Forecasting Cross-Sections of Frailty-Correlated Default *

Siem Jan Koopman[(a,c)] *Andre Lucas*[(b,c)] *Bernd Schwaab*[(b,c)]

[(a)] Department of Econometrics, VU University Amsterdam

[(b)] Department of Finance, VU University Amsterdam

[(c)] Tinbergen Institute

20 February 2008

Abstract

We propose a novel econometric model for estimating and forecasting cross-sections of time-varying conditional default probabilities. The model captures the systematic variation in corporate default counts across e.g. rating and industry groups by using dynamic factors from a large panel of selected macroeconomic and financial data as well as common unobserved risk factors. All factors are statistically and economically significant and together capture a large part of the time-variation in observed default rates. In this framework we improve the out-of-sample forecasting accuracy associated with conditional default probabilities by about 10-35% in terms of Mean Absolute Error, particularly in years of default stress.

Keywords: *Non-Gaussian Panel Data; Common Factors; Unobserved Components; Forecasting Conditional Default Probabilities*

JEL classification: *C33; G21.*

*We are grateful to Peter Boswijk, Darrel Duffie and Michel van der Wel for their comments on different versions of this paper. We further would like to thank for the comments of participants in presentations at the BIS Research Task Force workshop 'Stress Testing of Credit Portfolios', Tinbergen Institute Amsterdam, the NAKE 2007 day in Utrecht, and VU University Amsterdam. All errors are our own.

1 Introduction

Modeling dependence between default events is considered to be one of the major challenges in modern credit risk management. To understand and price the risk of a loan portfolio it is necessary to have reliable estimates of current default probabilities and default correlations for the obligors in the portfolio. Default probabilities may depend on firm specific information as well as the general macroeconomic conditions, see *inter alia* the recent papers by Das et al. (2007), Duffie et al. (2007), Pesaran et al. (2006), and Figlewski et al. (2006).

In this paper we develop a model targeted towards estimation and out-of-sample forecasting of conditional default probabilities. We include a very large array of selected macro variables by focusing on what they have 'in common'. In effect, the proposed model combines the non-Gaussian panel data approach of Koopman and Lucas (2008) with the main features of Stock and Watson's (2002a) approximate dynamic factor model. To our knowledge, this article is the first to nest these two strands of literature on high-dimensional multivariate time series modeling. As a result, the final model accommodates common factors from observed data as well as unobserved dynamic factors. For ease of reference we will refer to our model as the Common Factor Panel (CFP) model.

While very popular, the Stock and Watson methodology is typically not applied outside of a linear regression framework. We show that principal components can be used in a nonlinear non-Gaussian model to address the important problem of estimating and forecasting time-varying default probabilities. The main novelty is the development of a framework in which default conditions depend on both unobserved components and common factors from a large set of selected macro and financial data.

Following Das et al. (2007), we refer to such a situation as 'frailty' correlated defaults. The task of estimating and forecasting conditional default probabilities is not standard when default conditions depend on unobserved serially correlated risk factors in addition to observed firm characteristics and macroeconomic variables. The econometric literature which can allow for unobserved risk factors is fairly recent. Most notably it includes Duffie et al. (2006), McNeil and Wendin (2007), Koopman et al. (2008), and Koopman and Lucas (2008). When default events depend on unobserved components, advanced econometric techniques based on simulation methods are required. For example, Duffie et al. (2006) employ a Sim-

ulated EM approach with Gibbs sampling, while Koopman et al. (2008) use importance sampling techniques derived for non-Gaussian state space models. The dependence on simulation methods is one reason why unobserved component models typically allow for only a limited number of observable macro variables alongside the unobserved factor.

This paper makes three contributions to the econometric credit risk literature. First, we show theoretically how a non-Gaussian panel data specification for default risk can be combined with an approximate dynamic factor model. The resulting model inherits the best of both worlds. Factor models readily permit the use of information from very large arrays of relevant predictor variables. The non-Gaussian panel structure in addition captures the cross-sectional heterogeneity of firms, allows for unobserved 'frailty' factors, and can easily accommodate missing values. The missing values arise easily if we consider default counts at a highly disaggregated level.

Second, we show that common factors from a panel of selected macroeconomic and financial variables capture a statistically and economically significant part of the time-variation in observed default rates. Thus, macroeconomic risk and systematic default risk conditions are closely linked. By decomposing overall default risk into a systematic and idiosyncratic part we follow the credit risk literature on latent variable models as given by Wilson (1998), Gordy (2000), and Lando (2003). For the computation of common macro factors we draw from the extensive and growing literature on large N, large T dynamic factor models, most notably Stock and Watson (2002a, 2002b, 2005), and Bai and Ng (2002, 2007).

Third, we show that common factors are useful for out-of sample forecasting of default risk conditions. In a forecasting experiment we find that adding common factors to an unobserved component specification improves forecasting accuracy. Feasible improvements are substantial, in particular in years of high default stress such as 2001. The extent of the improvements depend mainly on firm's rating classes and prevailing macro conditions. Improved forecasts of conditional default probabilities over a large cross-section of firms are relevant to credit risk management in financial institutions, banking supervision, asset management, and potentially for institutional investors in credit derivatives markets. The forecasted probabilities can be used as input for the calculation of one-year ahead Value-at-Risk levels as well as for stress testing loan portfolios.

We proceed as follows. In Section 2 we introduce the econometric framework of the Com-

mon Factor Panel model and show how the non-Gaussian panel and the approximate dynamic factor model are combined. In Section 3 we discuss the estimation of the model. Section 4 shows that there exists a one-on-one correspondence between the proposed econometric model and a multi-factor firm value model for dependent defaults. Section 5 introduces the two panel data sets used in this article, presents the empirical findings and the forecasting results. Section 6 concludes.

2 The econometric framework

In this section we present the full set of model equations. We denote the default counts of cross section j at time t as y_{jt}, where $j = 1, \ldots, J$, and $t = 1, \ldots, T$. The index j denotes a combination of firm characteristics, such as industry specification, current rating class, or company age. Defaults are assumed to be correlated in the cross-section through risk factors. We distinguish two different sets of risk factors, i.e., an unobserved factor f_t^{uc} and exogenous factors F_t which we construct from a large panel of macroeconomic and financial time series. The default counts are modeled as Binomially distributed after conditioning on these factors,

$$y_{jt}|f_t^{uc}, F_t \sim \text{Binomial}(k_{jt}, \Pi_{jt}), \tag{1}$$

where y_{jt} is the number of default 'successes' from k_{jt} independent Bernoulli-trials, each with probability Π_{jt}. In our case, k_{jt} denotes the number of firms in cell j that are active at the beginning of period t and can default with probability Π_{jt}. The conditional independence assumption is standard in the credit risk literature on latent variable models, see for instance the CreditMetrics (2007) framework as well as the textbook exposition of Lando (2003, Chapter 9).

The conditional default probabilities Π_{jt} are specified as the logistic transform of an index function θ_{jt},

$$\Pi_{jt} = (1 + e^{-\theta_{jt}})^{-1}, \tag{2}$$

$$\theta_{jt} = \lambda_j + \beta_j f_t^{uc} + \gamma_j' F_t, \tag{3}$$

where λ_j constitutes a fixed effect for each cross section, and coefficients β_j and γ_j capture risk factor sensitivities which may depend on firm characteristics such as industry specification or rating class. This specification is analogous to a standard logit model commonly used in micro-econometrics to model discrete events. Estimation and forecasting Π_{jt} is the main focus of this paper. The conditional default probabilities may vary over time due to either variation in the unobserved component, f_t^{uc}, or variation in the common factors F_t from a large set of macroeconomic and financial data.

The dynamics of the unobserved component f_t^{uc} are specified as a stationary autoregression of order 1,

$$f_t^{uc} = \phi f_{t-1}^{uc} + \sqrt{1-\phi^2}\eta_t, \quad \eta_t \sim \text{NID}(0,1), \tag{4}$$

where $0 < \phi < 1$. Other dynamic specifications for f_t^{uc} can also be considered. The autoregressive process is normalized such that $\text{E}[f_t^{uc}] = 0$, $\text{Var}[f_t^{uc}] = 1$, and $\text{Cov}[f_t^{uc}, f_{t-h}^{uc}] = \phi^h$. It follows that coefficient β_j can be interpreted as the standard deviation (volatility) of the unobserved factor f_t^{uc} for the firms of cross section j.

Finally, we collect a large number of macroeconomic and financial variables into a panel of time series x_{it} for $i = 1, \ldots, N$. This large array of macroeconomic predictor variables is assumed to contain information about economy-wide default risk conditions, and adhere to a factor structure such as

$$x_{it} = \Lambda_i F_t + e_{it}, \tag{5}$$

where F_t is a vector of factors, Λ_i is a row vector of factor loadings, and e_{it} is an idiosyncratic error term which satisfies the weak regularity conditions of Stock and Watson (2002b, Assumptions F1 and M1). Equation (5) gives the static representation of an approximate dynamic factor model, see Stock and Watson (2002a). Intuitively, (5) states that a large

part of the variation in macroeconomic and financial data may be traced back to only a few common factors. This idea has a long tradition in macro-econometrics, dating back to Sargent and Sims (1977) and Geweke (1977). The static representation (5) can be derived easily from a dynamic specification such as $x_{it} = \upsilon_i(L)f_t + e_{it}$ by assuming that the lag polynomials $\upsilon_i(L)$ operating on the factors f_t are of finite (low) order, see Stock and Watson (2002b). The coefficients in υ_i can be stacked in Λ_i, while the contemporaneous and lagged factors can be stacked in F_t. The estimated F_t represent current and lagged forces in the economy. This methodology has proven to be effective in forecasting inflation or industrial production, see Massimiliano, Stock, and Watson (2003).

The main advantage of the static representation (5) is that F_t can be estimated consistently using the method of principal components. This method is convenient for several reasons. First, dimensionality problems do not occur even for very large values of N and T. All computations remain tractable. Second, the method works under relatively weak assumptions. Finally, the obtained factors can be used directly for forecasting purposes. Equations (1) to (5) combine the approximate dynamic factor model with the non-Gaussian panel data model by inserting the elements of F_t from (5) into the signal equation (3). Statistical model formulation and estimation is discussed below.

3 Estimation and state space form

In this section we provide the details of the estimation of the parameters and factors in model (1) to (5). We first estimate the macro factors using the method of Stock and Watson (2002a) as discussed in Section 3.1. Next, we cast the complete model in state space form with the details provided in Section 3.2. We estimate the parameters using computationally efficient (Monte Carlo) Maximum Likelihood and Signal Extraction techniques based on Importance Sampling. A brief outline of the procedure is given in Section 3.3. We perform all computations using the Ox programming language and the associated set of state space routines from SsfPack, see Koopman et al. (1998), and Doornik (2002).

3.1 Estimation of the macro factors

The common factors F_t from the macro data are estimated by minimizing the objective function given by

$$\min_{\{F_1,\ldots,F_T,\Lambda\}} V(F,\Lambda) = (NT)^{-1}\sum_{t=1}^{T}(X_t - \Lambda F_t)'(X_t - \Lambda F_t), \tag{6}$$

where X_t is of dimension Nx1 and contains stationary macroeconomic variables. Concentrating out F_t and rearranging terms shows that (6) is equivalent to

$$\max \bar{V}(\Lambda) = \text{tr}\left(\Lambda'\left[\sum_{t=1}^{T} X_t X_t'\right]\Lambda\right) = T\ \text{tr}\left(\Lambda' S_{X'X}\Lambda\right) \tag{7}$$

subject to $\Lambda'\Lambda = I_r$, and where $S_{X'X} = T^{-1}\sum_t X_t X_t'$ denotes the covariance matrix of the data, see Stock and Watson (2002a). The principal components estimator of F_t is given by $\hat{F}_t = X_t'\hat{\Lambda}$, where $\hat{\Lambda}$ collects the normalized eigenvectors associated with the R largest eigenvalues of $S_{X'X}$.

In case variables are not completely observed, we employ the Expectation Maximization (EM) procedure as devised in the appendix to Stock and Watson (2002a). This iterative procedure takes a simple form under the assumption that $x_{it} \sim \text{NID}(\Lambda_i F_t, 1)$, where Λ_i denotes the ith row of Λ. In this case $V(F,\Lambda)$ from (6) is affine to the complete data log-likelihood $L(F,\Lambda|X)$, where X denotes the missing parts of the data. Since $V(F,\Lambda)$ is proportional to $-L(F,\Lambda|X)$, the minimizers of $V(F,\Lambda)$ are also the maximizers of $L(F,\Lambda|X)$.

The procedure for obtaining the principal components in case of missing data is as follows. The objective function (6) is given by

$$\min_{\{F_1,\ldots,F_T,\Lambda\}} V^*(F,\Lambda) = \sum_{i=1}^{N}\sum_{t=1}^{T} I_{it}(x_{it} - \Lambda_i F_t)^2, \tag{8}$$

where $I_{it} = 1$ if x_{it} is observed, and zero otherwise. Equation (8) is minimized iteratively, using the following two step EM algorithm:

1. For the Expectation-step, take as given $\hat{F}_t, \hat{\Lambda}$. In the first round we use the estimates

from the balanced panel as starting values. The complete panel is balanced as follows:

$$\hat{x}_{it}^{bal} = \begin{cases} x_{it} & \text{if } x_{it} \text{ is observed,} \\ \hat{\Lambda}_i \hat{F}_t & \text{if } x_{it} \text{ is missing.} \end{cases}$$

Thus, missing values are replaced by their expectations given the smaller set of observed data points, which we denote as X^*.

2. In the Maximization-step, $\hat{F}_t, \hat{\Lambda}$ are updated by performing the eigenvalue/-vector decomposition on the estimated covariance matrix of the balanced data, $S_{X'X}^{bal} = T^{-1}(\hat{X}_t^{bal\prime}\hat{X}_t^{bal})$. Since $V(F,\Lambda) \propto const - L(F,\Lambda|X)$, the pair $\hat{F}_t, \hat{\Lambda}$ also maximizes $\mathrm{E}_{\hat{F}_t,\hat{\Lambda}}[L(F,\Lambda)|X^*]$.

We iterate the two E/M steps until convergence has taken place. To formulate a stopping criterion, the objective function $V(F,\Lambda)$ can be computed as the squared Frobenius matrix norm of the TxN error matrix $E = \hat{X}^{bal} - \hat{F}\hat{\Lambda}'$, since $V(F,\Lambda) = (NT)^{-1}\mathrm{tr}(E'E)$. The iterations stop when the changes in the objective function become negligible, say smaller than 10^{-7}.

3.2 The Common Factor Panel model in state space form

In this subsection we formulate the model (1) to (4) in state space form where F_t is treated as given. In practise, F_t will be replaced by $\hat{F}_t$.

The conditionally Binomial log-density function of the model (1) is given by

$$\log p(y_{jt}|\Pi_{jt}) = y_{jt} \log\left(\frac{\Pi_{jt}}{1-\Pi_{jt}}\right) + k_{jt}\log(1-\Pi_{jt}) + \log\binom{k_{jt}}{y_{jt}}.$$

By substituting (2) for Π_{jt} we obtain the log-density in terms of the log-odds ration θ_{jt} as

$$\log p(y_{jt}|\theta_{jt}) = y_{jt}\theta_{jt} + k_{jt}\log(1+e^{\theta_{jt}}) + \log\binom{k_{jt}}{y_{jt}}. \tag{9}$$

The 'signal' is given by

$$\theta_{jt} = Z_{jt}\alpha_t,$$

where

$$Z_{jt} = (e_j', \quad F_t' \otimes e_j', \quad \beta_j),$$

and e_j denotes the jth column of the unit matrix of dimension j. The system matrices Z_{jt} are time-varying due to the inclusion of F_t.

The state equation is given in its general form as

$$\alpha_{t+1} = T_t \alpha_t + B_t \xi_t, \qquad \xi_t \sim \mathrm{N}(0, Q_t), \tag{10}$$

where $\alpha_t = (\lambda_1, \ldots, \lambda_J; \gamma_{1,1}, \ldots, \gamma_{R,J}, f_t^{uc})'$ collects the fixed effects λ_j, all macro factor sensitivities $\gamma_{r,j}$ as well as the unobserved component, and where R denotes the dimension of F_t. The initial elements of the state vector are set to zero with a diffuse prior distribution, except for f_t^{uc} whose prior is given by N(0,1). The state equation system matrices are given by

$$T_t = \mathrm{diag}(I, \phi), \quad B_t = \begin{bmatrix} 0 \\ \sqrt{1-\phi^2} \end{bmatrix}, \quad Q_t = 1.$$

Equations (9) and (10) form a non-Gaussian state space model as discussed in Durbin and Koopman (2001) part II, and Koopman and Lucas (2008). We note that equation (9) replaces the more familiar observation equation associated with a linear Gaussian model. In this formulation, most unknown coefficients are part of the state vector α_t and are estimated as part of the filtering and smoothing procedures described in Section 3.3. This increases the computational efficiency of our estimation procedure. The remaining parameters are collected in a coefficient vector $\psi = (\phi, \beta_1, \ldots, \beta_J)'$ and are estimated by the Monte Carlo Maximum Likelihood methods of Section 3.3.

3.3 Estimation for the Common Factor Panel model

Parameter estimation for a non-Gaussian model in state space form proceeds in two steps. First, the coefficients in ψ are estimated by Monte Carlo maximum likelihood. Second, we obtain conditional mean and variance estimates of the state vector α_t. Both steps make use of importance sampling.

In the presentation of Monte Carlo maximum likelihood estimation, we suppress the

dependence of the density $p(y;\psi)$ on ψ and express the likelihood as

$$\begin{aligned} p(y) &= \int p(y,\theta)d\theta = \int p(y|\theta)p(\theta)d\theta \\ &= \int p(y|\theta)\frac{p(\theta)}{g(\theta|y)}g(\theta|y)d\theta = \mathrm{E}_g\left[p(y|\theta)\frac{p(\theta)}{g(\theta|y)}\right], \end{aligned} \tag{11}$$

where $y = (y_{11}, y_{21}, \ldots, y_{JT})'$, $\theta = (\theta_{11}, \theta_{21}, \ldots, \theta_{JT})'$, $p(\cdot)$ is a density function, $p(\cdot,\cdot)$ is a joint density, $p(\cdot|\cdot)$ is a conditional density, $g(\theta|y)$ is a Gaussian importance density, E_g denotes expectations with respect to $g(\theta|y)$, and

$$p(y|\theta) = \prod_{t,j} p(y_{jt}|\theta_{jt}).$$

Using Bayes' identity $g(\theta|y)g(y) \equiv g(y|\theta)g(\theta)$, where $g(y)$ denotes the likelihood associated with an approximating linear Gaussian model, (11) can be rewritten as

$$\begin{aligned} p(y) &= \mathrm{E}_g\left[p(y|\theta)\frac{p(\theta)}{g(y|\theta)}\frac{g(y)}{p(\theta)}\right] \\ &= \mathrm{E}_g\left[g(y)\frac{p(y|\theta)}{g(y|\theta)}\right] = g(y)\mathrm{E}_g\left[w(y,\theta)\right], \end{aligned}$$

where $w(y,\theta) = p(y|\theta)/g(y|\theta)$. The Monte Carlo likelihood is thus estimated as

$$\hat{p}(y) = g(y)\bar{w},$$

where

$$\bar{w} = M^{-1}\sum_{m=1}^{M} w^m = M^{-1}\sum_{m=1}^{M}\frac{p(y|\theta^m)}{g(y|\theta^m)},$$

where θ^m is a draw of θ from $g(\theta|y)$, and M is the number of importance draws of θ. The simulated draws are obtained using the simulation smoothing algorithm of Durbin and Koopman (2002). We estimate the log-likelihood as $\log\hat{p}(y) = \log\hat{g}(y) + \log\bar{w}$, and include the bias correction term discussed in Durbin and Koopman (1997).

The approximating Gaussian model is found by matching the first and second derivative of $\log p(y|\theta)$ and $\log g(y|\theta)$ with respect to the signal θ. This matching takes place around a current guess of the mode of θ. The next guess of the mode is then obtained as the smoothed

estimate of θ from a linear model which relates y and θ. Iterations proceed until convergence to the final approximating linear Gaussian model is achieved, which usually occurs in less than 10 iterations. A new approximating model is constructed for each trial evaluation of $\log p(y)$ for a different value of parameter vector ψ.

Standard errors for the parameters in ψ are constructed from the numerical second derivatives of the log-likelihood,

$$\hat{\Sigma} = \left[-\frac{\partial^2 \log p(y)}{\partial \psi \partial \psi'} \right]^{-1} \Bigg|_{\psi=\hat{\psi}}.$$

For signal extraction, we require the estimation of the conditional mean of an arbitrary function of θ, say $x(\theta)$, as given by

$$\begin{aligned} \bar{x} &= \mathrm{E}\left[x(\theta)|y\right] = \int x(\theta) p(\theta|y) d\theta \\ &= \int x(\theta) \frac{p(\theta|y)}{g(\theta|y)} g(\theta|y) d\theta = \mathrm{E}_g \left[x(\theta) \frac{p(\theta|y)}{g(\theta|y)} \right]. \end{aligned}$$

Using Bayes' identities and the fact that $p(\theta) = g(\theta)$ we obtain

$$\bar{x} = \frac{\mathrm{E}_g \left[x(\theta) w(\theta, y)\right]}{\mathrm{E}_g \left[w(\theta, y)\right]},$$

where $w(\theta, y)$ are the importance sampling weights as defined above, see also Durbin and Koopman (2001), p. 190.

Given these results, we estimate the conditional mean as

$$\hat{\theta} = \mathrm{E}[\theta|y] = \left[\sum_{m=1}^{M} w^m \right]^{-1} \sum_{m=1}^{M} \theta^m w^m,$$

where $w^m = w(\theta^m, y)$ denotes the importance weight associated with the m-th draw θ^m from $g(\theta|y)$. The associated conditional variances are given by

$$\mathrm{Var}[\theta_{it}|y] = \left(\left[\sum_{m=1}^{M} w^m \right]^{-1} \sum_{m=1}^{M} (\theta_{it}^m)^2 w^m \right) - \left(\hat{\theta}_{it} \right)^2.$$

4 The financial framework

In this section we discuss the connection between the above econometric model and a multi-factor firm value model for dependent defaults, see e.g. Tasche (2006) and Lando (2003, Chapter 9). The financial framework of the firm value model gives economic meaning to the statistical estimates and clarifies the economic mechanism at work. Single- and multi-factor models for firm default risk are widely used in risk management practice.

In a standard static one-factor credit risk model for dependent defaults the values of the obligors' assets, V_i, are usually driven by a common, standard normally distributed factor Y, and an idiosyncratic standard normal noise term ϵ_i, $i = 1, \ldots, I$,

$$V_i = \sqrt{\rho} Y + \sqrt{1-\rho}\epsilon_i.$$

A dynamic version of the single-factor specification would specify how V_i varies over time. Since we would in addition also like to allow for multiple factors, we generalize the model to

$$\begin{aligned} V_{it} &= \delta_{0i} f_t^{uc} + \delta_{1i} F_{1,t} + \cdots + \delta_{Ri} F_{R,t} + \sqrt{1 - (\delta_{0i})^2 - (\delta_{1i})^2 - \cdots - (\delta_{Ri})^2}\epsilon_{it} \\ &= \delta_i' f_t + \sqrt{1 - \delta_i'\delta_i}\epsilon_{it}, \end{aligned} \tag{12}$$

where $f_t := (f_t^{uc}, F_{1,t}, \ldots, F_{R,t})'$, and $\delta_i := (\delta_{0i}, \delta_{1i}, \ldots, \delta_{Ri})'$. In the remainder we assume that the δ_i parameters are common to all firms with characteristic j, and denote this vector δ_j.

The $F_{1,t}, \ldots, F_{R,t}$ are by construction uncorrelated principal components. The unobserved component f_t^{uc} serves to pick up credit cycle conditions which are not captured by the first R macro factors. We thus proceed by assuming, for this section only, that all factors in the model are unconditionally uncorrelated and normally distributed, such that $f_t = (f_t^{uc}, F_{1,t}, \ldots, F_{R,t})' \sim \mathrm{N}(0, I_{R+1})$. This in turn implies that $\mathrm{E}[V_{it}] = 0$ and $\mathrm{Var}[V_{it}] = 1$, regardless of the assumed distribution for the idiosyncratic noise component ϵ_{it}.

Following Merton's (1974) firm value-model, we assume that a default occurs as soon as a firm's net asset value V_{it} drops below a specified default barrier, say c_j. This default barrier may depend on the current rating class, industry specification, or time from initial rating

assignment. With these assumptions a default of firm i with firm characteristic j occurs as soon as

$$\begin{aligned} V_{it} < c_j \quad &\Leftrightarrow \quad \delta_j' f_t + \sqrt{1-\delta_j'\delta_j}\epsilon_{it} < c_j \\ &\Leftrightarrow \quad \epsilon_{it} < \left(\frac{c_j - \delta_j' f_t}{\sqrt{1-\delta_j'\delta_j}} \right). \end{aligned}$$

Denoting information up to time t as $\mathcal{F}_t$ we obtain,

$$\Pi_{jt} = \Pr\left(\epsilon_{it} < \frac{c_j - \delta_j' f_t}{\sqrt{1-\delta_j'\delta_j}} \middle| \mathcal{F}_t \right) = F_\epsilon \left(\frac{c_j - \delta_j' f_t}{\sqrt{1-\delta_j'\delta_j}} \right), \tag{13}$$

where F_ϵ denotes the cumulative distribution function for ϵ_{it}.

Equation (13) is intuitive. Good credit cycle conditions, i.e. high values of f_t are associated with low default probabilities Π_{jt}. The choice of F_ϵ as logistic allows to express the structural parameters of the firm value model from (13) in terms of the coefficients from the econometric specification. Specifically,

$$\begin{aligned} c_j &= \lambda_j \sqrt{1-a_j}, \\ \delta_{0,j} &= -\beta_j \sqrt{1-a_j}, \\ \delta_{r,j} &= -\gamma_{r,j} \sqrt{1-a_j}, \end{aligned}$$

where

$$a_j = \frac{\beta_j^2 + \gamma_{1,j}^2 + \gamma_{2,j}^2}{1 + \beta_j^2 + \gamma_{1,j}^2 + \gamma_{2,j}^2}.$$

5 Estimation results and Forecasting Accuracy

5.1 The Data: Macro Variables and Default Counts

We estimate the CFP model using data from two main sources. First, a large panel of time series is constructed from the Federal Reserve Economic Database FRED.[1] In total 120 vari-

[1] http://research.stlouisfed.org/fred2

ables are selected from about 3000 available US variables in the complete database. The goal is to select series which contain information about systematic credit risk conditions. The variables are sorted into five broad categories, see Table 1. These are (1) bank lending conditions and the extend of problematic loans, (2) macroeconomic and business cycle indicators, including measurements of general economic activity, labor market conditions and monetary policy instruments, (3) Open Economy macroeconomic indicators from the balance of payments and terms of trade, (4) Micro-level business conditions such as wage rates, cost of capital and cost of resources, and finally (5) stock market returns and volatilities. Thus, the panel contains both current information indicators (such as real GDP, unemployment rate, new orders, etc.) as well as forward looking variables (such as stock prices, interest rates, inflation expectations, etc.). As is common in factor analysis, each variable from this panel is transformed to covariance stationarity, either by (log-)differencing the original series once or twice, as appropriate, or by alternatively employing a suitable filter to remove the stochastic trend. Each series is standardized to zero mean and unit variance. We remove outliers by winsorizing the stationary series. This implies that observations larger than 3.5 in absolute value are adjusted to either 3.5 or -3.5.

[insert Table 1 around here]

[insert Figure 1 around here]

A second set of data comes from the Standard and Poor's CreditPro 7.0 database. The latter contains a full set of rating transition histories and (possibly) a default date for all S&P-rated firms from 1980:1 to 2005:2. This set allows us to calculate the required values of y_{jt} and k_{jt} in (1). We distinguish 13 industries which we pool into $D = 7$ industry groups. These are the consumer goods, financials, transport and aviation, leisure, utilities, high tech and telecom, and health care sector. We further consider $A = 4$ 'age' cohorts. These indicate less than 3, 3 to 6, 6 to 12, and more than 12 years from the time of initial rating. The rationale for this distinction is that default probabilities may depend on the age of a company, which we proxy here by the time since the initial rating assignment. Finally, there are $S = 4$ rating groups, specifically one investment grade group $AAA - BBB$, and three speculative grade groups BB, B, CCC. Pooling over investment grade firms is necessary

since defaults are very rare for this segment. The disaggregated default fractions can be observed from Figure 2. Default fractions cluster most visibly around the recession years of 1991 and 2001, and are most visible in the BB and B rating class.

[insert Figure 2 around here]

5.2 The macro factors

We first report the results from applying principal components to the macro panel introduced in Section 5.1. We employ the EM procedure from Section 3.1 to iteratively balance the panel before estimating the factors. Figure 3 shows the first four principal components from this panel. It can be seen that the first PC exhibits clear peaks around NBER US business cycle troughs located around 1969/70, 1973/75, 1980, 1981/82, 1990/91, and 2000/01, see www.nber.org. This would suggest that it mainly loads from macro data and business cycle indicators, which is confirmed below. The second factor also appears to exhibit peaks around these times, but the association with a business cycle is less strong. Factors three and four do not exhibit the clear cyclical swings present in the first two factors.

[insert Figure 3 around here]

To determine a good value for R - the dimension of F_t - we compute the panel information criteria (IC) suggested by Bai and Ng (2002) in Table 2. We evaluate the IC for both the balanced subset of the data as well as the full panel. The criterion function $IC_{p1}(r)$ is minimized for $r = 2$, indicating two common factors. This finding is not robust, as $IC_{p2}(r)$ indicates only one factor, and $IC_{p3}(r)$ decreases monotonously over a range of plausible values. We interpret these results as evidence that most information is contained in the first two factors. These factors capture about 44% of the total variation in the macro panel.

[insert Table 2 around here]

To further illustrate the empirical economic underpinnings of the two common factors we regress each macro variable on each of the two factors separately. Figure 4 depicts the R-squared from these regressions. We observe that the first PC mainly loads mainly from macro and employment data, as well as business cycle indicators and interest rates. According to its associated eigenvalue, the first factor accounts for about 30% of the data variance.

[insert Figure 4 around here]

The second principal component loads mainly from series associated with firm profit margins, such as the price of intermediate inputs and resources, the cost of energy, and prices of final goods. It accounts for about 14% of data variance. Without presenting the respective graph, we report that the third factor loads from series related to financing conditions and from variables indicating the extent of problematic banking loans (7%). The fourth factor explains relatively little, and the loadings do not appear to be concentrated in a particular field (6%).

5.3 The complete CFP model

We now turn to the estimates of the complete non-Gaussian model. Since defaults are rare events we cannot freely and reliably estimate all parameters λ_j, β_j and $\gamma_{r,j}$ for each cross section j. Instead we propose a parsimonious model structure that allows enough flexibility to address the key issues. We do so by setting

$$\begin{aligned} \lambda_j &= \lambda_0 + \lambda_{1,d_j} + \lambda_{2,a_j} + \lambda_{3,s_j}, \\ \beta_j &= \beta_0 + \beta_{1,d_j} + \beta_{2,s_j}, \\ \gamma_{r,j} &= \gamma_{r,s_j}, \end{aligned}$$

where $d_j = 1, \ldots, 7$, $a_j = 1, \ldots, 4$ and $s_j = 1, \ldots, 4$ are the industry index, rating age index, and rating class index of cross section j, respectively. For identification, we set $\lambda_{1,7} = \lambda_{2,4} = \lambda_{3,4} = \beta_{1,7} = \beta_{2,4} = 0$. Baseline intensities λ_j and factor sensitivities β_j and $\gamma_{r,j}$ thus depend on industry, rating, and rating age in a well-defined and parsimonious way.

We report three different specifications of the model in Table 3. Model 1 contains the first two common factors (principal components) from the macro panel, and no unobserved risk factor. Conversely, Model 2 contains an unobserved risk factor, but no common macro factors. Finally, Model 3 combines both specifications. In Model 1 and 3, the macro factor sensitivities $\gamma_{1,s}$ and $\gamma_{2,s}$ depend on the firm's current rating class. In Model 2 and 3, the β coefficients depend on industry and rating class. Rating dependent factor sensitivities

capture the notion that exposure to systematic risk may be less pronounced for lower rating classes. Similarly, industry specific sensitivities capture the notion that some industries may be more sensitive to macro risk than others.

[insert Table 3 around here]

The fixed effects λ_j are similar across models. There is a highly significant monotonic pattern in the coefficients for the rating classes $\lambda_{3,s}$. This pattern indicates that lower ratings are more likely to default. The coefficients indicating the age cohort $\lambda_{2,a}$ show a similar pattern. This suggests that a firm which has just recently acquired access to the capital market is less likely to default. This initial effect appears to subside over time. Finally, there is considerable heterogeneity across industry groups $\lambda_{1,d}$. Firms categorized as being part of the financial or leisure industry are less likely to default than for instance firms from the transport and aviation segment.

We now address the time varying part of the models. It is useful to recall that $F_{1,t}, F_{2,t}$, and f_t^{uc} have zero mean and unit unconditional variance by construction. This implies that all factor sensitivities can also be interpreted in terms of factor standard deviations for these firms. The estimated β-coefficients indicate an important role for the unobserved component even after the first two common macro factors are included. The impact of the unobserved component differs considerably across rating and industry groups. For example, financial firms are found to have much lower systematic risk than firms from the high tech or transport and aviation sector. We report t-statistics for the β-coefficients, but note that they are not asymptotically normal. The null-hypothesis $\beta_0 = 0$ entails a restriction on the rank of the covariance matrix of the signal. Such tests have non-standard properties, cf. for instance Nyblom and Harvey (2000). Similarly, the large increase in likelihood from Model 1 to 2 cannot be used in a formal Likelihood Ratio test. However, the increase by more than 70 points is indicative of a large improvement in model fit. The further increase in likelihood from Model 2 to 3 by 10 points is statistically significant at a 5% level. Thus, all factors are both statistically and economically significant and help to explain the systematic co-movement in the cross section. For scaled estimates of the risk factors we refer to Figure 5.

[insert Figure 5 around here]

The factor sensitivities $\gamma_{1,s}, \gamma_{2,s}$ also differ considerably across rating groups. In all specifications, investment grade firms appear to have high systematic risk. Conversely, defaults from the lowest rating class appear to be largely unrelated to the current macroeconomic climate.

5.4 Out of sample forecasting accuracy

In this subsection we estimate a number of competing model specifications and compare them in terms of their out of sample forecasting accuracy. This is achieved by forecasting conditional default probabilities for a cross-section of firms one year ahead. Measuring the forecasting accuracy of time-varying default probabilities is not straightforward. The basic reason for this is that observed default fractions are only a crude measure of the 'true' default probability pertaining to a certain cross section at a given time. To see this most clearly, consider a cell with, say, 5 firms. Even if the default probability for this cell is forecast perfectly, it is unlikely to coincide with the observed default fraction of either 0, 1/5, 2/5, etc. The forecast error may be large but does not indicate a bad forecast.

Observed default fractions are a useful measure only for a sufficiently large number of firms per cell. For this reason we pool default and exposure counts over the four age cohorts and consider only two rating groups, i.e., firms rated $AAA - BB$ (IG), and B speculative grade (SG). Furthermore, we focus on predicting an annual quantity instead of quarterly fractions. A mean absolute error (MAE) and root mean squared error statistic (RMSE) is computed as follows.

$$
\begin{aligned}
MAE(t) &= \frac{1}{D}\sum_{d}\left|\hat{\Pi}^{an}_{d,t+4|t} - \bar{\Pi}^{an}_{d,t+4}\right|, \\
RMSE(t) &= \left(\frac{1}{D}\sum_{d}\left[\hat{\Pi}^{an}_{d,t+4|t} - \bar{\Pi}^{an}_{d,t+4}\right]^2\right)^{\frac{1}{2}},
\end{aligned}
$$

where

$$\begin{aligned}\hat{\Pi}^{an}_{d,t+4|t} &= 1-\prod_{h=1}^{4}\left(1-\hat{\Pi}^{qu}_{d,t+h|t}\right),\\ \bar{\Pi}^{an}_{d,t+4} &= 1-\prod_{h=1}^{4}\left(1-\frac{y_{d,t+h}}{k_{d,t+h}}\right).\end{aligned}$$

There are several ways to forecast the required default signals. In this paper we first forecast all factors jointly using a vector autoregression. This approach takes into account that the factors are conditionally correlated. We then predict the conditional default probabilities using equations (2) and (3).

Table 4 reports the forecast error statistics for five competing models. Model M0a does not contain any common factors. It thus corresponds to the practice of forecasting the time-varying probabilities using long-term historical averages. This yields relatively small forecast errors when the risk factors happen to be close to their unconditional averages, such as in the years 1998 and 2004, see Figure 5. However, there are substantial forecast errors when this is not the case. Model M0a thus constitutes a lower benchmark. As an upper bound, Model M4 uses the true (estimated) factors, and holds the model parameters fixed at their end-of-sample values. This constitutes an infeasible best case. This upper bound for improvements on average over the years 1997 to 2004 is about 26% for both rating groups. The reductions in MAE are largest when risk factors are far from their long-term averages. For instance, the MAE associated with the year 2000 'forecast' of the recession year 2001 is 67% lower for investment grade firms, and about 48% lower for speculative grade firms.

[insert Tables 4 and 5 around here]

Model M0b uses three observed regressors instead of common factors to forecast conditional default probabilities. These are the HP-filtered US unemployment rate, percentage change in filtered unemployment, and the Baa corporate yield spread over treasuries. Similar regressors are found to have a good in-sample fit, see Metz (2007). This set of regressors turns out to improve out-of-sample forecasting accuracy only very slightly by about 1-2% on average in terms of MAE.

Models M1, M2, and M3 from Table 4 correspond to out of sample forecasts using the models estimated in Table 3. M1 contains only the common macro factors, with rating

dependent factor sensitivities. Model M2 contains one unobserved component only, and allows its sensitivity to vary over both rating classes and industry groups. Model M3 contains both types of factors. We note that the common macro factors $F_{1,t}$ and $F_{2,t}$ are helpful in out of sample forecasting. The observed reduction in MAE is about 1-7%. Forecasts improve when an unobserved component is added to about 11-18% on average. Reductions in MAE are again highest when risk factors are far from their long term averages. The MAE associated with the year 2000 forecast of 2001 default conditions is reduced by about 37% (IG) and 26% (SG) when compared to Model M0b which contains only observable macro variables. When compared to Model M0a the reduction is 38% for investment grade firms and 27% for speculative grade firms. Such improvements are substantial and have clear practical implications for the computation of capital requirements.

6 Conclusion

We propose and motivate a novel times series panel data model to estimate and forecast large cross-sections of time varying conditional default probabilities. The model is the first to combine the non-Gaussian panel data model of Koopman and Lucas (2007) with Stock and Watson's (2002a) approximate dynamic factor model. The final model accommodates two different types of factors, both of which are statistically and economically significant and capture a large part of the time-variation in observed default rates.

In this paper we can overcome a number of complications that arise naturally when modeling firm defaults. For instance, we consider a 'frailty' setting in which all risk factors are unobserved and need to be estimated. We take into account the information from a large array of relevant macroeconomic and financial variables without running into dimensionality problems. Finally, the panel data specification allows to efficiently capture the heterogeneity in the cross-section of firms at any point in time. We focus on combinations of the current rating class, industry specification and time from initial rating as characterizing the cross-section. Other dimensions of firm heterogeneity such as firm size or geographical location can be addressed in exactly the same way.

In an out-of-sample forecasting experiment we improve forecasts of time-varying conditional default probabilities. Out-of-sample reductions are greatest when risk factors are

far from their unconditional averages. Improvements range up to 25% compared to models which only use observable variables, and up to 27-30% when compared to models that disregard changes in systematic risk conditions. The largest improvements on average are achieved for a model specification which contains both unobserved components as well as common factors from macro data.

References

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70(1)*, 191–221.

Bai, J. and S. Ng (2007). Determining the number of primitive shocks in factor models. *Journal of Business and Economic Statistics 25(1)*, 52–60(9).

CreditMetrics (2007). CreditMetrics (TM) - Technical Document, RiskMetrics Group. www.riskmetrics.com/pdf/dnldtechdoc/CMTD1.pdf.

Das, S., D. Duffie, N. Kapadia, and L. Saita (2007). Common Failings: How Corporate Defaults Are Correlated. *The Journal of Finance 62(1)*, 93–117(25).

Doornik, J. A. (2002). *Object-oriented Matrix Programming using Ox 3.0*. Timberlake Consultants Press, London.

Duffie, D., A. Eckner, H. Guillaume, and L. Saita (2006). Frailty Correlated Default. *Working paper, Stanford University*.

Duffie, D., L. Saita, and K. Wang (2007). Multi-Period Corporate Default Prediction with Stochastic Covariates. *Journal of Financial Economics 83(3)*, 635–665.

Durbin, J. and S. J. Koopman (1997). Monte Carlo Maximum Likelihood estimation for non-Gaussian State Space Models. *Biometrica 84(3)*, 669–684.

Durbin, J. and S. J. Koopman (2001). *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Durbin, J. and S. J. Koopman (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika 89(3)*, 603–616.

Figlewski, S., H. Frydman, and W. Liang (2006). Modeling the Eect of Macroeconomic Factors on Corporate Default and Credit Rating Transitions. *New York University Discussion Paper*.

Geweke, J. (1977). The Dynamic Factor Analysis of Economic Time Series. *in Aigner, D and Goldberger, A, eds.: Latent variables in socio-economic models, North-Holland*.

Gordy, M. (2000). A comparative anatomy of credit risk models. *Journal of Banking and Finance 24*, 119–149.

Koopman, S., N. Shephard, and J. Doornik (1998). Statistical algorithms for models in state space using ssfpack 2. *Econometrics Journal 2*, 113–66.

Koopman, S. J. and A. Lucas (2008). A Non-Gaussian Panel Time Series Model for Estimating and Decomposing Default Risk. *Journal of Buisness and Economic Statistics, forthcoming*.

Koopman, S. J., A. Lucas, and A. Monteiro (2008). The Multi-Stage Latent Factor Intensity Model for Credit Rating Transitions. *Journal of Econometrics 142(1)*, 399–424.

Lando, D. (2003). *Credit Risk Modelling - Theory and Applications*. Princeton University Press.

Massimiliano, M., J. H. Stock, and M. W. Watson (2003). Macroeconomic forecasting in the Euro area: Country specific versus area-wide information . *European Economic Review 47(1)*, 1–18.

McNeil, A. and J. Wendin (2007). Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance 14(2)*, 131–149.

Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance 29(2)*, 449–470.

Metz, A. (2007). Credit ratings-based multiple horizon default prediction. *Moody's Investors Services, mimeo*.

Nyblom, J. and A. Harvey (2000). Tests of common stochastic trends. *Econometric Theory 16*, 176199.

Pesaran, H., T. Schuermann, B. Treutler, and S. Weiner (2006). Macroeconomic dynamics

and credit risk: A global perspective. *Journal of Money, Credit, and Banking 38, No. 5*, 1211–1261.

Sargent, T. and C. Sims (1977). Business cycle modeling without pretending to have too much a priori economic theory. *Federal Reserve Bank of Minneapolis, working paper No 55*.

Stock, J. and M. Watson (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association 97(460)*, 1167–1179.

Stock, J. and M. Watson (2002b). Macroeconomic forecasting using diusion indexes. *Journal of Business and Economic Statistics 20(2)*, 147–162.

Stock, J. and M. Watson (2005). Implications of dynamic factor models for var analysis. *NBER working paper*.

Tasche, D. (2006). Measuring sectoral diversification in an asymptotic multi-factor framework. *Journal of Credit Risk 2(3)*, 33–55.

Wilson, T. (Oktober 1998). Portfolio credit risk. *Federal Reserve Board New York Economic Policy Review*, 71–82.

Table 1: Predictor Time Series in the Macro Panel

Main category, sub-category	Summary of time series in category		Total no
Bank lending conditions			
Size of overall lending	Total Commercial Loans Total Real Estate Loans Total Consumer Credit outstanding Debt of Domestic Sector Commercial &Industrial Loans Bank loans and investments Household obligations/income	Household Debt/Income-ratio Federal Debt of Non-fin. sector Excess Reserves Depository Inst. Total Borrowing from Fed Reserve Household debt service payments Total Loans and Leases, all banks	13
Extend of problematic banking business	Non-performing Loans Ratio Net Loan Losses Return on Bank Equity Non-perf. Commercial Loans	Non-performing Total Loans Total Net Loan Charge-offs Loan Loss Reserves	7
Macro and BC conditions			
General macro indicators	Real GDP Industr. Production Index Private Fixed Investments National Income Manuf. Sector Output Manuf. Sector Productivity Government Expenditure	ISM Manufacturing Index Uni Michigan Consumer Sentiment Real Disposable Personal Income Personal Income Consumption Expenditure Expenditure Durable Goods Gross Private Domestic Investment	14
Labor market conditions	Unemployment rate Weekly hours worked Employment/Population-Ratio Unemployed, more than 15 weeks	Total No Unemployed Civilian Employment Unemployed, less than 5 weeks	7
Business Cycle leading/ coinciding indicators	New Orders: Durable goods New orders: Capital goods Capacity Util. Manufacturing Capacity Util. Total Industry Light weight vehicle sales Housing Starts New Building Permits Final Sales of Dom. Product	Retail sales and Food services Inventory/Sales-ratio Change in Private Inventories Inventories: Total Business Non-farm housing starts New houses sold Final Sales to Domestic Buyers	15
Monetary policy indicators	M1 Money Stock M2 Money Stock M3 Money Stock UMich Infl. Expectations Personal Savings Gross Saving	CPI: All Items Less Food CPI: Energy Index Personal Savings Rate GDP Deflator, chain type GDP Deflator, implicit	11
Corporate Profitability	Corp. Profits Net Corporate Dividends	After Tax Earnings Corporate Net Cash Flow	4
Intern'l competitiveness			
Terms of Trade	Trade Weighted USD USD/GER Exchange Rate	FX index major trading partners USD/GBP Exchange Rate	4
Balance of Payments	Current Account Balance Balance on Merchandise Trade Real Exports Goods, Services	Balance on Services Real Imports Goods & Services	5
Micro-level conditions			
Labour cost/wages	Unit Labor Cost: Manufacturing Total Wages & Salaries Wholesale Trade Wages Management Salaries Technical Services Wages Wages & Salaries: Other Employee Compensation Index	Unit Labor Cost: Nonfarm Business Non-Durable Manufacturing Wages Durable Manufacturing Wages Employment Cost Index: Benefits Employment Cost Index: Wages & Salaries Employee Compensation: Salary Accruals	13
Cost of capital	1Month Commerical Paper Rate 3Month Commerical Paper Rate Effective Federal Funds Rate AAA Corporate Bond Yield BAA Corporate Bond yield	Treasury Bond Yield, 10 years Term Structure Spread Corporate Yield Spread 30 year Mortgage Rate Bank Prime Loan Rate	10
Cost of resources	PPI All Commodities PPI Interm. Energy Goods PPI Finished Goods PPI Crude Energy Materials	PPI Industrial Commodities PPI Fuels and Related Products PPI Intermediate materials	7
Equity market conditions			
Equity Indexes and respective volatilities	S&P 500 Nasdaq 100 S&P Small Cap Index	Dow Jones Industrial Average Russell 2000	10

Table 2: Panel Information Criteria by Bai and Ng (2002)

The table reports three different information criteria for both the balanced subset of the data as well as the full panel. We calculate $IC_{p1}(r)$, $IC_{p2}(r)$, and $IC_{p3}(r)$, $r = 1, \ldots, 5$. Bold print indicates minimal reported values.

$IC_{p1}(r)$	$r = 1$	$\mathbf{r = 2}$	$r = 3$	$r = 4$	$r = 5$
Bal. Subset	−0.0344	**−0.0429**	−0.0306	−0.0101	0.0073
Full Panel	−0.2860	**−0.2892**	−0.2795	−0.2669	−0.2630
$IC_{p2}(r)$	$\mathbf{r = 1}$	$r = 2$	$r = 3$	$r = 4$	$r = 5$
Bal. Subset	**−0.0004**	0.0252	0.0716	0.1261	0.1777
Full Panel	**−0.2725**	−0.2621	−0.2388	−0.2127	−0.1852
$IC_{p3}(r)$	$r = 1$	$r = 2$	$r = 3$	$r = 4$	$\mathbf{r = 5}$
Bal. Subset	−0.0938	−0.1616	−0.2086	−0.2476	**−0.2894**
Full Panel	−0.3180	−0.3531	−0.3753	−0.3946	**−0.4127**

Table 3: Estimation results

Note: The factor sensitivity parameters β pertain to the unobserved component and depend on the firm's industry specification. The sensitivity parameters γ of the common macro factors depend on the firm's current rating class. Numbers in bold print are significant at a 5% significance level. The groups mnemonics are given by fin: financial, tra: transport and aviation, lei: leisure, utl: utilities, hte: high tech and telecom, hea: health care. The consumer goods industry constitutes the reference group. The results are calculated using 1000 importance samples. Estimation period is 1981:1 to 2005:2.

	Model 1: Only F_t			Model 2: Only f_t^{uc}			Model 3: All Factors		
par	val	se	t-val	val	se	t-val	val	se	t-val
λ_0	-1.47	0.13	11.47	-1.50	0.19	8.10	-1.50	0.17	8.91
$\lambda_{1,fin}$	-0.47	0.13	3.57	-0.38	0.15	2.43	-0.40	0.14	2.84
$\lambda_{1,tra}$	-0.11	0.09	1.26	-0.18	0.09	1.95	-0.12	0.09	1.39
$\lambda_{1,lei}$	-0.48	0.11	4.30	-0.51	0.10	4.98	-0.67	0.17	3.98
$\lambda_{1,utl}$	-0.37	0.10	3.58	-0.39	0.11	3.58	-0.43	0.10	4.34
$\lambda_{1,hte}$	-0.20	0.11	1.84	-0.46	0.14	3.20	-0.34	0.12	2.75
$\lambda_{1,hea}$	-0.34	0.13	2.70	-0.44	0.13	3.35	-0.55	0.17	3.28
$\lambda_{2,0-3}$	-0.73	0.12	6.24	-0.64	0.12	5.46	-0.68	0.13	5.35
$\lambda_{2,4-5}$	-0.33	0.12	2.77	-0.36	0.13	2.85	-0.38	0.13	2.94
$\lambda_{2,6-12}$	-0.43	0.12	3.49	-0.36	0.13	2.82	-0.39	0.13	3.13
$\lambda_{3,IG}$	-6.22	0.21	29.81	-6.35	0.26	24.24	-6.40	0.26	24.80
$\lambda_{3,BB}$	-3.96	0.13	29.52	-4.15	0.19	22.20	-4.21	0.21	19.88
$\lambda_{3,B}$	-2.37	0.08	30.50	-2.51	0.12	20.51	-2.63	0.18	14.72
ϕ				0.87	0.08	11.70	0.85	0.08	10.40
β_0				0.70	0.18	3.72	0.64	0.17	3.81
$\beta_{1,fin}$				-0.20	0.17	1.19	-0.14	0.19	0.78
$\beta_{1,tra}$				0.06	0.11	0.53	0.01	0.11	0.05
$\beta_{1,lei}$				0.00	0.13	0.01	0.24	0.17	1.43
$\beta_{1,utl}$				0.02	0.14	0.16	0.09	0.15	0.61
$\beta_{1,hte}$				0.27	0.16	1.70	0.18	0.16	1.12
$\beta_{1,hea}$				0.10	0.16	0.63	0.26	0.19	1.40
$\beta_{2,IG}$				0.28	0.22	1.27	-0.14	0.26	0.55
$\beta_{2,BB}$				0.14	0.15	0.95	0.00	0.19	0.02
$\beta_{2,CCC}$				-0.32	0.12	2.72	-0.43	0.15	2.94
γ_1^{IG}	0.76	0.17	4.59				0.57	0.18	3.10
γ_1^{BB}	0.73	0.12	5.97				0.38	0.15	2.59
γ_1^{B}	0.44	0.06	7.14				0.07	0.12	0.56
γ_1^{CCC}	0.42	0.06	6.74				0.24	0.07	3.30
γ_2^{IG}	0.28	0.16	1.74				0.37	0.17	2.19
γ_2^{BB}	-0.05	0.12	0.43				0.12	0.14	0.89
γ_2^{B}	0.21	0.06	3.77				0.40	0.09	4.28
γ_2^{CCC}	-0.02	0.06	0.26				0.05	0.06	0.78
LogLik	-2994.74			-2942.64			-2929.52		

Table 4: Out-of-sample Forecasting Accuracy

The table reports Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) statistics associated with one-year ahead out of sample forecasts of conditional default probabilities. We report the statistics separately for investment grade (AAA-BBB) and speculative grade (BB) firms. The last two columns indicate the percentage changes in MAE with respect to the benchmark model M0a and M0b, respectively.

		1997	1998	1999	2000	2001	2002	2003	2004	average	change M0a	change M0b
M0a, No factors	MAE, IG	0.18	0.24	0.29	0.84	1.06	1.32	0.46	0.29	0.59		
	MAE, SG	3.05	2.08	2.24	3.49	7.18	4.47	2.12	3.76	3.55		
	RMSE, IG	0.18	0.27	0.46	1.03	1.22	1.78	0.51	0.35	0.72		
	RMSE, SG	3.65	2.50	3.14	4.05	7.93	4.83	2.48	3.93	4.07		
M0b, Observables	MAE, IG	0.18	0.24	0.29	0.85	1.03	1.31	0.47	0.27	0.58	-1.1%	-
	MAE, SG	3.18	1.98	2.25	3.74	7.03	4.24	2.24	3.22	3.48	-1.9%	-
	RMSE, IG	0.18	0.26	0.46	1.04	1.19	1.76	0.50	0.33	0.72		
	RMSE, SG	3.81	2.32	3.11	4.28	7.76	4.30	2.60	3.39	3.94		
M1, Only F_t	MAE, IG	0.16	0.22	0.28	0.88	1.01	1.32	0.47	0.26	0.57	-1.8%	-0.7%
	MAE, SG	2.72	1.86	2.37	4.13	6.76	4.00	2.28	2.46	3.32	-6.4%	-4.6%
	RMSE, IG	0.17	0.27	0.48	1.07	1.17	1.79	0.52	0.33	0.72		
	RMSE, SG	3.23	2.12	3.37	4.68	7.43	4.09	2.67	2.66	3.78		
M2, Only f_t^{uc}	MAE, IG	0.15	0.20	0.29	0.85	0.90	1.11	0.45	0.26	0.53	-9.9%	-8.9%
	MAE, SG	2.41	1.27	2.19	3.51	5.75	4.67	2.07	2.44	3.04	-14.4%	-12.8%
	RMSE, IG	0.17	0.28	0.46	1.04	1.07	1.42	0.50	0.33	0.66		
	RMSE, SG	2.66	1.39	3.06	4.07	6.35	5.59	2.45	2.60	3.52		
M3, Both F_t, f_t^{uc}	MAE, IG	0.13	0.19	0.28	0.87	0.65	1.30	0.48	0.26	0.52	-11.1%	-10.1%
	MAE, SG	2.12	1.29	2.79	4.03	5.24	3.97	2.19	1.88	2.94	-17.2%	-15.6%
	RMSE, IG	0.17	0.28	0.49	1.06	0.82	1.76	0.54	0.34	0.68		
	RMSE, SG	2.32	1.45	3.85	4.57	5.64	4.18	2.53	2.09	3.33		
M4, 'True' Factors	MAE, IG	0.13	0.21	0.30	0.62	0.34	1.17	0.45	0.26	0.43	-26.0%	-25.1%
	MAE, SG	2.40	1.38	2.53	2.00	3.64	4.48	3.04	1.34	2.60	-26.8%	-25.4%
	RMSE, IG	0.18	0.27	0.46	0.80	0.43	1.52	0.48	0.35	0.56		
	RMSE, SG	2.68	1.66	3.19	2.54	5.27	5.88	3.33	1.52	3.26		

Table 5: Changes in MAE

The table reports changes in mean absolute error for out of sample forecasting of the years 1997-2004 (left) and the year 2000 forecast of the recession year 2001.

		Reduction in MAE, years 1997 - 2004		Reduction in MAE, year 2001	
		M0a "no factors"	M0b "observables"	M0a "no factors"	M0b "observables"
M1, only $\hat{F}_t$	IG	-1.8%	-0.7%	-5.1%	-2.8%
	SG	-6.4%	-4.6%	-6.0%	-4.0%
M2, only $\hat{f}_t^{uc}$	IG	-9.9%	-8.9%	-14.9%	-12.8%
	SG	-14.4%	-12.8%	-20.0%	-18.3%
M3, $\hat{F}_t$ and $\hat{f}_t^{uc}$	IG	-11.1%	-10.1%	-38.2%	-36.7%
	SG	-17.2%	-15.6%	-27.1%	-25.6%
M4, F_t and f_t^{uc}	IG	-26.0%	-25.1%	-68.0%	-67.2%
	SG	-26.8%	-25.4%	-49.4%	-48.3%

Figure 1: Aggregated Default Data and Default Fractions
The graph exhibits total default counts, the total number of firms, and total default fractions after aggregation over all cells in the cross section.

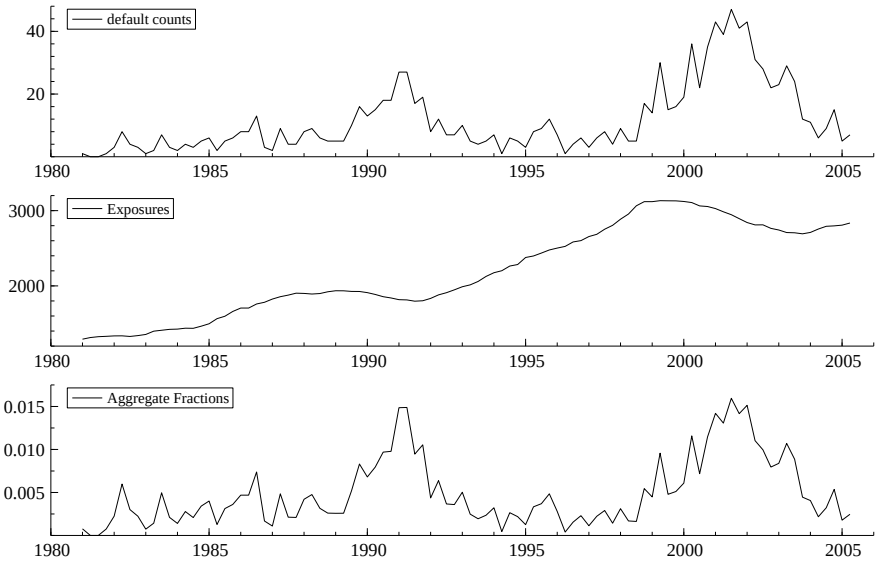

Figure 2: Default Fractions Scatterplot
The graph shows the full cross section of default fractions y_{jt}/k_{jt} over time t (where observed). The second figure shows disaggregated default fractions for rating groups $AAA - BBB$, BB, B, and CCC.

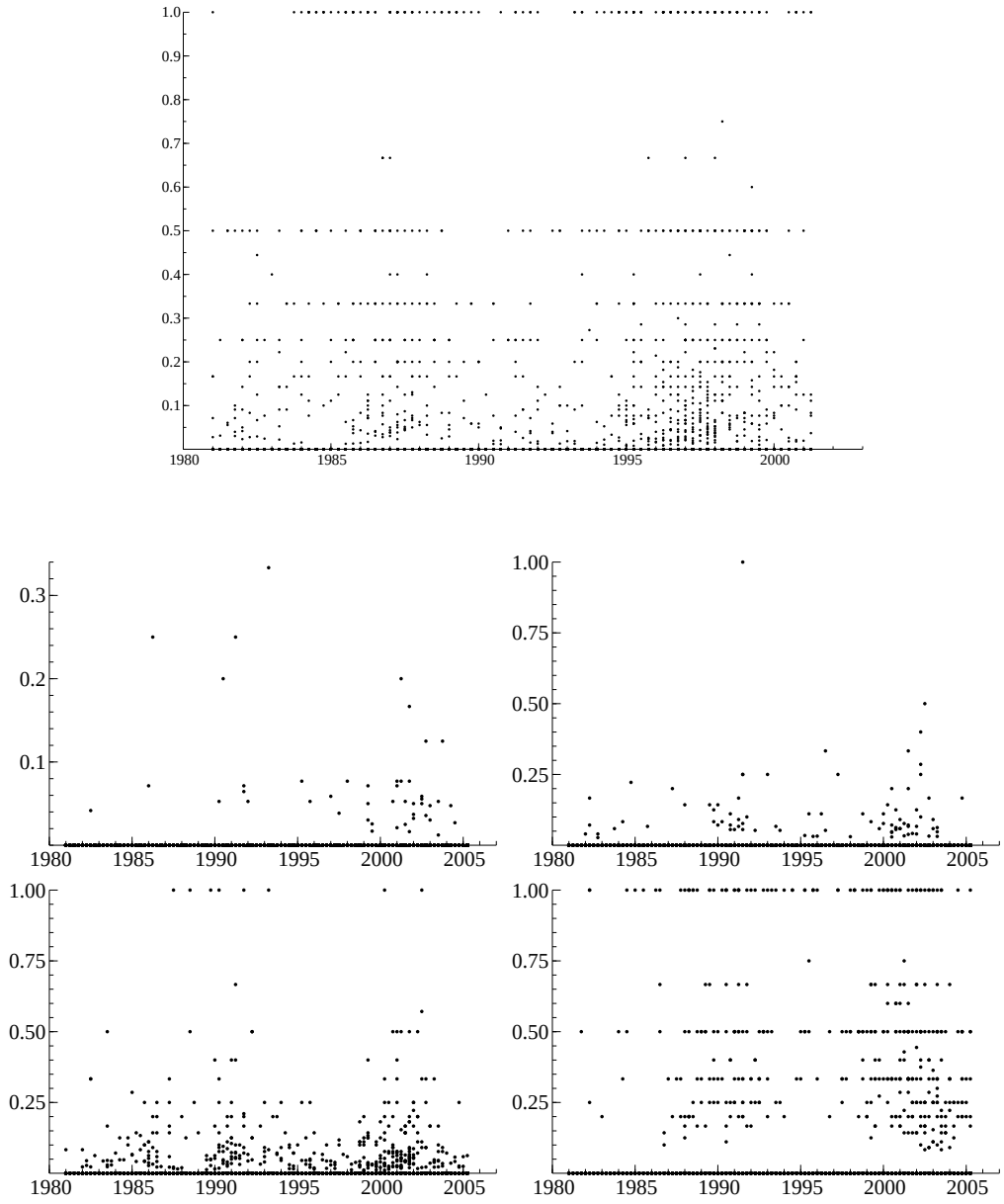

Figure 3: Principal components from unbalanced macro data

The first four principal components are calculated from unbalanced macro data (N=1,..,120) using the EM algorithm of Stock and Watson (2002b).

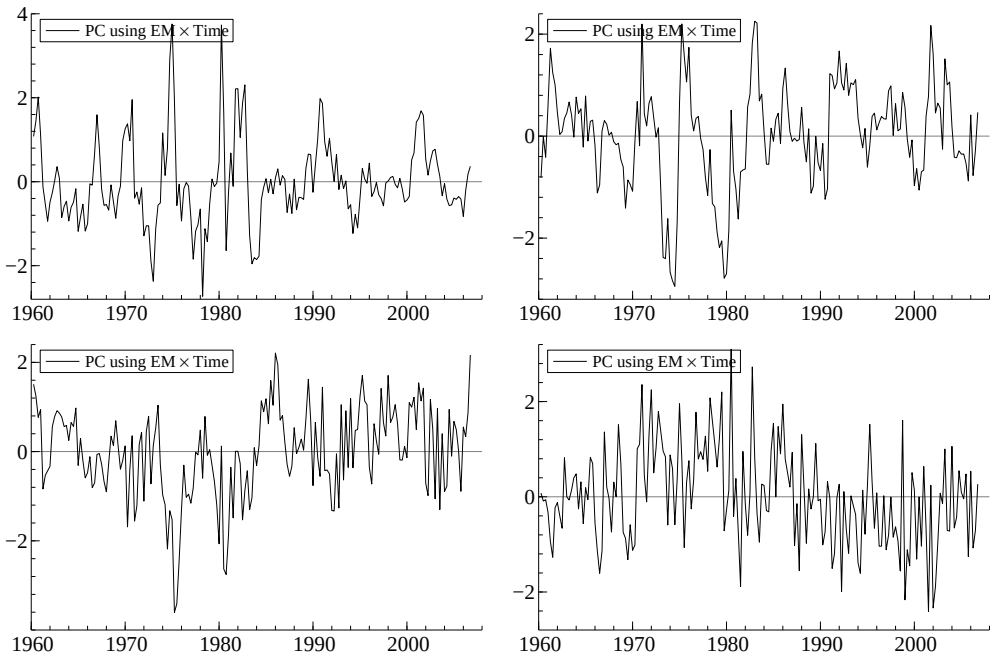

Figure 4: Factor loadings pertaining to the first two principal components
Each macroeconomic and financial variable from Table 1 is regressed on $F_{1,t}$ and $F_{1,t}$, respectively. The R-squared from these regressions are presented in the figures. The mnemonics are Bank: Bank lending conditions, PrLn: Extent of problematic loan business; Macro&Empl: Macro indicators and labor market conditions; BCInd: Leading and coinciding business cycle indicators; MP: monetary policy indicators; Intl: Balance of Payments and Terms of Trade variables; Wag: Wages and Salaries data; IR&CR: Interest rates and cost of intermediate goods and resources; Stocks: Equity indices and respective volatilities.

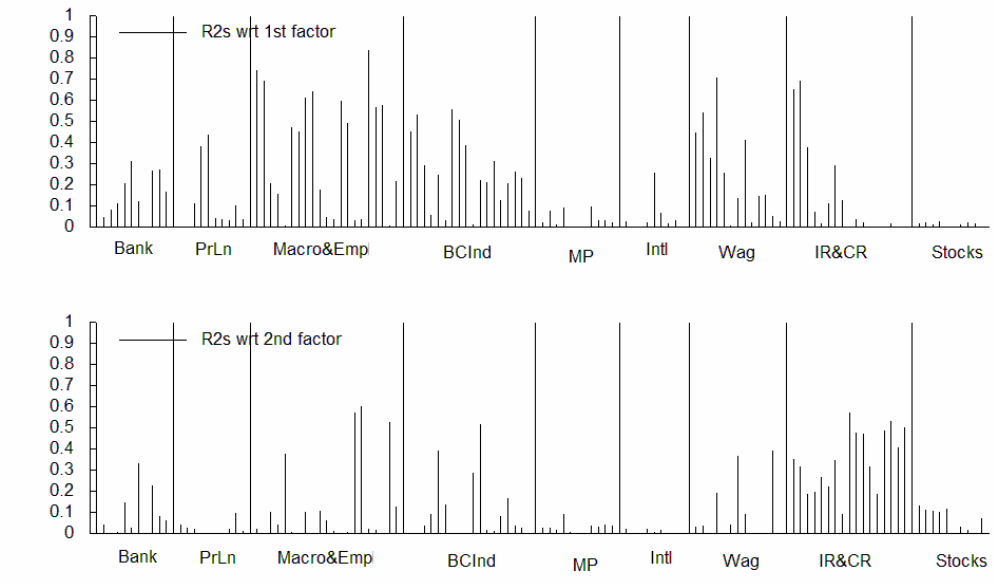

Figure 5: Economic Significance of Common Factors

The figure indicates how the time variation of the total signal can be decomposed into variation of f_t^{uc} and of the first two principal components $F_{1,t}$, $F_{2,t}$. All factor sensitivities depend only on a firm's current rating class. The first figure shows the signal for investment grade firms. The second figure plots the three series scaled by their respective factor standard deviations (sensitivity coefficients).

$$UC = (\beta_0 + \beta_{1,IG}) f^{uc} \tag{14}$$

$$UCF1 = (\beta_0 + \beta_{1,IG}) f^{uc} + \gamma_{1,IG} F_1 \tag{15}$$

$$UCF12 = (\beta_0 + \beta_{1,IG}) f^{uc} + \gamma_{1,IG} F_1 + \gamma_{2,IG} F_2 \tag{16}$$