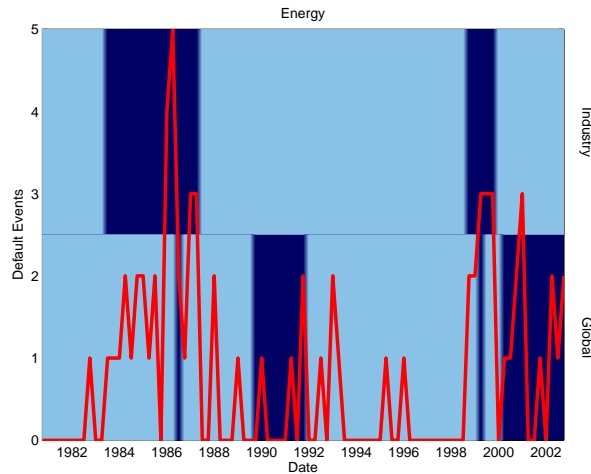


Figure 10. Same as fig. 9, for the Energy sector.

obvious that, in principle, the two risk states cannot be fully independent. In other words, each sector contributes to the global economy. However, we have relied on the assumption that none of the 13 sectors dominates, in terms of population (the larger sector containing $\sim 15\%$ of the total number of bonds) and number of defaults. In addition, our estimated global risk sequence is almost coincident with the business cycle (see Fig. 8), which is totally independent of the S&P's database. If we had used the yellow curve in Fig. 8 as our global risk sequence, instead of the estimated hidden sequence, the results in this section would have not changed appreciably.

8. Concluding remarks

The model we have introduced is certainly very simple, but our empirical analysis shows that it has good explanatory power. We have already mentioned several extensions that could be pursued, for example explicit inclusion of separate industry-specific and general economic variables. One could also consider a hidden process with more than two states, though an interpretation of this might be problematic.

One area that we have so far ignored entirely is changes of rating. We have only used information about realized default events, but in fact the database contains far more. A more sophisticated model might posit interaction effects in rating changes as well, something more in the spirit of the Credit Metrics approach [7].

Acknowledgments

We are very grateful to Standard & Poor's, and particularly to our colleagues there Arnaud de Servigny and Olivier Renault, for allowing us access to S&P's database of default history. We also thank Jonathan Staples for helpful discussions. The work of Giacomo Giampieri was supported by the UK EPSRC under Grant GR/R80131/01.

References

- [1] Moody's KMV Company 1997 *The Binomial Expansion Technique* Technical document
<http://www.moodyskmv.com>
- [2] Duffie D and Singleton K J 2003 *Credit Risk: Pricing, Measurement and Management* (Princeton: Princeton University Press)
- [3] Schönbucher P 2003 *Credit Derivatives Pricing Models* (Chichester: Wiley Finance)
- [4] Davis M and Lo V 2001 Infectious defaults *Quant. Finance* **1** 382-386
- [5] Davis M and Lo V 2001 Modelling default correlation in bond portfolios, in *Mastering Risk Volume 2: Applications* ed Carol Alexander (Financial Times Prentice Hall) pp 141-151
- [6] Jarrow R and Yu F 2001 Counterparty risk and the pricing of defaultable securities *J. Finance* **56** 1765-1799
- [7] Risk Metrics Group 1997 *Credit Metrics Technical Document*
<http://www.riskmetrics.com/cmtdovv.html>
- [8] MacDonald I L and Zucchini W 1997 *Hidden Markov and Other Models for Discrete-valued Time Series* (London: Chapman and Hall)
- [9] Rabiner L R 1989 A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition *Proc. IEEE* **77** 257-286
- [10] National Bureau of Economic Research, *US Business Cycle Expansions and Contractions*
<http://www.nber.org/cycles/cyclesmain.html>