

Confidence Intervals for Probabilities of Default¹

Samuel Hanson²

Til Schuermann²

Federal Reserve Bank of New York
33 Liberty St.
New York, NY 10045
samuel.hanson@ny.frb.org, til.schuermann@ny.frb.org

First Draft: July 2004
This Draft: July 2005
This Print: July 19, 2005

Abstract: In this paper we conduct a systematic comparison of confidence intervals around estimated probabilities of default (PD) using several analytical approaches as well as parametric and nonparametric bootstrap methods. We do so for two different PD estimation methods, cohort and duration (intensity), with 22 years of credit ratings data. We find that the bootstrapped intervals for the duration based estimates are relatively tight when compared to either analytic or bootstrapped intervals around the less efficient cohort estimator. We show how the large differences between the point estimates and confidence intervals of these two estimators are consistent with non-Markovian migration behavior. Surprisingly, even with these relatively tight confidence intervals, it is impossible to distinguish notch-level PD s for investment grade ratings, e.g. a PD_{AA-} from a PD_{A+} . However, once the speculative grade barrier is crossed, we are able to distinguish quite cleanly notch-level estimated PD s. Conditioning on the state of the business cycle helps: it is easier to distinguish adjacent PD s in recessions than in expansions.

Keywords: Risk management, credit risk, bootstrap

JEL Codes: G21, G28, C16

¹ We would like to thank Halina Frydman, Mark Levonian, Thomas Mählmann, Tony Rodrigues, Joshua Rosenberg, Marc Saidenberg and two anonymous referees, as well as seminar participants at the Federal Reserve Bank of New York for their insightful comments. An earlier version of this paper had the title “Estimating Probabilities of Default.” All remaining errors are ours.

² Any views expressed represent those of the authors only and not necessarily those of the Federal Reserve Bank of New York or the Federal Reserve System.

1. Introduction

Credit risk is the dominant source of risk for banks and the subject of strict regulatory oversight and policy debate (BCBS, 2001a, 2004).¹ Credit risk is commonly defined as the loss resulting from the failure of obligors to honor their payments. Arguably a cornerstone of credit risk modeling is the probability of default (*PD*). Two other components are loss-given-default or loss severity and exposure at default.² In fact, these are three of the four key parameters that make up the internal ratings based (IRB) approach that is central to the New Basel Accord (BCBS, 2001b, 2004).³ In this paper we address the issue of how to obtain confidence intervals for *PDs* using estimates computed from publicly available credit rating histories. We systematically compare two well known estimation methods, the cohort and duration based, and their corresponding confidence intervals. Confidence intervals for cohort *PDs* can be obtained either analytically or by bootstrapping, while confidence intervals for duration *PDs* must be obtained by bootstrapping; the latter turn out to be relatively tight.

Regulators are, of course, not the only constituency interested in the properties of *PD* estimates. *PDs* are inputs to the pricing of credit assets, from bonds and loans to more sophisticated instruments such as credit derivatives, and they are needed for effective risk and capital management. However, default is (hopefully) a rare event, especially for high credit quality firms which make up the bulk of the large corporate segment in any large bank. Thus estimated *PDs* are likely to be very noisy. Moreover, *PDs* may vary systematically with the business cycle and are thus unlikely to be stable over time. There may also be other important sources of heterogeneity such as country or industry that might affect rating migration dynamics

¹ The typical risk taxonomy includes market, credit and operational risk. See, for instance, discussions in Crouhy, Galai and Mark (2001) or Marrison (2002).

² For a review of the *LGD* literature, see Schuermann (2004).

generally (i.e. not just the migration to default), as documented by Altman and Kao (1992), Nickell, Perraudin and Varotto (2000) and others. For instance, Cantor and Falkenstein (2001), when examining rating consistency, document that sector and macroeconomic shocks inflate PD volatilities.

We estimate PD s using publicly available data from rating agencies, in particular credit rating histories. In this way we do not attempt to build default or bankruptcy models from firm observables but take the credit rating as a sufficient statistic for describing the credit quality of an obligor. For discussions on bankruptcy and default modeling, see for instance Altman (1968), Shumway (2001), and Hillegeist, Keating, Cram and Lundstedt (2004).

Our main contribution is a systematic comparison of confidence intervals using several analytical approaches as well as small-sample confidence intervals obtained from parametric and nonparametric bootstrapping. We do so for two different PD estimation methods, cohort and duration (intensity). We find that the bootstrapped intervals for the duration based estimates are surprisingly tight and that the less efficient cohort approach generates much wider intervals.

We then use these confidence intervals to analyze ratings migration behavior and to conduct policy-relevant analysis. In particular, even with the tighter bootstrapped confidence intervals for the duration based estimates, it is impossible to distinguish notch-level PD s for neighboring investment grade ratings, e.g. a PD_{AA-} from a PD_{A+} or even a PD_A . However, once the speculative grade barrier is crossed, we are able to distinguish quite cleanly notch-level estimated default probabilities. The New Basel Accord sets a lower bound of 0.03% on the PD estimate which may be used to compute regulatory capital (§285, BCBS, 2004). Our results indicate that 0.03% is above the upper limit of the bootstrapped 95% confidence interval for the

³ The fourth parameter is maturity.

top three rating grades, AAA through A, using the duration approach, but within the 95% confidence interval of the AA rating using the cohort approach.

When we condition on a common factor, namely the state of the business cycle (recession vs. expansion), we find that bootstrapped *PD* densities overlap significantly for investment grades, even at the whole grade level (e.g. the density for PD_A estimated during a recession vs. expansion). For the speculative grades the densities are cleanly separated, suggesting that firms with these lower credit ratings are more sensitive to systematic business cycle effects. Moreover, we find that these densities are surprisingly close to normal (Gaussian).

Our approach is closest to a recent study by Christensen, Hansen and Lando (2004) who use simulation-based methods, a parametric bootstrap, to obtain confidence intervals for *PDs* obtained with the duration (intensity) based approach. Their results are similar in that the confidence intervals implied by their simulation technique for duration *PDs* are also tighter than those implied by analytical approaches for cohort *PDs*. Our resampling-based approach may arguably be better able to pick up any small sample properties of these estimators. Moreover, we consider the impact of sample length on the ability to conduct inference on *PD* estimates. Finally, we take into account recent results in the statistics literature which document erratic behavior of the coverage probability of the standard Wald confidence interval (Brown, Cai and Dasgupta, 2001, Vos and Hudson, 2005) by also including an alternative, the Agresti-Coull confidence interval (Agresti and Coull, 1998).

The efficiency gains from using duration based approaches are well known; see Lando and Skødeberg (2002) and Jafry and Schuermann (2004). The cost, however, is imposing an assumption that the ratings are governed by a Markov process, and there is considerable evidence that this assumption may be unrealistic. A prime example is non-Markov ratings drift,

first documented by Altman and Kao (1992); recent papers include Fledelius, Lando and Nielsen (2004) and Hamilton and Cantor (2004). The latter study, for instance, finds that once the rating outlook is controlled for, e.g. whether the obligor has been placed on the watch list for possible downgrade, it becomes much harder to find evidence of non-Markovian behavior. This non-Markovian behavior is not limited to ratings by the rating agencies, as documented by Trück and Rachev (2005) who look at ratings histories of a bank's borrowers where the ratings are generated by internal models. In computing confidence intervals for PDs , our nonparametric bootstrap is able to relax this assumption somewhat by resampling directly from the observed histories rather than using a fitted Markov process as the basis for generating synthetic histories. Indeed, we show how the large differences between the point estimates and associated intervals of the cohort and duration estimators are consistent with a particular form a non-Markovian migration behavior that has received considerable attention in the literature: downward persistence or momentum.

The rest of the paper proceeds as follows. In Section 2 we discuss the estimation of transition matrices and default probabilities as well as methods for obtaining confidence intervals for PDs . Section 3 discusses properties of empirical estimates of default probabilities; here we compare analytical approaches with the bootstrap. In Section 4 we make use of the confidence interval results to conduct policy-relevant analysis, and Section 5 provides some final comments.

2. Credit ratings and transitions

Credit migration or transition matrices characterize past changes in credit quality of obligors (typically firms) using ratings migration histories. We focus our attention on the last column of this matrix which captures the probability of default. It is customary to use a one-year

horizon in credit risk management, and we follow suit.⁴ Lando and Skødeberg (2002) present and review several approaches to estimating these migration matrices which are compared extensively in Jafry and Schuermann (2004). Broadly there are two approaches, cohort and two variants of duration (or hazard) – parametric (imposing time homogeneity or invariance) and nonparametric (relaxing time homogeneity). In this section we provide a brief sketch of these approaches; interested readers seeking details should consult the references provided.

In simple terms, the cohort approach just takes the observed proportions from the beginning of the year to the end (for the case of annual migration matrices) as estimates of migration probabilities. Suppose there are $N_i(t)$ firms in rating category i at the beginning of the year t , and $N_{ij}(t)$ migrated to grade j by year-end. An estimate of the transition probability for

year t is $P_{ij}(t) = \frac{N_{ij}(t)}{N_i(t)}$. For example, if two firms out of 100 went from grade ‘AA’ to ‘A’, then

$P_{AA \rightarrow A} = 2\%$. Any movements within the year are not accounted for. Typically firms whose ratings were withdrawn or migrated to Not Rated (NR) status are removed from the sample.⁵ It is straightforward to extend this approach to multiple years. For instance, suppose that we have data for T years, then the estimate for all T years is:⁶

$$(2.1) \quad P_{ij} = \frac{N_{ij}}{N_i} = \frac{\sum_{t=1}^T N_{ij}(t)}{\sum_{t=1}^T N_i(t)}$$

⁴ Extensions to multi-year horizons, relevant for instance to buy and hold investors, while important are beyond the scope of this paper.

⁵ The method which has emerged as an industry standard treats transitions to NR as non-informative. The probability of transitions to NR is distributed among all states in proportion to their values. This is achieved by eliminating companies whose ratings are withdrawn. We use this method, which appears sensible and allows for easy comparisons to other studies.

⁶ Indeed this is the MLE of the transition probability under a discrete time-homogeneous Markov chain.

By contrast, the parametric duration approach counts *all* rating changes over the course of the year (or multi-year period) and divides by the number of firm-years, N_R^* , spent in each state or rating to obtain a matrix of migration intensities which are assumed to be time homogenous. Under the assumption that migrations follow a Markov process, these intensities can be transformed to yield a matrix of migration probabilities. To illustrate some of the differences between the two methods, consider a firm that begins the year in A, transitions mid-year to BBB, before ending the year in BB. In the duration approach, both transitions (A→BBB and BBB→BB) as well as the portion of time spent in each of the three states would contribute to the estimated probabilities. In the cohort approach, the mid-year transition to BBB as well as the time spent in BBB would have been ignored. Moreover, firms which end the period in an NR status still contribute to the estimated probabilities up until the date when they transition to NR.⁷

The migration matrix can also be estimated using nonparametric methods such as the Aalen-Johansen estimator which relaxes the assumption of time homogeneity while maintaining the above Markov assumption.⁸ Jafry and Schuermann (2004) find that relaxing the time homogeneity assumption by using this nonparametric estimator generates annual transition matrices that are statistically indistinguishable from their parametric counterparts. For this reason we focus our modeling efforts just on the parametric duration approach.

⁷ There is a range of differences between the number of firm-years spent in rating R under the duration approach, N_R^* , and N_R from the cohort approach range. For instance, the total number of firm-years spent in ‘BBB’ during 2002 was $N_{BBB}^* = 857$ whereas $N_{BBB} = 804$ under the cohort approach. The difference is driven by time spent in ‘BBB’ by firms in mid-year transit and by firms whose ratings were withdrawn. By contrast, the difference for the ‘A’ rating was much smaller: $N_A^* = 695$ against $N_A = 694$.

⁸ For details, see Aalen and Johansen (1978) and Lando and Skødeberg (2002).

2.1. Estimating confidence intervals for PDs

Once we obtain estimates of the default probabilities, we can discuss several approaches for inference and hypothesis testing. Denote PD_R as shorthand for the one-year probability of default for a firm with rating R . We seek to construct a $(1-\alpha)\%$ confidence interval for PD_R , e.g. $\alpha = 5\%$, given an estimate of PD_R , \widehat{PD}_R :

$$(2.2) \quad \Pr \left[PD_R^{\text{low}} < PD_R < PD_R^{\text{up}} \right] = 1 - \alpha$$

As default rates are very small for high quality borrowers, PD_R^{low} may be zero, and in this way the interval may not be symmetric about \widehat{PD}_R .

2.2. Analytical confidence intervals for cohort based PDs

If default is taken to be a binomial random variable, as is the underlying assumption for the cohort approach, then the standard Wald confidence interval CI_W is

$$(2.3) \quad CI_W = \widehat{PD}_R \pm \kappa \sqrt{\frac{\widehat{PD}_R (1 - \widehat{PD}_R)}{N_R}},$$

where N_R is the total number of firms that began the year in rating R , and κ is the $1 - \alpha/2$ quantile of the standard normal distribution. Equation (2.3) follows from the standard asymptotic results for a binomial random variable. For example, in the case of $\alpha = 5\%$, $\kappa = 1.96$. Naturally this assumes that \widehat{PD}_R is estimated from a set of *iid* draws, meaning, for instance, that the probability of default does not vary systematically across time or industry, and that the likelihood of default for firm i in year t is independent of firm j in the same year. This clearly seems unreasonable as there are likely to be common factors such as the state of the economy

which affect all firms, albeit differently, in a given year t . For this reason the Wald confidence interval described by (2.3) may be too narrow.⁹

Brown, Cai and DasGupta (2001) show persuasively that the coverage probability of the standard Wald interval can be significantly less than its nominal value not just for cases when the true (but unknown) probability is near the $[0,1]$ boundary but throughout the unit interval. Moreover, when no outcomes (defaults) are observed at all, the resulting confidence interval is degenerate, a problem not suffered by the methods outlined below.¹⁰ Among the many alternative methods for computing a confidence interval, their final recommendation for cases where the number of observations is at least 40 is the Agresti-Coull interval, from Agresti and Coull (1998). Instead of using the simple sample proportion, namely $\widehat{PD}_R = \frac{N_{R,D}}{N_R}$, as the center of the confidence interval, Agresti and Coull suggest

$$(2.4) \quad \widetilde{PD}_R = \frac{\widetilde{N}_{R,D}}{\widetilde{N}_R}, \text{ where } \widetilde{N}_{R,D} = N_{R,D} + \kappa^2 / 2 \text{ and } \widetilde{N}_R = N_R + \kappa^2.$$

The corresponding confidence interval for one year is

$$(2.5) \quad CI_{AC} = \widetilde{PD}_R \pm \kappa \sqrt{\frac{\widetilde{PD}_R (1 - \widetilde{PD}_R)}{\widetilde{N}_R}}.$$

Agresti and Coull (1998) describe this as “add 2 successes and 2 failures” if one uses 2 instead of 1.96 for κ in the case of $\alpha = 5\%$. Brown et al. (2001) show that the coverage probability for the Agresti-Coull interval is far closer to its nominal $(1-\alpha)\%$ value.

Both the Wald and Agresti-Coull intervals depend on asymptotic theory. Alternatively, one can compute the Clopper-Pearson exact interval, exact because it is derived from the (finite

⁹ See also Stein (2003) for a related discussion on sample size with dependence.

¹⁰ For an alternative approach to estimating PD s when no defaults are available, see Pluto and Tasche (2005).

sample) binomial distribution. For a given α , this confidence interval has endpoints PD_R^{low} and PD_R^{up} that are solutions in PD to the equations:

$$(2.6) \quad \begin{aligned} \sum_{k=N_{R,D}}^{N_R} \binom{N_R}{k} PD^k (1-PD)^{N_R-k} &= \alpha/2 \\ \sum_{k=0}^{N_{R,D}} \binom{N_R}{k} PD^k (1-PD)^{N_R-k} &= \alpha/2, \end{aligned}$$

except that $PD_R^{\text{low}} = 0$ when $N_{R,D} = 0$. In other words, PD_R^{low} is the PD so low that the probability of observing $N_{R,D}$ or more defaults is $\alpha/2$. Similarly, PD_R^{up} is the PD so large that the probability of observing $N_{R,D}$ or fewer defaults is $\alpha/2$. Although Brown et al. (2001) claim that the Clopper-Pearson interval is “wastefully conservative” (p. 113), it is used by Christensen, Hanson and Lando (2004) as a comparison to their parametric bootstrap and thus serves as a useful baseline comparison to their results.¹¹

2.3. Confidence intervals based on bootstrapping

An alternative approach to obtaining confidence intervals for default probability estimates is via the bootstrap method. As it is not clear how to obtain analytical confidence intervals for PD s obtained via the duration or intensity approach, this is our preferred method for constructing confidence intervals for these PD s. By resampling on the firm rating-histories, we create B bootstrap samples¹² of size N_t each, where N_t is the number of firm-histories over some time interval which could be a year or multiple years, compute the entire migration matrix

$\{\mathbf{P}(t)^{(j)}\}_{j=1}^B$ and then focus our attention just on the last vector, $\{\mathbf{PD}(t)^{(j)}\}_{j=1}^B$, where $j = 1, \dots, B$

¹¹ The debate on the proper choice of confidence intervals for a binomial proportion is ongoing. For a recent discussion on this topic, see Vos and Hudson (2005).

denotes the number of bootstrap replications. Efron and Tibshirani (1993) suggest that for obtaining standard errors for bootstrapped statistics, bootstrap replications of 200 are sufficient. For confidence intervals, they suggest bootstrap replications of 1000.¹³ To play it safe we set $B = 10,000$. Note that this bootstrap methodology is model-independent or *nonparametric* in that the resampling is not based on a specific parametric data generating process.

The nonparametric bootstrap based on resampling the data presumes that the data are serially uncorrelated or independent as the resampling process naturally reshuffles the data. It is difficult to impose independence across multiple years, but easier at shorter horizons such as one year. By conditioning on economic regimes (i.e. expansions versus recessions) or by focusing on shorter time horizons, firm defaults may approach conditional independence, an issue to which we return in Sections 4.2 and 4.5.¹⁴ In addition we are able to control for some but not all of the factors relating to cross-sectional (as opposed to temporal) dependence. For instance, we restrict our analysis to U.S. firms, i.e. no government entities (municipal, state or sovereign), and no non-U.S. entities, but do not perform separate analysis by industry for reasons of sample size. By mixing industries together, the resulting bootstrap samples will likely be noisier than they would be otherwise. To the degree that such factors matter, they will be picked up by the nonparametric but ignored by the parametric bootstrap. In addition, firm business relationships

¹² A bootstrap sample is created by sampling *with replacement* from the original sample. For an excellent exposition of bootstrap methods, see Efron and Tibshirani (1993).

¹³ Andrews and Buchinsky (1997) explore the impact of non-normality on the number of bootstraps. With multimodality and fat tails the number of bootstrap replications often must be increased two or three fold relative to the Efron and Tibshirani benchmarks.

¹⁴ Similarly Christensen, Hansen and Lando (2004) perform their bootstrap simulations by dividing their sample into multi-year “stable” and “volatile” periods. See also Lopez and Saidenberg (2000) for a related discussion on evaluating credit models.

(either within or between industries) may lead to correlated defaults, a problem that we do not address here.¹⁵

Our method contrasts with the parametric bootstrap approach put forth in Christensen, Hansen and Lando (2004) who estimate an intensity-based migration matrix using all the available data and then generate many, say B , synthetic rating histories for each firm.¹⁶ These synthetic histories are generated using standard results on continuous time Markov chains under the assumption that the estimated intensities describe the true data generating process. From these B synthetic data sets they compute B intensity based migration matrices and thus are able to compute a simulation-based confidence interval from the default columns of the B migration matrices. In this way their parametric bootstrap approach may be thought of as simulation-based whereas ours is resampling-based. Below in Section 3.2 we compare the two approaches.

For our nonparametric bootstrap the unit of resampling is a realized firm-history, and since these histories are of irregular length, the total number of firm-years N^* may differ slightly across bootstraps samples. It turns out, however, that this variation is quite small. The coefficient of variation, $(\hat{\sigma}/\hat{\mu})$, of N^* across B bootstrap replications is just under 1%. Alternatively one could cut off the marginal resampled history so that N^* would be identical across all B bootstrap replications, but obviously at the cost of not preserving the basic data unit from the perspective of PD estimation, i.e. the firm-history.

¹⁵ See Egloff, Leippold and Vanini (2004) for a model of credit portfolio losses that explicitly takes such firm-level linkages into account.

¹⁶ Christensen, Hansen and Lando (2004) pay close attention to the issue of censoring in carrying out their parametric bootstrap. Naturally, all of the B synthetic histories for a given firm have the same initial state as the realized firm-history. In addition, they require that the observation period for each synthetic history be no greater than the observation period for the realized firm-history. The observation period for each firm is the time from when the firm is first observed to the time its history is right censored. Christensen et al. consider transitions to NR, the end of the observation window, and defaults as right censoring events. It is not clear that defaults should be treated as right censored since the firm might not have defaulted in some of the synthetic histories. However, the

It is worth pointing out that there will also be variation in the number of firm-years for the parametric bootstrap. Due to the possibility of default, a synthetic firm-history may have a shorter observation period than the realized firm-history to which it is paired. As a result, there will again be a distribution of total firm-years, N^* , across the synthetic data sets with the realized number of firm-years now serving as an upper bound. As before, the variation in firm-years across the synthetic data sets is quite small.

3. Comparing Confidence Intervals for *PDs*

To compare these various confidence intervals we make use of credit rating histories from Standard & Poor's where the total sample ranges from January 1, 1981 to December 31, 2002. Our data set is very similar to the data used in Bangia et al. (2002) and Jafry and Schuermann (2004). The universe of obligors is mainly large corporate institutions. In order to examine the effect of business cycles, we restrict ourselves to U.S. obligors only; there are 6,776 unique U.S. domiciled obligors in the sample. The resulting database has a total of $N^* = 50,611$ firm-years of data, excluding withdrawn ratings, and a total of 842 rated defaults, yielding an average annual default rate of 1.66% for the entire sample.¹⁷

In Table 1 we present *PD* estimates across notch-level credit ratings using the entire sample period, 1981-2002, for both the cohort and the duration based methods with the last

choice has minimal impact on the resulting confidence intervals, so we have followed Christensen et. al. for the sake of comparison.

¹⁷ These measures are based on the duration estimator so that number of firm-years includes the time that firms were rated prior to transitioning to NR within a given year. Similarly, the 842 *rated* defaults necessarily excluded cases where a firm transitions to NR and then to D. For the cohort estimator, $N = 46,814$ which is noticeably less than $N^* = 50,611$ since we no longer count firms that end a year in NR. In addition, for the cohort estimator, we observe an additional 13 defaults for a total of 855 since cases where a firm transitions to NR and then D in a single year are now counted.

column comparing the two PD estimates by grade.¹⁸ Since no defaults over one year were witnessed for firms that started the year with a AAA, AA+ or AA ratings, the cohort estimate is identically equal to zero, in contrast to the duration estimate where $PD_{AAA} = 0.02\text{bp}$, $PD_{AA+} = 0.05\text{bp}$ and $PD_{AA} = 0.93\text{bp}$.

3.1. Comparing confidence intervals for cohort PD s

We start our empirical discussion by considering the different confidence intervals for cohort PD s, both analytical as discussed in Section 2.2 and nonparametric bootstrap. These results are summarized in Table 2; all numbers are in basis points. The PD point estimates by grade are given in column four, and for each set we show the upper and lower limit of the 95% confidence interval as well as the interval length. The top panel contains first the Wald interval, obtained using (2.3), and the nonparametric bootstrap, while the bottom panel shows first the preferred analytic alternative, the Agresti-Coull, computed using (2.5), and finally the Clopper-Pearson exact interval, computed using (2.6). As expected, for each grade the Wald CI is the shortest of the four (though for single-A the nonparametric bootstrap is 0.07bp shorter), and the Clopper-Pearson interval is the longest with the exception of the top rating, AAA. Having said that, none of the four are very different with the exception of the top two grades, AAA and AA, where only one actual default (AA) was observed during the sample period. For AAA the Wald and nonparametric bootstrap intervals are degenerate, as they should be, whereas the Clopper-Pearson is more than 15bp and the Agresti-Coull more than 19bp in length. Since in the case of zero defaults PD_R^{up} depends only on N_R , the two latter methods generate wider confidence intervals for AAA than for AA due to the smaller number of AAA firms. Moreover, for all

¹⁸ All credit ratings below CCC are grouped into CCC for reasons of few observations.

methods the confidence intervals for the top three ratings are highly overlapping, implying that it is practically not possible to distinguish statistically PD_{AAA} from PD_{AA} or PD_A .

3.2. Comparing bootstrap confidence intervals for duration PD s

Next we compare confidence intervals for duration based PD s using the nonparametric and parametric bootstrap methods discussed in Section 2.3. We summarize the results in Table 3 where we report the PD point estimates in the second column, followed by the lower and upper limit of the 95% CI, as well as its length, first for the nonparametric and then for the parametric bootstrap. Both are obtained using 10,000 bootstrap replications. We notice that the differences between the two approaches are quite modest; only in the last two grades are differences more than a basis point. For the lowest grade, CCC, the nonparametric bootstrap generates a confidence interval that is one-third longer than the parametric bootstrap. The latter imposes the Markov assumption at the (re)sampling stage, an assumption which is relaxed by the former (though, to be sure, the estimation of the migration matrix itself for each bootstrap replication imposes the Markov assumption). Our evidence is consistent with results in Frydman and Schuermann (2005) who find that the CCC rating in particular is likely generated by a mixture of two distinct Markov processes, reflecting in part those firms which started with the CCC rating and those which were downgraded into it. Perhaps not surprisingly, the default probability for the first group is less than half that of the second: firms which are downgraded into CCC have a one-year PD of 67%. For firms which start with a CCC rating, their implied default probability is similar to the overall cohort estimate of around 30%. Overall, however, it seems that not much is lost by imposing the parametric assumption for the duration approach.

3.3. Comparing confidence intervals across estimators

We now go on to compare the bootstrap confidence intervals for the duration PD s with analytical and bootstrap confidence intervals for the cohort PD s. Since the three analytical CI estimates are rather similar, and following the results of Brown, Cai and DasGupta (2001), in what follows we present only the Agresti-Coull CI as the “analytical” CI. For duration PD s we present confidence intervals using the nonparametric bootstrap approach; we add the nonparametric bootstrap confidence intervals for the cohort approach when we examine whole grades below.

The results, using the entire sample period, are presented in Figure 1 in logs for easier cross-grade comparison. The top panel shows the investment grade, and the bottom panel the high yield or speculative grade PD s and their 95% confidence intervals. Here we present results at the notch level, meaning that for the AA category, for example, we show 95% bars for AA+, AA and AA-. The top and bottom grades, AAA and CCC, do not have these modifiers. The first set of bars for each pair is the interval implied by the bootstrap, centered on the duration PD estimate, the next set is the analytic Agresti-Coull interval, for the cohort PD estimate.

Several aspects of the results are striking. First, for nearly every rating, the bootstrapped confidence interval for the duration based estimate is tighter than the one implied by the Agresti-Coull interval for the cohort estimate. For the lower bound this may not be surprising. \widehat{PD}_R is small enough for the investment grades that the lower limit of the confidence interval hits the zero boundary. For example, for grade AA-, $\widehat{PD}_{AA-}^{coh} = 3.84\text{bp}$, $\hat{\sigma}_{AA-}^{coh} = 3.84\text{bp}$ so that $\widehat{PD}_{AA-}^{coh} - 1.96\hat{\sigma}_{AA-}^{coh} = -3.68\text{bp}$ which is clearly not possible. Notice also that even though no defaults were observed for AAA, AA+ and AA ratings, and hence the corresponding cohort PD

estimate is equal to zero, their corresponding Agresti-Coull intervals have different lengths since the number of firm-years differs across ratings.

Second, most of the confidence intervals, be they for the duration or cohort estimates, overlap within a rating category for investment grades. In the speculative grade range, the bottom panel in Figure 1, one is much more clearly able to distinguish default probability ranges at the notch level. For example, the bootstrapped 95% confidence intervals for the AA- through A- ratings almost completely overlap, implying that the estimated duration *PDs* for the three ratings are statistically indistinguishable *even with 22 years of data*. This is not the case for the B ratings, for example. Whether one uses intervals for duration or cohort based estimates, all the ratings, B+, B and B- are clearly separated.

At the whole grade level, default probabilities become somewhat easier to distinguish, as can be seen from Figure 2. Here we add the nonparametric bootstrap confidence intervals for the cohort estimate (see also Table 2). The first three grades are not statistically distinguishable using the cohort method with either the analytical Agresti-Coull or the nonparametric bootstrapped confidence intervals. However, using the bootstrap for duration *PDs* we can distinguish AAA from the next two ratings, but the confidence intervals for grades AA and A, whether analytic or bootstrapped, still largely overlap. Thus even at the whole grade level, dividing the investment grade into four distinct groups seems optimistic from the vantage point of *PD* estimation.

Several studies, including this one, have consistently shown that *PD* estimates using the cohort approach are higher for most grades than *PD* estimates generated from the duration approach (Lando and Skødeberg, 2002, Jafry and Schuermann, 2004, Christensen, Hansen and Lando, 2004). The exceptions are the top and bottom grades. There is no mystery for the top

grades: since no actual defaults have been observed for AAA-rated firms over the course of any one year, the cohort estimates must be identically equal to zero. The difference for the CCC rating has been discussed by Lando and Skødeberg (2002) who observe that the majority of firms default after only a brief stop in the CCC rating state. The mystery lies in the intermediate grades. However, if ratings exhibit downward persistence (firms that enter a state through a downgrade are more likely to be downgraded than other firms in the state), as shown among others by Nickell, Perraudin and Varotto (2000), Lando and Skødeberg (2002), and Bangia et al. (2002), one would expect *PDs* from the duration-based approach, which assumes that the migration process is Markov, to be downward biased. Such a bias would arise because the duration estimator ignores downward ratings momentum, and consequently underestimates the probability of a chain of successive downgrades ending in default.

One way to investigate this hypothesis is by comparing both the parametric and non-parametric bootstrap confidence intervals for the cohort and duration estimators. Recall from Sections 2.3 and 3.2 that the parametric bootstrap generates B sets of synthetic ratings histories from the estimated duration migration intensities under the Markov assumption. Using those synthetic histories, one can estimate *PDs* using either the cohort or duration approach and build up the corresponding confidence intervals. Under the null of Markov, the two sets of estimates ought to be relatively similar. By comparing the estimates and intervals obtained from the parametric bootstrap with those from the non-parametric bootstrap, we can assess how non-Markovian behavior contributes to the observed differences between the two estimators. We perform this comparison in Table 4. In this table, only the parametric cohort results in the top panel are new; the nonparametric cohort results are already in Table 2, and the duration-based results in Table 3. As a reference point we also present the *PD* point estimates using the two

approaches. Note again that for all categories except for the AAA and CCC ratings, the cohort point estimates exceed the duration point estimates.

Using the nonparametric bootstrap shown in the top panel, the 95% confidence intervals only overlap for the AA rating. This serves to highlight how differently the two estimators perform when confronted with data generated by the actual ratings migration process. However, using the parametric bootstrap, which assumes that the data is generated by a time-homogenous Markov chain, not only do the 95% confidence intervals overlap for every rating grade, but the mean estimates across 10,000 bootstraps, \overline{PD}_R , are very close.¹⁹ For instance, \overline{PD}_{AA} is 0.54bp for duration and 0.56bp for cohort. They diverge more at the lower end, with cohort generating higher PD estimates (e.g. for \overline{PD}_B , 500.42bp versus 470.48bp), but each mean PD estimate is contained in the other's 95% confidence interval. Thus, it appears that the differences between the empirical point estimates, especially for the middle grades, can be explained, at least in part, by the violation of the Markov assumption.

4. Using Confidence Intervals for Policy-Relevant Analysis

We now proceed to illustrate how the confidence intervals and more generally the nonparametric bootstrapping techniques introduced above can be used to conduct policy-relevant analysis.

¹⁹ Because only the duration approach can properly account for censored observations, we would expect to see some differences in \overline{PD}_R between the two approaches.

4.1. Can we tell if *PDs* are monotonic?

At a minimum, a rating system should be ordinally consistent or monotonic meaning that *PDs* should be increasing as one moves from higher to lower ratings.²⁰ Returning to Table 1, notice that the notch-level *point estimates* for both duration and cohort *PDs* are not even monotonically increasing. To evaluate the issue of monotonicity more formally, we perform one-tailed tests using the bootstrap results along the following lines. For ratings $k < j$, where rating k is of better credit quality (e.g. A+) than j (e.g. A), we compute the one-tailed test

$$(4.1) \quad \Pr[PD_j(\Delta t) < PD_k(\Delta t)] = \alpha\%.$$

In Table 5 we report the fraction of replications for which the duration based $\widehat{PD}_j(\Delta t) < \widehat{PD}_k(\Delta t)$ over $B = 10,000$ (nonparametric) bootstrap replications; this should be no greater than $\alpha\%$. We find, in fact, that the nominal p-value often exceeds 5% for the investment grades. This is the case, for instance, with the first test, $\Pr[PD_{AA+} < PD_{AAA}] = 9.16\%$. The nominal p-value is especially poor for the range of AA ratings; see Section 4.4 for more discussion on behavior of this particular grade. Even the BBB grades have trouble meeting this monotonicity criterion. For example, $\Pr[PD_{BBB} < PD_{BBB+}] = 6.72\%$ and $\Pr[PD_{BBB-} < PD_{BBB}] = 31.90\%$. Only at the non-investment grade end of the rating spectrum can we reliably state that notch level *PDs* are indeed monotonically increasing. Similar calculations for grade levels *PDs* to those shown in Table 5 reveal that the only violation of monotonicity is between AA and A.

²⁰ It is quite difficult to see how a set of estimated *PDs* that failed monotonicity could be consistently employed in either regulatory, risk management, or pricing applications.

4.2. Common factors: recession vs. expansion

The analysis above made the arguably unrealistic assumption that all rating histories from the whole 22-year sample period were draws from the same *iid* process. However, it is likely that systematic risk factors affect all firms within a year. A simple approach may be to condition on the state of the economy, say expansion and recession, so that defaults are *conditionally* independent. Nickell, Perraudin and Varotto (2000) were perhaps the first to formally test for business cycle dependence in credit rating dynamics, and they did so using an ordered probit model. Our goal is to examine the degree of divergence between the small sample \widehat{PD}_R distributions, conditioning on the state of the business cycle. For instance, if monotonicity of estimated *PDs* is often violated in the unconditional estimates, does conditioning on the business cycle help to differentiate *PD* estimates, as previous research would suggest?

Using the business cycle dates from the NBER,²¹ in the 22 years of our sample only 1982 was a “pure” recession year. The years 1981, 1990, 1991 and 2001 experienced a mix of recession and expansion states. All other years are “pure” expansion years. The NBER delineates peaks and troughs of the business cycle at monthly frequencies. Since rating histories are available at a daily frequency, insofar as rating changes are dated at that level, we pick the middle of a month as the regime change from expansion to recession or vice versa and re-estimate duration *PDs* on this basis, i.e. using “recession days” and “expansion days.”

We repeat the monotonicity experiment as above, but this time we compute (nonparametric) bootstrapped p-values separately for expansions and recessions. The results are summarized in Table 6 where we repeat in the first column labeled 1981-2002 the p-values for the whole sample range. Conditioning on the state of the economy appears to help in

differentiating *PDs* in adjacent credit ratings. Looking at the first column, half of the 16 bootstrapped p-values exceed 5%, meaning that we would have to reject (at the 95% level) that the two adjacent *PDs* are monotonic (or ordinally consistent). The proportion is the same in expansions, but conditioning on recessions reduces this proportion to 25% (4 out of 16). For example, the unconditional $\Pr[PD_{A-} < PD_{A+}] = 17.96\%$, and during an expansion it is even worse at 19.35%, but it drops to less than 0.01% during a recession. A similar pattern can be observed for the next pair, $\Pr[PD_{A-} < PD_A]$. Interestingly there are some instances when monotonicity is violated in a recession but not in an expansion: $\Pr[PD_{BBB+} < PD_{A-} | \text{expansion}] = 0.87\%$ and $\Pr[PD_{BBB+} < PD_{A-} | \text{recession}] = 28.00\%$. Speculative grade ratings are monotonic in both recessions and expansions, implying that these firm ratings are more business cycle sensitive than their investment grade counterparts.

4.3. Empirical densities of *PDs*

It may also be of interest to see how much the empirical (bootstrapped) *PD* distributions for recession and expansion periods overlap. Although the rating agencies strive to achieve a “cycle neutral” credit rating, the speculative grades tend to be more sensitive to business cycle conditions, both empirically and by design of the rating agencies (Moody’s, 1999). Thus we would expect that the conditional *PD* distributions would be farther apart for speculative than for investment grades. This is seen quite clearly in Figure 3 where we include the unconditional density for each grade for easy comparison.

²¹ See <http://www.nber.com/cycles/cyclesmain.html>.

For speculative grades the recession and expansion densities show very little overlap as expected, in contrast to investment grade PDs . The multi-modality in BBB and AA ratings is a result of default clustering from the bootstrap; see also the discussion in Section 4.4. The unconditional and expansion densities are very close, especially for investment grade. This makes sense since we have been in an expansion most of the time (88%) since 1981. As a result the distributions for recessions are also wider than for expansions. For the A through AAA ratings, it seems that the recession densities are to the left of the expansion densities, implying that defaults may actually be lower in recessions. Overall we find that speculative grade PDs are more business cycle sensitive than the investment grades which is consistent with the rating agencies' own view.

Finally, it is striking just how close to normal most of the \widehat{PD}_R densities appear to be, especially for the speculative grades. In Figure 4 we display kernel density plots of the bootstrapped default probabilities using the nonparametric bootstrap, overlaid against a normal density with the same mean and variance (as the nonparametric bootstrap) as a visual guide. In addition we overlay the \widehat{PD}_R densities obtained using the parametric bootstrap, although it is visually difficult to distinguish the densities of the two bootstrap approaches. The AA grade is a glaring exception to the general pattern, and we discuss this below. The proximity to the normal density is perhaps especially striking for the high credit quality grades since their estimated default probabilities are so low. The means of our estimates of annual PD_R across the 10,000 nonparametric bootstrap replications are 0.03bp for AAA, 0.54bp for AA, 0.87 for A and 10.44 for BBB.²² Of course, \widehat{PD}_R can not fall below 0, so the density has a natural left boundary (and

²² Note those means are not necessarily identical to the point estimates (see Table 1) since the densities are slightly skewed.

right at 1, of course) to which the investment grade densities are very close indeed. One would expect probability mass to pile up against that boundary, and we see this in the slight right skew of the investment grade densities, but this skew is indeed slight, even for AAA and A whose estimated PD s are under a basis point.

Our initial conclusions about the differences between the parametric and nonparametric bootstrap from Section 3.2 are confirmed with the charts in Figure 4 (see also Table 3): for most grades, the densities are quite close. They do diverge for the lowest two ratings, especially for the CCC rating, where the parametric bootstrap generates a narrower density of \widehat{PD}_R ; recall from Table 3 that the parametric bootstrap yields a 95% confidence interval which is one-third shorter than the nonparametric bootstrap.

4.4. Multi-modality of PD_{AA}

Since we never observe a direct transition from AA+ or AA to default, our estimated PD for AA under the duration approach primarily reflects the probability of experiencing a sequence of successive downgrades that ends in default. Thus, transitions far from the diagonal, such as downgrades from AA to B, play a key role in determining estimated PD s for investment grade ratings. It turns out that the multi-modal kernel density plot for AA is being driven by a single firm, TICOR Mortgage Insurance, which transitioned from AA to CCC in December of 1985. In Figure 5 we display kernel densities for \widehat{PD}_{AA} (nonparametric bootstrap) estimated with and without TICOR. The modes in the density plot correspond to the number of times TICOR appears in the bootstrap sample and hence the number of observed AA to CCC transitions. Note also that this multi-modality is not an artifact of the nonparametric bootstrap as the same pattern is exhibited by the parametric bootstrap density in the top right chart of Figure 4.

This pattern is not a peculiarity specific to the duration method, but is due to the general difficulty of estimating probabilities for such rare events. For instance, there is a single instance of a firm beginning a year in AA- and ending the year in default (General American Life Insurance Co. in 1999) and bootstrapped \widehat{PD}_{AA} 's from the cohort approach show a similar type of clustering. More broadly, it points to limitations of the bootstrap approach for computing confidence intervals as the bootstrap samples will never contain any downgrades worse than those observed in the real dataset, and thus imaginable or likely events yet to happen historically are simply not taken into account.

4.5. Comparing conditional and unconditional PD s

We now examine the effects of varying T , the length of the estimation window, on grade-level PD estimates, an issue particularly relevant to practitioners. There is a trade-off between parameter uncertainty and heterogeneity, proxied here simply by economic regime. The longer T , the more accurate the estimates \widehat{PD}_R are likely to be. However, one will invariably mix recessions (higher average \widehat{PD}_R) and expansions (lower average \widehat{PD}_R). If one is interested in a long run or unconditional estimate, one would explicitly be interested in mixing these regimes. Since the average post-war recession is slightly more than one year, and since the most recent two recessions have each lasted less than one year, it seems reasonable to impose conditional independence over a one year period. Thus, comparing conditional PD s using rolling one-year windows to the unconditional (i.e. full sample length) estimate seems reasonable.

In Figure 6 we compare duration (top panel) and cohort based (bottom panel) PD estimates using a one-year rolling estimation window by grade with the unconditional estimate

(reported in log basis points, bp).²³ The CCC chart is repeated at the end in levels. The 95% confidence interval for both approaches are computed using the nonparametric bootstrap. Focusing first on the top panel, for most grades we are able to reliably determine that the annual PD using just one year of data is significantly different from the estimated long-run average for a surprisingly large number of years. For instance, with 95% confidence we can say that PD_B was above its estimated long-run average in 5 of the 22 years and below its long run average in 9 of 22 years. Specifically, we see that PDs for BB and B were significantly above their estimated long-run averages during 1990-1991 (there was a recession from July 1990 to March 1991), while all grades *except* for AAA and AA were above their unconditional levels in 2001 (the most recent recession lasted from March to November 2001). We also point out that during the mid-1990s conditional PDs were below their estimated unconditional levels across most ratings, consistent with the business cycle.

Looking at the top panel of Figure 6 we note that for the top two grades (and to some extent A as well) there seems to be a regime shift around 1989. Prior to that year the conditional PD estimates were occasionally above the long run average, but since then the entire 95% interval has been below with the single exception of AA in 2002. As discussed in Section 4.4, estimated PDs for these grades are significantly impacted by the number of transitions far from the diagonal, particularly by downgrades of three or more grade levels, e.g. AAA \rightarrow BBB. However, large migrations like that have become extremely rare since 1989. This observation may be consistent with an increasing desire on the part of the rating agencies to limit ratings volatility and move towards more gradual rating adjustments (Hamilton and Cantor, 2004,

²³ In this discussion we abstract from sampling variation of the unconditional PD estimate.

Altman and Rijken, 2004). However, we cannot rule out the possibility that AAA and AA firms were simply subject to larger shocks during the earlier period.

The bottom panel in Figure 6 shows the one-year cohort estimates with their 95% confidence intervals based also on the nonparametric bootstrap. The information loss incurred by applying the cohort instead of duration based method is again striking. No defaults from AAA occurred at all in these 22 years, and only one default from AA (specifically AA- in 1999). In addition, we note that it is more difficult to distinguish the conditional from the unconditional *PD* using this estimation method.

The New Basel Accord stipulates that banks have at least five years of data on hand in order to be eligible for the advanced IRB approach (FRB, 2003). In Figure 7 we show duration based \widehat{PD}_s (in logs) by rating with five-year rolling estimation windows against the unconditional estimate, i.e. using the entire sample length. Again the CCC graph is repeated at the end in levels. In each case we accompany the yearly point estimates with their 95% (nonparametric) bootstrapped confidence intervals. The unconditional estimates naturally are just a straight line across time (the x-axis). Even though we are mixing recession and expansions, nonetheless with the additional data from the wider estimation window we are still able to distinguish the five-year conditional *PD* from the unconditional estimate for many of the sub-periods. The regime shift for the highest grades mentioned above is even more pronounced in this chart, although with the wider estimation window it now appears around 1993.

5. Concluding remarks

Using credit rating histories from S&P, we estimate probabilities of default using two estimation techniques, cohort and duration (or intensity), and compare confidence intervals based

on both analytical as well as parametric and nonparametric bootstrap approaches. For the duration based estimates, we find that confidence intervals from bootstrapping are significantly tighter than either the bootstrapped or standard analytical intervals for cohort based estimates, which reflects the greater efficiency of the duration approach. However, we also show how the large differences between the point estimates and associated intervals of the cohort and duration estimators are consistent with downward persistence or momentum, a clear violation of the underlying Markov assumption needed for the duration estimator. But even those tighter bootstrapped confidence intervals overlap considerably for investment grades, making it difficult if not impossible to distinguish them. Moreover, our results indicate that the lower bound of 0.03% imposed on any *PD* used to compute regulatory capital by the New Basel Accord is above the upper limit of the bootstrapped 95% confidence interval for the top three rating grades, AAA through A using the duration approach, but within the 95% confidence interval of the AA rating using the cohort approach.

We next consider the effect of varying the number of grades in the rating system. We propose that rating systems should satisfy monotonicity and test this requirement formally. Using notch level *PD* estimates from the duration approach, we cannot conclude that monotonicity holds for most investment grade ratings, although this criterion is generally met for speculative grade ratings. Conditioning on the state of the business cycle helps: it is easier to distinguish adjacent *PDs* in recessions than in expansions.

We also consider the effects of varying the length of the estimation window to consider conditional, i.e. time-varying, *PD* estimates. We compute bootstrapped confidence intervals for intensity-based *PDs* estimated using one and five-year rolling windows, allowing for comparisons between *PDs* estimated over these shorter intervals and their long-run averages.

For both the one and five-year windows, we are able to determine that the conditional *PD* differs from the unconditional estimate for a large number of years.

Our findings have significant implications for regulators and credit risk practitioners alike. In a survey of internal rating systems at the fifty largest U.S. banking organizations, Treacy and Carey (2000) report that the median banking organization had five pass grades with a range from two to the low twenties. The authors also report that many banks expressed interest in increasing the number of internal grades either through the addition of \pm modifiers or by splitting riskier grades while leaving low-risk grades intact. Our results suggest that the latter approach is to be preferred from the vantage point of *PD* estimation. The addition of \pm modifiers to existing low-risk ratings could result in non-monotonic *PD* estimates, whereas it appears likely that meaningful estimates for additional high-risk grades could be obtained. To be sure, our analysis and hence the conclusions are limited to the one-year horizon. Although this is the standard horizon used by industry practitioners, regulators and academics in the analysis of credit risk, further work is required to extend the analysis to longer horizons which are relevant market participants such as buy-and-hold investors.

References

- Aalen, O.O. and S. Johansen, 1978, "An Empirical Transition Matrix for Nonhomogeneous Markov Chains Based on Censored Observations," *Scandinavian Journal of Statistics* 5, 141-150.
- Agresti, A. and B.A. Coull, 1998, "Approximate is Better Than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician* 52, 119-126.
- Altman, E.I., 1968, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance* 23, 589-609.
- Altman, E.I. and D.L. Kao, 1992, "Rating Drift of High Yield Bonds," *Journal of Fixed Income*, March, 15-20.
- Altman, E.I. and H.A. Rijken, 2004, "How Rating Agencies Achieve Rating Stability," *Journal of Banking & Finance* 28, 2679-2714.
- Andrews, D.W.K. and M. Buchinsky, 1997, "On the Number of Bootstrap Repetitions for Bootstrap Standard Errors, Confidence Intervals, and Tests," Cowles Foundation Paper 1141R.
- Bangia, A., F.X. Diebold, A. Kronimus, C. Schagen and T. Schuermann, 2002, "Ratings Migration and the Business Cycle, With Applications to Credit Portfolio Stress Testing," *Journal of Banking & Finance* 26, 445-474.
- Basel Committee on Banking Supervision, 2001a, *The New Basel Capital Accord*, <http://www.bis.org/publ/bcbsca.htm>, January.
- Basel Committee on Banking Supervision, 2001b, *The Internal Ratings Based Approach*, <http://www.bis.org/publ/bcbsca.htm>, May.
- Basel Committee on Banking Supervision, 2004, *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*, <http://www.bis.org/publ/bcbs107.htm>, June.
- Brown, L.D., T. Cai and A. Dasgupta, 2001, "Interval Estimation for a Binomial Proportion," *Statistical Science* 16, 101-133.
- Cantor, R. and E. Falkenstein, 2001, "Testing for Rating Consistency in Annual Default Rates," *Journal of Fixed Income*, September, 36-51.
- Christensen, J. E. Hansen and D. Lando, 2004, "Confidence Sets for Continuous-Time Rating Transition Probabilities," *Journal of Banking & Finance* 28, 2575-2602.
- Crouhy, M., D. Galai, and R. Mark (2001), *Risk Management*, New York, NY: McGraw Hill.
- Efron, B. and R.J. Tibshirani, 1993, *An Introduction to the Bootstrap*, New York, NY: Chapman & Hall.
- Egloff, D., M. Leippold, and P. Vanini, 2004, "A Simple Model of Credit Contagion," EFA 2004 Maastricht Meetings Paper No. 1142; EFMA 2004 Basel Meetings Paper

- Federal Reserve Board, 2003, "Supervisory Guidance on Internal Ratings-Based Systems for Corporate Credit," Attachment 2 in <http://www.federalreserve.gov/boarddocs/meetings/2003/20030711/attachment.pdf>.
- Fledelius, P., D. Lando and J. P. Nielsen, 2004, "Non-Parametric Analysis of Rating Transition and Default Data," *Journal of Investment Management* 2, 71-85.
- Frydman, Halina and Til Schuermann, 2005, "Credit Rating Dynamics and Markov Mixture Models," Wharton Financial Institutions Center Working Paper #04-17.
- Hamilton, D. and R. Cantor, 2004, "Rating Transitions and Defaults Conditional on Watchlist, Outlook and Rating History," *Special Comment*, Moody's Investor Service, New York.
- Hillegeist, Stephen A., Elizabeth K. Keating, Donald P. Cram and Kyle G. Lundsted, 2004, "Assessing the Probability of Bankruptcy," *Review of Accounting Studies* 9 (1), 5-34.
- Jafry, Yafry and Til Schuermann, 2004, "Measurement, Estimation and Comparison of Credit Migration Matrices," *Journal of Banking & Finance* 28, 2603-2639.
- Lando, D. and T. Skødeberg, 2002, "Analyzing Ratings Transitions and Rating Drift with Continuous Observations," *Journal of Banking & Finance*, 26, 423-444.
- Lopez, J.A. and M. Saidenberg, 2000, "Evaluating Credit Risk Models," *Journal of Banking & Finance* 24, 151-165.
- Marrison, C. (2002), *The Fundamentals of Risk Management*, New York: McGraw Hill.
- Moody's Investors Services (1999), *Rating Methodology: The Evolving Meanings of Moody's Bond Ratings*.
- Nickell, P. W. Perraudin and S. Varotto, 2000, "Stability of Rating Transitions," *Journal of Banking & Finance* 24, 203-227.
- Pluto, Katja and Dirk Tasche, 2005, "Estimating Probabilities of Default for Low Default Portfolios," working paper; available at http://www.defaultrisk.com/pp_score_45.htm.
- Schuermann, Til, 2004, "What Do We Know About Loss Given Default?" ch. 9 in David Shimko (ed.) *Credit Risk: Models and Management*, 2nd Edition, London, UK: Risk Books.
- Shumway, Tyler, 2001, "Forecasting Bankruptcy more Accurately: A Simple Hazard Model," *Journal of Business* 74, 101-124.
- Stein, Roger M., 2003, "Are the Probabilities Right?" Moody's | KMV Technical Report #030124.
- Treacy, W.F. and M. Carey, 2000, "Credit Risk Rating Systems at Large US Banks," *Journal of Banking & Finance* 24, 167-201.
- Trück, Stefan and Svetlozar T. Rachev, 2005, "Credit Portfolio Risk and PD Confidence Sets through the Business Cycle," University of Karlsruhe Working Paper.
- Vos, P.W. and S. Hudson, 2005, "Evaluation Criteria for Discrete Confidence Intervals: Beyond Coverage and Length," *The American Statistician* 59, 137-142.

Rating Categories	Cohort	Duration	% $\frac{\text{Cohort}}{\text{Duration}}$
AAA	0.00	0.02	0.0%
AA+	0.00	0.05	0.0%
AA	0.00	0.93	0.0%
AA-	3.84	0.44	863.4%
A+	5.20	0.46	1130.0%
A	6.99	0.84	834.2%
A-	5.99	1.00	597.7%
BBB+	31.37	4.67	671.1%
BBB	36.23	11.65	311.0%
BBB-	40.12	14.53	276.1%
BB+	55.01	33.01	166.7%
BB	116.33	45.64	254.9%
BB-	207.18	88.51	234.1%
B+	349.80	175.41	199.4%
B	982.01	758.33	129.5%
B-	1,430.16	1,343.30	106.5%
CCC	2,853.54	4,249.04	67.2%

Table 1: Estimated annual probabilities of default (*PDs*) across methods. All numbers in basis points. The CCC rating category includes CC and C rated obligors due to small sample size.²⁴ The final column compares the cohort with the duration point estimates. If the entry exceeds 100%, then the cohort *PD* exceeds the duration *PD* estimate, and vice versa. S&P rated U.S. obligors, 1981-2002.

²⁴ This table is similar to Table 2 in Jafry and Schuermann (2004) who report *PD* estimates for all (global) S&P rated obligors.

Rating	N_R	$N_{R,D}$	\widehat{PD}_R	<u>Wald 95% CI</u>			<u>Nonparametric Bootstrap 95% CI</u>		
				Lower	Upper	Length	Lower	Upper	Length
AAA	2,417	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00
AA	6,690	1	1.49	0.00	4.42	4.42	0.00	4.68	4.68
A	12,907	8	6.20	1.90	10.49	8.59	2.31	10.84	8.53
BBB	9,794	35	35.74	23.92	47.55	23.64	24.53	47.92	23.38
BB	6,681	94	140.70	112.46	168.94	56.48	113.73	170.69	56.96
B	7,533	491	651.80	596.06	707.54	111.48	597.51	706.77	109.26
CCC	792	226	2,853.54	2,539.03	3,168.04	629.00	2,500.00	3,234.31	734.31

Rating	N_R	$N_{R,D}$	\widehat{PD}_R	<u>Agresti-Coull 95% CI</u>			<u>Clopper-Pearson 95% CI</u>		
				Lower	Upper	Length	Lower	Upper	Length
AAA	2,417	0	0.00	0.00	19.15	19.15	0.00	15.25	15.25
AA	6,690	1	1.49	0.00	9.37	9.37	0.04	8.33	8.29
A	12,907	8	6.20	2.90	12.46	9.56	2.68	12.21	9.53
BBB	9,794	35	35.74	25.55	49.81	24.26	24.90	49.67	24.76
BB	6,681	94	140.70	114.98	172.00	57.02	113.84	171.91	58.06
B	7,533	491	651.80	598.20	709.83	111.63	597.08	709.91	112.82
CCC	792	226	2,853.54	2,549.81	3,177.98	628.17	2,541.20	3,181.94	640.74

Table 2: Four confidence intervals for PDs obtained using the cohort approach. All numbers in basis points. Wald confidence interval (CI) computed using (2.3), the nonparametric bootstrap is discussed in Section 2.3, Agresti-Coull 95% CI computed using (2.5), and Clopper-Pearson 95% CI computed using (2.6). S&P rated U.S. obligors, 1981-2002.

Rating Category	\widehat{PD}_R	Nonparametric 95% CI			Parametric 95% CI		
		Lower	Upper	Length	Lower	Upper	Length
AAA	0.03	0.01	0.07	0.06	0.01	0.07	0.06
AA	0.54	0.11	1.32	1.20	0.11	1.34	1.23
A	0.86	0.55	1.32	0.77	0.54	1.32	0.78
BBB	10.43	6.09	15.60	9.51	6.05	15.79	9.75
BB	62.62	51.11	75.44	24.34	51.15	75.59	24.44
B	470.19	430.12	511.30	81.18	431.66	510.92	79.26
CCC	4,228.42	3,879.11	4,597.62	718.51	3965.75	4500.40	534.65

Table 3: Duration based PD estimates with 95% confidence intervals obtained with nonparametric and parametric bootstrap, using in each case 10,000 bootstrap replications. All numbers in basis points. S&P rated U.S. obligors, 1981-2002.

Nonparametric Bootstrap										
Rating	Point Estimates		Duration				Cohort			
	Duration	Cohort	Mean Estimate	95% CI		Length	Mean Estimate	95% CI		Length
				Lower	Upper			Lower	Upper	
AAA	0.03	0.00	0.03	0.01	0.07	0.06	0.00	0.00	0.00	0.00
AA	0.54	1.49	0.54	0.11	1.32	1.20	1.50	0.00	4.68	4.68
A	0.86	6.20	0.86	0.55	1.32	0.77	6.21	2.31	10.84	8.53
BBB	10.43	35.74	10.44	6.09	15.60	9.51	35.65	24.53	47.92	23.38
BB	62.62	140.70	62.70	51.11	75.44	24.34	140.88	113.73	170.69	56.96
B	470.19	651.80	470.49	430.12	511.30	81.18	652.12	597.51	706.77	109.26
CCC	4,228.42	2,853.54	4,230.35	3879.11	4597.62	718.51	2,854.51	2,500.00	3,234.31	734.31

Parametric Bootstrap										
Rating	Point Estimates		Duration				Cohort			
	Duration	Cohort	Mean Estimate	95% CI		Length	Mean Estimate	95% CI		Length
				Lower	Upper			Lower	Upper	
AAA	0.03	0.00	0.03	0.01	0.07	0.06	0.03	0.00	0.00	0.00
AA	0.54	1.49	0.54	0.11	1.34	1.23	0.56	0.00	3.24	3.24
A	0.86	6.20	0.86	0.54	1.32	0.78	0.88	0.00	2.52	2.52
BBB	10.43	35.74	10.43	6.05	15.79	9.75	10.75	4.36	17.85	13.49
BB	62.62	140.70	62.66	51.15	75.59	24.44	65.27	46.02	85.83	39.81
B	470.19	651.80	470.48	431.66	510.92	79.26	500.42	449.35	554.08	104.73
CCC	4,228.42	2,853.54	4,230.07	3,965.75	4,500.40	534.65	4,453.80	4,110.10	4,813.10	703.00

Table 4: Side by side comparison of parametric and nonparametric bootstrap confidence intervals, by credit rating grade, for two estimation approaches: cohort and duration. All numbers in basis points. Top panel shows the parametric bootstrap, bottom panel the nonparametric bootstrap. Columns two and three contain the point estimates from actual data using the two approaches. S&P rated U.S. obligors, 1981-2002.

	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	BBB-	BB+	BB	BB-
AAA	x	9.16%	0.23%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
AA+	x	x	0.48%	0.50%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
AA	x	x	x	69.63%	68.47%	50.11%	42.23%	4.44%	0.01%	0.00%	0.00%	0.00%	0.00%
AA-	x	x	x	x	46.33%	18.41%	6.04%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%
A+	x	x	x	x	x	17.96%	4.57%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
A	x	x	x	x	x	x	33.99%	1.37%	0.00%	0.00%	0.00%	0.00%	0.00%
A-	x	x	x	x	x	x	x	0.63%	0.00%	0.00%	0.00%	0.00%	0.00%
BBB+	x	x	x	x	x	x	x	x	6.72%	2.04%	0.00%	0.00%	0.00%
BBB	x	x	x	x	x	x	x	x	x	31.90%	0.63%	0.01%	0.00%
BBB-	x	x	x	x	x	x	x	x	x	x	1.75%	0.02%	0.00%
BB+	x	x	x	x	x	x	x	x	x	x	x	13.82%	0.00%
BB	x	x	x	x	x	x	x	x	x	x	x	x	0.03%
BB-	x	x	x	x	x	x	x	x	x	x	x	x	x
B+	x	x	x	x	x	x	x	x	x	x	x	x	x
B	x	x	x	x	x	x	x	x	x	x	x	x	x
B-	x	x	x	x	x	x	x	x	x	x	x	x	x
CCC	x	x	x	x	x	x	x	x	x	x	x	x	x

Table 5: Bootstrapped p-values. Proportion when $\widehat{PD}_{row} > \widehat{PD}_{col}$ across $B = 10,000$ bootstrap replications. For example, taking the A+ row, the first entry is 17.96% which is the proportion of replications where $\widehat{PD}_{A+} > \widehat{PD}_A$. The columns for B+ to CCC are omitted since the p-values were less than 0.0001.

Rating Category	1981-2002	Expansion	Recession
AA+ minus AAA	9.16%	12.45%	1.34%*
AA minus AA+	0.48%**	0.35%**	30.75%
AA- minus AA	69.63%	70.59%	1.23%*
A+ minus AA-	46.33%	45.05%	59.28%
A minus A+	17.96%	19.35%	0.00%**
A- minus A	33.99%	39.73%	0.19%**
BBB+ minus A-	0.63%**	0.87%**	28.00%
BBB minus BBB+	6.72%	12.33%	0.04%**
BBB- minus BBB	31.90%	21.78%	65.68%
BB+ minus BBB-	1.75%*	2.57%*	0.03%**
BB minus BB+	13.82%	25.84%	0.00%**
BB- minus BB	0.03%**	1.50%*	0.02%**
B+ minus BB-	0.00%**	0.00%**	2.34%**
B minus B+	0.00%**	0.00%**	0.00%**
B- minus B	0.00%**	0.00%**	0.13%**
CCC minus B-	0.00%**	0.00%**	0.00%**

* and ** denote one-tailed significance of 5% and 1% respectively.

Table 6: Testing for monotonicity. % of bootstrap replications for rating $k < j$, where k is of better credit quality (e.g. A+) than j (e.g. A) in which $\widehat{PD}_j < \widehat{PD}_k$. S&P credit rating histories of U.S. firms from 1981-2002. PD s are taken from the last column of the migration matrix estimated using the parametric intensity approach. The number of bootstrap replications $B = 10,000$.

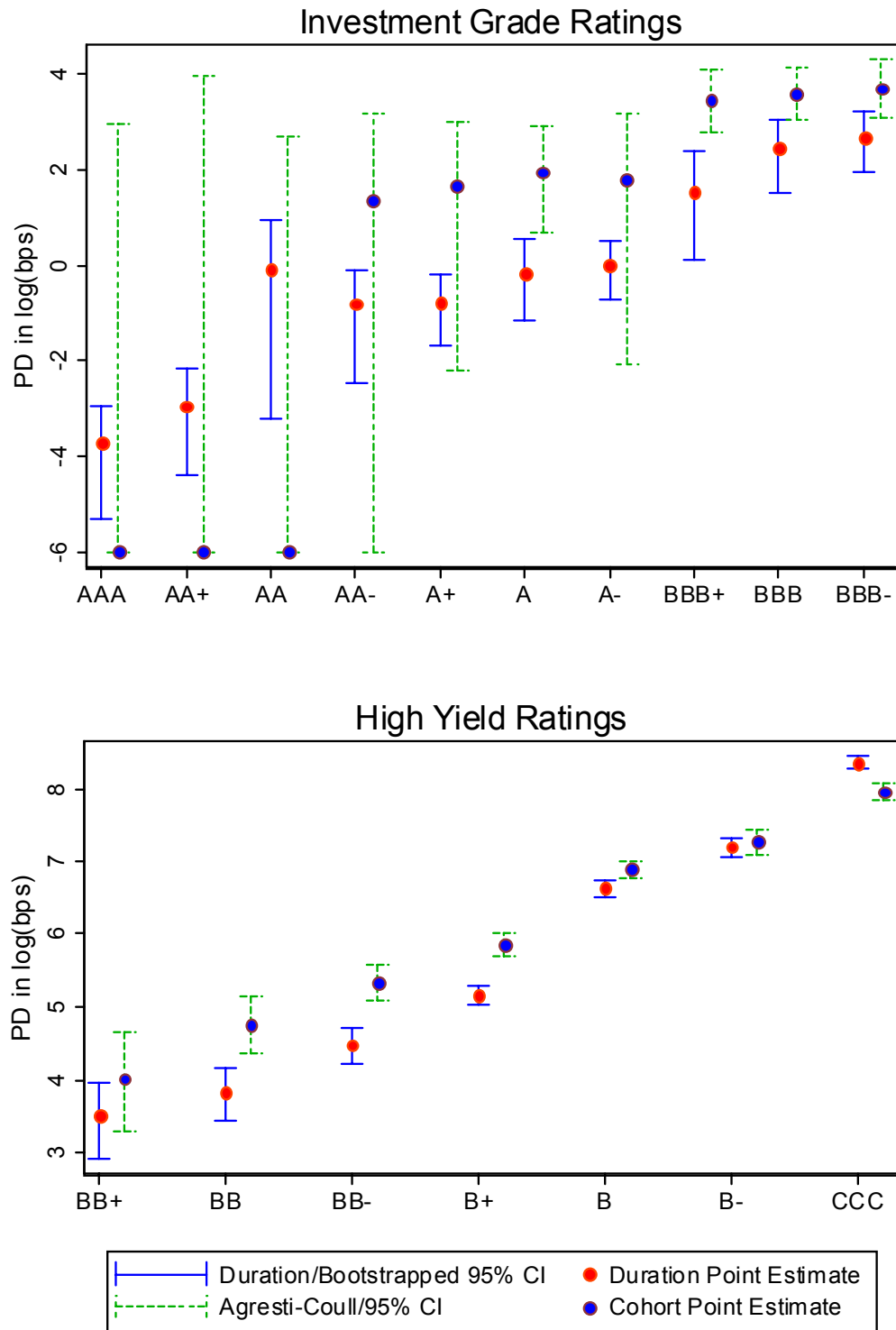


Figure 1: Comparing (nonparametric) bootstrapped 95% confidence intervals for notch level probabilities of default (PD s) obtained using the duration methodology with analytical (Agresti-Coull) confidence intervals for PD s obtained using the cohort approach. PD s are estimated using S&P credit rating histories of U.S. firms from 1981-2002. Note that the results are presented in $\log(PD)$ for easier comparison.

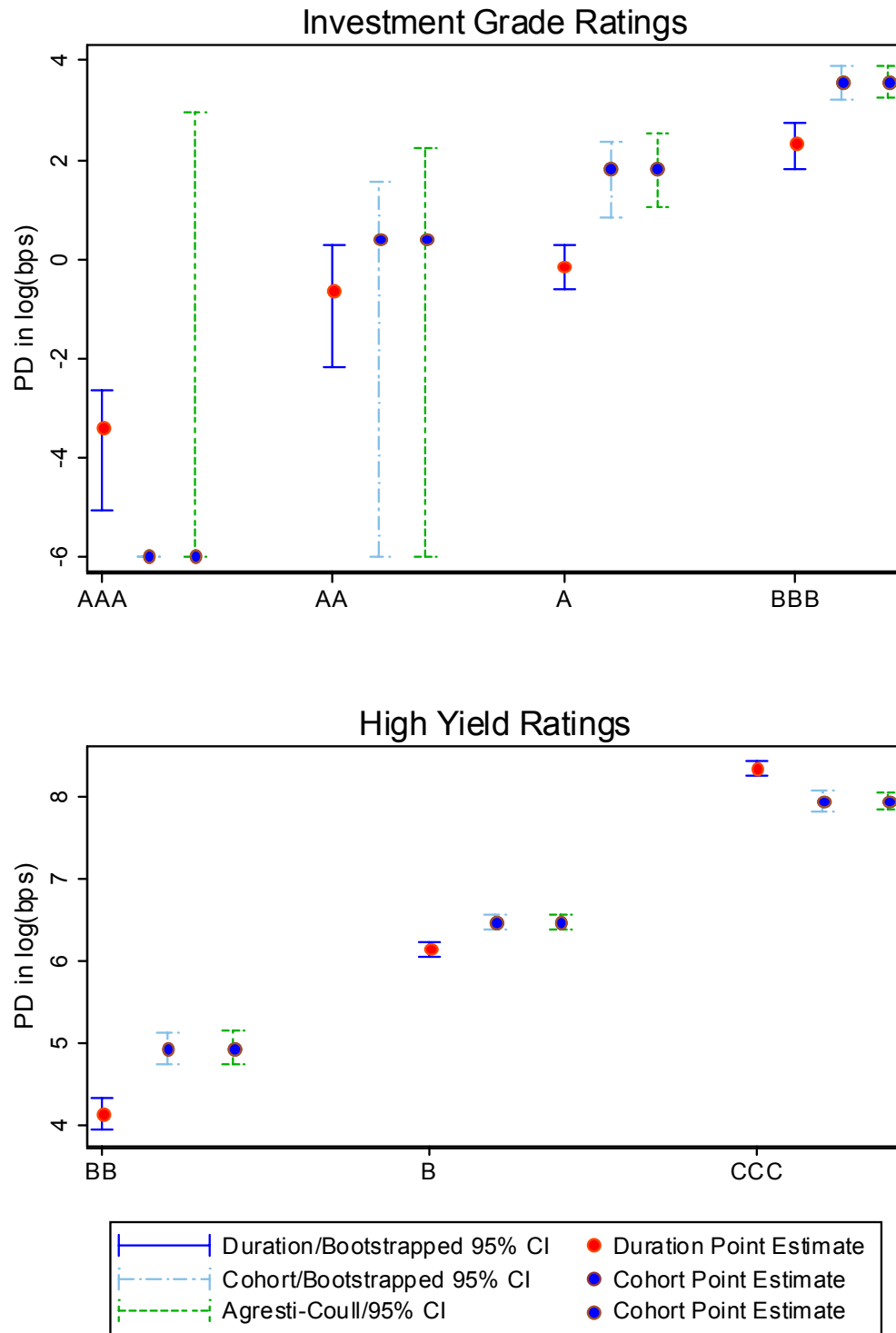


Figure 2: Comparing (nonparametric) bootstrapped 95% confidence intervals for grade level probabilities of default (PDs) obtained using the duration methodology with bootstrapped and analytical (Agresti-Coull) confidence intervals for PDs obtained using the cohort approach. PDs are estimated using S&P credit rating histories of U.S. firms from 1981-2002. Note that the results are presented in $\log(PD)$ for easier comparison.

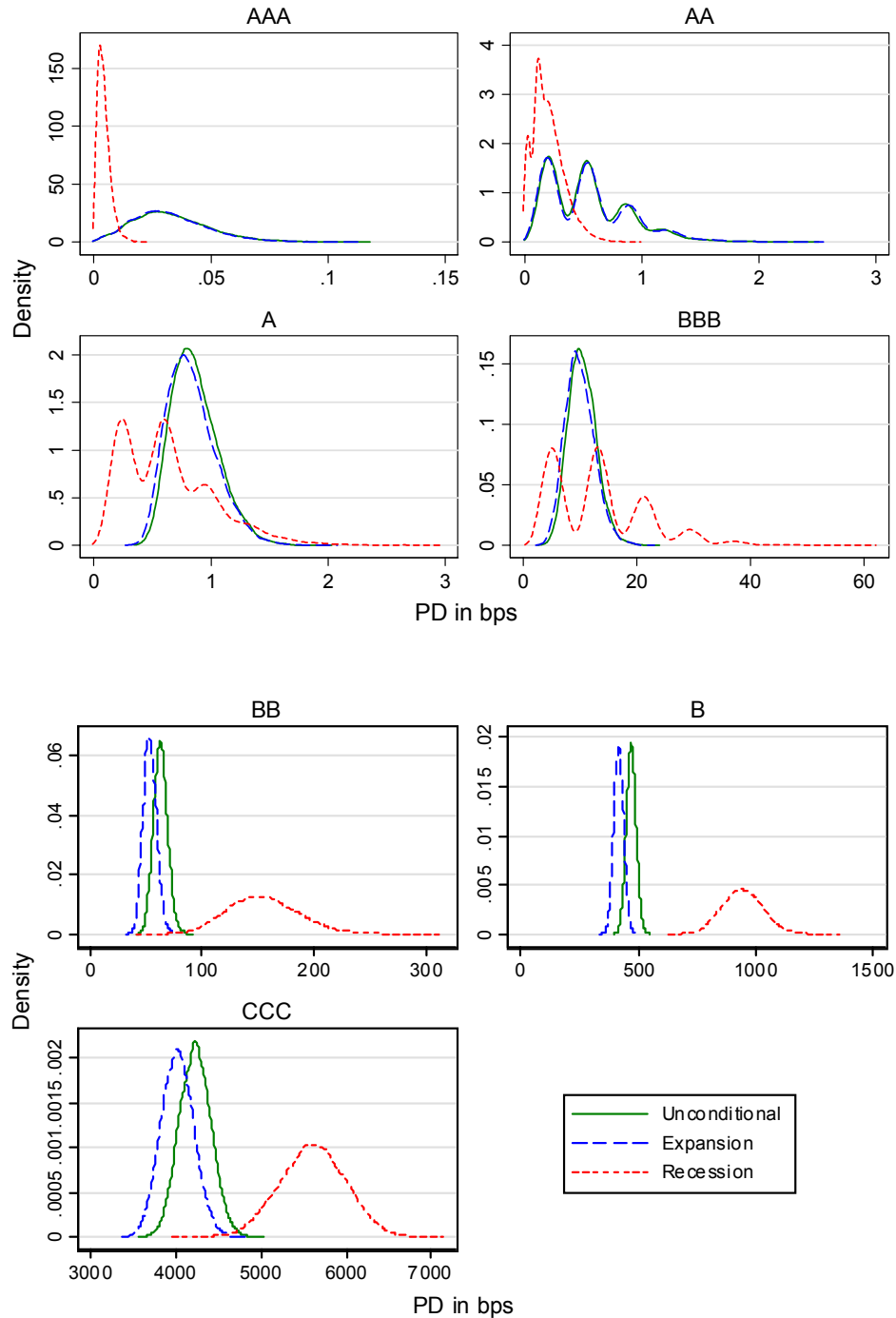


Figure 3: Kernel density plots of (nonparametric) bootstrapped probabilities of default (*PDs*) using S&P credit rating histories of U.S. firms from 1981-2002, split by recession and expansion. The red line denotes recession, blue expansion, and green the unconditional density (as in Figure 4). *PDs* are taken from the last column of the migration matrix estimated using the duration approach. $B = 10,000$ bootstrap replication, Epanechnikov kernel using Silverman's optimal window.

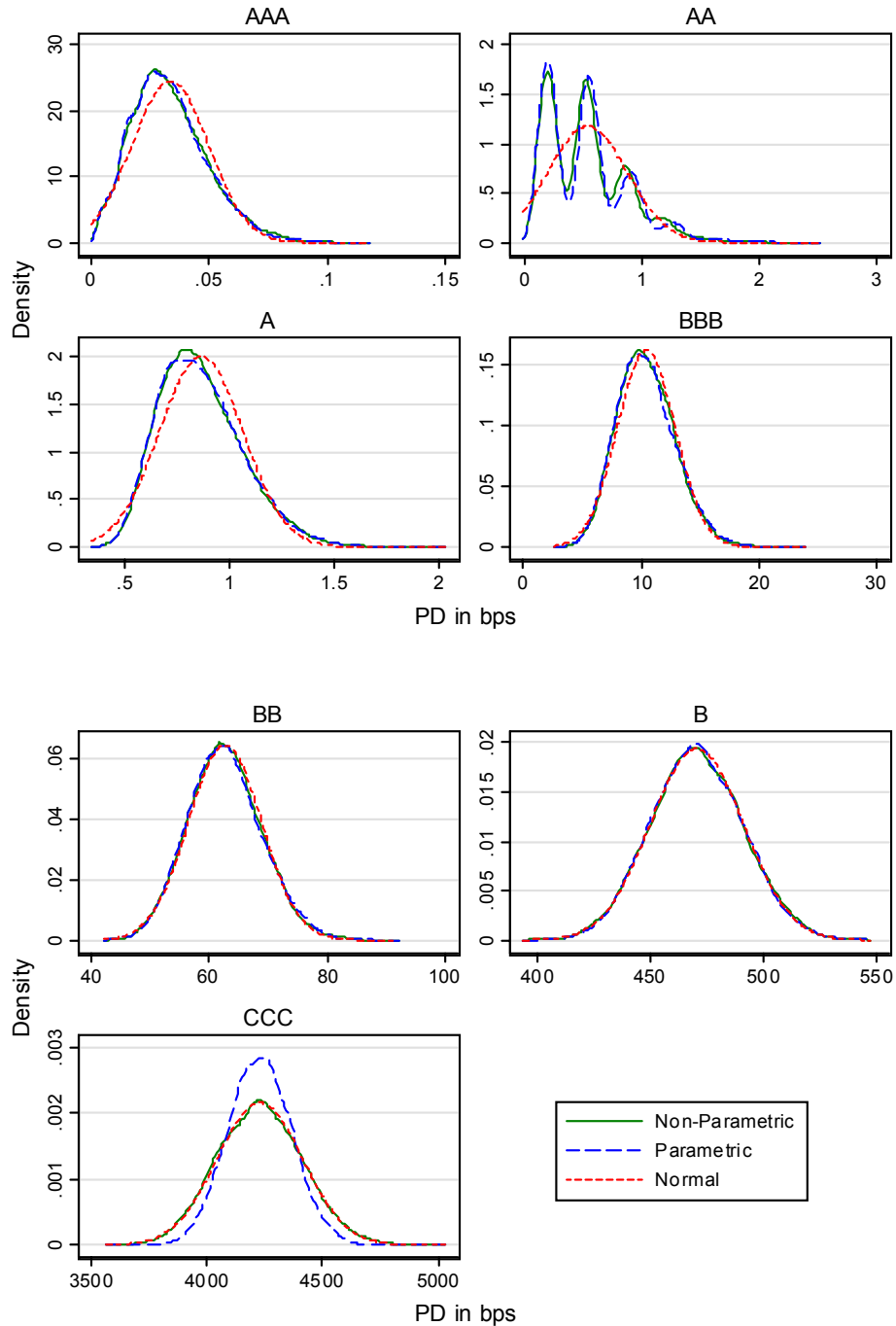


Figure 4: Kernel density plots of bootstrapped probabilities of default (PDs) using S&P credit rating histories of U.S. firms from 1981-2002. The solid green line represents the density of PDs from the nonparametric bootstrap and the dashed blue line represents the PDs from the parametric bootstrap. The dashed red line is the implied normal density with the same mean and variance as the empirical density of nonparametric PDs , plotted as a visual guide. Estimated PDs are taken from the last column of the migration matrix estimated using the duration approach. $B = 10,000$ bootstrap replication, Epanechnikov kernel using Silverman's optimal window.

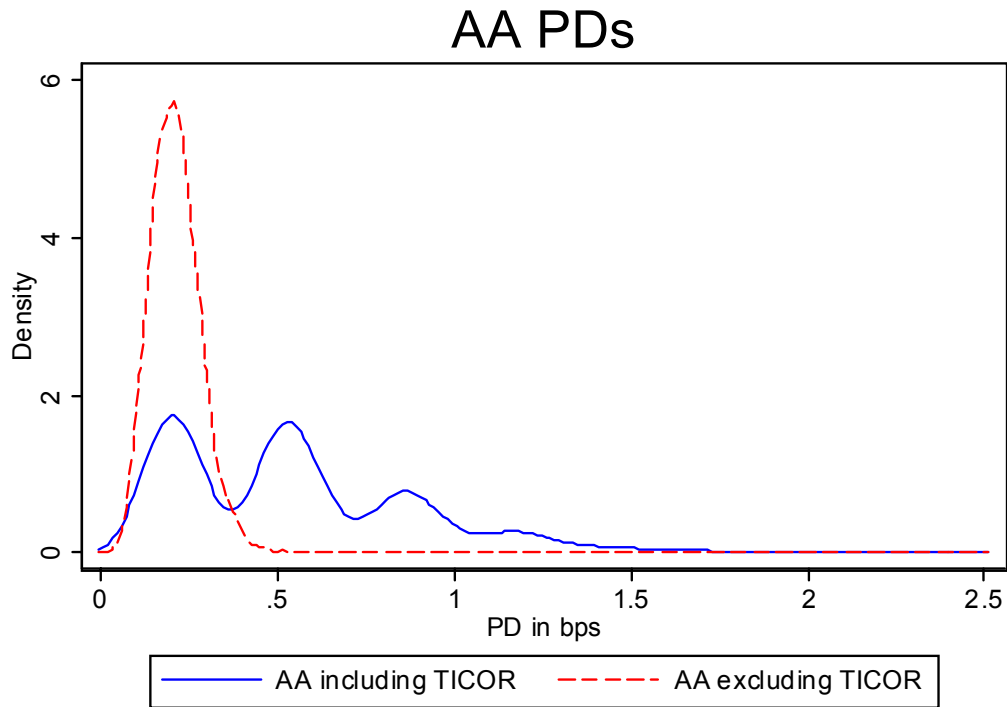
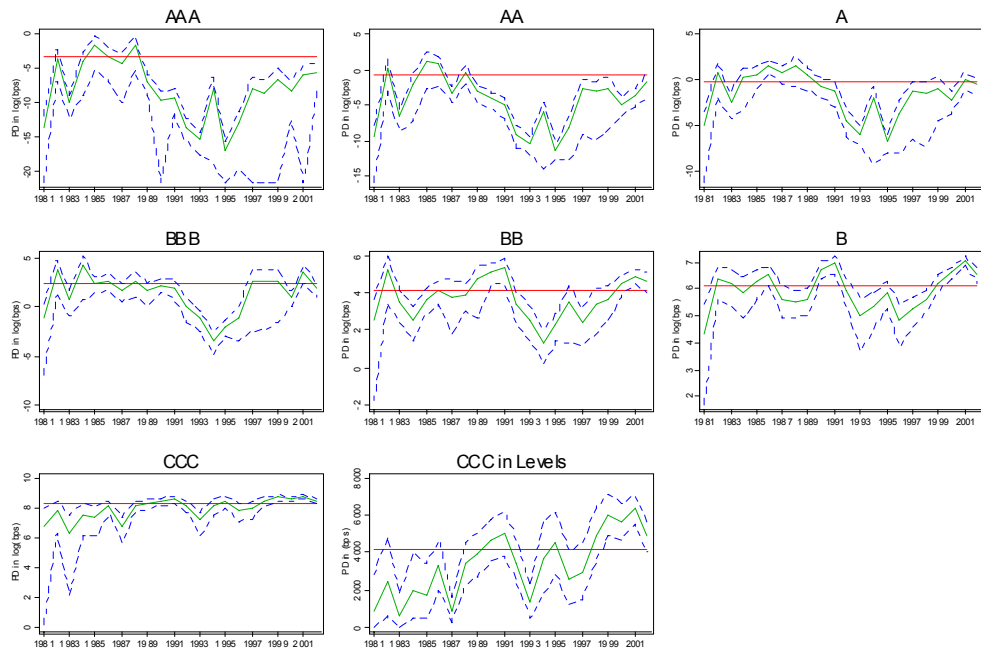


Figure 5: Kernel density plots of nonparametric bootstrapped AA probabilities of default (*PDs*) using S&P credit rating histories of U.S. firms from 1981-2002. The solid blue line includes the single firm, TICOR Mortgage Insurance, that migrated from AA \rightarrow CCC and the modes correspond to the number of times TICOR appears in the bootstrap sample. The dashed red line repeats the same calculation excluding TICOR. *PDs* are taken from the last column of the migration matrix estimated using the parametric intensity approach. $B = 10,000$ bootstrap replication, Epanechnikov kernel using Silverman's optimal window.

Duration



Cohort

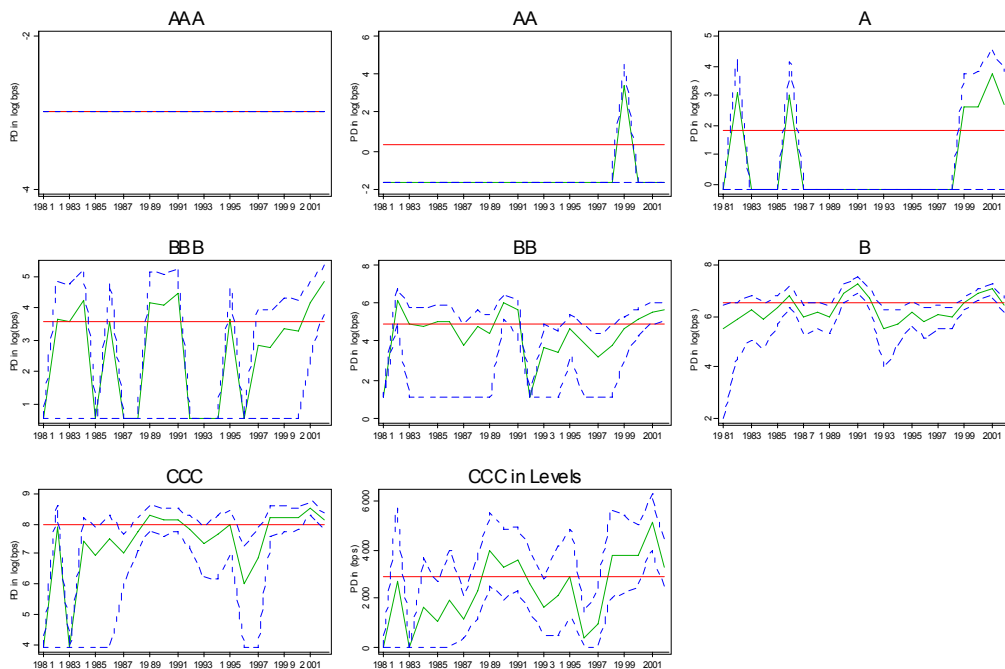


Figure 6: Comparing duration (top panel) and cohort based (bottom panel) estimates of default probabilities by credit grade using a 1-year rolling estimation windows by grade with the unconditional estimate (reported in log basis points, bp) using S&P credit rating histories of U.S. firms from 1981-2002. The CCC chart is repeated at the end in levels. The dashed lines are confidence intervals which are estimated using the nonparametric bootstrap for both the duration based PD estimates (top panel) and the cohort based PD estimates (bottom panel).

Duration

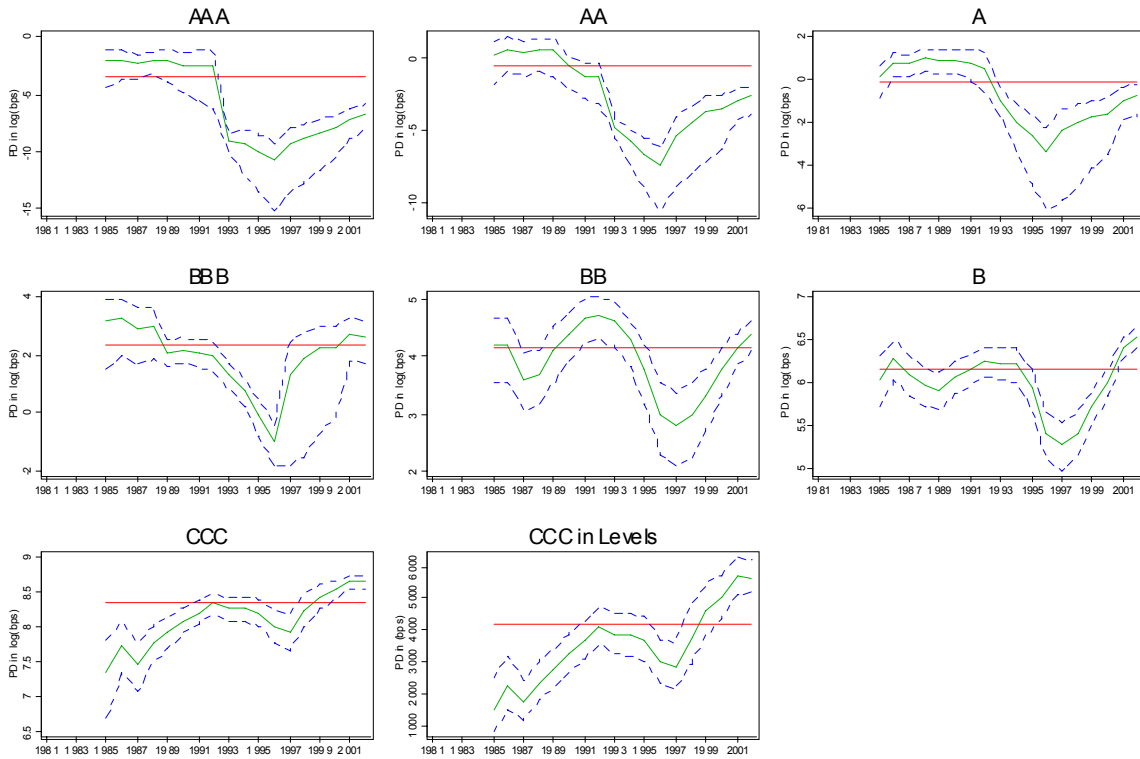


Figure 7: Comparing the 5-year rolling estimation windows by grade with the unconditional estimate (reported in log basis points, bp) using S&P credit rating histories of U.S. firms from 1981-2002. The CCC chart is repeated at the end in levels. The dashed lines are confidence intervals which are estimated using the nonparametric bootstrap for the duration based *PD* estimates.