International Journal of Financial Engineering
Vol. 10, No. 2 (2023) 2250037 (25 pages)

DOI: 10.1142/S2424786322500372



# On Wasserstein distances, barycenters, and the cross-section methodology for proxy credit curves

Matteo Michielon[*,†,§], Asma Khedher[†,¶] and Peter Spreij[†,‡,‖]

[*]*Quantitative Analysis and Quantitative Development*
*ABN AMRO Bank N.V., Gustav Mahlerlaan 10*
*1082 PP Amsterdam, The Netherlands*

[†]*Korteweg-de Vries Institute for Mathematics*
*University of Amsterdam, Science Park*
*105-107 1098 XG Amsterdam, The Netherlands*

[‡]*Institute for Mathematics*
*Astrophysics and Particle Physics*
*Radboud University Nijmegen Huygens Building*
*Heyendaalseweg 135,*
*6525 AJ Nijmegen*
*The Netherlands*

Received: 22 July 2022; Accepted: 26 November 2022
Published: 2 February 2023

**Abstract**

The credit default swap (CDS) market plays an important role for financial institutions. This is not only for their trading activities, but also as it provides a source of information to extract default probabilities to be used for (counterparty) credit risk purposes, as for instance in credit valuation adjustment calculations. Nonetheless, the number of entities for which liquid single-name CDSs are traded is of the order of a few thousands. This requires financial institutions to employ proxy methodologies to estimate the credit risk they face when trading with counterparties for which no (liquid) CDSs are available in the market. In this paper, we propose and compare different approaches to take into account counterparty-specific information in terms of rating, region, sector, etc. at cross-sectional level to strip risk-neutral default probabilities from CDSs. This is achieved by taking into account the intrinsic probabilistic information characterizing each CDS by means of suitably-defined Wasserstein distances and barycenters. The results suggest that default probabilities are likely to be overestimated if the construction of the proxy credit curves overlooks the probability

[§]Corresponding author.
*Email addresses*: [§]matteo.michielon@nl.abnamro.com, [¶]a.khedher@uva.nl, [‖]p.j.c.spreij@uva.nl

structure underlying the CDS market, potentially resulting in a too conservative counterparty credit risk pricing framework.

*Keywords*: Credit default swap; credit valuation adjustment; cross-section methodology; proxy credit curve; Wasserstein barycenter; Wasserstein distance.

## 1. Introduction

In the paper at hand, we develop a novel approach for constructing proxy credit curves from credit default swaps (CDSs) starting from the *cross-section methodology.* In particular, we investigate how to embed the concepts of *Wasserstein distance* and *Wasserstein barycenter* between implied CDS probability distributions in a cross-sectional framework.

Credit risk and counterparty credit risk are key aspects financial institutions. Quantifying (and, thus, risk-managing) these risks involves procedures aiming to estimate the credit worthiness of given counterparties, which necessitate the calculation of their default probabilities within the relevant portfolio-specific horizons. For this reason techniques to compute the likelihood of a default given the observable quantities available in the financial markets need to be in place. A common procedure to estimate risk-neutral default probabilities leverages on CDS quotes: from these observations it is possible to compute default probabilities in a simple and nowadays practically standardized manner (White, 2014). These probabilities are particularly important, amongst other areas, for valuation adjustments calculations; see for instance Brigo and Pallavicini (2014) and Abbas-Turki *et al*. (2018). However, the liquidity of the CDS market is often below-par compared to that of other markets (Junge and Trolle, 2015), which makes the estimation of default risk starting from these instruments not a straightforward task. This is mainly due to a couple of key factors. On the one hand, at global level, the number of counterparties for which CDS quotes are available is somewhat limited, i.e., in the order of a couple of thousands; see Green (2016, Sec. 4.6) and Ruiz (2015, Sec. 7.4). On the other hand, an important distinction needs to be made between liquid and illiquid CDS names: in the former case the procedure to compute default risk for counterparties with liquid CDSs quoted in the market is straightforward, while in the latter case it might happen that either quoted CDSs for the given counterparties are highly illiquid (and, as a consequence, not reliable) or, even worse, no CDSs for the relevant names are traded in the market at all. This is a very common situation faced by several financial institutions (especially banks) which do business with small retail companies and enterprises, for which in the vast majority of cases no CDSs are traded. Therefore, under these circumstances, procedures and methodologies need to be set up to indirectly calculate the default probabilities for these companies.

There are procedures in place to extract the aforementioned default probabilities in the case the relevant CDS market data is unreliable or not directly observable. Other than mapping the illiquid counterparties to specific single-name or index CDSs (see Green (2016, Secs. 4.6.1 and 4.6.2), respectively), the easiest way of doing so is by means of the so-called *intersection methodology* (EBA, 2013) (see also Chourdakis *et al*. (2013) and Sourabh *et al*. (2018)). This involves proxying the unobservable quotes for a counterparty with the average of the available CDS quotes of companies with the same characteristics (in terms of rating, region, sector, etc.). This approach has been looked at from another angle in Michielon *et al*. (2022), where the intersection methodology has been enhanced by computing the Wasserstein barycenter of the probability distributions implied from the CDSs belonging to the same bucket (i.e., with the same rating, region, sector, etc.). Nonetheless, the intersection methodology has two drawbacks. That is, in some cases, the number of liquid CDS quotes in a given bucket can be low or even zero, which might make some proxy curves unreliable or, in the worse-case scenario, impossible to compute. Further, the intersection methodology often seems to fail to guarantee, given buckets with different ratings but the same region, sector, etc., that the higher the rating the lower the likelihood of default is. To remedy these shortcomings, Chourdakis *et al*. (2013) propose a new approach, also known as *cross-section methodology*, which assumes that a given CDS spread can be split as the product of multiple factors (either global or depending on rating, region, sector, etc.). These factors can be then optimally calibrated in a least-squares sense by means of a linear regression procedure with categorical variables. From a data perspective, the approach of Chourdakis *et al*. (2013) has been further extended in Sourabh *et al*. (2018), where it is illustrated how to include additional information (i.e., equity data such as returns and volatilities) in the regression equations to further enhance the explicative power of the technique.

In this paper, we propose a hybrid calibration procedure to proxy CDS spreads. We propose different variations for this methodology and highlight their advantages and drawbacks by comparing them to the cross-section approach. We show how, overall, the best amongst the approaches proposed in this paper shares the main qualitative features of the cross-section methodology and, at the same time, tends to produce lower default probabilities. This is in line with the results outlined in Michielon *et al*. (2022) as far as the intersection methodology is concerned: ignoring the structure of the probability distributions implied from CDSs might result in overly-conservative counterparty credit risk pricing.

The main contribution of this paper is that of developing a new framework to construct proxy credit curves starting from CDS quotes. In particular, the methodology proposed allows us to build credit curves by means of the Wasserstein barycenters of the (probability distributions implied from the) available CDS buckets, as well as the cross-sectional information between them. The resulting approach allows us to explicitly take into account, in a sensible and intuitive manner, the implicit probabilistic information hidden in CDS quotes by means of Wasserstein distances, and to easily "interpolate" this information using a simple regression step based on categorical variables. The paper at hand clearly illustrates how Wasserstein distances can be considered as suitable tools for modeling credit curves and, to the best of our knowledge, it is one of the few exploring the possibility of modeling credit risk using distances defined on probability spaces.

This paper is arranged as follows. Section 2 illustrates the methodologies currently available in the literature for proxying CDS spreads from liquid market quotes. In Sec. 3, different novel approaches to incorporate Wasserstein distances and barycenters within the cross-sectional framework are proposed. Further, Sec. 4 compares the different methodologies given a real market data set, while Sec. 5 concludes. In Appendix A, a remark concerning the possibility of applying the cross-section methodology at hazard rate level is provided.

## 2. Literature Review

In this section, we provide an overview of the main methodologies currently available in the literature to build proxy credit curves from CDS data. This is necessary to make the article self-contained, as the methodologies proposed in Sec. 3.1 are based on the main concepts that are outlined in the current section. We highlight that, at the time of writing, the number of articles available in the literature concerning proxying credit curves from CDSs is quite limited, as only two main approaches are available, i.e., the intersection and the cross-section methodologies. Section 2.1 focuses on the former, while Sec. 2.2 concerns with the latter.[1]

### 2.1. The intersection methodology

A way to construct proxy CDS curves that does not involve simply mapping missing CDS data to that of other companies or indices is given by the so-called

[1]For completeness, we highlight that, in the sections that follows, default probabilities are always calculated under risk-neutral settings and within the framework of structural credit models, to which the specifications available in White (2014) apply. For methodologies to compare actual and risk-neutral default probabilities the interested reader can refer, e.g., to and Heynderickx *et al*. (2016) and Zou and Li (2022), while for alternative approaches to calculate hazard rates to Lee and Kuo (2015), amongst others.

intersection methodology (EBA, 2013). This approach suggests to bucket liquid CDSs according to their rating, region, sector, etc. By averaging the CDS quotes in each bucket, one can then define proxy CDS data per bucket and maturity; see also Chourdakis *et al*. (2013) and Sourabh *et al*. (2018). This methodology is very intuitive and easy to implement. Also, should one need to make the approach more granular, then this could be easily achieved by considering a finer bucketing procedure, despite this might results in more buckets with few or no observations. Despite the simplicity of the approach, the intersection methodology exhibits two main drawbacks, as highlighted in Sec. 3.1.

The intersection methodology has been investigated from a different angle in Michielon *et al*. (2022). In particular, therein the technique is revised from a probabilistic perspective. That is, instead of simply arithmetically averaging the market quotes within a given bucket, proxy CDS quotes for a given group are constructed by finding the (unique) hazard rate which defines the Wasserstein barycenter of the probability distributions implied from the considered CDSs. In particular, in Michielon *et al*. (2022) it is assumed that, for a fixed maturity, default distributions implied from CDSs are of the form

$$\mathbb{Q}(\tau \leq t) := 1 - e^{-\lambda t}, \tag{1}$$

where $\tau$ denotes the (random) default time of a given entity, and where $\lambda \in (0, +\infty)$ denotes the (constant) hazard rate. A simple manner to estimate $\lambda$ is given by the well-known *credit triangle* relationship (White, 2014), which reads

$$\lambda \approx \frac{s}{1-R}, \tag{2}$$

where $s$ refers to the par spread of the CDS taken into account, while $R$ to its recovery rate.[2]

We denote with $W_2(\mathbb{P}, \mathbb{Q})$ the square Wasserstein distance between the two absolutely continuous probability distributions on $\mathbb{R}$, i.e.,

$$W_2(\mathbb{P}, \mathbb{Q}) := \left( \inf_{J \in \mathcal{J}(\mathbb{P}, \mathbb{Q})} \int_{\mathbb{R}^2} \| x - y \|^2 J(dx \times dy) \right)^{\frac{1}{2}},$$

with $\mathcal{J}(\mathbb{P}, \mathbb{Q})$ denoting the set containing all the joint distributions which have $\mathbb{P}$ and $\mathbb{Q}$ as marginals, and where $\| \cdot \|$ stands for the Euclidean norm. For more details concerning Wasserstein distances refer, amongst others, to Villani (2003, Sec. 7), Villani (2009, Sec. 6), or to Panaretos and Zemel (2020, Secs. 1 and 2).

[2]Observe that one could alternatively imply hazard rates numerically, but the credit triangle relationship (2) provides an easy approximation formula that is very useful for comparison purposes and for the explanation of the results obtained. Therefore, this approach is followed throughout this paper (see also Michielon *et al*. (2022)).

Assume now that $N$ CDSs with the same maturity and with spreads denoted as $s_1, \dots, s_N$ are bucketed together. The Wasserstein barycenter of the $N$ CDS-implied exponential distributions $\mathbb{Q}_1, \dots, \mathbb{Q}_N$ is any probability distribution $\tilde{\mathbb{Q}}$ solution of

$$\inf\left\{\sum_{i=1}^{N} W_2^2(\mathbb{Q}_i, \mathbb{Q})\text{: the probability distribution } \mathbb{Q} \text{ has a finite second moment}\right\}. \tag{3}$$

It can be proven, see Michielon *et al*. (2022), that the solution of (3) is unique. In particular, it results that the unique solution of (3) is still of the form (1), and that its hazard rate can be computed analytically by finding the minimum of

$$\sum_{i=1}^{N}\left(\frac{1}{\lambda_i} - \frac{1}{\tilde{\lambda}}\right)^2 \tag{4}$$

with respect to $\tilde{\lambda}$. This hazard rate coincides with the harmonic mean of the implied hazard rates $\lambda_1, \dots, \lambda_N$ associated with $\mathbb{Q}_1, \dots, \mathbb{Q}_N$; see Michielon *et al*. (2022). For completeness, observe that

$$W_2^2(\mathbb{Q}_i, \tilde{\mathbb{Q}}) = 2\left(\frac{1}{\lambda_i} - \frac{1}{\tilde{\lambda}}\right)^2. \tag{5}$$

Assume that recovery rates within the same bucket are close to each other (usually this is the case as CDSs belonging to the same bucket share the same characteristics in terms of rating, region, sector, etc., and, thus, also similar recovery rates), and that the recovery rate for the proxy CDS is defined as their (harmonic) average. Then, again by means of (2), it results that the proxy CDS spread corresponding to $\tilde{\lambda}$ is close to the harmonic mean of the CDS spreads $s_1, \dots, s_N$. This leads to CDS proxy spreads that are smaller than their counterparts calculated by simply averaging the available CDS quotes within the bucket considered, as illustrated in Michielon *et al*. (2022), as the harmonic mean of $s_1, \dots, s_N$ is bounded from above by their arithmetic counterpart. As default probabilities are strictly increasing with respect to CDS spreads, the intersection methodology with Wasserstein barycenter produces in general lower default probabilities than in the standard case.

We highlight that the idea outlined in Michielon *et al*. (2022) was that of investigating how proxy credit curves (therein within the intersection approach) can take into account the distance between the probability distributions implied by the different CDSs. This is because just averaging CDS quotes does not explicitly accounts for the probabilistic information between default times. In particular, in order to do so, it is advisable to define a metric between probability distributions.

One convenient (and nowadays very popular) choice due to its properties is the Wasserstein distance as, e.g., it produces geodesics which are shape preserving and, further, it takes into account the "shape" of the probability distributions considered. These properties are not in general satisfied by other metrics, and in Michielon *et al*. (2022), see Sec. 4 therein, these concepts are described and examples are provided to endorse the choice made.

Further, we now provide some considerations concerning the intuitive interpretation of the Wasserstein barycenter as highlighted in Michielon *et al*. (2022). To do so, we assume to have a default time $\tau$ exponentially distributed under $\mathbb{Q}$ with rate parameter $\lambda$. It results, given a time interval $\delta t$ and assuming $\lambda$ "small", that

$$\frac{\mathbb{Q}(t \leq \tau \leq t+\delta t)}{\delta t}\left(=\frac{\mathbb{Q}(\tau \leq t+\delta t)-\mathbb{Q}(\tau \leq t)}{\delta t}\right)=\frac{\int_t^{t+\delta t} \lambda e^{-\lambda s}\, ds}{\delta t} \approx \lambda. \quad (6)$$

Note that the numerator in the first member of (6) measures by how much the default probability changes during the time interval $\delta t$. Therefore, (6) measures a change in probability during the time frame $\delta t$, i.e., $\lambda$ can be intuitively interpreted as the "default velocity".

We take into account a very simple framework where we have only two CDSs, denoted as $\mathrm{CDS}_1$ and $\mathrm{CDS}_2$, respectively. We also assume that both $\mathrm{CDS}_1$ and $\mathrm{CDS}_2$ have similar specifications, i.e., that they are linked to entities with the same rating, region, sector, etc., so that it makes sense considering $\mathrm{CDS}_1$ and $\mathrm{CDS}_2$ both in the same bucket. We denote with $s_1$ the CDS spread for $\mathrm{CDS}_1$, and with $s_2$ the CDS spread for $\mathrm{CDS}_2$ (we assume $s_1 \neq s_2$ to avoid trivial cases). We assume that both $\mathrm{CDS}_1$ and $\mathrm{CDS}_2$ have the same recovery rate $R$. Further, we approximate the hazard rates corresponding to $\mathrm{CDS}_1$ and $\mathrm{CDS}_2$, denoted with $\lambda_1$ and $\lambda_2$, respectively, by means of the credit triangle relationship (2). That is, $\lambda_1 \approx \frac{s_1}{1-R}$, and $\lambda_2 \approx \frac{s_2}{1-R}$. We denote with $\tau_1$ and $\tau_2$ the default times linked to $\mathrm{CDS}_1$ and $\mathrm{CDS}_2$, respectively.

We consider a fixed time $\bar{t}$ and a sufficiently small amount $\delta\mathbb{Q} \in (0,1)$ such that $\mathbb{Q}(\tau_1 \leq \bar{t}+\delta t_1) = \mathbb{Q}(\tau_1 \leq \bar{t}) + \delta\mathbb{Q}$ and $\mathbb{Q}(\tau_2 \leq \bar{t}+\delta t_2) = \mathbb{Q}(\tau_2 \leq \bar{t}) + \delta\mathbb{Q}$, for $\delta t_1 > 0$ and $\delta t_2 > 0$. That is, during the time intervals $\delta t_1$ and $\delta t_2$ the probabilities of default of $\tau_1$ and $\tau_2$ both increase by the same amount $\delta\mathbb{Q}$, i.e., $\delta\mathbb{Q}$ represents the change in probability for each default time during the two time frames considered. Therefore, due to (6), it results that $\delta t_1 \approx \frac{\delta\mathbb{Q}}{\lambda_1}$ and $\delta t_2 \approx \frac{\delta\mathbb{Q}}{\lambda_2}$. Thus, the average default velocity of $\tau_1$ and $\tau_2$ results being equal to $\frac{\text{total decrease in probability}}{\text{total time}} = \frac{2\delta\mathbb{Q}}{\frac{\delta\mathbb{Q}}{\lambda_1}+\frac{\delta\mathbb{Q}}{\lambda_2}} = \frac{2\lambda_1\lambda_2}{\lambda_1+\lambda_2} =: \hat{h}$, which corresponds to the harmonic mean of the default velocities $\lambda_1$ and $\lambda_2$. Therefore, the proxy hazard rate relevant for the bucket, which is calculated as the harmonic mean of the relevant hazard rates, can be naturally interpreted as "average default velocity". We also note, see Theorems 1 and 2 in Michielon *et al*. (2022), that the Wasserstein barycenter between exponential distributions is still exponentially distributed, with rate parameter equal to the harmonic mean of the relevant rate parameters. This makes Wasserstein distances also convenient from a computational perspective.

## 2.2. The cross-section methodology

The cross-section methodology, introduced in Chourdakis *et al*. (2013), is based on the idea that the logarithm of a CDS spread can be decomposed as the product of one global and different bucket-dependent factors. These factors can then be estimated in a least-squares sense by taking into account all the CDS spreads available in the market simultaneously; see also Gregory (2020, Sec. 12.4.4). By doing so, the approach by Chourdakis *et al*. (2013) allows to determine CDS spreads also for buckets for which no data is directly observable in the market. The cross-section technique has been enhanced in Sourabh *et al*. (2018), where equity data is used simultaneously to credit data to take into account even more information in the least-squares optimization procedure. We now describe the cross-section methodology more in detail. For consistency with the rest of the paper, see Sec. 3, in our exposition we follow a similar approach to that outlined in Sourabh *et al*. (2018), but without taking into account equity data.

Assume, given a CDS dataset for a fixed maturity, that we are able to distinguish between $N_{\text{Rating}}$ possible ratings, $N_{\text{Region}}$ possible regions, $N_{\text{Sector}}$ possible sectors, and $N_{\text{Seniority}}$ possible seniorities. This identifies a total of $N_{\text{Bucket}} := N_{\text{Rating}} \cdot N_{\text{Region}} \cdot N_{\text{Sector}} \cdot N_{\text{Seniority}}$ possible rating-region-sector-seniority combinations. Having defined an enumeration for the ratings, one for the regions, one for the sectors and one for the seniorities, we fix a rating, a region, a sector and a seniority to be used as a baseline. We are thus left with $N_{\text{Bucket}} - 1$ possible combinations different from the base one. Given a total of $K$ CDS quotes available for a fixed maturity, denoted as $s_1, \dots, s_K$, it is assumed that, for every $i \in \{1, \dots, K\}$, the following relationship holds:

$$\ln(s_i) = x_0^{\text{Base}} + \sum_{j=1}^{N_{\text{Rating}}-1} x_j^{\text{Rating}} \mathbb{1}_j(i) + \sum_{k=1}^{N_{\text{Region}}-1} x_k^{\text{Region}} \mathbb{1}_k(i) + \sum_{l=1}^{N_{\text{Sector}}-1} x_l^{\text{Sector}} \mathbb{1}_l(i) + \sum_{m=1}^{N_{\text{Seniority}}-1} x_m^{\text{Seniority}} \mathbb{1}_m(i). \tag{7}$$

In (7), $\mathbb{1}_j(i)$ denotes the indicator function that equals one if the rating of the $i$th CDS coincides with the $j$th rating, while similar conventions apply for $\mathbb{1}_k(i)$, $\mathbb{1}_l(i)$

and $\mathbb{1}_m(i)$ as far as region, sector and seniority are concerned. Therefore, this means that (7) relates to a linear regression problem involving categorical variables. For this reason, in the first of the four summations in (7), $j$ ranges from one to $N_{\text{Rating}} - 1$ rather than to $N_{\text{Rating}}$, and similar considerations hold for $k$, $l$, and $m$, respectively. From an interpretation angle, see Sourabh *et al.* (2018), $x_0^{\text{Base}}$ can be understood as the logarithm of the proxy spread of the base bucket, with the other $N_{\text{Total}} - 4$ unknowns representing the changes in the logarithm when migrating from the base bucket to a different one. Estimating the unknowns in (7) can be done by means of a simple least-squares minimization. As we use the right-hand side of (7) also in the rest of this paper, for conciseness and readability we define it as $\xi_i(\boldsymbol{x})$, where the first component of the vector $\boldsymbol{x}$ coincides with $x_0^{\text{Base}}$, followed by $x_1^{\text{Rating}}$, and so on until $x_{N_{\text{Seniority}}-1}^{\text{Seniority}}$.

The cross-section methodology provides a framework to proxy credit curves that is very intuitive and computationally inexpensive. Therefore, for this reason, the aim of this paper is that of extending this approach by also taking into account the underlying probability distributions of the considered CDSs by means of Wasserstein barycenters as per (Michielon *et al.*, 2022).

## 3. Wasserstein Distances and the Cross-Section Methodology

Now that the general aspects of the cross-section methodology have been outlined in Sec. 2.2, we describe alternative approaches allowing to embed Wasserstein distances within this framework. In particular, we will outline three different possibilities before providing a comparison between them in Sec. 4. For the ease of readability and clarity, in the sections that follow we will adopt the following convention: quantities with the tilde superscript always indicate that a bucketing procedure has already taken place, while the absence of the tilde superscript is used in the opposite situation.

### 3.1. Wasserstein cross-intersection methodology on spreads

Assume that, given the CDSs observed in the market with the same maturity, it is possible to bucket them in $N_{\text{Bucket}}^{\text{Market}}$ distinct rating-region-sector-seniority buckets (in general, it holds that $N_{\text{Bucket}}^{\text{Market}} \leq N_{\text{Bucket}}$, as CDS quotes are not usually available for every possible rating-region-sector-seniority combination). In Michielon *et al.* (2022), see also Sec. 2.1, it has been described how, for each bucket, it is possible to define the unique optimal CDS quote corresponding to the Wasserstein barycenter. This is done by reversing the credit triangle relationship (2), where the recovery rate therein is estimated as the (harmonic[3]) average of the recovery rates of the CDSs in the bucket (later in this section an alternative approach to estimate the recovery rate for a given bucket is illustrated). Assume that the $N_{\text{Bucket}}^{\text{Market}}$ proxy CDS quotes $\tilde{s}_1, \ldots, \tilde{s}_{N_{\text{Bucket}}^{\text{Market}}}$ have been defined for each of the market-observable buckets by means of Wasserstein barycenters following the framework outlined in Sec. 2.1. It is then assumed that, for every $i \in \{1, \ldots, N_{\text{Bucket}}\}$, the following relationship, like in (7), holds:

$$\ln(\tilde{s}_i) = \xi_i(\boldsymbol{x}). \tag{8}$$

We point out that also appropriate recovery rates need to be calculated. Otherwise, unless default values are assigned to them, it would not be possible to strip credit curves solely from the spreads. Therefore, in order to do so, observe that given a recovery rate $R \in (0, 1)$, then $-\ln(R) \in (0, +\infty)$.[4] Hence, a similar approach to that just outlined can be followed to proxy recovery rates by substituting the opposite of the logarithm of the $i$th (harmonically) averaged recovery rate $\tilde{R}_i$ in an expression as (8). This time, however, one needs to be aware that the regressed variables still need to be positive, otherwise negative recovery rates could be calculated. For this reason, it is needed to either perform a constrained regression where the components of $\boldsymbol{x}$ have to be positive or, otherwise, to perform an unconstrained regression and to eventually floor the negative recovery rates (if any) with a reasonable positive constant. Note that, as the step to perform the regression on the recovery rates should be performed before the regression on spreads takes place, the regression on recovery rates can potentially also be performed without taking into account any bucketing.[5]

### 3.2. Wasserstein cross-intersection methodology on hazard rates

We have assumed, throughout this paper, that the default probability distribution implied from a CDS is of the form (1). Therefore, the survival probability implied

[3]In Michielon *et al*. (2022), it is shown that considering the arithmetic average rather than the harmonic one has a negligible impact on the magnitude of the proxy CDS quotes constructed. Notwithstanding, as here we are dealing with the averaging of (recovery) rates, the choice of the harmonic average seems the natural one to consider.

[4]By considering $R \in (0, 1)$ we assume that it is not possible to recover all the notional given a default and, further, that always at least a (very) small fraction of the invested notional, which corresponds to a small $R$, can be recovered if a default event is triggered.

[5]Observe that, in principle, one can also perform an unconstrained regression after having transformed the (potentially harmonically-averaged) hazard rates by means of the map such that $x \mapsto -\ln(\frac{1}{x} - 1)$, or by any other function that maps $(0, 1)$ to $(-\infty, +\infty)$ in a smooth and strictly increasing manner, as for instance via the trigonometric functions such that $x \mapsto \cot(\pi x)$ or $x \mapsto \tan(\pi(x - \frac{1}{2}))$, to name but a few.

by means of Wasserstein barycenters for the $i$th bucket, see Sec. 2.1, is given by

$$\mathbb{Q}(\tilde{\tau}_i > t) = e^{-\tilde{\lambda}_i t}.$$

We can assume that, for every $i \in \{1, \ldots, N_{\text{Bucket}}\}$, the following relationship holds:

$$\tilde{\lambda}_i = \xi_i(\boldsymbol{x}).$$

Also in this case, a procedure as that outlined in Sec. 3.1 should be followed to compute proxy recovery rates for each bucket. Note that, as hazard rates need to be positive, then either a constrained regression should be performed, or the computed hazard rates should be properly floored in the case of negative regressed quantities.

### 3.3. Global Wasserstein distance optimization

We now look at the cross-section methodology from a different angle compared to Secs. 3.1 and 3.2. That is, we consider all the $K$ CDS quotes available for the chosen maturity simultaneously. Here, no preliminary bucketing takes place. By considering the flat hazard rate hypothesis (1), we assume that the default distribution implied from the $i$th CDS can be written as

$$\mathbb{Q}(\tau_i \leq t) := 1 - e^{-(\lambda_{\text{Rating}(i)} + \lambda_{\text{Region}(i)} + \lambda_{\text{Sector}(i)} + \lambda_{\text{Seniority}(i)})t}, \tag{9}$$

where the subscript Rating$(i)$ denotes the rating of the $i$th CDS, Region$(i)$ its region, and so forth. Note that if we consider four independent exponentially distributed random variable with rate parameters $\lambda_{\text{Rating}(i)}$, $\lambda_{\text{Region}(i)}$, $\lambda_{\text{Sector}(i)}$, and $\lambda_{\text{Seniority}(i)}$, respectively, then their minimum is still exponentially distributed with rate parameter given by the sum $\lambda_{\text{Rating}(i)} + \lambda_{\text{Region}(i)} + \lambda_{\text{Sector}(i)} + \lambda_{\text{Seniority}(i)}$. Therefore, (9) can be interpreted as the probability distribution that corresponds to the "first-to-default" among the four aforementioned independent exponential random variables. Given the implied probability distribution $\mathbb{Q}_i$ of the $i$th CDS calculated from the market data assuming a flat hazard rate, we denote with $\mathbb{Q}_i^*$ its counterpart defined via (9). For every quoted CDS with distribution $\mathbb{Q}_i$ and flat hazard rate $\lambda_i$ we can then calculate, via (5) (see also Michielon *et al.* (2022) for further details), the amount

$$W_2^2(\mathbb{Q}_i, \mathbb{Q}_i^*) = 2\left(\frac{1}{\lambda_i} - \frac{1}{\lambda_{\text{Rating}(i)} + \lambda_{\text{Region}(i)} + \lambda_{\text{Sector}(i)} + \lambda_{\text{Seniority}(i)}}\right)^2. \tag{10}$$

Thereafter, one can attempt to minimize the aggregate amount

$$\frac{1}{2}\sum_{i=1}^{K} W_2^2(\mathbb{Q}_i, \mathbb{Q}_i^*) \tag{11}$$

with respect to the $N_{\text{Total}} := N_{\text{Rating}} + N_{\text{Region}} + N_{\text{Sector}} + N_{\text{Seniority}}$ hazard rates (see (4)). Note, however, that (11) is not strictly convex, as the function $f(x) := (\frac{1}{\lambda} - \frac{1}{x})^2$ has zero second derivative at $\frac{3}{2}\lambda$. Therefore, the existence of a unique global minimum for (11) is in principle not necessarily guaranteed. Also, note that for common datasets there will be several CDS which share similar features (e.g., they might have the rating and different region, sector and seniority, etc.). This obviously means that equations of the form (10) are not, irrespectively on $i$, independent from each other. Thus, this implies that finding the minimum (or minima) of (11) analytically is not possible in general and, therefore, numerical techniques should be employed in this respect. As the uniqueness of a global minimum for (11) is not guaranteed, one would need to properly assess which area of the domain of the function to be minimized should be taken into account for the optimization procedure. This suggests, therefore, the use of stochastic optimization algorithms as Basin-hopping (Wales and Doye, 1997) or Dual Annealing (Xiang *et al*., 2013), at least to determine a sensible estimate for the initial point for the minimization of (11). As a consequence, given the high number of parameters involved in the optimization, we now propose a modified version of the optimization problem (11) to guarantee strict convexity.

Given $\epsilon > 0$ small and fixed *a priori*, define the function

$$g_\epsilon^\lambda(x) := \begin{cases} f(x) \quad \text{if} x \le \frac{3}{2}\lambda - \epsilon, \\ f\left(\frac{3}{2}\lambda - \epsilon\right) + f'\left(\frac{3}{2}\lambda - \epsilon\right)\left(x - \frac{3}{2}\lambda + \epsilon\right) & , \\ +f''\left(\frac{3}{2}\lambda - \epsilon\right)\left(x - \frac{3}{2}\lambda + \epsilon\right)^2 \qquad \text{if } x > \frac{3}{2}\lambda - \epsilon \end{cases} \tag{12}$$

i.e., $g_\epsilon^\lambda(\cdot)$ coincides with $f(\cdot)$ until just before its inflection point, and afterwards it coincides with its second-order Taylor approximation. By construction, this function is twice continuously differentiable and strictly convex on the positive real line. We can therefore consider, once a fixed value for $\epsilon$ has been chosen, instead of (11) its approximated counterpart given by

$$\sum_{i=1}^{K} g_\epsilon^{\lambda_i}\left(\frac{1}{\lambda_{\text{Rating}(i)} + \lambda_{\text{Region}(i)} + \lambda_{\text{Sector}(i)} + \lambda_{\text{Seniority}(i)}}\right). \tag{13}$$

We now show that the alternative function to be minimized, i.e., (13), is strictly convex on $(0, +\infty)^{N_{\text{Total}}}$.

**Theorem 1.** *The function* (13) *is strictly convex on* $(0, +\infty)^{N_{\text{Total}}}$.

**Proof.** Fix an index $i$ and observe that the map $\pi_i : (0, +\infty)^{N_{\text{Total}}} \to (0, +\infty)$ such that $\pi_i(\boldsymbol{\lambda}) := \lambda_{\text{Rating}(i)} + \lambda_{\text{Region}(i)} + \lambda_{\text{Sector}(i)} + \lambda_{\text{Seniority}(i)}$ is linear. Given $\theta \in [0, 1]$ and $\boldsymbol{\lambda}_*$ and $\boldsymbol{\lambda}^*$ in $(0, +\infty)^{N_{\text{Total}}}$, it results that

$$\begin{aligned} g_\epsilon^\lambda \circ \pi_i(\theta\boldsymbol{\lambda}_* + (1-\theta)\boldsymbol{\lambda}^*) &= g_\epsilon^\lambda(\pi_i(\theta\boldsymbol{\lambda}_* + (1-\theta)\boldsymbol{\lambda}^*) \\ &= g_\epsilon^\lambda(\theta\pi_i(\boldsymbol{\lambda}_*) + (1-\theta)\pi_i(\boldsymbol{\lambda}^*)) \qquad (14) \\ &< \theta g_\epsilon^\lambda(\pi_i(\boldsymbol{\lambda}_*)) + (1-\theta) g_\epsilon^\lambda(\pi_i(\boldsymbol{\lambda}^*)) \\ &= \theta g_\epsilon^\lambda \circ \pi_i(\boldsymbol{\lambda}_*) + (1-\theta) g_\epsilon^\lambda \circ \pi_i(\boldsymbol{\lambda}^*), \qquad (15) \end{aligned}$$

where equality (14) comes from the linearity of $\pi_i(\,\cdot\,)$, while the strict inequality in (15) is a consequence of the strict convexity of $g_\epsilon^\lambda(\cdot)$. Recalling that the sum of strictly convex functions is still strictly convex, if follows that (13) is strictly convex on its (convex) domain $(0, +\infty)^{N_{\text{Total}}}$. □

We highlight that, from a practical perspective, it is not necessary to perform a minimization for each variable in (13) on the whole positive line, as setting a sufficiently large upper bound suffices. This is because, despite in principle hazard rates can reach extremely large values in $(0, +\infty)$, in practice this is usually not the case. For example, on 12 September 2008 for Lehman Brothers, just before its bankruptcy was announced on 14 September 2008, the implied hazard rate for its one-year CDS was of the order of 230,000 bps while the other hazard rates corresponding to longer-dated maturities were all below 10,000 bps; see Brigo *et al*. (2010). Also note that, for close-to-default names, often in practice no proxy CDS curves are calculated, and more conservative approaches are used, instead. Therefore, minimizing (13) can be performed on a suitable compact set of $(0, +\infty)^{N_{\text{Total}}}$, which guarantees existence and unicity of a global minimum in virtue of Theorem 1. The fact that it is possible to approximate the minimum of (11) by minimizing (13) by means of a twice continuously differentiable function guarantees that quasi-Newton methods such that the Broyden–Fletcher–Goldfarb–Shanno (Press *et al*., 2007, Sec. 10.9), often used in financial practical applications, can be employed directly. Note that there are other minimization techniques that can be used as, for instance, the Powell's method (Press *et al*., 2007, Sec. 10.7). For more details about optimization methods used in different financial applications, the interested reader can refer to Cornujols *et al*. (2018).[6]

## 4. Comparison of the Different Methodologies

In this section, we provide a comparison between the different methodologies proposed in this paper to assess how they perform and compare with each other

[6]Observe that many optimization algorithms are nowadays available in standard libraries for various programming languages as, e.g., QuantLib (www.quantlib.org), or SciPy (www.scipy.org).

under real market circumstances. Before doing so in Sec. 4.2, we describe the data preprocessing steps we have followed in Sec. 4.1.

### 4.1. Data preprocessing

The data preprocessing steps we have followed are quite similar to those considered in Sourabh *et al*. (2018). As CDS quotes that are related to the “CCC” or “D” rating are often not reliable or simply not available, we exclude those ratings from the dataset. Further, we only consider, as regions, “Asia”, “Europe” and “North America”, and label the remaining ones as “Other”. For “Europe” we consider only CDSs in EUR currency, while for all the other regions we only consider USD CDSs, as they are more liquid. As far as the sector is concerned, we remove all the “Government” quotes, since there seems to be evidence that those should be analyzed separately; see Longstaff *et al*. (2011) and Pan and Singleton (2008). We also label the “Basic materials”, “Consumer services”, “Energy”, “Industry”, “Technology” and “Telecommunication services” sectors as “Cyclical”, while “Consumer goods”, “Healthcare”, “Utilities” as “Noncyclical”.[7] The “Financials” sector is left unchanged. The choices and mappings for the ratings, regions, sectors and seniorities have been summarized in Table 1, for convenience. The mapping performed allows to define six possible ratings, six possible regions, three possible sectors and two possible seniorities, leading to a total of 144 buckets. We also keep a distinction between senior and subordinated credit, and we remove all the CDS quotes with spread larger than 1,000 bps. Further, we only take into account CDS names for which data was provided by at least two market participants.

### 4.2. Results

Before providing a comparison between the different methodologies outlined in this paper, we introduce some abbreviations for the ease of clarity and readability. In particular, we refer to the cross-section methodology described in Sec. 2.2 as CSM, to the Wasserstein cross-intersection methodology on spreads of Sec. 3.1 as WCIMS, to the Wasserstein cross-intersection methodology on hazard rates of Sec. 3.2 as WCIMHR and, further, to the global Wasserstein distance optimization methodology of Sec. 3.3 as GWDOM. Further, again for comprehensibility purposes, if the approach outlined in Sec. 3.1 is applied to the arithmetically-averaged

[7]We recall that cyclical and noncyclical sectors refer, in this case, to how correlated the credit quality of a given entity is with the relevant macroeconomic factors: cyclical sectors are closely related to the current economic trend, while the opposite holds for the noncyclical ones.

Table 1. The mapping performed in terms of rating, region, sector and seniority for the datasets considered in this paper. The dash symbol "—" means that all the CDSs linked to that category have been removed from the dataset: the left column collects the raw labels, while the right one contains the preprocessed ones.

| | From | To |
|---|---|---|
| Rating | AAA<br>AA<br>A<br>BBB<br>BB<br>B | AAA<br>AA<br>A<br>BBB<br>BB<br>B |
| | CCC<br>D | — |
| Region | Asia | Asia |
| | Europe | Europe |
| | North America | North America |
| | Africa<br>Eastern Europe<br>India<br>Latin America<br>Middle East<br>Oceania<br>Offshore | Other |
| Sector | Basic materials<br>Consumer services<br>Energy<br>Industrials<br>Technology<br>Telecommunications services | Cyclical |
| | Consumer goods<br>Healthcare<br>Utilities | Noncyclical |
| | Financials | Financials |
| | Government | — |
| Seniority | Senior | Senior |
| | Subordinated | Subordinated |

CDS quotes for each observable bucket (i.e., the standard intersection methodology is applied before the regression takes place), then we name this variation as average cross-intersection methodology on spreads, and refer to it as ACIMS. This last approach will be solely used in this section for benchmarking purposes with the WCIMS.

As a first step in our analysis, we start by considering the CSM given the dataset taken into account and the preprocessing of the data outlined in Sec. 4.1. In particular, for comparison purposes we consider the spreads computed for the base bucket using the CSM during the yearly period considered in this paper. Moreover, keeping region, sector and seniority unchanged, we consider all the possible ratings, and depict the results in Fig. 1. As expected, Fig. 1 illustrates how the CSM properly differentiates amongst ratings, in the sense that to lower (higher) CDS spreads correspond higher (lower) ratings.

As a second step, as this will be useful for benchmarking purposes, we implement the ACIMS. We do this as, by the results available in Michielon *et al*. (2022), see also Sec. 2.1, the proxy spreads generated by means of Wasserstein barycenters should be lower than their counterparts computed by averaging the spreads in each bucket. Therefore, we would overall expect a similar behavior also in the cases considered here when the regression takes place. For the base bucket, and keeping its region, sector and seniority fixed while varying the rating, we obtain the results illustrated in Fig. 2. Also in this case, in line with what observed in Fig. 1, we see a clear distinction between ratings, with the expected


Fig. 1. Proxy CDS spreads computed using the CSM, given different ratings, for the base bucket during the time period spanning from 1 November 2020 to 1 November 2021.

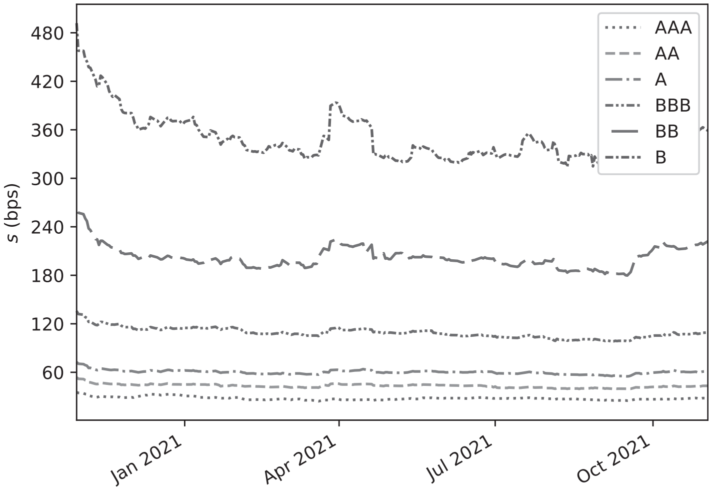



Fig. 2. Proxy CDS spreads computed using the ACIMS, given different ratings, for the base bucket during the time period spanning from 1 November 2020 to 1 November 2021.

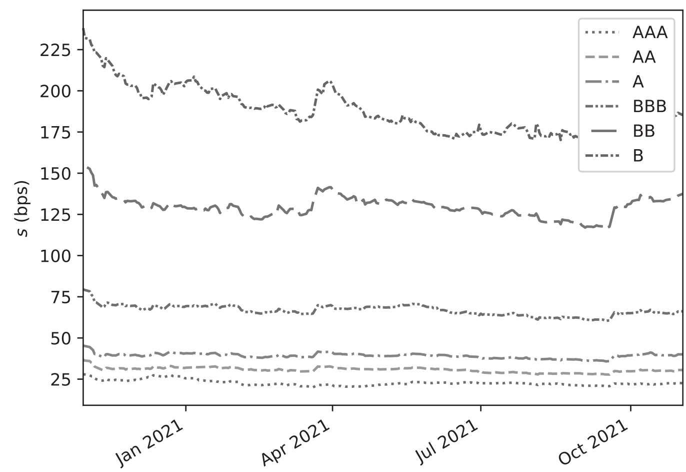
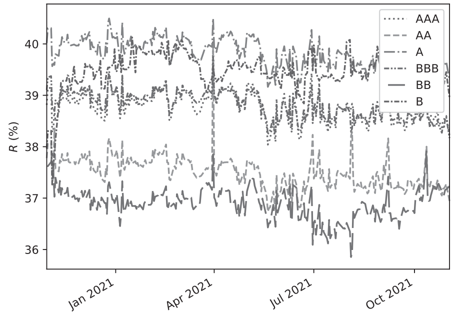



Fig. 3. Proxy CDS spreads computed using the WCIMS, given different ratings, for the base bucket during the time period spanning from 1 November 2020 to 1 November 2021.

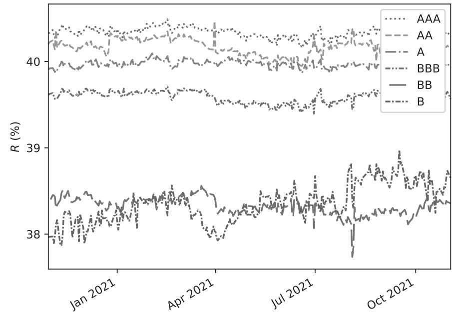
monotonicity property being accurately reflected in the graphs, that is, to higher (lower) ratings correspond lower (higher) credit spreads.

We now analyze the approach proposed in Sec. 3.1, i.e., the WCIMS. The results are shown in Fig. 3.

As expected, in this case, we obtain proxy spreads which are below their counterparts depicted in Fig. 2. We also recall that, once proxy spreads are computed with the WCIMS, then in order to imply proxy credit curves, proxy




(a)




(b)

Fig. 4. Proxy recovery rates computed using the cross-intersection methodology for recovery rates available in Sec. 3.1 (based on regressing minus the logarithms of the recovery rates) with and without harmonically averaging the recovery rates before performing the regression, given different ratings, for the base bucket during the time period spanning from 1 November 2020 to 1 November 2021; see panels (a) and (b), respectively.

recovery rates are also necessary. For this reason, as outlined in Sec. 3.1, we have proposed different approaches (which share the same base rationale) for doing so. We provide Figs. 4(a) and 4(b) for two examples. As expected, the regressed recovery rates all have an order of magnitude neighboring 40%. Also, as it can be seen for instance in Varna *et al*. (2003), in practice there is not always monotonicity in recovery rates with respect to ratings (i.e., higher credit ratings are expected to have higher recovery rates), which is what we can see in Figs. 4(a) and 4(b) (despite regressing without harmonically averaging the recovery rates first, see Fig. 4(b), produces results which better satisfy the monotonicity with respect to ratings compared to Fig. 4(a)). For additional details concerning the link between ratings and recovery rates refer, e.g., to Altman *et al*. (2005).

The results of Figs. 4(a) and 4(b) indicate that, ball-park, just always using a recovery rate of 40% would be, in practice, an acceptable choice. In order to assess the impact on proxy default probabilities of a 40% recovery rate rather than any of the other values observed in Figs. 4(a) and 4(b), we consider a CDS belonging to the base bucket. For each of the dates in the period taken into account, we calculate the default probability for the relevant entity within a five-year horizon. The results are available in Fig. 5.

It is a well-known fact that default probabilities are not very sensitive to the value chosen for the recovery rate, and this is confirmed by Fig. 5. Figure 4(a) indicates that the uncertainty in recovery rate values is of the order of 4%, while Fig. 4(b) seems to indicate a slightly lower uncertainty, of the order of a couple of percentage points. Taking into account recovery rates which differ from the somewhat standard value of 40% as done in Fig. 5 results in default probabilities

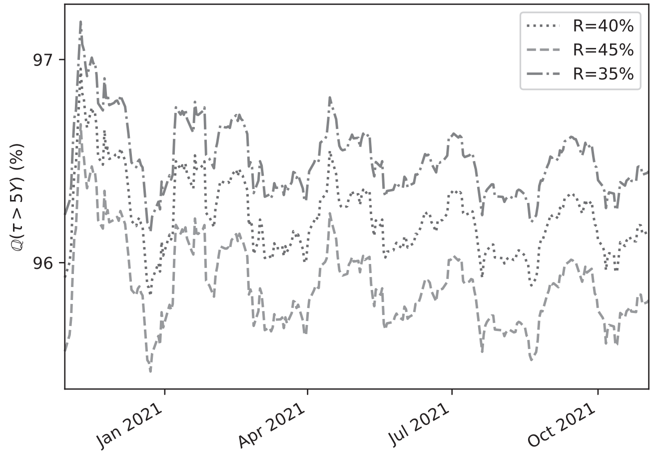
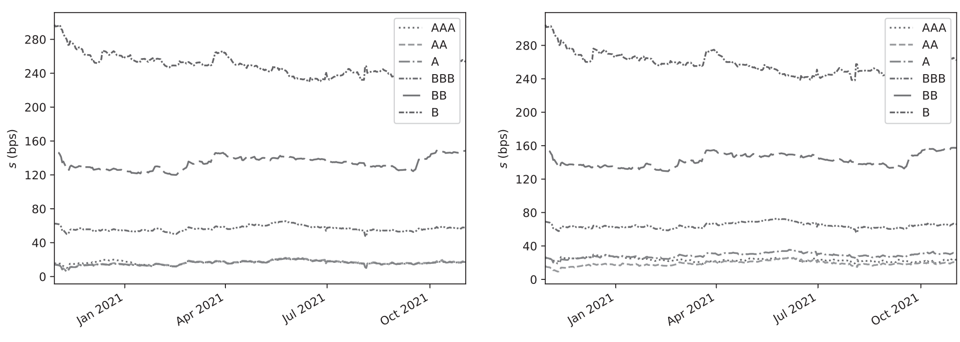



Fig. 5. Five-year default probabilities from an entity belonging to the base bucket during the time period spanning from 1 November 2020 to 1 November 2021 calculated using different recovery rates.

which do not differ much more than 1%. Given that recovery rates are not quoted, simply using the (often) commonly-used value of 40% seems a justified choice from a practitioner's angle in this respect.

We now show how the proxy CDS spreads for the base bucket look like if they are proxied by means of the WCIMHR. The results are shown in Figs. 6(a) and 6(b), which display a similar behavior.

We observe that the WCIMHR seems to perform less well than the WCIMS. In particular, the WCIMHR fails to properly distinguish between the top three investment-grade ratings. For the remaining three ratings, the results seem slightly more conservative than their counterparts in Fig. 3, and the monotonicity property



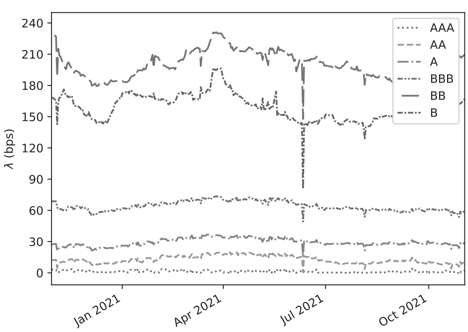

(a) (b)

Fig. 6. Proxy CDS spreads computed using the WCIMHR with and without constraints to guarantee positiveness, given different ratings, for the base bucket during the time period spanning from 1 November 2020 to 1 November 2021; see panels (a) and (b), respectively.

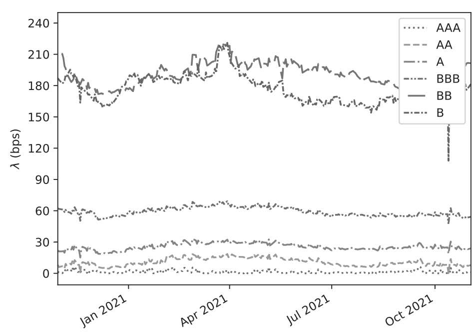
with respect to the ratings remains satisfied in this case. Note that, when constraints to guarantee the positiveness of the hazard rates are imposed as in Fig. 6(a) (otherwise, this might in principle produce negative hazard rates), having the regression coefficients less degrees of freedom causes the results to be less accurate in the area where the spreads are small (i.e., for entities with high credit worthiness), as in this case we are close to the boundary with zero. This is confirmed by the fact that, for lower credit ratings, the distinction between credit spreads linked to different ratings is crystal clear and, ballpark, in line with what observed in Fig. 3. If the constraints are removed, see Fig. 6(b), the distinction between the three top ratings is clearer than in Fig. 6(a), but the desirable monotonicity property is not always guaranteed. We highlight that recovery rates are not market-observable quantities, and that the same applies to hazard rates. Therefore, as it is also suggested by Figs. 6(a) and 6(b) compared to Fig. 3, the choice of the WCIMHR would be sub-optimal compared to the WCIMS.

We now look at the third option analyzed in this paper, i.e., the GWDOM. In particular, we look at this approach from two angles, i.e., by means of performing the minimization of (11) directly, but also by considering its approximated counterpart (13). The results, in terms of hazard rates, are depicted in Figs. 7(a) and 7(b), respectively.

First of all, we notice how, for investment-grade ratings, approaching the optimization problem in its original form (11) or via its approximation (13) produce similar results. This means that setting up the approximated error function (12) (here, we have chosen $\epsilon$ equal to 1bp) produces results in line with expectations. We notice in the results outlined in Figs. 7(a) and 7(b) that the hazard rates are monotonic with respect to the rating, expect for the “B” and “BB” cases. Here, the approximated approach (13) seems to perform slightly better than its counterpart




(a)




(b)

Fig. 7. Proxy CDS spreads computed using the GWDOM in its original form (11) and via its approximation (13), given different ratings, for the base bucket during the time period spanning from 1 November 2020 to 1 November 2021; see panels (a) and (b), respectively.

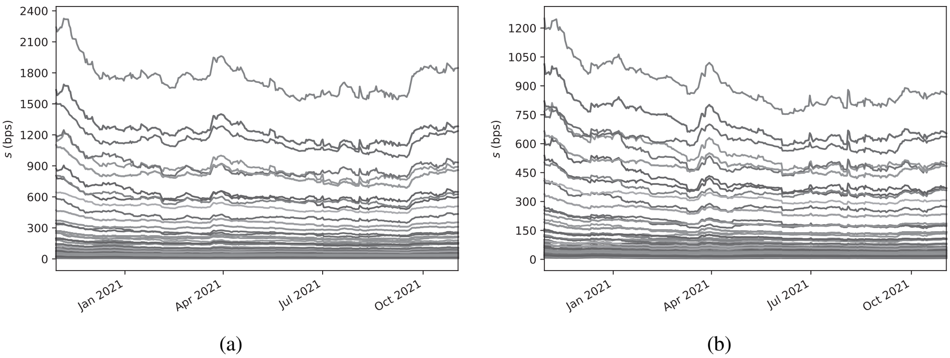



Fig. 8. Proxy CDS spreads computed using the WCIMHR with and without constraints to guarantee positiveness, given different ratings, for the base bucket during the time period spanning from 1 November 2020 to 1 November 2021; see panels (a) and (b), respectively.

(11). Nonetheless, we note that in both cases, and independently on the credit worthiness level, the trend of the hazard rates exhibits some evident sawtooth patterns. This is due to the fact that both optimization problems require several parameters to be estimated and, therefore, seem to be quite sensitive to the market data used as input. We observe this especially for the worst two ratings where, due to the data containing CDS quotes of badly-rated entities, the results show even less smooth features. Therefore, despite sound from a theoretical perspective, the GWDOM seems to underperform compared to both the WCIMS and the WCIMHR. We also empathize that the GWDOM is computationally much more intensive than the other two.

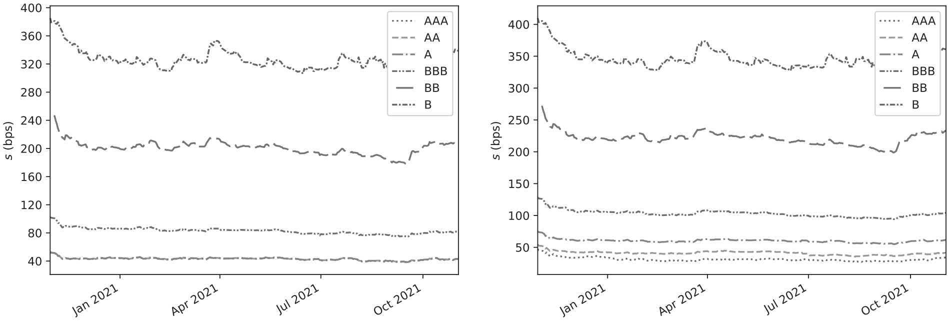
From the analysis we have carried out in this section, we can conclude that amongst the three approaches analyzed, the WCIMS clearly outperforms the other two. Therefore, we make a final comparison between the WCIMS and the CSM. In particular, for all the 144 possible buckets that can be defined given the choices and mappings described in Table 1, we plot the proxy spreads for the time period considered. The results are illustrated in Figs. 8(a) and 8(b), respectively.

We first highlight how both methodologies provide stable results for all the buckets with the same qualitative features (i.e., curve shapes). This also illustrates the robustness of the methodology originally proposed in Chourdakis *et al*. (2013) and further enhanced in Sourabh *et al*. (2018). We observe that the WCIMS produces, for rating, region, sector and seniority where CDS spreads are high, hazard rates that are below those of their counterparts calculated using the CSM. For combinations where, on the other hand, CDS are low, the two approaches tend to be closer to each other, or the other relationship might hold. This is because, as far as the intercept of the model is concerned, the CSM is consistent with the average of the (log) CDS spreads observed, while the WCIMS with the mean of the (log) harmonic averages of the bucketed spreads. This means that, in the WCIMS, lower weights (higher) are given to high (low) credit spreads. During the time frame taken into consideration, on average approximately 77% of the curves produced by the WCIMS are below their counterparts constructed by means of the CSM, with a standard deviation of roughly 3%.

As an example, we consider the second highest curves in Figs. 8(a) and 8(b). In the former case, at the beginning of the period, the CDS spread is approximately 1,600 bps, while in the second case it is roughly 1,000 bps. In the first case, by using the credit triangle relationship (2) and considering a recovery rate of 40%, it results that the hazard rate approximately equals 2,777 bps, while in the second case it roughly equals 1,667 bps. These hazard rates lead to five-year default probabilities of approximately 74% and 57%. However, for lower CDS spreads the magnitude of the differences between hazard rates would be milder, if not slightly lower in the case the CSM is used, in some cases.

## 5. Conclusion

The main contribution available in this paper is the development of a novel methodology for constructing proxy credit curves starting from observable CDS market quotes. In particular, we have investigated three possible approaches to take into account the default probability structure of CDSs within the cross-sectional framework of Chourdakis *et al*. (2013) and Sourabh *et al*. (2018), with the addition of explicitly including probability-related considerations by means of Wasserstein distances. Amongst the three possibilities envisaged, we have assessed how regressing proxy CDS spreads computed by means of Wasserstein barycenters for the observable buckets, in order to interpolate and extrapolate the available information to the empty buckets as well, seems to be the best strategy to follow. We have compared this methodology to the cross-section approach and observed that, in line with Michielon *et al*. (2022), neglecting the default probability distribution of the underlying CDSs might lead to overestimating credit risk. The analysis performed on the chosen methodology also confirms how stable, robust and flexible the cross-section methodology of Chourdakis *et al*. (2013) and Sourabh *et al*. (2018) is. By means of the hybrid curve construction framework outlined in the article at hand we have shown how Wasserstein distances can be employed to model default risk, and this hopefully paves the way to further developments towards this direction.




(a) (b)

Fig. 9. Proxy CDS spreads computed by applying the cross-section methodology to hazard rates with and without constraints to guarantee positiveness, given different ratings, for the base bucket during the time period spanning from 1 November 2020 to 1 November 2021; see panels (a) and (b), respectively.

## Appendix A. The Cross-Section Methodology on Hazard Rates

In Sec. 2.2, the basic features of the cross-section methodology (Choudakis *et al*., 2013; Sourabh *et al*., 2018) have been highlighted. Despite this paper dealing with the cross-section methodology from a different angle (i.e., by employing Wasserstein distances and barycenters), we want to highlight that, even if the regression equations in (7) are based on the logarithms of the CDS spreads, one could look at this approach from an hazard rate perspective. That is, one could reinterpret (7) by considering, for every $i \in \{1, \ldots, K\}$, a relationship of the form

$$\lambda_i = \xi_i(\boldsymbol{x}), \tag{A.1}$$

where each hazard rate is computed by means of the credit triangle (2).[8] Also in this case, due to the fact that the regressed hazard rates are expected to be positive, one could either impose the regression coefficients to be positive *a priori*, or floor the results, if needed, by a sufficiently small positive constant. By also regressing for every $i \in \{1, \ldots, K\}$ the recovery rates in a similar manner as Sec. 3.1 outlines, one could then employ the credit triangle (2) again to back out the proxy CDS spreads. For completeness, in Figs. A.1(a) and A.1(b), we illustrate how the proxy CDS spreads will look like in this case.

As Fig. A.1(a) shows, restraining the regression coefficients to be positive imposes too strict conditions on the regression results, leading to the ratings equal to "A" or better to be almost indistinguishable from each other (a similar behavior has been observed in Fig. 6(a)). On the other hand, not imposing the constrains (in

[8]As an additional variation, one could also perform the regression on the logarithms of the hazard rates.

this case it is not flooring of the regressed variables is needed) allows to clearly distinguish amongst all the ratings and, in addition, to guarantee the desired monotonicity property with respect to the credit worthiness. We also observe that regressing at hazard rate level results in proxy CDS spreads that are more conservative than their counterparts calculated using the cross-section methodology depicted in Fig. 1. Also in this case, we highlight that recovery rates are derived quantities and, therefore, they are not quoted (and the same applies to hazard rates). So, despite in this case the regression methodology applied to hazard rates provides stable results which are monotonic with respect to credit ratings, regressing the spreads directly (which are observable) should still be the preferred choice.

## Acknowledgment

Matteo Michielon is very grateful to Raoul Pietersz for highly valuable discussions. The authors are grateful to an anonymous referee for providing valuable comments on the first version of this paper. The opinions expressed in this paper are solely those of the authors and do not necessarily reflect those of their current and past employers.

## References

Abbas-Turki, LA, S Crépey and B Diallo (2018). XVA principles, nested Monte Carlo strategies and GPU optimization, *International Journal of Theoretical and Applied Finance*, 21(6), 1850030.

Altman, EI, B Brady, A Resti and A Sironi (2005). The link between default and recovery rates: Theory, empirical evidence and implications, *The Journal of Business*, 78(6), 2203–2228.

Brigo, D, M Morini and M Tarenghi (2010). Credit calibration with structural models: The Lehman case and equity swaps under counterparty risk. In eds. T Bielecki, D Brigo and F Patras, *Recent Advancements in Theory and Practice of Credit Derivatives*, Bloomberg Press.

Brigo, D and A Pallavicini (2014). Nonlinear consistent valuation of CCP cleared or CSA bilateral trades with initial margins under credit, funding and wrong-way risks, *Journal of Financial Engineering*, 1(1), 1450001.

Chourdakis, K, E Epperlein, M Jeannin and J McEwen (2013). A cross-section across CVA, Nomura.

Cornujols, G, J Peña and R Tütüncü (2018). *Optimization Methods in Finance*, Cambridge University Press, Cambridge.

EBA (2013). EBA final draft regulatory technical standards on credit valuation adjustment risk for the determination of a proxy spread and the specification of a limited number of smaller portfolios under article 383(7) of regulation (EU) No 575/2013 (capital requirements regulation – CRR).

Green, A (2016). *XVA. Credit, Funding and Capital Valuation Adjustments*, Wiley, Chichester.

Gregory, J (2020). *The xVA Challenge: Counterparty Risk, Funding, Collateral, Capital and Initial Margin*, Wiley, Chichester.

Heynderickx, W, J Cariboni, W Schoutens and B Smits (2016). The relationship between risk-neutral and actual default probabilities: The credit risk premium, *Applied Economics*, 48(42), 4066–4081.

Junge, B and AB Trolle (2015). Liquidity risk in credit default swap markets, Swiss Finance Institute Research Paper, 13–65.

Lee, C and C Kuo (2015). Combining hazard rates with the CreditGrades model: A hybrid method to value CDS contracts, *International Journal of Financial Engineering*, 2(4), 1550037.

Longstaff, FA, J Pan, LH Pedersen and KJ Singleton (2011). How sovereign is sovereign credit risk? *American Economic Journal: Macroeconomics*, 3(2), 75–103.

Michielon, M, A Khedher and P Spreij (2022). Proxying credit curves via Wasserstein distances, *Annals of Operations Research*.

Pan, J and KJ Singleton (2008). Default and recovery implicit in the term structure of sovereign CDS spreads, *The Journal of Finance*, 63(5), 2345–2384.

Panaretos, VM and Y Zemel (2020). *An Invitation to Statistics in Wasserstein Space*, Springer, Cham.

Press, WH, SA Teukolsky, WT Vetterling and BP Flannery (2007). *Numerical Recipes. The Art of Scientific Computing*, Cambridge University Press, Cambridge.

Ruiz, I (2015). *XVA Desks – A New Era for Risk Management*, Palgrave, New York.

Sourabh, S, M Hofer and D Kandhai (2018). Liquidity risk in derivatives valuation: An improved credit proxy method, *Quantitative Finance*, 18(3), 467–481.

Varna, P, R Cantor and D Hamilton (2003). Recovery rates on defaulted corporate bonds and preferred stocks, 1982–2003, Moody's Investors Service.

Villani, C (2003). *Topics in Optimal Transportation*, American Mathematical Society, Providence.

Villani, C (2009). *Optimal Transport: Old and New*, Springer-Verlag, Berlin.

Wales, DJ and JPK Doye (1997). Global optimization by Basin-hopping and the lowest energy structures of Lennard–Jones clusters containing up to 110 atoms, *The Journal of Physical Chemistry A*, 101(28), 5111–5116.

White, R (2014). Pricing and risk management of credit default swaps, *OpenGamma Quantitative Research*, 16.

Xiang, Y, S Gubian, B Suomela and J Hoeng (2013). Generalized simulated annealing for global optimization: The GenSA package, *The R Journal*, 5(1), 13–29.

Zou, L and W Li (2022). Estimating actual probability of default from structural models. *International Journal of Financial Engineering*, 9(1), 2150045.