BANK OF CANADA
BANQUE DU CANADA

Staff Working Paper/Document de travail du personnel—2024-46

Last updated: November 26, 2024

Is This Normal? The Cost of Assuming that Derivatives Have Normal Returns

by Radoslav Raykov

Financial Stability Department
Bank of Canada
rraykov@bankofcanada.ca

DOI: https://doi.org/10.34989/swp-2024-46 | **ISSN** 1701-9397

Acknowledgements

The author would like to thank Lerby Ergun, Javier Ojea Ferreiro, Yaz Terajima, Christian Friedrich, colleagues and seminar participants at the Bank of Canada for helpful comments. Mila Nikolova provided research assistance.

Abstract

Derivatives exchanges often determine collateral requirements, which are fundamental to market safety, with dated risk models assuming normal returns. However, derivatives returns are heavy-tailed, which leads to the systematic under-collection of collateral (margin). This paper uses extreme value theory (EVT) to evaluate the cost of this margin inadequacy to market participants in the event of default. I find that the Canadian futures market was under-margined by about $1.6 billion during the Great Financial Crisis, and that the default of the highest-impact participant generates a cost of up to $302 million to be absorbed by surviving participants. I show that this cost can consume the market's entire default fund and result in costly risk mutualization. I advocate for the adoption of EVT as a benchmarking tool and argue that the regulation of exchanges should be revised for financial products with heavy tails.

Topics: Financial institutions; Financial stability
JEL codes: G10, G11, G20

Résumé

Les bourses de produits dérivés établissent souvent leurs exigences en matière de garantie – qui sont essentielles à la sûreté des marchés – en utilisant de vieux modèles de risque qui supposent une distribution normale des rendements. Cependant, la distribution des rendements sur les produits dérivés présente des queues épaisses. Résultat : les garanties (marges) perçues sont systématiquement insuffisantes. Dans cette étude, j'utilise la théorie des valeurs extrêmes pour évaluer le coût de cette insuffisance de marges qu'assument les participants au marché en cas de défaillance. Je constate que le déficit de marges dans le marché canadien des contrats à terme était d'environ 1,6 milliard de dollars pendant la grande crise financière. Je conclus également que si le participant dont la défaillance aurait la plus grande incidence en dollars faisait effectivement défaut, les coûts pour les participants restants pourraient atteindre 302 millions de dollars. Je montre que ce coût pourrait vider la totalité du fond de défaillance du marché et entraîner une mutualisation des risques dispendieuse. Je préconise l'adoption de la théorie des valeurs extrêmes comme outil de référence, et soutiens que la réglementation des bourses concernant les produits financiers dont les distributions de rendement ont des queues épaisses devrait être revue.

Sujets : Institutions financières; Stabilité financière
Codes JEL : G10, G11, G20

Disclaimer

This work does not represent a supervisory assessment or a risk assessment of any financial entity mentioned therein. The analysis presents strictly hypothetical risk scenarios reflecting only the views and opinions of the author. These views and opinions cannot be attributed to the Bank of Canada, to any of its departments, or to its Governing Council.

1 Introduction

Derivatives exchanges play a fundamental role for the safety and efficiency of financial markets. They establish the rules of the marketplace, manage counterparty risk, and collect collateral, also referred to as *margin*, to buffer that risk. The amount of margin pledged for a transaction plays a key role in whether institutions are affected by counterparty defaults, whether default stress spills over to healthy entities, and whether distressed institutions go out on the market to seek liquidity, potentially worsening market disruptions (Brunnermeier and Pedersen, 2009). Hence, adequately computed margin requirements are key for the smooth functioning of financial markets (CPMI-IOSCO, 2012).

Many exchanges calculate margin requirements with standard but dated risk models assuming normal returns, contrary to the heavy-tailed nature of financial return distributions. This leads to the systematic undercollection of margin (Cotter and Dowd, 2017; Longin, 1999), which can be costly to market participants in the event of default. This paper revisits the tail risk adequacy of the most prevalent normal-based risk model for futures margining, SPAN, and quantifies the cost to market participants from the inadequate normal distribution assumption. To do so, I use recent advances in extreme value theory (EVT), specifically designed for heavy tail estimation, and leverage on proprietary futures positions data from the Montreal Exchange between 2003 and 2011.

SPAN was developed by the Chicago Mercantile Exchange in the 1980s and adopted by 54 leading exchanges and clearing organizations, such as the London Stock Exchange (via LCH.Clearnet), Eurex Exchange, NYMEX exchange, the Montreal Exchange, and Options Clearing Corporation, to name a few. The *de facto* standard for futures margining, SPAN estimates margin requirements based on a parametric distribution of returns to achieve computational simplicity, transparency, and ease of use. For futures margining, the normal distribution is most often assumed. However, this simplifying assumption comes at a cost. The more heavy-tailed the actual distribution of risks is, the larger is the deviation from normality, and, as a consequence, the lower the level of margin charged relative to the model's stated nominal coverage – a phenomenon I henceforth refer to as *under-margining*.[1] If a market participant defaults, this under-margining can facilitate risk spillovers above posted margin. The paper quantifies these spillovers and shows that they are large

[1]This term is not an indication of an exchange's actual compliance with relevant risk management regulations or targets. Exchanges can, for example, set the target coverage high enough above the regulatory minimum so that empirically, they still meet regulation even if the risk model makes unrealistic theoretic assumptions.

enough to consume not only the defaulter's pre-pledged resources, but also those of remaining market participants, effectively disabling the market's default fund as a risk mitigation device. This calls for a fresh rethink of the risk models in place and their underlying assumptions.

The paper quantifies the cost of under-margining conditional on the default of the highest-impact participant by simulating the Montreal Exchange's actual default management process. To gauge the degree of under-margining, the paper estimates margin requirements with EVT, which allows more accurate estimation of tail risk quantiles based on power function approximations. This approach, which relies on estimating a parameter α, called a tail index, permits inference about "the tail of the tail" based on the fact that heavy-tailed distributions exhibit scaling behavior described by Pareto power laws (De Haan and Ferreira, 2006). Thus, EVT can predict the severity of a given financial event and its frequency even outside the range of observed data without restrictive distributional assumptions; this feature makes EVT a prime candidate for examining heavy-tailed data. Recent developments in EVT have resulted in more frequent finance applications, which have provided significant insights for financial risk management and market behavior (e.g., Jansen and de Vries, 1991; Van Oordt and Zhou, 2016; Davydov, Vähämaa, and Yasar, 2021. For a compendium, see Longin (editor), 2017).

By contrast, SPAN was designed under the limitations of the 1980s statistical theory and computing technology, which sometimes required significant computational shortcuts. Since margin requirements aim to protect market participants with a high level of statistical confidence (often as high as 99.87%), SPAN's designers addressed the need to deal with events outside of the range of observed data by adopting a parametric distribution – the normal distribution – to compute extreme quantiles. However, derivatives returns and futures returns in particular are heavy-tailed rather than normal (Cotter and Dowd, 2017; Longin, 1999).

The main problem with the normality assumption in this context is that it can be financially costly. The first cost is the potentially inadequate margin level and reduced protection of the rest of the market against counterparty risk. However, in most markets served by exchanges, default risks above a certain level are mutualized,[2] which causes a second cost: default risk spillovers to non-defaulting entities, referred to as *survivors*, which need to absorb residual default costs according to the exchange's rules. This risk-propagation potential is a major concern in a number of studies

[2]Post-2008, most standardized derivatives markets are centrally cleared consistent with the G20 mandate on central clearing.

(Duffie and Zhu, 2011; Menkveld, 2017; Raykov, 2022; Paddrick, Rajan, and Young, 2020). The default of an under-margined participant therefore poses not only the risk that collateral may be insufficient to address the default, but also that survivors end up paying for the inadequate margin. At the same time, centrally cleared markets also feature elaborate mechanisms for the diffusion of spillovers among participants using a complex sequence of defaulter- and survivor-pledged resources such as *default funds*, reviewed in Section 2.3. Thus the effect of under-margining on surviving participants is far from apparent.

To estimate more accurate margins and quantify risk spillovers from under-margining, I rely on recent advances in EVT in the estimation of a distribution's tail index – the shape parameter determining the heaviness of the tail. Tail index estimation is traditionally performed with the Hill (1975) quasi maximum-likelihood estimator. Two traditional obstacles in this literature are the large number of observations required by the Hill estimator as well the lack of a universally agreed method to determine the tail threshold. To overcome these challenges, I use three cutting-edge tail threshold selection methods, some of which are geared specifically towards small samples: the Huisman et al. (2001) small-sample method, the Danielsson et al. (2019) KS-quantile method, and an automated version of the heuristic "eyeballing" method.

My results point to three major findings. Firstly, under-margining due to SPAN's normality assumption is pervasive and results in significant costs to market participants. In the aggregate, the Canadian futures market under-collected margin amounting to 1.6 billion CAD during the Great Financial Crisis, with the default of the highest-impact participant generating a maximal cost of 254 to 302 million solely due to the difference between EVT and SPAN margins. In a centrally cleared market, this cost would have to be mutualized by surviving participants; I show that this would be sufficient to deplete the survivors' entire default fund, thereby exhausting the market's second line of defense and resulting in replenishment calls averaging at about half of the representative participant's default fund deposit.

Secondly, I find SPAN's under-margining is not limited to crises and occurs despite the risk model largely meeting the 99% confidence interval required by financial market regulation; this happens because, far into the tail, small differences in probability coverage translate into large differences in margin. As a consequence, I recommend that regulators should reconsider if the 99% minimum threshold is adequate for heavy-tailed data, and that exchanges should consider using

EVT to benchmark existing risk models.

Thirdly, my results are not sensitive to the choice of period (crisis versus calm period) and are not driven by differences in the transitional dynamics of SPAN versus EVT following a shock. I find that the aggregate margin shortfall peaked at 1.6 billion during 2008, and at 1.06 billion in the calm period. Similarly, the default cost for the highest-impact defaulter was 254 million at the crisis peak and 148 million in the calm period, consuming 110% and 122% of the survivors' default fund, respectively. In this light, the ad hoc margin add-ons implemented by some exchanges may have been justified, but their absence in normal periods still leaves some uncovered exposures.

These results are consistent with Cotter and Dowd (2017), who use EVT to find under-margining in 12 European stock futures, and those of Longin (1999), who reports similar findings for silver futures. My work improves on theirs in two major ways: (1) by calculating time-varying, rather than unconditional, margins with EVT and (2) by quantifying the cost of under-margining to survivors in centrally cleared markets by leveraging on actual trading positions data. To my knowledge, the latter improvement is entirely new to the literature.

Therefore, the paper adds insight to several streams of literature. On the one hand, it contributes to the literature on the safety of derivatives markets, which has featured lively recent discussions about the desirability of central clearing, margining and systemic risk, crowded trades, and the joint exposures they create. This literature has debated the costs and benefits of central derivatives clearing (Duffie and Zhu, 2011) and the risks associated with it. For instance, Menkveld (2017) discusses the risk of crowded trades, while Cruz Lopez et al. (2017) is concerned with the adequacy of initial margin to absorb joint defaults. Raykov (2022) shows that SPAN, although dated, buffers joint risks adequately because of the loss-equals-default assumption. Paddrick, Rajan, and Young (2020) discuss risk propagation in CDS markets via the variation margin channel and debate the adequacy of default funds calibrated to withstand the default only of the largest one or two participants.

On the other hand, this paper also offers an econometric application: a comparison of the performance of EVT tail threshold selection methods on futures returns data. Earlier EVT applications by Longin (1999), Cotter (2001), and Cotter and Dowd (2017) tend to select thresholds based on minimizing the asymptotic mean squared errors, which has subsequently been shown to perform poorly in finite samples. Offering three tail selection methods, one of which specifically geared

towards small samples, provides valuable comparison of their performance. Most importantly, the paper shows the economically significant costs that can be incurred by market participants relying on normality-based risk models.

2 Market background and data

2.1 The SPAN model

The SPAN risk model, introduced by the Chicago Mercantile Exchange in 1988, is the *de facto* industry standard for futures margining. SPAN margins are designed to cover future price moves with 99.87% statistical confidence by calculating the maximum likely loss that a portfolio might sustain daily, based on a measure of how difficult it is to liquidate a defaulting position of an asset with a given volatility. (The 99.87% target aims to meet and exceed the 99% minimum confidence level mandated by the international regulation applicable to exchanges.)[3] The formula for *base initial margin* (IM) determines the fraction of an open position's market value that must be pledged to the exchange as collateral. SPAN divides the portfolio into contract families with the same underlying asset, called combined commodity groups, and independently estimates a margin charge for each combined commodity group (Chicago Mercantile Exchange, 2019). For futures, the SPAN base initial margin during our sample period (2003–2011) was calculated by the Montreal Exchange as

$$IM = \sqrt{\tau}\ 3\sigma, \tag{1}$$

where IM stands for initial margin (the fraction of the position value that needs to be collateralized), τ is the number of liquidation days (for futures, the model assumes $\tau = 2$), and σ is a volatility estimator of the contract's daily price variation. Since volatility can be estimated over different time horizons, SPAN conservatively takes volatility σ as the maximum standard deviation of the contract return over 20, 90, and 260 business days (corresponding respectively to approximately 4 weeks, 4 months, and 1 year calendar time). This is done in an effort to reflect the maximum severity of a tail event across different horizons (however, as I show, even taking the maximum is not adequate when normality is involved). Increases in volatility, which is recalculated monthly,

[3] According to CPMI-IOSCO's (2012) Principles for Financial Market Infrastructures, Principle 6, Key consideration 3: "Initial margin should meet an established single-tailed confidence level of at least 99 percent with respect to the estimated distribution of future exposure."

directly lead to higher initial margin. This setup reflects the original CME implementation effective at the Montreal Exchange during the sample period (2003–2011); post-2011 improvements to risk management are therefore not reflected.[4] Likewise, following Cruz Lopez et al. (2017), I do not calculate several discretionary margin adjustments unrelated to the normality of the tail, called intra-commodity and inter-commodity spread charges, which constitute a very small fraction of margin.[5]

The base initial margin, calculated by SPAN is subsequently fed into an array of 16 predetermined risk scenarios, each varying the price of the underlying asset up and down by a fixed multiplicative factor. These scenarios aim to reflect any non-linearities between the derivative's price and that of its underlying asset for instruments such as options. However, since futures are linear in the price of the underlying, for the pure futures portfolios in my data, the base initial margin is given directly by equation (1).

The literature on SPAN has historically interpreted SPAN's base initial margin formula as a normal distribution-based estimate of the 0.13% quantile corresponding to 99.87% coverage, since, for a normal distribution, this quantile is located exactly 3 standard deviations away from the mean.[6] Using value-at-risk terminology, the 0.13% quantile can be denoted as $\text{VaR}_{0.13\%}$. The adjustment $\sqrt{\tau}$ with $\tau = 2$ is applied to the 1-day loss above to produce the 2-day loss, as the exchange estimates it may take up to 2 days to liquidate a defaulting futures position (Canadian Derivatives Clearing Corporation, 2012). In line with this, I follow Longin (1999) and Cotter and Dowd (2017) to dispense with the normality assumption and generalize SPAN's formula to

$$IM^{EVT} = \sqrt{\tau}\ \text{VaR}_{0.13\%}, \tag{2}$$

where the corresponding tail quantile $\text{VaR}_{0.13\%}$ is estimated with EVT, instead of being assumed to equal 3σ as dictated by the normality assumption. Under these authors' approach, the difference between this quantile's EVT estimate and the normal-based estimate provides a measure of under-

[4] Post-2011, the Montreal Exchange amended its SPAN implementation with margin floors, procyclicality-mitigating smoothing of the volatility estimator σ, and various other enhancements.

[5] These charges adjust the base margin for existing price correlations between same-type contracts expiring in different months, or for mutually offsetting positions of correlated but non-identical contracts expiring in the same month. They constitute a small percentage of margin and are not consistently implemented across exchanges.

[6] For other products, such as options, the theoretical t distribution is also frequently used. I focus on futures because their tails are sufficiently heavy to deviate from normality without being as heavy as those of the t distribution.

margining from assuming normality.

2.2 Data

I use data on contract returns, prices, and proprietary market positions in the three most liquid futures contracts traded at the Montreal Exchange between January 2, 2003, and March 31, 2011.[7] These contracts are the BAX future over the 3-month CDOR interest rate, the CGB future over the 10-year Government of Canada bond rate, and SXF future over the S&P/TSX 60 index capturing the 60 most liquid Canadian stocks. Collectively, these three contracts represent over 90% of the value of futures positions and over 75% of all Canada-issued derivatives, and are highly representative of Canada's derivatives market (Campbell and Chung, 2003; TMX Montreal Exchange, 2013a, 2013b, 2013c). The institutions with open futures positions are listed in Table 1. The positions and price data, provided at daily frequency, contains a total of 117 contracts of the above three types with 39 expiry dates for BAX contracts and 34 expiry dates for CGB and SXF contracts. Contracts with different expiry dates allow market participants to bet on or hedge against movements in interest rates and the stock market in concrete future periods. I calculate monthly SPAN base initial margins during the sample period using equation (1), along with their EVT analogs, matching the Montreal Exchange's actual update schedule. Table 2, Panel A, shows that, overall, the BAX future over the 3-month rate was by far the most popular with a market share of 71%, followed by the government bond future CGB at 16%, and the stock market future SXF at 13%.

Since margins are calculated by contract type, contract volatility matters for a portfolio's margin adequacy. Table 2, Panel A, shows that BAX was the least volatile contract, with an average return of 0.003%, standard deviation of 0.07%, and a maximum price move of 0.55%; the CGB future was more volatile, with an average return of 0.02%, standard deviation of 0.35, and largest price move of -1.99%; and SXF was the most volatile, with mean return 0.05%, standard deviation of 0.27%, and largest price move of -10.20%. Quantile normal plots give a visual indication of how the three distributions' tails differ from the respective normal. Figure 1 displays a quantile normal plot for each contract type; deviations from normality are plotted as deviations from the main diagonal.

The plots show all three contracts' returns are symmetric[8] and heavy-tailed. For BAX, for

[7]I thank the Canadian Derivatives Clearing Corporation (CDCC) for providing the data.

[8]In section 3.2.4, I confirm with a formal test that the left and right tails are symmetric using the split-sample statistic of Loretan and Phillips (1994).

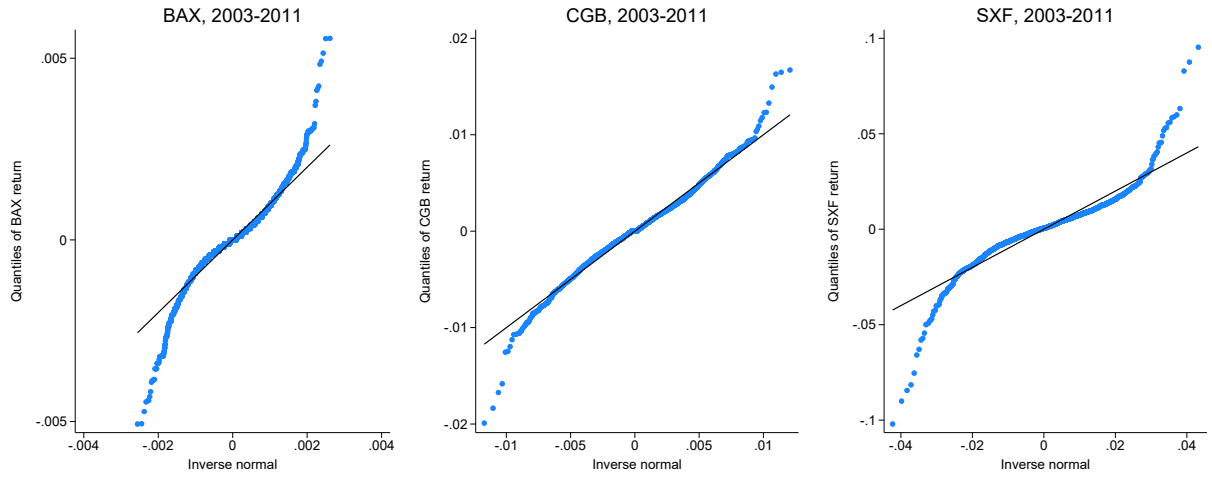

Figure 1: Quantile normal plots for the returns of the BAX, CGB, and SXF contracts between January 1, 2003, and March 31, 2011.

example, the maximum deviation from normality in the left tail amounts to almost half of the maximum price move (0.0025), with similar results for the right tail. For CGB, the extreme left tail falls by 0.008 below the left tail of the respective normal quantile, also close to half of the maximum price move; the right tail is a bit thinner, but still symmetric. For SXF, the maximum deviations from normality, in both the left and the right tail, amount to about 0.006, more than half of the maximum price move. Assuming normality for these heavy tails therefore biases tail quantile estimates towards the center, resulting in under-margining. In centrally cleared markets, such as those for standardized derivatives, this can create default risk spillovers.

2.3 The risk spillover architecture in centrally cleared markets

Institutional setup. Modern exchanges buffer and diffuse default risk using a complex sequence of defaulter and survivor resources.[9] In a centrally cleared exchange, all trades are submitted to an entity, called a central counterparty (CCP), which becomes the buyer to every seller and the seller to every buyer, thereby *novating* the trade. Novation replaces the original counterparties with the exchange's CCP, a centralized entity with regulated risk management, thereby subjecting all market participants to equal and rigorous standards in terms of collateral requirements. In the event of default by one of the original transactors, the central counterparty is responsible for

[9]This section is loosely based on Raykov (2022).

making good on the trade. For this purpose, a CCP uses a sequence of pre-pledged defaulter and survivor resources, known as the *default waterfall*, to cover default costs.[10] Figure 2 shows the actual default waterfall used by the Montreal Exchange's affiliated CCP, the Canadian Derivatives Clearing Corporation (CDCC). It is highly representative of the order of resources used in CCPs worldwide and is used in my subsequent calculations on risk mutualization.

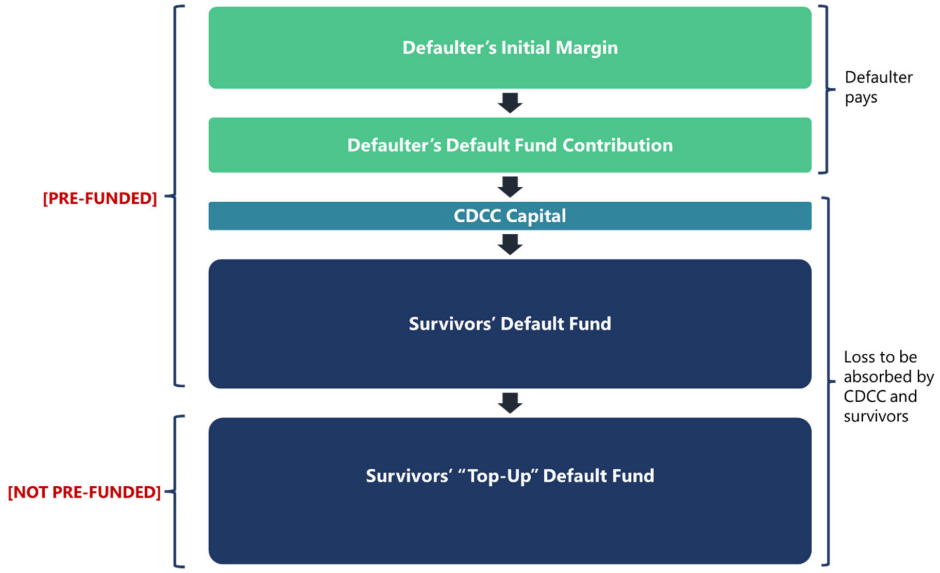

Figure 2: A schematic showing the default waterfall of CDCC, the Montreal Exchange CCP.

Figure 2 shows that the first and most important pre-funded resource to absorb an uncovered default is the defaulter's initial margin. If the defaulted amount is larger than posted margin, the next tranche of resources used is the defaulter's contribution in the CCP's default fund – a pre-funded resource pool shared by market participants, where each member contributes proportional to its positions. The default fund is not transaction-specific and is maintained at all times as a second line of pre-funded resources. Exchanges use various methodologies to size their default funds, but they all typically involve a stress test on the one or two participants most likely to infringe on the default fund in a hypothetical stress scenario.[11] Defaults exceeding the defaulter's margin and default fund spill over to survivors, including the exchange's CCP. In such cases, the CCP first applies a capped amount of its own capital (in the case of the Montreal Exchange, $5 million CAD, labeled "CDCC capital"[12]), after which survivors' default fund contributions

[10]For instance, see Canadian Derivatives Clearing Corporation (2012), Schedule A, Appendix 1, pp. 1–5.
[11]See CPMI-IOSCO (2012), Principle 7, Key Consideration 4.
[12]This amount is projected to be increased to $15 million after 2024.

are consumed pro rata, resulting in risk mutualization. Since the consumed amounts need to be subsequently replenished, resulting in payment obligations for members, the bottom rectangle in Figure 2 is labeled "survivors' top-up default fund." These mandatory top-ups, or replenishments, if overly large, can be instrumental in propagating financial distress from a default; therefore, many exchanges cap them. Nonetheless, despite this, Paddrick, Rajan, and Young (2020) argue that current default fund sizing practices are inadequate, while Raykov (2022) shows that, when default risk is properly measured, the effect on financial stability is not as extreme (although replenishment calls can still be in the millions). The extent of survivors' default fund utilization is therefore a key measure of risk spillovers to surviving participants.

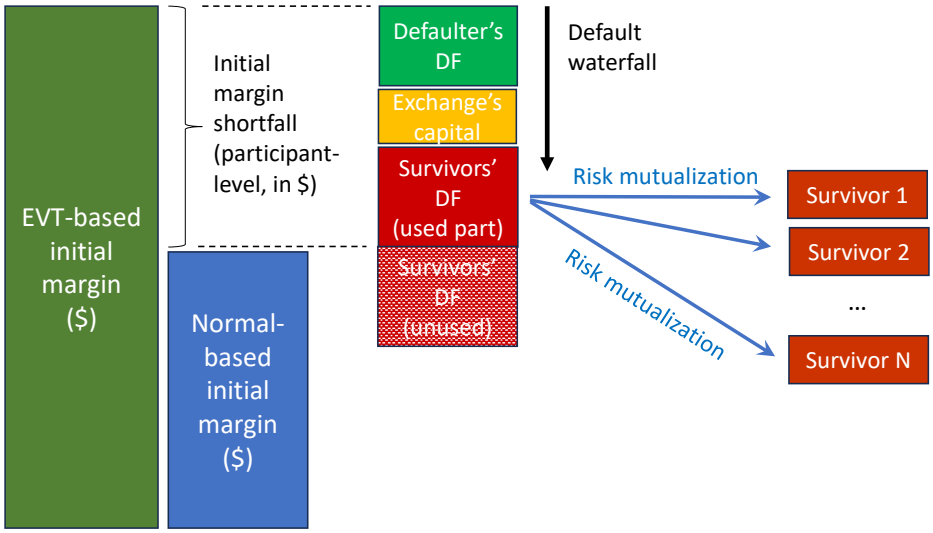

Figure 3: A roadmap of this paper. EVT stands for extreme value theory and DF stands for default fund.

Roadmap. The roadmap of the paper in calculating risk spillovers is visualized in Figure 3. First, I use EVT to calculate a margin requirement targeting the same tail quantile as SPAN for each contract in each participant's portfolio and arrive at an EVT-based margin requirement in dollars at the market participant level. To gauge the cost of the normality assumption in the event of default, I stress each contract with a price move corresponding to its 0.13% empirical quantile, that is, just large enough to consume the participant's EVT-based initial margin corresponding to 99.87% coverage.[13] Since the data is heavy-tailed, this quantile is typically farther out in the tail than in

[13]That is, I deliberately implement a multi-asset shock with the same relative severity in each asset. This shock is likely more severe than the historical stress scenarios used by the CDCC to calibrate its default fund, hence the results should not be interpreted as implying a deficiency in CDCC's risk management.

the normal distribution; hence, this price move breaches the SPAN margin. The dollar amount of this spillover is then buffered by the pre-pledged resources from the default waterfall (Figure 2). When the breach is large enough to exceed the defaulter-pays and the exchange's resources, it spills over to the survivors' portion of the default fund, thereby causing risk mutualization. I label the participant causing the largest dollar infringement on the survivors' default fund as *the largest-impact participant* and assume this participant defaults in the event of stress, consistent with the approach of international regulation for default fund stress testing.[14] The risk mutualization consumes each survivor's default fund deposit pro rata according to the Montreal Exchange's actual default management process, thereby imposing costs on affected participants. My end goal is to quantify these costs and the extent of risk mutualization they cause.

3 Empirical strategy

3.1 Risk measurement

To measure the risk from inadequate SPAN margining, I introduce contract-level and aggregate risk measures. On the individual contract level, Cotter and Dowd (2017) and Longin (1999) propose a measure of under-margining based on the difference between the EVT estimate and the SPAN estimate of the tail quantile corresponding to a fixed value at risk (here, 0.13%, corresponding to 99.87% coverage). I therefore compute the margin shortfall for each contract c on day t as

$$Shortfall_{c,t} = IM^{EVT}_{c,t} - IM_{c,t}, \tag{3}$$

where $IM^{EVT} = \sqrt{2}\ \mathrm{VaR}^{EVT}_{0.13\%,c,t}$ and $IM_{c,t}$ is the SPAN initial margin defined in equation (1). This shortfall captures the fraction of the position value that needs to be further collateralized (in addition to SPAN) to reach the desired single-tail confidence of 99.87%. The shortfall can be converted to dollars when applied to the position of a market participant i in the relevant contract c:

$$Shortfall^{\$}_{i,c,t} = |Netposval_{i,c,t}| \cdot \left(IM^{EVT}_{c,t} - IM_{c,t}\right), \tag{4}$$

[14] See CPMI-IOSCO's (2012) Principles of Financial Market Infrastructures, Principle 4 (Credit risk).

where $|Netposval_{i,c,t}|$ is participant i's net position value in contract c, taken in absolute terms (since, by convention, short positions are entered as negative numbers). This measure can further be aggregated across contracts on the market participant level as

$$Shortfall^{\$}_{i,t} = \sum_{c} |Netposval_{i,c,t}| \cdot \left(IM^{EVT}_{c,t} - IM_{c,t}\right). \tag{5}$$

From this it is clear that the extent to which a participant is under-margined depends on both its position size and its portfolio composition. One can also aggregate margin shortfalls at the total market level as

$$MktShortfall^{\$}_{t} = \sum_{i} \sum_{c} |Netposval_{i,c,t}| \cdot \left(IM^{EVT}_{c,t} - IM_{c,t}\right). \tag{6}$$

Conditional on a price move equal to or exceeding $\text{VaR}^{EVT}_{0.13\%}$, if member i defaults, its dollar shortfall will spill over to be absorbed by the next tranche of defaulter-pays resources: i's own default fund deposit DF_i. According to the default management process, beyond this deposit, risk mutualization begins, consuming the CCP's \$5 million pledge and then the surviving members' default fund deposits. Since the assumed stress scenario corresponds exactly to $\text{VaR}^{EVT}_{0.13\%,c,t}$ for each contract, this ensures that any such spillovers are entirely due to the difference between EVT and SPAN.[15]

To gauge how institutions are affected by risk mutualization on the individual level, I follow Raykov (2022) and Brennan et al. (2024) to calculate the risk mutualization amounts to be levied at each surviving member j after default of participant i according to Montreal Exchange's actual default management process. According to the rules of the exchange, default costs in excess of defaulter-pays resources are distributed among survivors proportional to their respective 60-day shares of initial margin (used as a proxy for risk-weighted activity):[16]

$$Cost_{j,t} = IMshare^{60}_{j,t} \cdot \max\left\{0, \left(Shortfall^{\$}_{i,t} - DF_{i,t} - 5\right)\right\}, \tag{7}$$

where $IMshare^{60}_{j,t}$ is participant j's 60-day share of initial margin, and the term inside the braces

[15] Since this is a multi-asset shock with equal relative severity across all contracts, this scenario is likely more severe than the historical scenarios used by exchanges to size their default funds.

[16] Canadian Derivatives Clearing Corporation (2012), pp. 27–28.

is the dollar amount that needs to be mutualized by survivors. It equals the risk spillover above the defaulter i's pre-pledged resources (margin and default fund deposits) and the exchange's own capital of $5 million, after which risk mutualization begins (cf. Figure 2). The dollar amount $Cost_{j,t}$ therefore is a direct measure of the dollar cost to a surviving participant j from i's default. Here and going forward, I will always assume the defaulter i is the highest-impact participant in dollar terms, i.e. the one generating the largest spillover to survivors. This measure provides a direct idea of the dollar cost to each survivor conditional on the worst-case single default scenario.

These individual costs do not reflect the overall rate of default fund usage, which is key to the stability of the market (Paddrik, Rajan, and Young, 2020). I therefore also introduce an aggregate market measure of risk mutualization: the survivors' default fund utilization rate (SDFUR) after a default of the highest-impact participant:[17]

$$SDFUR_t = \frac{\max_i \left\{ \max \left\{ 0, (Shortfall^{\$}_{i,t} - DF_{i,t} - 5) \right\} \right\}}{\sum_{j \neq i} DF_{j,t}}, \tag{8}$$

where the numerator is the dollar amount that needs to be mutualized by survivors, and DF_j are the default fund deposits of those survivors. This type of resource metric is commonly calculated by exchanges and their regulators for internal risk management, particularly when sizing default funds (see the regulations in CPMI-IOSCO (2012), Principle 7, Key Consideration 4.) The numerator equals the risk spillover above the defaulter's pre-pledged resources (margin and default fund deposits) and the exchange's own capital of $5 million, after which risk mutualization begins (Figure 2). Taking the maximum over i finds the participant whose (presumed) default would result in the largest spillover of risk to the remaining participants on the given day. This participant is typically different across days, depending on the banks' current open positions, their sizes, and the difference between the SPAN margin and EVT margin for each contract held. The denominator is simply the survivors' portion of the default fund. This measure, computed daily and grounded in the Montreal Exchange's actual default management process, provides an effective way to gauge the risk propagation potential caused by under-margining. For example, an SDFUR of 0 would imply that the defaulter's margin shortfall is entirely contained by its own default fund deposit (plus the exchange's capital), and there is no risk propagation among survivors; an SDFUR of 1 implies that

[17] One can also calculate this measure conditioning on the default of the 2, 3, ..., k largest-impact participants, but this is not done here due to data disclosure requirements.

the exchange's pre-funded resources just barely contain the shock; and any larger shock with an SDFUR > 1 causes risk propagation, as the pre-pledged resources are insufficient. In situations like this, an exchange may be forced to start emergency recovery operations such as cash calls or profit gains haircutting (CPMI-IOSCO, 2017), which bring the exchange closer to either insolvency or resolution by authorities.

3.2 Econometric methodology: extreme value theory

Heavy tails feature different properties compared to the tail of the normal distribution assumed by SPAN. One can define a heavy-tailed distribution with cumulative density $F(x)$ as a distribution that satisfies[18]

$$1 - F(x) \sim x^{-\alpha} L(x), \qquad \text{as } x \to \infty \text{ for } \alpha > 0, \tag{9}$$

where L is a slowly varying function satisfying the limit $\lim_{t\to\infty} L(tx)/L(t) = 1$ for all $x > 0$. Heavy-tailed distributions are, to a first order, well approximated by the power functions family $Ax^{-\alpha}$, where A is a positive constant and the parameter α, known as a tail index, determines the tail's rate of convergence to zero or 'heaviness' (where a higher α produces a thinner tail). As a consequence, much of the literature has focused on how to correctly estimate α.[19]

The most widely used estimator of the tail index α is Hill's (1975) estimator, computed from the k upper-order statistics of the distribution of x.[20] This estimator is consistent for heavy-tailed data, including i.i.d. processes as well as many dependent stationary processes (Resnick and Starica, 1998). If one sorts the sample observations X_k from smallest to largest as $X_{1,n} < X_{2,n} < \ldots < X_{n,n}$, then the Hill estimator of α is computed from the largest k sample observations as follows:

$$\hat{\alpha} = \left[\frac{1}{k} \sum_{i=0}^{k-1} \log(X_{n-i,n}) - \log(X_{n-k,n}) \right]^{-1}. \tag{10}$$

(Since this estimation works only on positive numbers, when estimating negative (left) tails, by convention, negative observations are multiplied by -1.) A practical challenge is that the Hill estimator does not prescribe the number of tail observations k to be used in the estimation. The choice of k entails a tradeoff between the bias and variance of the estimator, with smaller k values

[18] This section is loosely based on Longin (editor), (2017) and De Haan and Ferreira (2006).
[19] The constant A is readily estimated when α is known.
[20] For a compendium of alternative estimators of α, see Berliant, Herrmann, and Teugels (2017).

producing less biased but higher-variance estimates, and large values of k introducing bias due to encroaching on the distribution's central region, where the approximation (9) no longer holds. This can be shown using the so-called Hall expansion

$$1 - F(x) = Ax^{-\alpha}[1 + Bx^{-\beta} + o(x^{\beta})], \tag{11}$$

which provides an approximation of the tail mass $1 - F(x)$. In it, $\alpha > 0$ is the tail index, $A > 0$ is a constant, $\beta > 0$ is a second-order shape parameter, and $B \in \mathbb{R}$ is another constant. Using this expansion, the bias and asymptotic variance of the Hill estimator (often inverted to $1/\hat{\alpha}$ for convenience) can be shown to be

$$A\mathbb{E}\left[\frac{1}{\hat{\alpha}} - \frac{1}{\alpha}\Big| X_{n-i,n} > s(k)\right] = \frac{\beta B s^{-\beta}}{\alpha(\alpha+\beta)} + o(s^{-\beta}(k)) \tag{12}$$

$$\text{AVar}\left(\frac{1}{\alpha}\right) = \frac{s^{\alpha}}{nA\alpha^2} + o\left(\frac{s^{\alpha}(k)}{n}\right), \tag{13}$$

where $s(k)$ is the tail threshold corresponding to the kth largest observation. From (12) it follows that as the tail threshold moves towards the center of the distribution, the bias of $\hat{\alpha}$ increases; at the same time, the asymptotic variance in equation (13) falls. For large s corresponding to thresholds far into the tail, the bias is small, but the variance dominates. This illustrates the tradeoff between bias and variance in the choice of k.

Much of the literature has therefore focused on how to select the 'optimal' tail threshold k^*. The theoretical-based approaches (Hall, 1990; Danielsson et al., 2001; Drees and Kaufmann, 1998) propose minimizing asymptotic mean squared errors (AMSE), but have been shown to feature unconvincing finite-sample performance (Danielsson et al., 2019). A second set of heuristic methods, widely used among practitioners, instead relies on graphically balancing the bias and variance of the estimator by plotting different estimates of $\hat{\alpha}$ as a function of the tail size k and seeking a stable region with a small k (see, e.g., Drees, de Haan, and Resnick, 2000). Such plots are called Hill plots. For instance, the Hill plot for the CGB contract illustrated in Figure 4, Panel A, shows the variance quickly stabilizing after $k > 25$, suggesting a stable region with $\hat{\alpha}$ around 3.

This method, labeled the Hill plot method or eyeballing method, can be helpful, but it introduces the challenges of subjectivity and non-replicability. Besides, what region is stable may often not

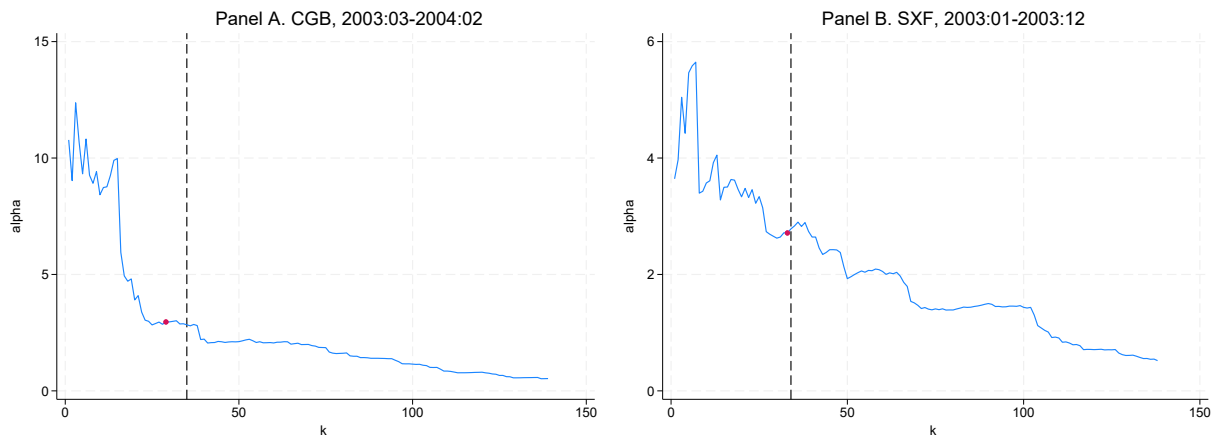

Figure 4: Panel A shows a Hill plot for CGB, data from 2003:03 to 2004:02. Panel B shows a Hill plot for SXF, data from 2003:01 to 2003:12. The red dot indicates the optimal $\hat{\alpha}$ selected by automated eyeballing. The vertical lines indicate the 10th percentile.

be apparent (Drees et al. 2000). For example, the Hill plot in Figure 4, Panel B, illustrates a relationship $\hat{\alpha}(k)$ without an obvious flat region in the 10% tail.

Another set of applications take a third, blunt approach by setting k^* at a fixed percentage of the total sample size, such as 4% or, in smaller samples, 10%. This approach is illustrated by the vertical lines in Figure 4 (here set at 10% because of the relatively small samples dictated by the SPAN's yearly rolling time window). Clearly, there is no guarantee that any fixed percentile will fall in a stable region, so this choice of k^* is atheoretic. Nonetheless, some practitioners use it (see, e.g., Van Oordt and Zhou, 2016, or Davydov, Vähämaa, and Yasar, 2021).

However, recent advances in tail threshold selection methods have scored considerable improvements in the automation, feasibility, and replicability of EVT estimates. Here I use three different threshold selection methods to estimate tail indices and use them for EVT-based margin requirements: an automated version of traditional eyeballing; Huisman et al.'s (2001) method geared specifically towards small samples; and Danielsson et al.'s (2019) method based on distance minimization on the quantile dimension. All three methods closely agree when applied on my data, with resulting margins being correlated around or above 0.90.

3.2.1 Automated eyeballing

Automated eyeballing, as the name suggests, automates the heuristic Hill plot procedure with an algorithm finding a stable region with relatively low k. The algorithm begins at $k = 2$ and sequentially considers successive windows of w observations $[\hat{\alpha}(k+1), \ldots, \hat{\alpha}(k+w)]$, moving k by 1 observation to the right at each iteration and checking whether at least $h\%$ of the observations in the current window fall within a band of $\pm\varepsilon$ around $\hat{\alpha}(k)$. The first window for which this is true is selected as a stable region, and its midpoint $k^* = k + \text{round}(w/2)$ is taken as the k^* that produces the optimal value of $\hat{\alpha}(k^*)$. This estimator of k^* can be formalized as

$$k_{eye} = \min \left\{ k \in S \;\middle|\; h < \frac{1}{w} \sum_{i=1}^{w} \mathbf{1}_{\{\alpha(k+i) < \alpha(k) \pm \varepsilon\}} \right\}, \tag{14}$$

where S is the set $\{2, 3, 4, ..., k_{\max}\}$, $k_{\max}$ is set to $0.10n$ (the maximum tail size, rounded to the nearest integer), w is the number of observations of the window, and h is the fraction of observations required to fall within a band of $\pm\varepsilon$ of $\hat{\alpha}(k)$. If the algorithm finds no stable region within the 10% tail, the search terminates and the estimator resorts to the blunt approach, setting $k^* = \text{round}(0.10n)$. (I chose 10% because the 5% tail is still within the high-variance region on some Hill plots.)[21] Following Danielsson et al. (2019), who test this estimator against several others with Monte Carlo simulations, I set $h = 0.9$ and $\varepsilon = 0.3$. The window size varies depending on the contract. For BAX, which on average has 2668 yearly observations, I set $w = 40$, or about 1.5% of the annual sample, close to Danielsson et al. (2019). For CGB and SXF, which feature fewer contracts active at the same point in time, and therefore have a lower average of 358 and 315 observations per year, I set the window w to 12 observations, or about 3.5% of the annual sample, since shorter windows appear too small to comprise a stable region. The value of 12 was chosen after test runs with w from 10 to 25 observations and manual reading of each Hill plot to make sure that the results are consistent with the heuristic Hill plot method; this step minimizes the results' sensitivity to the choice of window size, which is not theory-prescribed.[22] With these settings, automated eyeballing produces estimates highly comparable to the other two methods. However, since this method was not developed for small samples, its margin estimates sometimes

[21] The algorithm resorts to this blunt approach in about a quarter of the BAX estimates, and slightly more than half of CGB and SXF estimates. For relatively small sample sizes, this behavior is expected.

[22] The results $w = 13$ and $w = 14$ look very similar to those with $w = 12$.

look noisier than the Huisman et al. (2001) small-sample method. Hence, I use automated eyeballing predominantly as a benchmark or a proxy of what the traditional eyeballing method would do.

3.2.2 Husiman et al. (2001) small-sample method

Motivated by the deficiencies of traditional eyeballing in small samples, Husiman et al. (2001) propose a small-sample inference method for the selection of optimal k based on weighed least squares (WLS). Recalling that the bias of the Hill estimator is linear in k, whereas its variance is inversely related to k, these authors fit a WLS linear relationship between $\hat{\alpha}$ and k

$$\hat{\alpha}_m = \alpha_0 + b_1 k_m + \epsilon_m \tag{15}$$

with weights proportional to k, and show that, for a broad class of distributions, the intercept α_0 of this regression provides an unbiased estimate of the tail index α. This method leverages on a large number of values of k to reduce the sensitivity to the choice of a single k in the Hill procedure, which gives it advantages in small samples.[23] The value of k^* is determined as the k associated with the $\hat{\alpha}(k)$ closest to the intercept $\hat{\alpha}_0$. Among the three threshold selection methods used, this seems to produce the most plausible-looking and least noisy initial margins.

3.2.3 Danielsson et al. (2019) KS quantile method

Danielsson et al. (2019) propose an alternative method for the selection of optimal tail threshold based on optimization along the quantile dimension.[24] This method is a reaction to previous work attempting to fit the tail to the empirical distribution based on minimizing the Kolmogorov-Smirnov probability distance, $\sup_x |F_n(x) - F(x)|$, between an empirical CDF, $F_n(x)$, and the semi-parametric CDF $F(x) = 1 - Ax^{-\alpha}$ implied by the Hall expansion (11). These authors observe that "small deviations in the probability dimension lead to increasingly large distortions in the quantile dimension," because deep into the tail, differences in the probability dimension are of the order of $1/n$, whereas in the quantile dimension, they are of the order of $(n/k)^{1/\alpha}$. This introduces

[23] The authors argue their results are not sensitive to the choice of the tail cutoff point k_{max}, letting it go all the way to $n/2$, well into the distribution's center. In practice, given my data, I set $k_{max} = 0.35n$ before estimates become sensitive to the inclusion of large k's.

[24] I thank Lerby Ergun for generously sharing his code for this calculation.

distortions when tails are fit with the original Kolmogorov-Smirnov (KS) metric, leading to the choice of a suboptimally low k^* that produces high variance (Drees et al., 2018).

Instead, Danielsson et al. (2019) propose to minimize the distance between the empirical CDF $F_n(x)$ and the CDF $F(x) = 1 - Ax^{-\alpha}$ on the quantile dimension, choosing k so as to minimize the metric

$$\sup_{j=1,\ldots,T} |X_{n-j,n} - q(j,k)|, \tag{16}$$

where $q(j,k)$ is the EVT quantile estimator corresponding to cumulative probability $(n-j)/n$ (see Section 3.2.4). The authors label (16) the KS-quantile metric and show that this metric and the automated eyeball method outperform other heuristic and statistical methodologies in a series of Monte Carlo simulations performed on reasonably large samples. I use this method as an alternative to the other two methods, allowing k's up to $0.10n$ as candidates for the tail threshold. On my data, the KS quantile method performs well on the quantile dimension, but less so on the probability dimension, likely due to the smaller sample sizes in my study.

3.2.4 The EVT quantile estimator for $\mathbf{VaR}_{0.13\%}$

Having estimated the tail index α and the associated tail threshold k^*, it is now possible to estimate the tail quantile $q(p)$ corresponding to any given tail probability p using the tail approximation $p = 1 - F(q) = Aq^{-\alpha}$. At tail probability $p = 1 - F(q)$, the approximation implies $p = Aq^{-\alpha}$ and hence $q = (A/p)^{1/\alpha}$. Estimating $\hat{A} = X_{n-k^*,n}^{\hat{\alpha}} \left(\frac{k^*}{n}\right)$ as in Weissman (1978) yields the EVT quantile estimator

$$\hat{q} = X_{n-k^*,n}(k^*/pn)^{1/\hat{\alpha}}, \tag{17}$$

where $X_{n-k^*,n}$ is the observation corresponding to the tail threshold k^* and n is the sample size (the same estimator also appears in Cotter and Dowd, 2017, eq. (16.6)). This estimator is next applied to the left tail of each contract's return distribution.[25] To further increase comparability and match the monthly frequency of IM updates, I use the same annual rolling time window as SPAN, moving it in one-month increments beginning in January 2003 and ending in March 2011. I

[25] Futures' tails are most often symmetric (Cotter and Dowd, 2017), but I formally test tail stability with Loretan and Phillips' (1994) split-sample statistic $V = (\gamma_R - \gamma_L)^2 / \sqrt{\gamma_R^2/k_R + \gamma_L^2/k_L}$, where γ_L and γ_R are the reciprocals of the left and right tail index α_L and α_R, and k_L and k_R are the tail thresholds for the left and the right tail for each contract. This tail stability test fails to reject the null that the left and right tail have the same tail index, with a statistic of $V = 0.218$ and p-value of 0.83 for BAX, $V = 0.01$ and p-value of 0.99 for CGB, and $V = 0.04$ and p-value of 0.967 for SXF.

calculate $IM^{EVT} = \sqrt{2}\ \hat{q}(0.13\%)$ with each EVT method and consider the average EVT estimate as the benchmark against SPAN margins.

4 Results

4.1 Comparison of initial margin

Figure 5 shows a comparison of the initial margin computed with EVT methods against SPAN. The vertical distance between each EVT margin estimate and the SPAN initial margin in Panels A, C, and E is the contract-specific $Shortfall$ measure (3) introduced in Section 3.1. For clarity, I also plot the average shortfall by contract, taken across the three methods, in Panels B, D, and F. Before I discuss how EVT estimates compare to SPAN, I first gauge the robustness of the EVT approach by comparing the EVT estimates against each other.

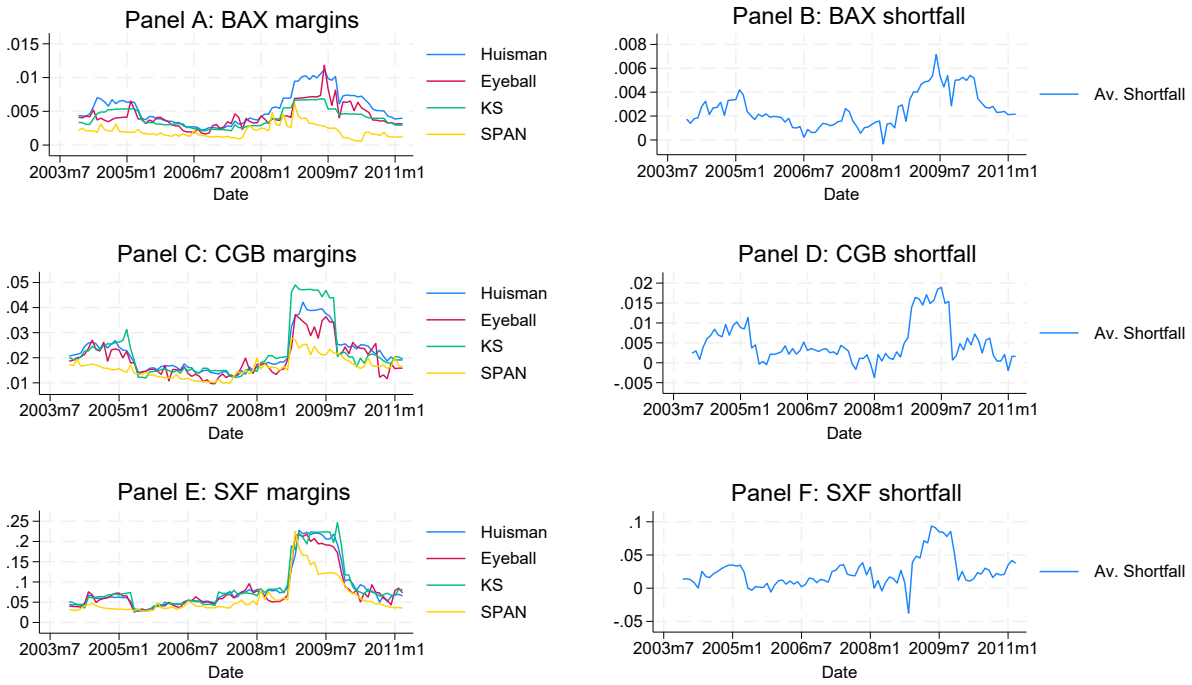

Figure 5: Panels A, C, and E show initial margins for BAX, CGB, and SXF contracts computed with three different tail threshold selection methods, plotted against SPAN. Panels B, D, and F show the risk measure $Shortfall$ from eq. (3) calculated as the average EVT initial margin minus the SPAN initial margin.

Relationship among EVT estimates. I observe that the three EVT initial margins are highly correlated for all three contract types. The Huisman and KS quantile method are the two

highest-correlated methods (at .95 for BAX and CGB, and .98 for SXF), followed by the automated eyeballing and Huisman (at .89, .92, and .97, respectively). The least correlated pair is eyeballing versus KS (at .79, .87, and .95 for each respective contract). Comparing the levels, one can see all three tend to rise and fall at the same times, but do not always agree on the levels, especially for BAX and CGB during the Great Financial Crisis. This illustrates the major influence of tail threshold selection on the tail index estimate. The Huisman estimates, followed by KS, appear the least noisy ones across contracts, while the automated eyballing method produces the choppiest estimates overall. This is explained by features in Huisman et al. (2001) and Danielsson et al. (2001) that circumvent stable region selection, which is more difficult in small samples; due to this difficulty, the automated eyeball method defaults to the blunt, fixed-size tail approach in roughly 25% of the months for CGB and in about 60% of the months for BAX and SXF. This appears to affect the noisiness of the estimates. Overall, however, the three methods paint a coherent picture of how margin should vary; going forward, I take the average estimate across the three methods.

Shortfall of SPAN margins. Panels B, D, and F show the shortfall between the average EVT initial margin across the three methods and SPAN, plotted by contract type. These contract-specific shortfalls provide a direct answer to the question whether the normality assumption results in the undercollection of margin.

On average, the least volatile BAX contract appears the most under-margined in percentage terms, commanding just 47% of the mean EVT margin, followed by SXF at 73%, and by the best-margined CGB contract at 80% of the corresponding EVT margin. These estimates are comparable with the numbers of Cotter and Dowd (2017), who calculate static EVT margins for 12 European stock futures and compare them to the normal method. The average future in their sample commands between 28% and 55% of its EVT margin, with a cross-sectional mean of 43%.[26] Longin (1999) reports similar numbers for silver futures. Cotter (2001), too, finds that normal-based margins fall below EVT estimates. A key difference from these studies is that I calculate time-varying margin, as in real exchanges, by leveraging newer EVT tail selection methods that work better on smaller samples, and, for the first time, quantify the economic cost of under-margining.

In absolute terms, BAX and CGB shortfalls are small. BAX's maximum shortfall is about 0.7 cents per dollar of open position, and for CGB, 1.9 cents per dollar; SXF has the largest peak

[26] Based on the quantile corresponding to 99.8% coverage.

shortfall, about 10 cents per dollar of open position. On the other hand, BAX and CGB have market shares of 71% and 16%, respectively, while SXF has only 13%, so it is not yet obvious which contract causes the biggest shortfall on the participant or the market level.

Panels B, D, and F of Figure 5 demonstrate how under-margining varies with time. In 2006–2007, just before the financial crisis, the margin shortfall for all three contracts is close to zero: in fact, even negative for BAX and CGB. However, this changes sharply in volatile periods. During the crisis (October 2008–December 2009), the margin shortfall clearly goes up with the onset of high volatility, although the pattern is different for each contract. For the less volatile interest rate futures BAX and CGB, this rise is immediate; SPAN and the EVT margins react at the same time, but SPAN doesn't catch up in terms of level, producing a relatively large shortfall from the very start of high volatility. In the case of the more volatile stock future SXF, by contrast, SPAN reacts vigorously and aligns closely with all EVT-based initial margins, even briefly outpacing the EVT estimates in October 2008. However, as the October shock subsides, SPAN goes down much faster than EVT, producing an increasingly large difference in 2009.

This has to do with the inherently different impulse response of SPAN versus EVT to one-off price shocks. Twenty business days after a large one-off shock, the SPAN estimator switches to the (lower) 90-day volatility, and after 90 business days, to the 260-day volatility; thus, in 91 business days, SPAN reverts to the annual volatility measure that is least sensitive to a single extreme observation. This taper-off rate is a design feature of SPAN; it does not appear dictated by theoretic considerations, but rather by the implicit assumption that, given a shock, the probability of a second shock subsides proportional to the above volatility measures. Given the ad hoc nature of this assumption, it is unrealistic that another approach to margining will reproduce SPAN's dynamics. Once the EVT estimator "sees" an extreme tail observation, it will continue to fit the tail to it (and produce high margins) as long as the observation remains in the sample window, since the observed risk distribution continues to have a long tail. Some newer implementations of SPAN, developed after the Great Crisis, feature slower dynamics based on features such as margin floors, integrated stress from past periods, or time-based weights to prolong the influence of past extreme observations on current volatility.[27] Since these differences are a matter of philosophy

[27] One example is CDCC's current (2024) modified SPAN implementation, which features a margin floor, a weighted past stress component which never expires, and exponential weights discounting past stress over a long time window. See Odabasioglu (2023).

about what constitutes the relevant risk distribution, I do not focus on matching model dynamics, as they are not reconcilable by mechanical means. Instead, I conduct separate margin adequacy comparisons *outside* of the periods with transitional dynamics and compare results for the crisis and normal periods, which I refer to as 'on-peak' and 'off-peak.' This is done in Section 4.5.

4.2 Statistical confidence of SPAN margins

An alternative way of gauging SPAN's margin adequacy is inspecting the statistical confidence afforded by SPAN margins, i.e. the probability that a potential price move falls within the SPAN initial margin, given the heavy-tailed distribution of returns. With the heavy-tailed probability estimator from De Haan et al. (1994), I calculate

$$1 - p^{SPAN} = 1 - \frac{k^*}{n}\left(\frac{X_{n-k^*,n}}{IM/\sqrt{2}}\right)^{\hat{\alpha}}, \tag{18}$$

which calculates the (single-tail) statistical confidence associated with SPAN's normal-based estimate of $\text{VaR}_{0.13\%}$. Figure 6 shows the results of this calculation by contract type.

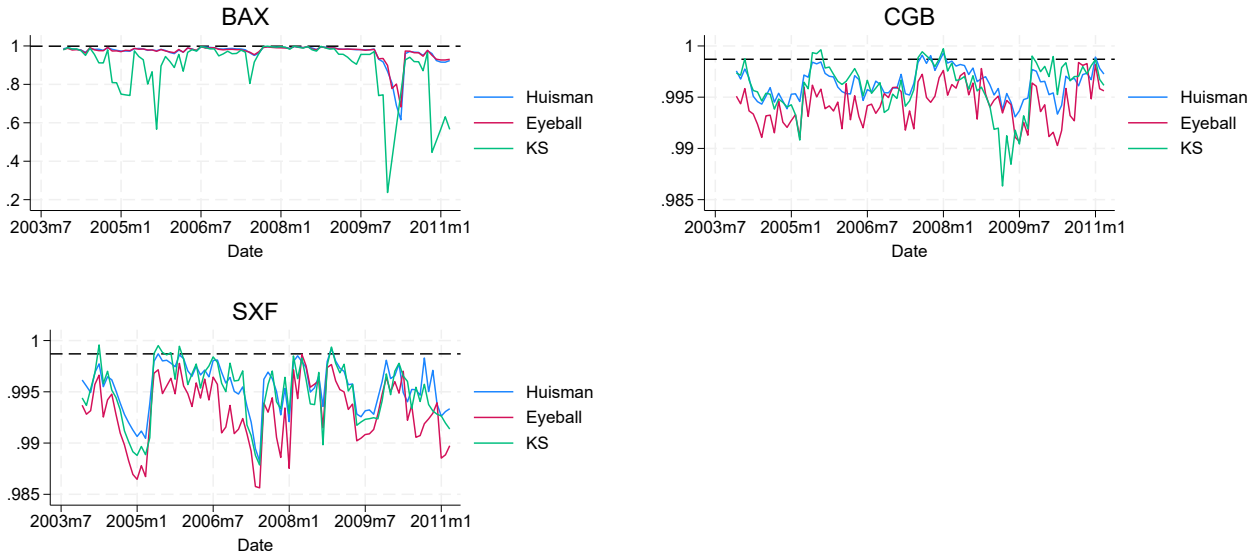

Figure 6: Single-tail confidence intervals implied by SPAN margins for BAX, CGB, and SXF contracts, calculated with 3 EVT methods. The dashed line corresponds to SPAN's target confidence level of 99.87%.

Figure 6 and Panel A of Table 2 confirm that, as a whole, the BAX margins provided the lowest coverage out of the three contract types: 96.6% and 96.8% according to the Huisman and automated eyeball methods, and 90.6% according to the KS method. Although short of the intended target

of 99.87%, most of the time the coverage was well above 95%, with the exception of the financial crisis recovery (2010), where SPAN and EVT margin dynamics diverged – a period I deal with separately. The Huisman and eyeball methods closely agree that coverage in that period briefly fell to 60%–70%. The KS method, by contrast, performs out of line here, predicting a drop in 2005:7 and 2009:12 far worse than (and unconfirmed by) the remaining methods. A similar distortion is also visible in the CGB plot in 2009:3. This poor performance on the probability dimension seems the result of a combination of two factors: firstly, the KS method optimizing on the quantile rather than probability dimension, resulting in probability distortions not observed with the eyeballing and Huisman methods; and secondly, the much smaller sample sizes compared to the original study (which uses $n = 10,000$ compared to an average n between 315 and 2,668 for my data). Smaller samples result in less densely populated tails, and therefore larger errors in matching empirical and predicted quantiles. Such distortions are not reported in the original study with large n; therefore, I interpret this as a small-sample shortcoming.

By contrast, the CGB and SXF contracts did much better in terms of statistical coverage: on average, between 99.4% and 99.6% for CGB, and 99.3%–99.5% for SXF. The CGB coverage stayed above 99% (the regulatory minimum) practically the entire time, with results only slightly worse for SXF, where coverage dropped a few times to 98.7%. These numbers, which fall well within regulatory expectations (see CPMI-IOSCO, 2012, Principle 4), demonstrate the tradeoff that SPAN's designers likely counted on: that aiming the confidence level at 99.87% under assumed normality would, in many cases, still provide coverage above 99% with heavy-tailed data. The key question going forward, therefore, is not one of regulatory compliance, but predominantly of cost: how costly is the normality assumption to market participants in terms of risk-sharing?

4.3 Margin shortfall on the contract and the participant level

To gauge the cost of the normality assumption to market participants, I calculate the dollar shortfalls from equations (5) and (6) at the participant and the market level and explore how the aggregate margin shortfall is distributed across contracts and among participants.

Figure 7 shows the distribution of the contract-level dollar shortfall (4) across participants over time. Panels A and B in conjunction with Table 2 show that the BAX contract, despite dominating the market with 71% market share and being the most under-margined in relative terms, generates

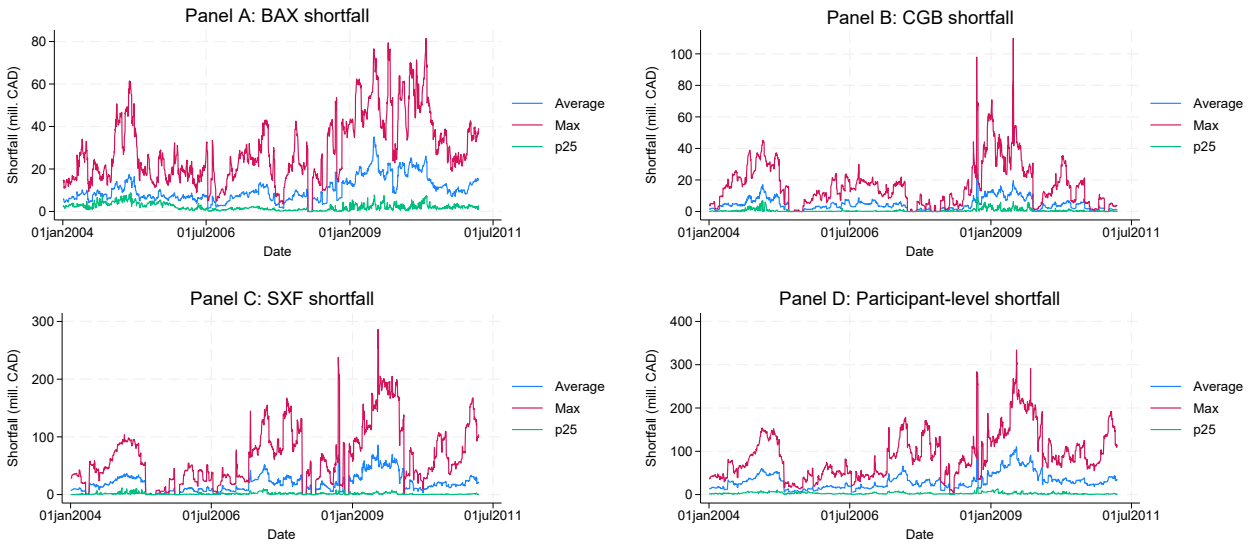

Figure 7: Panels A, B, and C show the distribution of the contract-level dollar shortfall (equation 4) across participants over time. Panel D shows the distribution of the participant-level shortfall (equation 5) over time.

a smaller average dollar shortfall than SXF (10.2 million versus 19 million) because of its low absolute margin level, which is in the thousandths of a cent per dollar of open position. (The CGB contract is in between, with a dollar shortfall of 4.5 million.) Similarly, the maximum BAX shortfall, aggregated across participants, is the smallest at 81.5 million, topped by 110 million for CGB and 286 million for SXF, because of the combined effect of portfolio composition and absolute margin levels.

Panel D shows the distribution of the participant-level shortfall (5) over time. The largest margin shortfall experienced by a single participant ranges from 4 million (in calmer times) to 334 million in mid-2009, with the average shortfall over time at 31 million. The second-largest peak in panel D is 284 million in October 2008, at the start of the Great Financial Crisis. Panel D also shows two lower peaks around 2005 and 2007, where the top under-margined participants realized margin shortfalls of 150 to 175 million with respect to the corresponding EVT benchmark.

Figures 6 and 7 point to three important conclusions. First, the normality assumption distorts margins for some contracts more than for others, and in some periods more than in others. Adopting a static-distribution approach to estimating margin requirements as in the previous literature masks this time heterogeneity, which has important economic consequences. Second, a participant's dollar shortfall depends both on (1) the contract's SPAN margin as a fraction of its EVT benchmark,

(2) the absolute levels of margin, and (3) the portfolio composition. These factors, acting jointly, determine a participant's margin shortfall compared to the EVT benchmark. And thirdly, the most under-margined participants can score shortfalls in the millions of dollars during both crises and calmer periods. Peak shortfalls in the order of 300 million become comparable with the exchange's default fund during the sample period (on average, 210 million in Table 2), which implies that risks that could have been covered by the defaulter's initial margin now become the responsibility of surviving participants. Paddrick, Rajan, and Young (2020) and Raykov (2022) show how this scenario can present a threat to financial stability; therefore, I turn to it next.

4.4 Risk-sharing costs to market participants

Existing margining studies do not quantify whether the under-margining caused by SPAN's normality assumption is costly for individual market participants, which mutualize default risk when initial margin is insufficient. The availability of daily positions data allows me to quantify these costs, which I measure with aggregate-level and participant-level metrics.

On the aggregate level, I calculate the aggregate initial margin shortfall (6) across all positions and market participants. It is shown in Panel A of Figure 8. This shortfall represents the dollar difference between the total initial margin calculated by EVT (as the average of the three methods) and the base initial margin collected by SPAN marketwide.[28] This difference is time-varying and can be substantial. The average aggregate shortfall is 484 million dollars, with standard deviation of 291 million, and the highest two peaks are 1.59 billion on October 1, 2008 (at the start of the 2008 crisis), and 1.66 billion on June 15, 2009, when SPAN and EVT margins diverged the most (mostly on account of SPAN's faster decay time). To distinguish under-margining stemming from these two sources, Section 4.5 conducts separate on-peak and off-peak cost comparisons circumventing the decay period after a shock.

The time pattern in Panel A of Figure 8 confirms that SPAN's normality assumption is problematic even outside financial crises – for instance, in late 2004 or early 2007, when the total market shortfall approached or exceeded 1 billion. This illustrates that the problem is not that SPAN underestimates tail quantiles just when extreme observations become more frequent; it underes-

[28] Consistent with the actual default management process at the Montreal Exchange, I do not allow a margin surplus for one participant to offset a margin shortfall for another member.

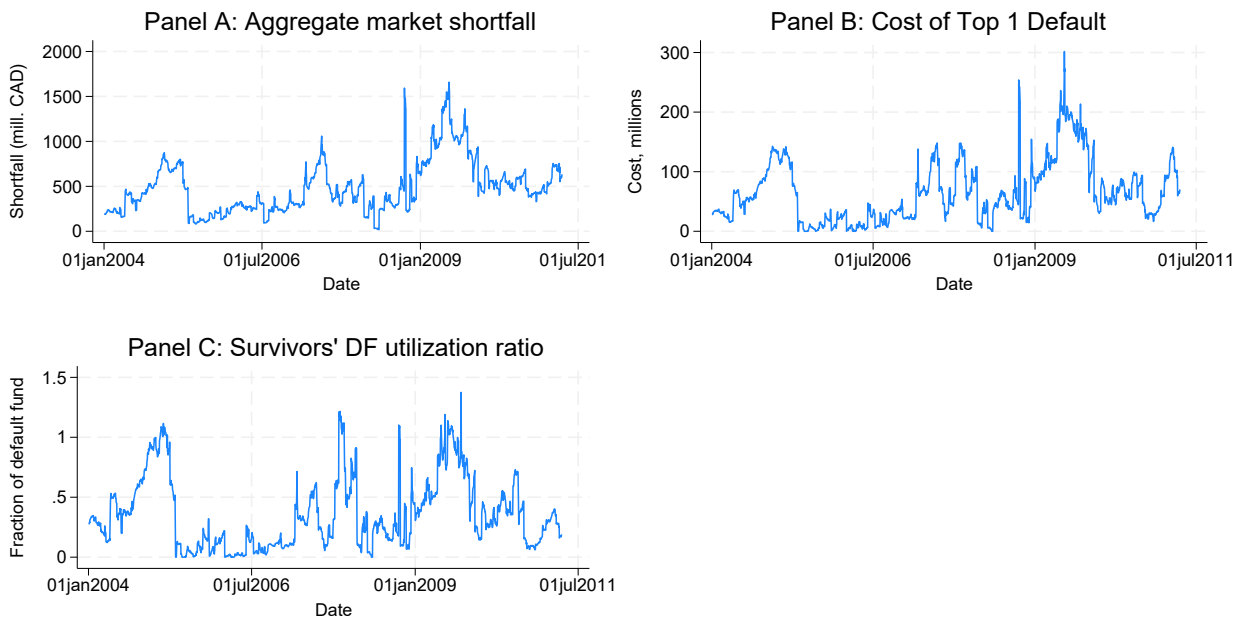

Figure 8: Panel A shows the aggregate undermargining on the market level relative to the average EVT margin. Panel B shows the total cost to be mutualized by surviving participants from the default of the highest-impact participant. Panel C shows the same cost as a fraction of the survivors' default fund.

timates extreme quantiles *in principle*, and it does so because the returns of these contracts are heavy-tailed even in normal times. This new result, enabled by EVT estimation of time-varying margins, leads to my first policy implication: that ad hoc margin add-ons imposed in crises (such as those implemented by some exchanges in 2020) do not consistently solve SPAN's under-margining problem, even though they are justified.

The entire market's margin vastly exceeds what would be used in the event of default of a single participant. To inform this latter case, Panel B shows the aggregate cost from the default of the single highest-impact participant (as defined in Section 3.1). This cost averages to 65.1 million and peaks at 253.9 million on October 1, 2008, and at 301.7 million on June 17, 2009. Conditional on default, these amounts would have to be absorbed by the survivors' default fund.

To gauge the effects of this process, I consider another aggregate metric, the survivors' default fund utilization rate (SDFUR) defined in Section 3, which is conditioned on the default of the participant with highest impact on the survivors' default fund. As before, I buffer this default through the pre-funded resources from the default waterfall in Figure 2: the defaulter's initial margin, the defauter's default fund deposit, and the exchange's 5 million pledge, after which any residual shortfall is shared among survivors. This residual shortfall, plotted as a fraction of the

survivors' default fund, is shown in Panel C of Figure 8, with summary statistics appearing in Table 3. The table shows that, on average, about 36% of survivors' default fund is used, but the figure shows at least four periods where this utilization rate exceeds 1, i.e. when the *entire* default fund is consumed,[29] with peaks of 1.22 on November 2, 2008, and 1.38 on September 15, 2009.

These numbers are sobering because they imply that there are multiple periods when the exchange's entire second line of defense – the default fund – is consumed by inadequacies in the first line of defense (initial margin). This way the burden of default costs not only shifts away from the defaulter, creating moral hazard, but also passes these costs on to survivors, potentially threatening their financial condition.

The pattern of variation of the SDFUR in Panel C is driven by two factors: the variation of the aggregate cost of under-margining in Panel B, and variation in the size of the survivors' default fund. Visually, the former factor dominates across the three panels, but some of the periods with high SDFUR are also due to low default fund levels following a calm period. This appears to be the case, for example, in late 2004 and early 2005, when the default fund fell to an average of 144 million, consistent with the low prior volatility, or in September 2007 to February 2008, when the default fund averaged 180.7 million.

As a next step, I examine how this risk mutualization burden affects individual participants. On the individual level, any uncovered default above the defaulter's initial margin and default fund deposit (and the exchange's 5 million pledge) is shared among surviving participants proportional to each one's 60-day share of initial margin (equation (7)). I calculate these risk mutualizations in dollar terms for each market member and normalize them by the member's default fund deposit to gauge the extent to which survivors bear costs that should have been absorbed by the defaulter's initial margin.[30] As before, the assumption is that the default comes after a price shock exactly equal to $\text{VaR}^{EVT}_{0.13\%}$ for each contract, such that all spillovers above initial margin are entirely due to the differences between EVT and SPAN. The results of this calculation are shown in Table 4, where market members are anonymized due to data disclosure requirements.

The table shows that SPAN's under-margining consumes nearly half (49%) of the average member's default fund deposit for the sample period, but there are periods during which individual participants exceed that percentage by a large scale. For instance, only three of the 19 institutions

[29] Because the defaulter's portion is already used up in stage 2 of Figure 2.

[30] I do not present these quantities in dollar terms due to data disclosure reasons.

in this market feature a maximum default fund deposit consumption ratio below 1; many exhibit ratios closer to 2 or 4 times their individual deposit, with a cross-sectional maximum of 7.48 times (assuming that the exchange does not cap that amount; such practices vary across jurisdictions). The reason for these high numbers is that the total default fund is calibrated for the default of the participant with the largest impact based on the exchange's internal stress test, which is also most likely to be the participant with the highest impact in dollar terms in my calculations. That participant's default fund deposit is proportionally the largest and is consumed first upon default, thereby automatically taking out the default fund's largest deposit. Because of this, smaller participants with minimal default fund balances can still face potential replenishment calls that exceed their default fund deposit by a large factor. In many exchanges, this factor is statutorily capped, so if the exchange cannot collect these assessments, it may need to enter recovery or resolution, or else exit the business. This underscores that the seemingly innocuous normality assumption in SPAN can have far-reaching consequences.

It is important to note that these results stand in spite of SPAN largely meeting the high bar of 99% for statistical confidence set in international regulation. Far into the tail, the small difference between 99% and 99.87% confidence corresponds to a large dollar difference in initial margin. In a market where the average bank's portfolio is worth 5.78 billion, the dollar difference between EVT and SPAN margins quickly becomes comparable to the total default fund size (80 to 501 million during the sample period). This is the core behind all results in this paper and is another example of how a small difference in probability coverage can translate into millions of uncovered financial losses.

4.5 Eliminating differences from model dynamics

The preceding discussions point to two clear peaks in under-margining and the costs associated with it, driven by separate causes. The 2008 peak is related to the onset of extreme volatility, which began in September 2008, whereas the mid-2009 peak stems from SPAN's faster decay rate. This difference in dynamics is inherent and cannot be reconciled by mechanical means. As long as an extreme observation remains in the data, the EVT model will continue to fit the tail to it, producing a higher margin; whereas the influence of a single extreme observation in SPAN fades as more non-extreme observations are added. This boils down to the different parts of the distribution

used by the two estimators (tail only versus the full distribution) and hence to the choice of whether the model should use a tail estimator or not – a normative question much broader than the scope of this paper. To avoid bias due to different model dynamics, I conduct separate margin comparisons outside of the periods of transitional dynamics. To do so, I first identify large isolated shocks in the underlying price data that could trigger transitional dynamics.

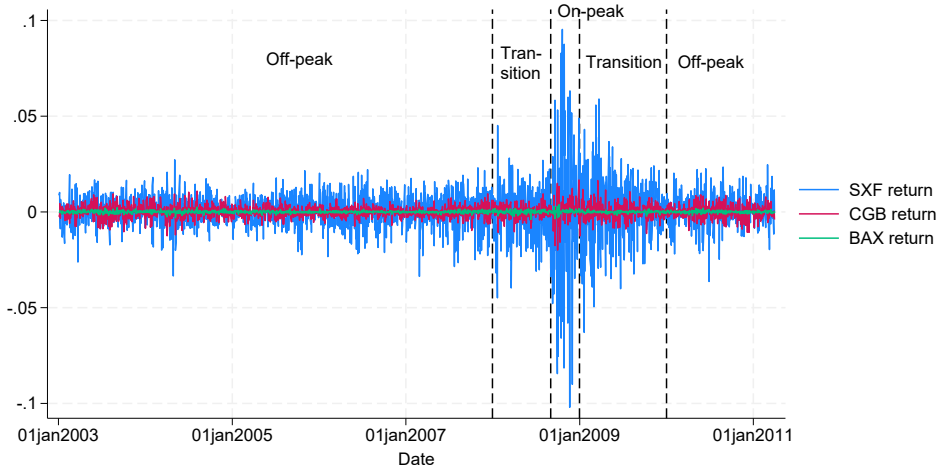

Figure 9: This figure shows realized contract returns between January 2003 and March 2011 by contract type. The dashed lines indicate, respectively, the end of the first off-peak period (Dec. 31, 2007), the beginning and the end of the on-peak period (Sept. 1, 2008–Dec. 31, 2008) and the start of the second off-peak period (Jan. 1, 2010).

Figure 9, which plots the returns underpinning each contract's margin, points to several distinct periods in the data. From 2003 to December 2007, the data is characterized by small, frequent shocks unlikely to trigger transitional dynamics. In January and March 2008, there are two price shocks that exceed the maximum for the previous period; such observations will register with the EVT quantile estimator, since EVT originates in the probabilities of exceeding existing maxima (Berliant, Herrmann, and Teugels, 2017). September to December 2008 feature the large shocks of the Great Financial Crisis, which certainly trigger transition dynamics in margins, clearly visible in the next months in Figure 5, with the volatility gradually subsiding till December 2009. From January 2010 on, the market seems to revert to the pre-crisis mode of small, frequent shocks.

Based on this, I divide the data into three periods: off-peak, spanning from January 2003 to December 2007 and from January 2010 to the end of the sample; on-peak, from September to December 2008; and two periods of transitional dynamics from January to August 2008 and from

January to December 2009 (inclusive). The off-peak periods are those with small, frequent shocks and without transitional dynamics. The on-peak period is the one where both EVT and SPAN margins peak in Figure 5, but before transitional dynamics sets in; and the transitional dynamics periods are those where sharp margin peaks recede. To isolate effects from transitional periods, I compare under-margining results in off-peak and on-peak periods, dropping the transitional periods from the sample. This comparison appears in Panels B and C of Tables 2 and 3.

Panels B and C of Table 2 show that eliminating transition periods doesn't change the overall results, but rather reinforces the fact that SPAN's underestimation bias is systematic and affects both crisis and normal periods. The ratio of the mean SPAN margin to average EVT margin is 0.58 on-peak and 0.43 off-peak for the BAX contract, 0.71 on-peak and 0.80 off-peak for CGB, and 0.94 on-peak and 0.70 off-peak for SXF, suggesting that sizeable under-margining still occurs relative to the EVT initial margins.[31]

In dollar terms, the picture is similar. The mean aggregate market shortfall, which comes to 484 million for the full sample, goes up to 545.3 million in the on-peak period and is 413 million in the off-peak period. The on-peak maximum is 1.6 billion, compared to 1.06 billion for off-peak and 1.66 billion for the full sample. Dropping the transition periods therefore does eliminate one large peak ascribable to different transition dynamics – that of 1.66 billion – but the remaining maximum is still close at 1.6 billion.

The cost of default for the highest-impact participant behaves in similar ways. Dropping transition dynamics eliminates the 2009 peak of 301.7 million, but the next highest one is 253.9 million in the on-peak period, which cannot be ascribed to transition dynamics. The off-peak maximum is 148.2 million. These numbers remain comparable to the size of the default fund.

To gauge the effects on the default fund, I once again calculate on-peak and off-peak statistics for SDFUR in Panels B and C of Table 4.[32] The on-peak average SDFUR and maximum SDFUR are very close to or the same as those for the full sample (0.50 and 7.47), thereby confirming that my results are not driven by transitional dynamics. The corresponding off-peak average is 0.42 and the maximum SDFUR is 5.62; thus, the issue of under-margining with SPAN's classical implementation

[31] These numbers are calculated as $IM/(IM + Shortfall)$, where $Shortfall$ is the shortfall between IM and the average EVT IM estimate in Table 2.

[32] I do not disaggregate these statistics by period in Table 4 to prevent disclosing any differences in the number of active market participants in each period.

is systematic, is consistent across calm and turbulent periods, and is not an artifact of different transitional dynamics of SPAN and EVT estimators.

5 Policy implications

The above findings provide several insights for policymaking and future research. First, EVT methods can be a valuable tool for benchmarking risk model performance in derivatives markets, since, on heavy-tailed data, EVT provides more plausible tail risk estimates than normality-based models. At the same time, it is clear that the different valid choices of tail threshold, and the resulting variation of estimates, all make EVT a less-than-ideal candidate for actual margin setting. Nonetheless, it is also clear that normal-based models underdeliver regardless of any intricacies with EVT, and the cost of this under-margining to surviving, healthy participants can be in the millions. Therefore, to avoid under-margining and risk propagation, one practical policy recommendation to exchanges could be to begin using EVT as a back-testing tool comparing recent historical margin levels to EVT estimates produced ex post. Insights from such back-testing could be used to improve the coverage of non-EVT margin models.

A second conclusion is that the under-margining inherent in the normal method can be large and is costly to market participants in both calm and turbulent periods. In at least four periods in my data, three of which fall outside of the Great Financial Crisis, the difference between the most impactful defaulter's SPAN margin and that commanded by EVT is of the same order as the market's entire default fund, effectively disabling it as a risk mitigation device. In a high-stress period, this might make the difference between the exchange surviving versus being placed into resolution or even exiting business. Moreover, this under-margining occurs despite SPAN largely meeting the mandatory 99% confidence minimum, since far into the tail, small changes in probability correspond to large changes in margin. Therefore, a policy recommendation to regulators might be to reconsider whether the 99% confidence threshold is adequate for heavy-tailed data.

Thirdly, discretionary margin add-ons imposed during crises (such as those implemented by some exchanges in 2020) are justified, but do not consistently solve SPAN's under-margining issue, because that is not limited just to crises and can be just as severe in normal periods. Therefore, a more comprehensive solution is needed. One practical solution could be moving away from

parametric, normal-based models towards, for example, historical value-at-risk (HVAR) models, perhaps back-tested with EVT to improve coverage. Many leading exchanges are moving towards HVAR, which, in addition, provides better anti-procyclicality properties.

6 Conclusion

Exchanges play a key role in the safety of financial markets by setting market-wide collateral requirements. Nonetheless, the most popular risk model used to set collateral levels counterfactually assumes that derivatives feature normal, rather than heavy-tailed, returns. This paper quantifies the cost of this assumption to market participants, using extreme value theory and proprietary positions data from the Montreal Exchange.

I find that the costs to market participants are significant. On the aggregate, the Canadian futures market was under-margined by up to 1.6 billion CAD during the 2008 crisis and 1.06 billion outside the crisis, suggesting that the normality assumption results in sizeable undercollection of collateral. This happens despite normal-based margins largely meeting the 99% confidence interval required by international regulation, because far into the tail, small changes in probability coverage translate into large changes in margin. This mechanism explains how, when aggregated across a large market, the dollar difference between EVT and normal-based margins quickly becomes comparable to the total default fund size.

I find that the default of the highest-impact participant generates a cost of up to 302 million CAD purely due to the difference between the EVT-based and the normal-based margin and is capable of consuming (or even exceeding) the entirety of the survivors' default fund, causing costly risk mutualization among participants. This level of risk mutualization would be avoided if margins were set closer to their EVT estimates. Moreover, I show that the under-margining problem associated with assumed normal returns is not limited to financial crises, is not driven by differences in model dynamics, and can be just as severe in normal times.

To resolve this, I offer three policy implications for exchanges and regulators: to adopt EVT as a back-testing method to improve coverage for non-EVT models; to consider revising upward the 99% minimum for statistical confidence in cases of heavy-tailed data; and to approach under-margining holistically, as it is not limited just to crises. The cost quantifications provided here aim

to serve as a positive motivation for this change, since, if anything, they only confirm one thing: assuming normality when things are not normal can be very costly.

7 Tables

Table 1. List of Institutions

Institution	Headquarters
BMO NESBITT BURNS INC.	Canada
CIBC WORLD MARKETS INC.	Canada
CREDIT SUISSE SECURITIES (CANADA), INC.	Switzerland
DESJARDINS SECURITIES INC.	Canada
FIDELITY CLEARING CANADA ULC.	USA
FRIEDBERG MERCANTILE GROUP LTD.	Canada
GOLDMAN SACHS CANADA	USA
J.P. MORGAN SECURITIES CANADA INC.	USA
LAURENTIAN BANK SECURITIES	Canada
MAPLE SECURITIES CANADA LIMITED	Canada
MERRILL LYNCH CANADA INC.	USA
MF GLOBAL CANADA CO.	USA
NATIONAL BANK OF CANADA	Canada
NEWEDGE CANADA INC.	France
PENSON FINANCIAL SERVICES CANADA	USA
RBC DOMINION SECURITIES INC.	Canada
SCOTIA CAPITAL INC.	Canada
TD SECURITIES INC.	Canada
TIMBER HILL CANADA COMPANY	USA

This table lists the institutions participating in the Canadian futures market during the sample period (January 2, 2003, to March 31, 2011) together with their country of headquarters.

Table 2: Summary statistics for contract-specific variables

Panel A: Full Sample		BAX					CGB					SXF				
Variable	Level of obs.	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N
Position mkt. value (mill.)	Day-participant	4069.5	4002.7	0.2	42025.4	30,646	947.3	1174.7	0.1	9563.5	29,655	760.7	921.1	0.1	5808.6	29,283
Market share	Daily	0.71	0.06	0.52	0.85	2,066	0.16	0.04	0.06	0.30	2,066	0.13	0.04	0.04	0.25	2,066
Daily contract return (%)	Daily	0.003	0.066	-0.507	0.555	21,798	0.017	0.350	-1.991	1.671	2,947	0.045	1.272	-10.203	9.534	2,579
IM, SPAN	Monthly	0.0020	0.0010	0.0006	0.0061	87	0.016	0.004	0.010	0.028	87	0.060	0.041	0.028	0.225	87
IM, Huisman	Monthly	0.0054	0.0024	0.0022	0.0112	87	0.022	0.008	0.013	0.042	87	0.085	0.055	0.031	0.228	87
IM, Eyeball	Monthly	0.0043	0.0018	0.0017	0.0118	87	0.019	0.007	0.010	0.037	87	0.080	0.052	0.026	0.218	87
IM, KS	Monthly	0.0040	0.0014	0.0021	0.0068	87	0.022	0.010	0.012	0.049	87	0.087	0.058	0.028	0.246	87
Mean Shortfall	Monthly	0.0025	0.0015	-0.0003	0.0072	87	0.005	0.005	-0.004	0.019	87	0.024	0.023	-0.037	0.094	87
Mean $ Shortfall (mill.)	Daily	10.18	5.79	0.00	35.11	1,819	4.49	3.90	0.00	28.40	1,819	19.04	15.37	0.00	85.69	1,819
SPAN coverage, Huisman	Monthly	0.966	0.056	0.616	0.997	87	0.996	0.001	0.993	0.999	87	0.995	0.002	0.988	0.999	87
SPAN coverage, Eyeball	Monthly	0.968	0.046	0.683	0.998	87	0.994	0.002	0.990	0.998	87	0.993	0.003	0.986	0.999	87
SPAN coverage, KS	Monthly	0.906	0.136	0.237	1.000	83	0.996	0.003	0.986	1.000	87	0.995	0.003	0.988	1.000	87
$\hat{\alpha}$, Huisman	Monthly	2.91	0.56	1.96	4.20	87	3.73	0.65	2.47	5.33	87	3.36	0.46	2.63	4.50	87
$\hat{\alpha}$, Eyeball	Monthly	2.62	0.55	1.46	4.04	87	2.95	0.64	1.94	5.30	87	2.65	0.55	1.82	4.29	87
$\hat{\alpha}$, KS	Monthly	5.20	1.00	3.67	7.99	87	4.19	1.14	2.19	7.29	87	3.64	0.95	2.30	6.67	87
Panel B: On-peak margin statistics (09/2008 – 12/2008)																
Variable	Level of obs.	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N
IM, SPAN	Monthly	0.0042	0.0014	0.0026	0.0061	4	0.0235	0.0059	0.0150	0.0279	4	0.1521	0.0720	0.0555	0.2245	4
IM, EVT (average)	Monthly	0.0072	0.0011	0.0055	0.0079	4	0.0338	0.0098	0.0199	0.0410	4	0.1613	0.0594	0.0823	0.2216	4
Mean Shortfall	Monthly	0.0030	0.0010	0.0016	0.0040	4	0.0104	0.0056	0.0049	0.0164	4	0.0091	0.0336	-0.0375	0.0390	4
SPAN coverage, Huisman	Monthly	0.9898	0.0058	0.9817	0.9941	4	0.9966	0.0003	0.9962	0.9970	4	0.9972	0.0025	0.9936	0.9992	4
SPAN coverage, Eyeball	Monthly	0.9888	0.0073	0.9789	0.9961	4	0.9950	0.0021	0.9929	0.9978	4	0.9956	0.0029	0.9915	0.9977	4
SPAN coverage, KS	Monthly	0.9861	0.0105	0.9731	0.9980	4	0.9952	0.0008	0.9941	0.9960	4	0.9961	0.0043	0.9898	0.9994	4
Panel C: Off-peak margin statistics (01/2004 – 12/2007 and 01/2010–03/2011)																
Variable	Level of obs.	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N	Mean	SD	Min	Max	N
IM, SPAN	Monthly	0.0017	0.0005	0.0006	0.0031	63	0.0144	0.0027	0.0099	0.0204	63	0.0410	0.0102	0.0280	0.0765	63
IM, EVT (average)	Monthly	0.0039	0.0011	0.0020	0.0061	63	0.0181	0.0043	0.0124	0.0259	63	0.0590	0.0158	0.0292	0.1017	63
Mean Shortfall	Monthly	0.0022	0.0012	0.0002	0.0054	63	0.0037	0.0030	-0.0020	0.0113	63	0.0181	0.0118	-0.0058	0.0417	63
SPAN coverage, Huisman	Monthly	0.9602	0.0643	0.6158	0.9961	63	0.9963	0.0014	0.9933	0.9991	63	0.9953	0.0024	0.9882	0.9987	63
SPAN coverage, Eyeball	Monthly	0.9629	0.0517	0.6826	0.9971	63	0.9942	0.0020	0.9903	0.9984	63	0.9926	0.0031	0.9856	0.9978	63
SPAN coverage, KS	Monthly	0.8868	0.1519	0.2373	0.9987	59	0.9965	0.0018	0.9908	0.9996	63	0.9949	0.0031	0.9879	0.9996	63

This table presents summary statistics for contract-specific variables and their level of observation. Summary statistics are taken with respect to the original data frequency and level of observation at which the variable is reported. Panel A covers the full sample period from January 2, 2003, to March 31, 2011. Panel B presents margin statistics for the on-peak period (09/2008 – 12/2008), and Panel C, for the off-peak periods (01/2004 – 12/2007 and 01/2010 – 03/2011). All margin and EVT-dependent data series start in 2004 to allow a leading 1-year estimation window consistent with SPAN.

Table 3: Summary statistics for participant-level and aggregate variables

Variable	Level of observation	Mean	SD	Min	Max	N
Participant-level shortfall (mill.)	Daily	31.0	19.4	1.2	110.6	1,819
Aggregate market shortfall (mill.)	Daily	484.0	291.1	18.4	1658.7	1,819
Cost of top 1 default (mill.)	Daily	65.1	51.1	0.0	301.7	1,819
Default fund size (mill.)	Monthly	210.0	91.4	79.5	501.3	99
Individual DF contribution (mill.)	Month-participant	4.7	11.6	0	124.7	4,455
SDFUR	Daily	0.36	0.30	0	1.38	1,819
Panel B: On-peak cost statistics (09/2008 – 12/2008)						
Aggregate market shortfall (mill.)	Daily	545.3	323.1	215.3	1592.9	84
Cost of top 1 default (mill.)	Daily	70.3	59.6	14.3	253.9	84
SDFUR	Daily	0.32	0.27	0.07	1.10	84
Panel C: Off-peak cost statistics (01/2004 – 12/2007 and 01/2010 – 03/2011)						
Aggregate market shortfall (mill.)	Daily	413.0	198.4	80.4	1058.4	1,318
Cost of top 1 default (mill.)	Daily	53.4	39.7	0	148.2	1,318
SDFUR	Daily	0.31	0.28	0	1.22	1,318

This table presents summary statistics for non–contract specific (participant-level and aggregate-level) variables and their level of observation. Summary statistics are taken with respect to the original data frequency and level of observation at which the variable is reported. Panel A covers the full sample period from January 2, 2003 to March 31, 2011. Panel B presents margin statistics for the on-peak period (09/2008 – 12/2008), and Panel C, for the off-peak periods (01/2004 – 12/2007 and 01/2010 – 03/2011). All margin and EVT-dependent data series start in 2004 to allow a 1-year leading estimation window consistent with SPAN.

Table 4: Risk Mutualization Costs for Individual Participants

Institution	Mean	SD	Min	Max
1	0.66	0.667	0	3.43
2	0.48	0.425	0	2.07
3	0.37	0.376	0	1.69
4	0.37	0.243	0.078	1.15
5	0.35	0.327	0	2.12
6	0.45	0.403	0	2.25
7	0.60	0.331	0.178	1.50
8	0.85	1.075	0	5.09
9	0.06	0.063	0	0.34
10	0.61	0.941	0	7.48
11	0.09	0.139	0	0.85
12	0.31	0.386	0	2.90
13	0.91	0.861	0	3.66
14	0.54	0.501	0	2.69
15	0.78	0.585	0.113	2.96
16	0.33	0.346	0	1.62
17	0.55	0.581	0	3.03
18	0.17	0.142	0	0.93
19	0.77	0.632	0	5.62
Cross-section	0.49	0.475	0	7.48

This table presents summary statistics for the risk mutualization cost for each market participant, measured as a fraction of the latter's default fund deposit, conditional on the default of the highest-impact participant. The institutions' identities and number of observations per institution are suppressed due to data disclosure requirements, and institutions are listed in random order. The sample period spans from January 2, 2003, to March 31, 2011, but all EVT and margin estimates (and resulting comparisons) start on January 2, 2004, to allow a 1-year margin estimation window consistent with SPAN.

References

[1] Berliant, J., Herrmann, K., and Teugels, J.L. 2017. Estimation of the Extreme Value Index. In: *Extreme Events in Finance*, Longin, F. (editor). Hoboken, NJ: John Wiley and Sons.

[2] Brennan, K., Chang, B.Y., Odabasioglu, A., and Raykov, R. (2024). Applying Extreme Value Theory to Central Counterparty Resolution Plans. Bank of Canada Staff Analytical Note, forthcoming.

[3] Brunnermeier, M., and Pedersen, L. H. 2009. Market liquidity and funding liquidity. *Review of Financial Studies* 22(6), 2201–2238.

[4] Campbell, B., and Chung, C. 2003. CGB: Poised for Takeoff. An Analysis of the Ten-Year Government of Canada Bond Future Based on Intraday Trading Data. Working Paper, CIRANO.

[5] Canadian Derivatives Clearing Corporation. 2012. Risk Manual. Montreal, QC: TMX Group.

[6] Chicago Mercantile Exchange. 2019. *CME SPAN: Standard Portfolio Analysis of Risk.* Chicago, IL: Chicago Mercantile Exchange.

[7] Cotter, J. 2001. Margin exceedences for European stock index futures using extreme value theory. *Journal of Banking & Finance* 25, 1475–1502.

[8] Cotter, J., and Dowd, K. 2017. Margin setting and extreme value theory. In: *Extreme Events in Finance*, Longin, F. (editor). Hoboken, NJ: John Wiley and Sons.

[9] CPMI-IOSCO. 2012. Principles for financial market Infrastructures. Basel, Switzerland: Bank of International Settlements.

[10] CPMI-IOSCO. 2017. Recovery of financial market infrastructures – revised report. Basel, Switzerland: Bank of International Settlements.

[11] Cruz Lopez, J., Harris, J., Hurlin, C, and Perignon, C. 2017. CoMargin. *Journal of Financial and Quantitative Analysis* 52(5), 2183–2215.

[12] Danielsson, J., Peng, L., de Vries, C., and de Haan, L., 2001. Using a bootstrap method to choose the sample fraction in tail index estimation. *Journal of Multivariate Analysis* 76, 226–248.

[13] Danielsson, J., Ergun, L., de Vries, C., and de Haan, 2019. Tail Index Estimation: Quantile-Driven Threshold Selection. Bank of Canada Staff Working Paper No. 2019-28.

[14] Davydov, D., Vähämaa, S., and Yasar, S. 2021. Bank liquidity creation and systemic risk. *Journal of Banking and Finance* 123, 1–17.

[15] De Haan, L., and A. Ferreira. 2006. *Extreme Value Theory: An Introduction.* New York, NY: Springer Verlag.

[16] De Haan, L., Jansen, D.W., Koedijk, K., de Vries, C.G. 1994. Safety first portfolio selection, extreme value theory and long run asset risks. In: Galambos, J., Lechner, J., Simiu, E. (Eds.), *Extreme Value Theory and Applications*, 471–487. Dordrecht: Kluwer Academic Publishers.

[17] Drees, H., de Haan, L., and Resnick, S. 2000. How to make a Hill plot. *The Annals of Statistics* 28(1), 254–274.

[18] Drees, H., Janssen, A., Resnick, S. I., and Wang, T. 2018. On a minimum distance procedure for threshold selection in tail analysis. *SIAM Journal on Mathematics of Data Science* 2(1), 75–102.

[19] Drees, H., and Kaufmann, E. 1998. Selecting the optimal sample fraction in univariate extreme value estimation. *Stochastic Processes and Their Applications* 75(2), 149–172.

[20] Duffie, D., and H. Zhu. 2011. Does a central clearing counterparty reduce counterparty risk? *Review of Asset Pricing Studies* 1(1), 74–95.

[21] Hall, P. 1990. Using the bootstrap to estimate mean squared error and selecting parameter in nonparametric problems. *Journal of Multivariate Analysis* 32, 177–203.

[22] Hill, B. 1975. A simple general approach to inference about the tail of a distribution. *Annals of Statistics* 3, 1163–1173.

[23] Huisman, R., K.G. Koedijk, C. Kool, and F. Palm. 2001. Tail-index estimates in small samples. *Journal of Business & Economic Statistics* 19(2), 208–216.

[24] Jansen, D. and de Vries, C. (1991). On the frequency of large stock returns: Putting booms and busts into perspective. *The Review of Economics and Statistics* 73(1), 18–24.

[25] Longin, F. 1999. Optimal margin level in futures markets: Extreme price movements. *Journal of Futures Markets* 19(2), 127–152.

[26] Longin, F. (editor). 2017. *Extreme Events in Finance.* Hoboken, NJ: John Wiley and Sons.

[27] Loretan, M., and Phillips, P.C.B. 1994. Testing the covariance stationarity of heavy-tailed time series. *Journal of Empirical Finance* 1, 211–248.

[28] Menkveld, A. 2017. Crowded positions: An overlooked systemic risk for central clearing counterparties. *Review of Asset Pricing Studies* 7, 209–242.

[29] Odabasioglu, A. 2023. Procyclicality in Central Counterparty Margin Models: A Conceptual Tool Kit and the Key Parameters. Bank of Canada Staff Discussion Paper No. 2023-34.

[30] Paddrick, M., Rajan, S., and Young, H. P. 2020. Contagion in derivatives markets. *Management Science* 66(8), 3603–3616.

[31] Raykov, R. 2022. Systemic risk and collateral adequacy: Evidence from the futures market. *Journal of Financial and Quantitative Analysis* 57(3), 1142–1173.

[32] Resnick, S. and Starica, C. 1998. Tail index estimation for dependent data. *The Annals of Applied Probability* 8(4), 1156–1183.

[33] TMX Montreal Exchange. 2013a. BAX three-month Canadian bankers' acceptance futures descriptive brochure. Montreal, QC: TMX Group.

[34] TMX Montreal Exchange. 2013b. *Government of Canada Bond Futures and Options on Futures Reference Manual.* Montreal, QC: TMX Group.

[35] TMX Montreal Exchange. 2013c. *Index Derivatives Reference Manual.* Montreal, QC: TMX Group.

[36] Van Oordt, M. and Zhou, C. 2016. Systematic tail risk. *Journal of Quantitative and Financial Analysis* 51(2), 685—705.

[37] Weissman, I. 1978. Estimation of parameters and large quantiles based on largest observations. *Journal of the American Statistical Association* 73, 812–815.