生成 AI を用いた仮想マーケットデータシナリオ生成

Syed Hashim SHAH[§,†]　Joel VIKLUND[§,†]　青木 稔 [§,‡]

[§] 株式会社日本証券クリアリング機構

[†]Aurora Solutions K.K.　[‡] 株式会社野村総合研究所

2025 年 8 月 4 日

概要

近年、生成 AI、AI は幅広い産業に革命を起こしている。しかし、主な使用事例は、チャットボット、コードアシスタント、およびテキスト分析などの大言語モデル (LLMS) に限定されている。本白書では、清算機関 (CCP) におけるリスク管理のための生成 AI の使用を検討する。金融の世界は厳密な数字によって支配されているので、数値データを扱い生成することができる生成モデルを検討したい。本稿では、まず、一般的な AI と機械学習の概念を紹介し、生成 AI モデルを用いて、TONA 3M 先物契約に焦点を当てて、合成であるが現実的な市場データを作成する方法を検討する。変分オートエンコーダ (VAE) と主成分分析 (PCA) を比較して、合成データを生成し、生成したシナリオを分析した。その後、生成された合成市場データを用いて、期待ショートフォール・バリュー・アット・リスク (ES-VaR) を計算することにより、様々なポートフォリオのリスク・プロファイルを推定する。結果は、VAE 生成シナリオは PCA 生成シナリオよりも多様であり、ES-VaR に影響することを示した。

JPX JAPAN SECURITIES CLEARING CORPORATION　AURORA SOLUTIONS　NRI

Disclaimer

1. JSCC は当初証拠金計算において生成 AI をもちいてシナリオ生成を行うことを意図しておりません。
2. 本ワーキング・ペーパーは英語版が原文であり、この日本語版は参考のために便宜上作成されたものです。英語版と日本語版に矛盾及び齟齬等がある場合は、英語版の内容が優先されます。

目次

1　はじめに

JSCC(日本証券クリアリング機構) は JPX グループにおける清算機関、CCP(central counterparty) である。株式の現物取引から、上場派生商品の取引所取引の他、金利スワップ、クレジットデフォルトスワップに店頭国債など OTC 商品の清算もおこなっている。

1.1　清算機関の役割とリスク管理

CCP は、FMI(Financial Market Infrastructures) の 1 つである。FMI は、金融市場の安全性と安定性にとって重要である。バイラテラル・エクスポージャの複雑な蜘蛛の巣状のネットワークは、CCP を介してより単純なネットワークに縮小され、マルチラテラル・ネッティングの利益を提供することができる。マルチラテラル・ネッティングは、決済リスクを低減する [1] 。

(a)　(b)

図 1.1: バイラテラル・ネットワーク (a) とマルチラテラル・ネッティング (b)

また、CCP は各取引当事者 (売り手や買い手) 取引に基づく債務を引き受ける保証人の役割ももっている。この役割は連鎖破綻を抑止する。債務引受により、一部の参加者が破綻を起こしたとしても、その破綻者が払う金額を CCP が肩代わりすることで危機の連鎖を防ぎ、金融リスクを低減している。

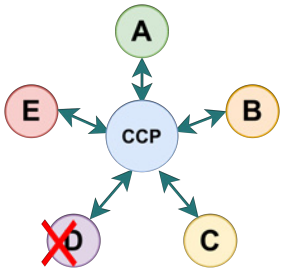


図 1.2: 債務引受により債務不履行者に代わって清算機関が受払を行う

1.2　CCP によるリスクの吸収

いずれかの清算参加者が CCP のほうで設定する時限までに支払いや担保の拠出といった義務を果たせないと破綻処理プロセス (DMP, default management process) が発動される。この破綻処理では破綻者が拠出していた担保、CCP が拠出する財源、生存参加者が拠出していた共同担保などの財源をつかって処理が進められていく。

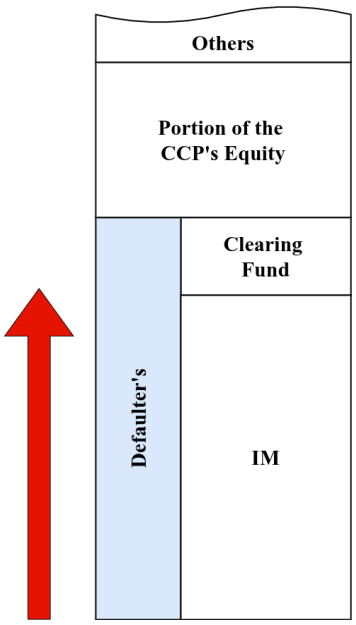


図 1.3: 破綻処理におけるリスクウォーターフォール

破綻処理で使用される財源リソースのうち主要な部分を占めるのが当初証拠金 IM (Initial Margin) である。IM が少なすぎると破綻時に不足が発生し、他の財源リソースを使用する可能性があるが、多すぎると各参加者の資本効率の悪化につながる。多すぎず少なすぎない IM 額というものは重要になってくる。一般に、ポートフォリオが有するリスクが大きいほど、IM はより多く必要とされる。

1.3 バリューアットリスク

IM とはデフォルト時における期待損失額である。多くの CCP はその IM の計算においてヒストリカルデータを用いた HS-VaR(Historical Simulation Value-at-Risk) と呼ばれる計算方式を採用している。その損失分布は特定の数学的分布を想定せず、経験分布をもとに計量されることが多い。IM とはその分布におけるパーセンタイルと呼ばれる信頼区間で示される期待損失額となっている。パーセンタイルで特定される損失金額は狭義 VaR(Value At Risk) とよばれる。一方パーセンタイルを超える損失をテールリスクと呼び、これらの平均をとることでテールリスクの金額まで加味した損失額を出すことがあり、これを Conditional VaR もしくは期待ショートフォールと呼んでいる。本稿では VaR といった場合にはこの期待ショートフォールでの算出を意味する。CCP では破綻処理において、その商品によって決められた日数で破綻処理を終える必要があり、この期間のことを MPOR(Margin Period Of Risk) と呼んでいる。

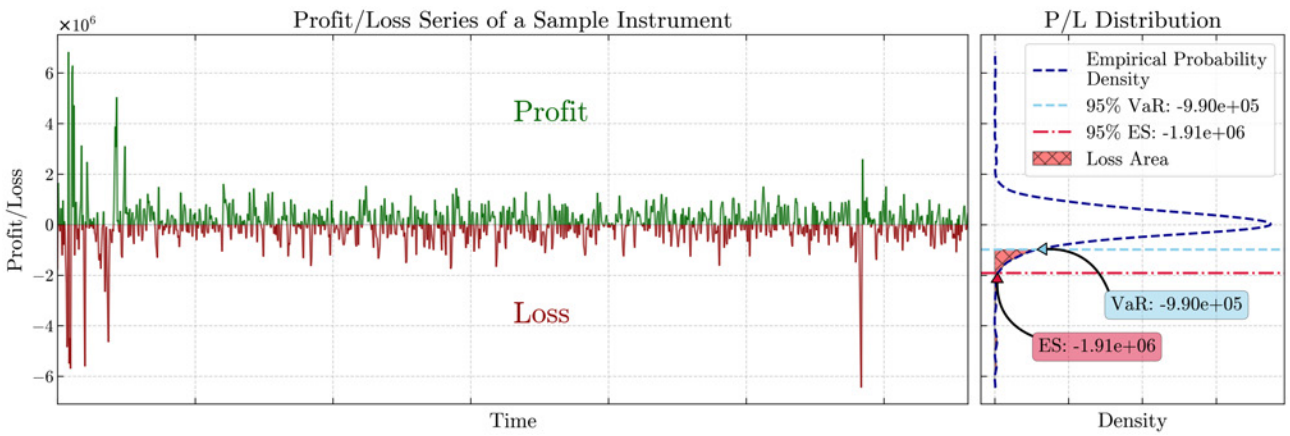


図 1.4: MPOR での損益分布

損益を決定する各リスクファクターの動きをリターンと呼んでいる。これらはかならず独立して動いているわけではなく、時としてほかのリスクファクターの動きと相関を持って動いている。過去の各リターンはほかのリスクファクターと相関のあるリターンとして、それらの確率に基づいた IM の計算に用いられる。現在の価格にリターンを適用することで、VaR 計算用の商品別 PL を計算できる。先物に関しては満期ごとの価格を反映した先物値段カーブとしてモデル化することで任意の満期の時系列リターンをつくることができる。計算上使われる実際に過去に起きたシナリオは、何より実際に起きたという説得力があり、起きうるシナリオといえる。この起きうるシナリオを集めて IM を計算するが、こうした過去に起きたシナリオはヒストリカルシナリオという分類に入る。これ以外にもヒストリカルシナリオでも観察期間を超えて大きく動いたシナリオをストレスシナリオと呼んでいる。

しかしながら、難しいのはまだ起きていないシナリオを想像することで、これらは仮想シナリオと呼ばれる。これらは過去に起きたものではないがある仮定をもとに作られたシナリオで極端ではあるがあり得るシナリオである。こうした各種シナリオをもとに IM は計算されることになる。起きうるシナリオというものは AI を使えば合成シナリオとして導出できると考えられる。このシナリオの質がより精緻なリスク管理の決定要素となる。

1.4　マーケットシナリオにおける生成 AI

生成 AI は巨大な過去データから学習でき、リスクファクター間の線形並びに非線形相関を見つけることができるかもしれない。その過程においてはデータの次元を減らし、圧縮されたデータポイントが生成される。次元を減らした断面は潜在空間 (Latent Space) とよばれる。潜在空間内の新しいデータ点をサンプリングし、それらを元の空間に復元することで合成データのサンプルを生成することができる。生成 AI を用いることで、他のリスクファクターの動きの発生確率に基づいて、もっともらしい合成シナリオを生成することが可能となる。

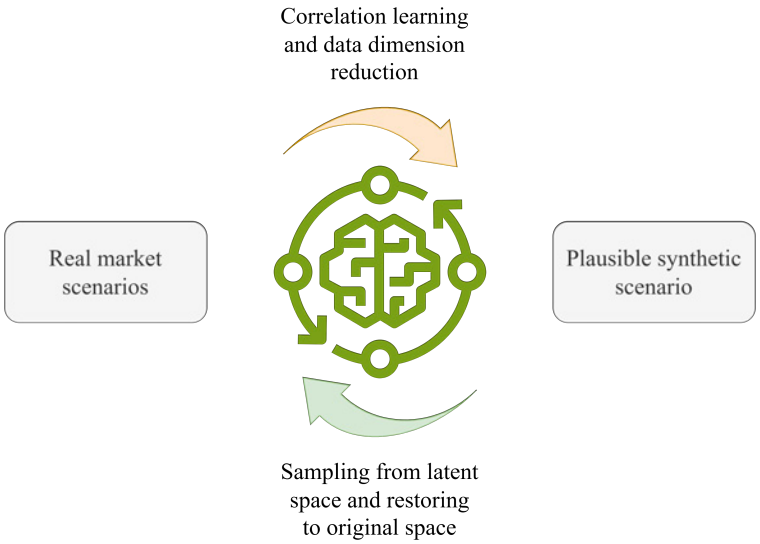


図 1.5: マーケットシナリオでの生成 AI ユースケース

今回生成 AI を使って検証してみようと思った動機はここにある。以下の章では、人工知能や機械学習、生成 AI の一般的な側面について説明する。そして生成 AI を合成データの生成に活用し、出てきたシナリオを VaR の計算に適用している。2023 年、JPX の大阪取引所では上場派生商品の新商品として金利先物の 1 つである TONA 3M 先物を上場した [2] 。この商品の MPOR は 2 日であり、今回の検証における VaR の計算例はこれに基づいている。

2　生命・機械学習・人工知能・生成 AI

人工知能、機械学習、および生成 AI は、広範な消費者領域および企業領域の両方に適用されており、我々が複雑な問題にアプローチし、解決する方法に革命をもたらしてきた。人工知能について議論する前に、少なくとも知能の実用的な定義を持つ必要がある。知能を定義することは難しいが、知能を「問題を解決する能力」と定義すると、知能の歴史は生命の歴史とよく一致することがわかる。我々人間は、我々の極めて複雑な脳のおかげで、極めて複雑な問題を解決し、抽象的な概念や言語を理解し、工学において芸術的で驚くべき成果を生み出すことができた。

2.1　定義

人工知能は、多くの方法、ヒューリスティック、アルゴリズムを含む広範な包括的用語である。それは、広範囲の分野や方法を有する。簡単に可視化した概要を図 2.1 に示す。

人工知能 (Artificial Intelligence: AI): AI は、機械、特にコンピュータシステムによる生物学的知能、特に人間に似た知能のシミュレーションとして定義することができる。

- 例としては、アリコロニー最適化、遺伝的アルゴリズムなどが挙げられる。

機械学習 (Machine Learning: ML): このような挙動をコンピュータに明示的にプログラミングすることなく、インテリジェントな挙動をシミュレーションする。コンピュータプログラムは、データから直接関係を自動的に学習する。

- 例としては、決定木、ランダムフォレスト、ディープニューラルネットワークなどが挙げられる。

生成 AI (GenAI): AI/ML を使用して、合成ではあるが現実的なデータを生成する。

- 例としては、ChatGPT などのテキストエンジン、Stable Diffusion や Dall-E などのリアルな画像生成モデルなどがある。[3], [4], [5] 。

テキスト、画像、音声などのコンピュータ上のデータは、2 進数の 0 と 1 のみを理解するコンピュータとして数値として表現しなければならない。金融データは数値で表現されるため、機械学習および生成 AI のための主要なターゲットである。

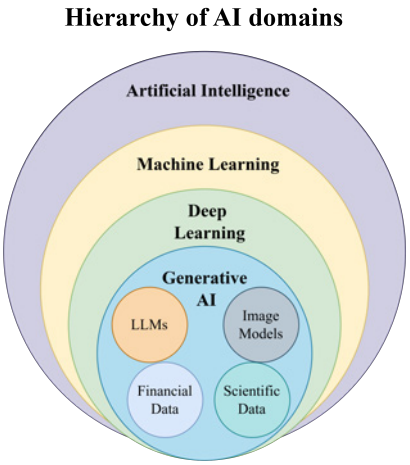


図 2.1: 人工知能とそのサブドメインの景観

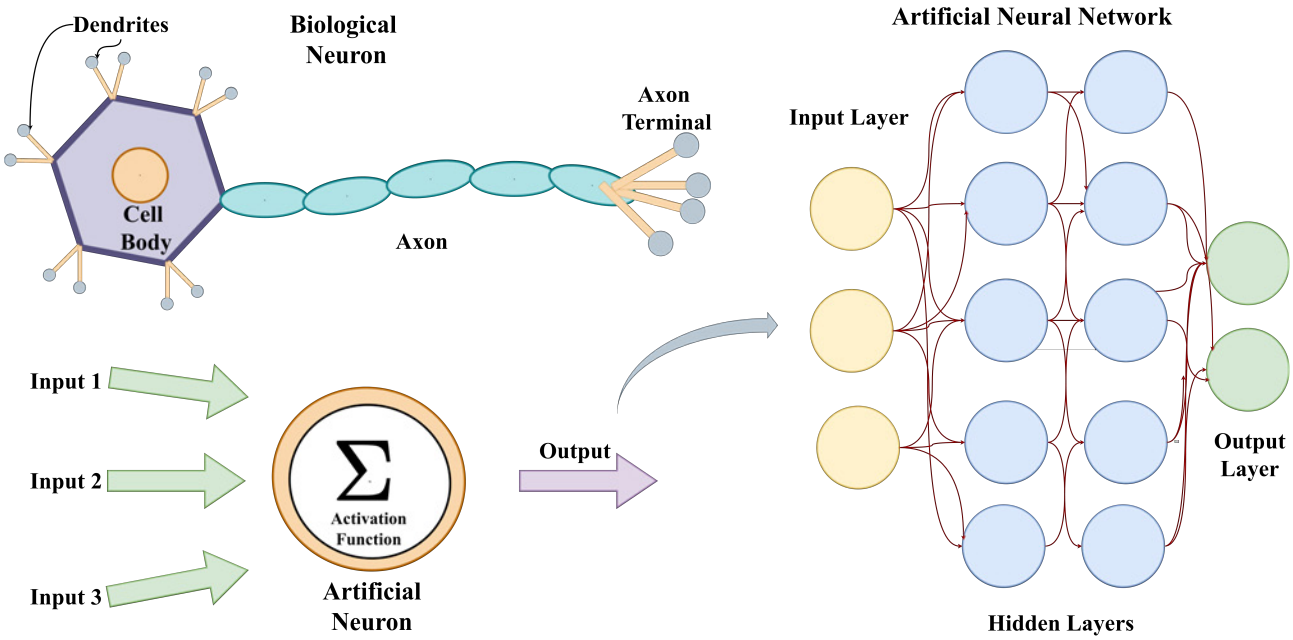


図 2.2: 生物学的ニューロンの模式図。樹状突起は、周囲の他のニューロンからの信号を受信し、次いで、細胞体内で処理され、電気信号が軸索を通って下方に送られる。軸索終末は下流の他のニューロンと連結している。「人工ニューロン」と呼ばれる生体ニューロンの数学的近似も示した。それは、複数の数値入力を受け取り、活性化関数によって単一の数値出力値に結合し、それを下流の他のニューロンに渡す。複数のそのようなニューロンがネットワーク内で接続されるとき、任意の複雑なパターンに適合することができる「人工ニューラルネットワーク」が得られる。入力層と出力層の間に複数の隠れ層を追加し、「ディープニューラルネットワーク」を得ることができる

複雑な規則を持つコンピュータをプログラムしてインテリジェントな想定動作をシミュレートすることができるが、これらのアルゴリズムは学習や適応をすることができない。世界のチャンピオンを打ち負かした初期のチェスやボードゲームのようなアルゴリズムは、この仕組みである。人工ニューラルネットワーク (ANN) と呼ばれる強力な AI アルゴリズムは、人間の脳のニューロンのようにデータから自動的に学習し、複雑なタスクを行うことができる [6], [7] 。ニューラルネットワークは、図 2.2 に示すように、脳内の生物学的ニューロンネットワークの数学的類似体である。我々の脳は、電気化学信号を流す非常に複雑なネットワークにおいて相互接続された数十億のニューロンを有することができるが、コンピュータソフトウェアにコード化することができる数学的ニューラルネットワークは、数値入力を取り込み、計算結果を出力するために複雑な数学的変換を行う。

ANN の成功の理由は「普遍的近似定理」と名付けられた数学的定理であった。簡単に言えば、普遍的近似定理は、十分に広いニューラルネットワークが任意の連続関数を任意の精度に近似できることを述べている [8] 。直観的には、ニューラルネットワークの内部には、独立して微調整することができる非常に多くの自由度 (パラメータ) が存在するので、それを用いて任意の複雑な関数またはパターンを近似することができる。

2.2 AI/機械学習を用いたカーブフィッティング

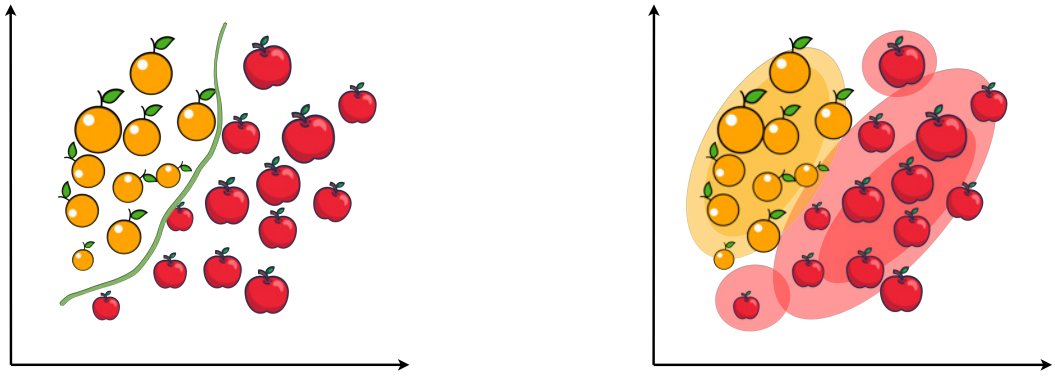
多くのリスクの専門家にとって、確率分布をカーブフィッティングまたはフィッティングするという考え方は、よく知られている仕事である。従来、スプラインは曲線の当てはめに使用され、コピュラはデータ分布の当てはめに使用されていた。すぐに、機械学習の核心は、長年の古典的な統計的方法では伝統的に見られない新しいアーキテクチャを使用することによって、曲線/表面適合または基礎となるデータ分布を捕捉することであることが分かるであろう。複雑な高次元データセットを持つ場合、決定境界と確率分布を学習するための汎用アーキテクチャが必要である。古典的な統計では、データに一定の仮定を置いて、ガウス分布、ガウス分布の混合分布、またはファットテール分布の組み合わせを用いて、データをフィッティングしようとする。このアプローチは、次の理由から限定的である

1. データに仮定を置いている (特性の独立性、非線形相関を無視)
2. 次元の問題。大きな次元ではデータ特性の組み合わせが大きく、我々の直観は容易に誤り、古典的なモデル化に失敗することが非常に多い

従って、ニューラルネットワークを用いることにより、単純化仮定を設けることなく、データに対して複雑なモデルを適合させることが可能となる。

AI モデルを構築するタスクに応じて、機械学習に使用される 2 つのパラダイムがある。

1. 識別
2. 生成



(a) 識別モデリング　　(b) ジェネレーティブ/生成モデリング

図 2.3: 二つの機械学習パラダイムの比較: 識別モデル（図 2.3a ）は決定境界を学習するのに対し、生成モデル（図 2.3b ）は確率密度を学習する

識別 AI/ML は、データサンプルが与えられたデータのラベルを分類または予測することに関するものである。データを異なるクラスに分類する場合、図 2.3a に示すように、データクラスを効率的に分離する特徴空間内の識別境界を見つけたい。この境界は、高次元データの場合には複雑な曲線または表面である可能性があるため、単純な線形境界または単純な関数形式では通常は十分ではない。境界を形成するためには、通常、柔軟な機能形態が必要とされる。識別モデルは、教師あり学習で訓練される。

データの生成モデリングでは、図 2.3b に示すように、データ特性とそのラベルの結合確率分布を学習することを試みる。つまり、データ特性とその範囲の異なる組み合わせを学習し、それらについての確率分布を学習する。たとえば、規模、色、形状パラメータ、リンゴとオレンジを定義する組み合わせを学習できる。この場合、生成的側面としては、この潜在的な分布から効率的にサンプリングし、それをデータ空間に変換することである。異なる特性は、複雑であり、通常は非ガウス型に分布することができ、また、互いに非線形依存性を有することができるので、それをモデル化するためには、柔軟性のある非パラメトリック機能が必要である。

2.3　教師あり、教師なし、強化学習

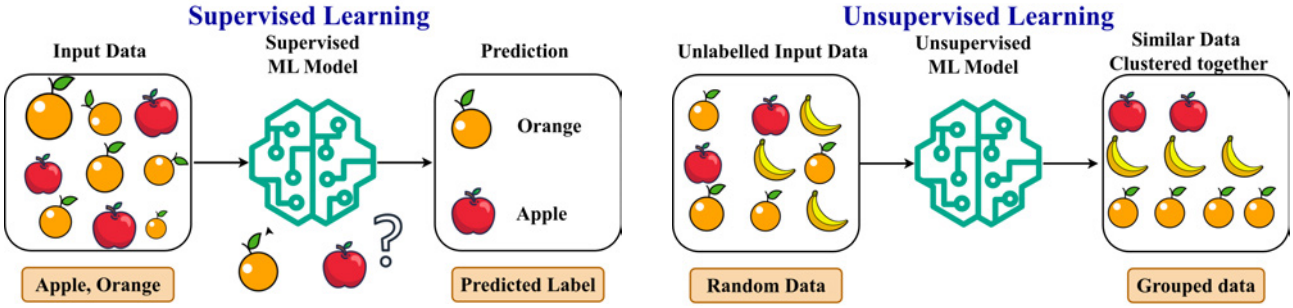
AI モデルと ML モデルをトレーニングまたはティーチングして、必要なタスクを実行するには、主に 3 つの方法がある。

1. 教師あり学習
2. 教師なし学習

3. 強化学習

教師あり学習は、質問またはデータと共に、そのデータのラベルの回答も提供する。たとえば、AI モデルにオレンジとリンゴの写真を十分に表示し、これらがリンゴであり、これらがオレンジであることを伝えると、図 2.4a に示すように、それらを区別することができる。

教師なし学習は、データを持つラベルを持たない場合である。データ中の複雑なパターンを学習するためにアルゴリズムを訓練する。図 2.4b に示すように、異なるデータサンプル間の類似性を見つけ、それらをまとめてクラスタリングするか、予測を行うことができる



(a) リンゴとオレンジの区別

(b) 類似した果物をグループ化

図 2.4: 教師あり学習 (図 2.4a) と教師なし学習 (図 2.4b) の方法論の概略比較

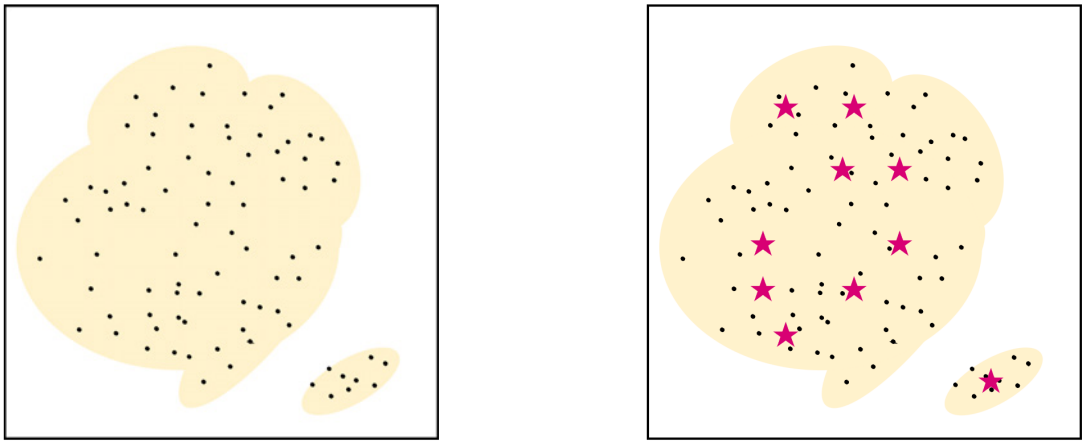
強化学習については、詳細には論じないが、ML モデルまたは知的エージェントが、報酬を最大化するために動的環境において行動または決定をとることを学習することである。それは正しい決定に対して報酬を与え、間違った決定に対してペナルティを与える。強化学習は、動的アセットアロケーション、トレーディング、ポートフォリオ管理、動的ヘッジなどにうまく適用されている。

実際に AI モデルを構築する場合、3 つのアプローチはすべて、以下の順序で組み合わされる:

1. 利用可能なデータのほとんどはラベルなしであるため、教師なし学習
2. 正確にラベル付けされたデータは限られているので、教師あり学習。これで、教師なしモデルを洗練することができる
3. 学習を強化し、モデルを微調整することで、非常に複雑なタスクを教える

2.4 生成 AI

画像、複数の金融時系列、またはテキストの言語資料のような大きなデータセットを有する場合、データの異なる側面と特性との間の複雑な関係を学習することができる。これは、生成 AI やモデリングからの手法を使用することによって行うことができる。生成 AI の核心は、いくつかの基本的な統計と確率理論の知識があれば、理解するのが非常に簡単である。図 2.5 のように本質的には、結合データおよび潜在的分布をできるだけ正確に学習し、次いでそれから適切にサンプリングして、元のデータ分布に従う新規な合成データを生成することである。



(a) データおよびその確率分布（黄色）　　(b) 確率分布から新しいデータサンプル（★）

図 2.5: 黒い点はデータサンプルを表し、黄色い雲はデータが生成された確率分布を近似している。星 (★) は、この確率分布からサンプリングされた新規サンプリングされたデータ点を示す。これらの新しいデータポイントは、同じ確率分布からのものであるが、元のデータセットには存在しない

例えば、2 次元または 3 次元といった低次元の場合、およびデータの同時分布がガウシアン関数またはガウシアン関数の混合のように単純である場合、そこからのモデリングおよびサンプリングは直感的で容易である。

しかしながら、実際のデータ分布は、複雑であり、高次元であり、疎である。ここで、疎とは、高次元特徴空間において、現実世界のデータサンプルの大部分が、空間の非常に小さな領域に集中することを意味する。特性値の組み合わせの利用可能なスペースは、特性の次元次第で指数関数的に増加するが、そのスペースのかなりの部分を満たすのに十分なデータがない可能性がある。どんな現実的なデータでも、われわれはすぐに次元の問題にぶつかることになる。この高次元空間においてランダムにサンプリングするだけでは、現実的なデータサンプルを作成することはできない。

第 2 に、複雑な実データ分布では、データ分布の異なる領域で様々な相関関係がある可能性がある。いくつかの領域は、低い確率を有し得るか、またはデータが疎であり得る。したがって、2 次元でも、確率密度関数 (PDF) とそれに付随する標本抽出法を持たない可能性がある。2 つの問題が手元にある:

1. データの連続的で滑らかな確率分布を学習する
2. その分布から効率的にサンプルを採取する

ほとんどすべての生成 AI 研究および方法は、これら 2 つの問題を解決することに関するものである。

2.4.1　確率分布とサンプリングの学習

実際のデータ分布は複雑であり、実際には、データを生成する真の答えとなる分布もわからないが、複雑で疎なデータサンプルを、連続的で単純な分布に変換する方法があれば処理が容易になり、十分に単純であればそれを簡単にサンプリングし、そのサンプルは逆変換を行うことで、元のデータに戻すことができる。このプロセスを図 2.6 に示す。我々は、我々のデータサンプルを、例えばガウス分布のような、より単純な潜在分布の領域にエンコードするエンコーダ機能を持っている。このガウス分布からランダムにサンプリングし、それをデコーダ機能に通すことにより、現実的にみえるデータ点を生成する。一般に、エンコーダ機能とデコーダ機能の両方がニューラルネットワークである。例えば、多変量正規分布から標本化し、その標本を現実的にみえるイメージまたは金融データのような実際のデータ点に変換することができる。

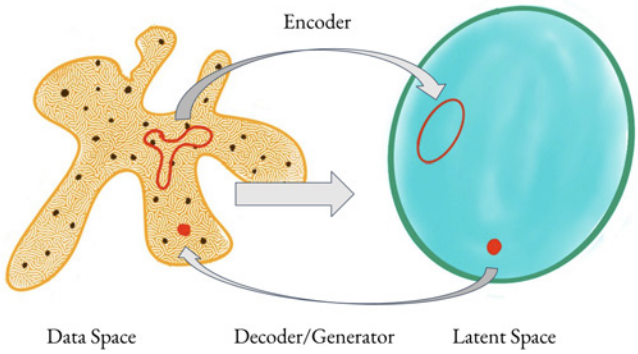


図 2.6: 複雑な実データ分布を単純な分布に変換し、元に戻す。エンコーダニューラルネットワークは、データサンプルをより単純な潜在的分布に変換し、デコーダは、潜在的分布のサンプルをデータ分布に変換する

それでは、複雑な高次元データ分布を単純な分布に変換する方法を検討する。本稿では 2 つの主な手法、これまでよく使われていた主成分分析と生成 AI での主たる手法である VAE を対象としてみることとする。

2.4.2 主成分分析

高次元データの主成分分析 (PCA) は、データの分散又は変化を最大にするような「主成分 (PC)」と呼ばれるデータ空間内の方向のシーケンスを発見する [9] 。主成分は、それらが説明する分散の量によって順番付けされる。これらの主成分方向は、直交し、直線的に無相関であり、新しい座標系を定義する。これらの PC は最大共分散を持つデータ特徴の線形結合であり、データ共分散行列の固有ベクトルとして得られる。データ分散の大部分を説明する最初の 2 つの主成分によって定義された平面上にデータを投影すると、高次元データの 2 次元表現が得られる。数学的には、主成分分析は、単なる線形変換であり、すなわち、これは、変換行列を用いたデータ行列の単純な行列乗算によって実行される。直観的には、図 2.7 に示すように、最も重要な線形構造をとらえるデータの観点を与えてくれる。

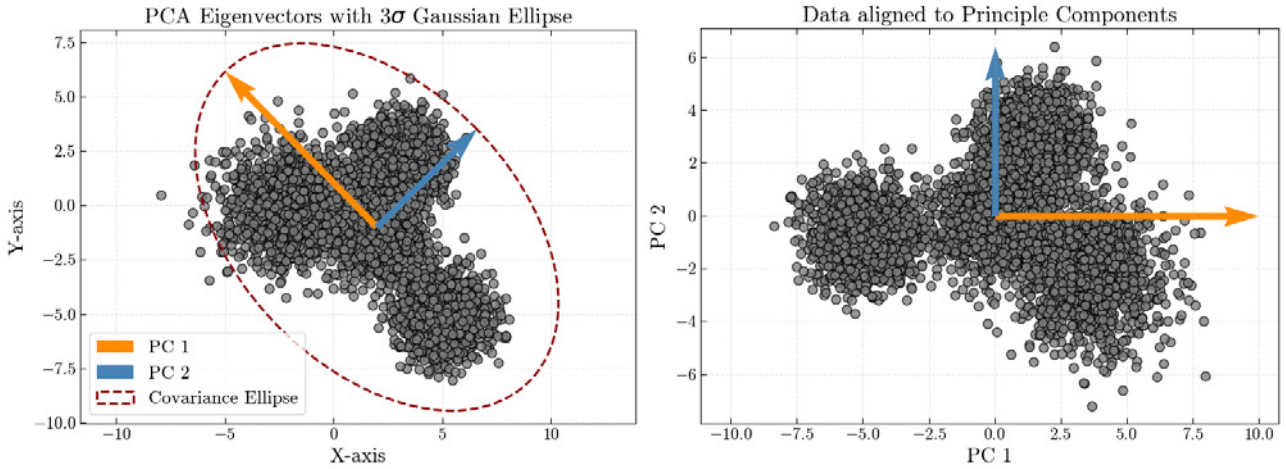


図 2.7: ガウス分布の混合から生成されたダミーデータの主成分分析 PC 1 の固有ベクトル点はデータの最大直線変化の指示にあり、PC 2 の固有ベクトル点は 2 番目に大きい直線移動の方向にあることがわかる。新しい座標のデータも表示される

主成分分析は、単にデータの線形構造を見て、データ内の非線形関係を完全に無視する。そのため、まったく異なるデータセットがまったく同じ主成分を持つことができる。例えば、図 2.7 のデータにガウス分布を当てはめ、それからサンプルを生成する場合、それは全く同じ固有ベクトルを有するが、完全に異なるデータ分布となる。

2.4.3 Variational Autoencoders 変分オートエンコーダ (VAE)

変分オートエンコーダと訳される VAE は、ニューラルネットワークおよび確率的グラフィカルモデルからのアイデアを使用する一種の生成モデルアーキテクチャである [10] 。VAE は、画像生成、データ圧縮、およびノイズ除去などのための一般的なモデルである。

VAE の段階的メカニズムは以下の通りである

1. 図 2.8 に示すように、データを複数の隠れ層を有するエンコーダと呼ばれる深層ニューラルネットワークに渡す
2. 各階層のニューロンの数は、潜在空間を表すボトルネック階層に達するまで、少なくなっていく。このボトルネックは、ニューラルネットワークにデータの最も重要な特徴を学習させ、非線形エンコーダニューラルネットワークは、データ間の非線形関係を捕捉する。この潜在空間 Z は、データの効率的な圧縮表現である
3. データの潜在的表現は、ガウスノイズでわずかに摂動され、デコーダニューラルネットワークに渡される
4. デコーダは、元の入力データを可能な限り正確に再構成しようとする。はじめは、再構成は不十分であり、再構成誤差は非常に大きくなる。これらの手順をループし、逆伝搬を用いてニューラルネットワークの両方をトレーニングすると、エンコーダおよびデコーダの両方の内部パラメータは、元のデータ再構成に近づくまで更新される。その結果、VAE の潜在空間も、多様性を持った下位データ表現に変換されることになる

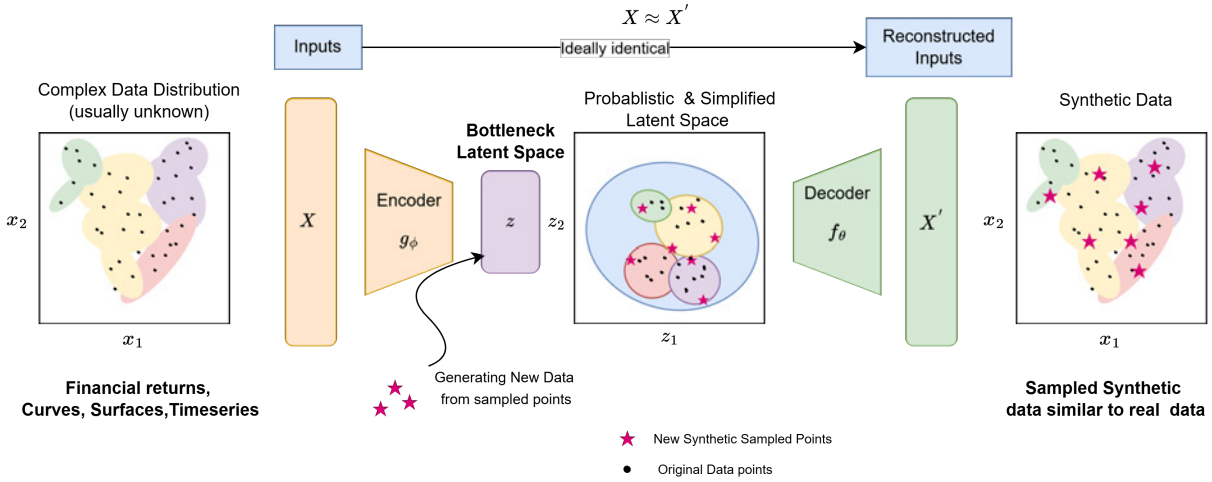


図 2.8: VAE のアーキテクチャ

VAE は、データの連続的なより低次元の潜在空間表現を学習していく。潜在分布は、通常、データよりも低い次元数であり、この潜在空間における方向は有意義である。この次元の削減は、冗長性を減らし、確率空間上での業務を行うための計算量を減らし、通常、サンプリングがはるかに容易になる。画像データのケースでは、数百万の次元 (各着色画素は 3 次元) を数百の次元に縮小することができる。これにより単純な分布からのサンプリングは、データ空間への復号化を可能にし、合成であるが現実的な画像を作成することを可能にする。

簡単に言えば、変分オートエンコーダは以下の問題を解決しようとする。

1. 高次元疎データを、構造化された低次元の単純な確率的で連続的な潜在空間に効率的に圧縮する
2. 潜在空間の簡単なサンプリングを有効にする

変分オートエンコーダーのアーキテクチャ (図 2.8) を用いることにより、滑らかで、構造的で、連続性の特性を持つ、データの圧縮潜在分布を学習することができる。例えば、数百次元のデータがあるかもしれないが、そのデータ

の大部分は、そのスペースに埋め込まれた数十個の内部次元のより単純な低次元の部分スペースにある。3 次元空間に埋め込まれた湾曲した紙片を想像してみよう。この紙は、(ほぼ)2 の次元を有するが、そのねじ曲げる曲率（カバチャー）は 3 次元であることを強いる。変分オートエンコーダを用いることにより、部分空間をもつこの低次元データを学習し、そこから点をサンプリングすることにより新しい合成データを効果的に生成することができる。

変分オートエンコーダーの潜在空間は、カルバック・ライブラー情報量を用いてなめらかで連続的にするためにさらに正則化される。これにより、変分オートエンコーダーの潜在空間における任意の線形補間が、図 2.9 に例示されているように、データ空間において意味のある再構成を有することが可能になる

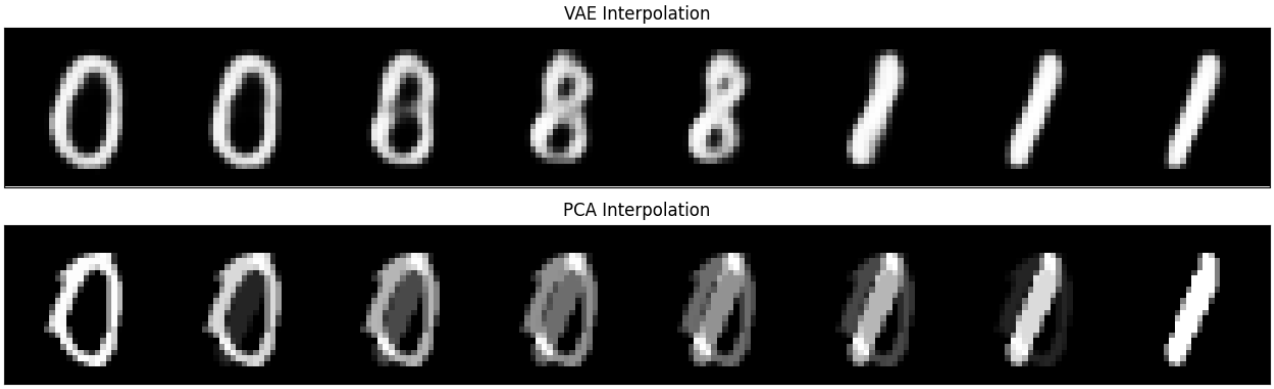


図 2.9: VAE と PCA の潜在空間における手書き 0 と 1 のイメージの補間。VAE 補間のすべての中間再構成は現実的で手書きに見えるが、PCA 空間補間は 1 つの数字がゆっくり消え、もう一方がゆっくり現れるような見え方につながる

VAE は、非常に拡張性があり、強力なフレームワークであり、多数の用途に使用され、大成功を収めている。PCA を VAE の両手法によりデータを 2 次元に圧縮し、それに基づく再構成を行うことで、両手法を比較することができる。VAE は、非線形マッピングを学習することに起因して、その潜在空間において有意により多くの情報を保存することが分かる。しかしながら、完全な再構成は VAE にデータを記憶させ、生成モデルとしての能力を低下させてしまうので VAE の目標ではない。自動車のように見える 3 次元ポイントクラウドがある場合、最初の 2 つの主成分は、深さ情報を無視しながら、自動車の側面図を与える。一方、VAE は、非線形変換を使用することで次元数を減らすというこの考え方を拡張している。PCA のように変換行列を用いてデータ行列を乗算する代わりに、我々が行うことは、より低い次元空間に非線形に圧縮するニューラルネットワークに我々のデータを渡すことである。例えば、車の 3 次元モデルは、車の全ての表面が剥がされ、2 次元表面上に平らに置かれた表現に変換される。

特に金融向け設定において生成 AI モデルを構築したい場合、生成プロセスを容易に制御できるように、VAE のような確率的で明示的な潜在空間を持つ生成 AI モデルが必要である。

2.5 その他の生成モデル

VAE は、データ分布を効果的に学習し、合成データを作成するための唯一の技術ではない。過去 10 年間に出現した他の複数の技法があり、それぞれが独自の強みと弱みを持って、効果的に生成モデリングを行うことを可能にする

1. Generative Adversarial networks (GAN)
2. フローベースモデル
3. 拡散モデル

2 つのニューラルネットワークが互いにミニマックス法のゲームを行う敵対的最適化技法を介して、ＧＡＮが訓練される [11] 。生成側ネットワークは、ランダムガウスノイズベクトルを、人間の顔の画像のような現実的な外観のデー

タサンプルに変換しようと試み、一方、識別側ネットワークは、偽の生成された画像またはデータを、現実のデータ点から区別しようと試みる。生成側の目的は、識別側が実際のデータから区別することができないようなサンプルを生成することである。識別側の目標は、実際のデータと偽のデータとを区別することができることであり、両方とも、それらが目標を達成できない場合には処罰される。生成側が生成したサンプルに対して、識別側が本物か偽物かを 50% の確率でしか割り当てられない場合 (つまり、ランダムな推測)、均衡状態に達します。この段階で、生成側は、真のデータ分布を学習している。ひとたびトレーニングされると、GAN は、標準多変量ガウスノイズベクトルを取り、それから現実的なサンプルを生成する。したがって、これは多変量ガウス分布からデータ分布へのマップである。多変量ガウス分布は、通常、実際のデータ次元よりも小さい次元である。しかし、GAN の欠点は、我々の潜在空間が単に標準的な正規ガウス分布であり、この潜在空間を制御できないことである。データを生成することはできるが、既存のデータを潜在空間に埋め込んだり、操作したりすることはできない。

同様に、フローベースモデル [12] は、データ分布を同じ次元のガウス分布に段階的に変換する可逆マッピングを学習する。この変換を逆にすることによって、任意のガウスサンプルを現実的なデータに変換することができる。これは、正確な尤度計算を可能にする。しかしながら可逆マッピングの必要性は利用できるニューラルネットワークアーキテクチャを制限する。さらに潜在空間の高次元性はトレーニングを困難にし、遅くする。

拡散モデルは [13], [14], [15]、厳密に物理学に基づくアプローチをとる。これらは、ガウス分布に基づくノイズを段階的にゆっくり加えることで、データサンプルをガウス分布ノイズにゆっくりと変換する。次いで、ニューラルネットワークが、サンプルのノイズ除去されたバージョンを予測するように訓練される。これらは、現実的なデータを生成するのに非常に良く、Stable Diffusion、Midjourney、DALL-E のようなモデルの核心にある [4], [5], [16]。理論的には、多次元ガウスノイズから現実的な見た目の画像を生成することができるので、潜在空間を有する必要はない。しかし、それらはサンプリングが非常に遅い。

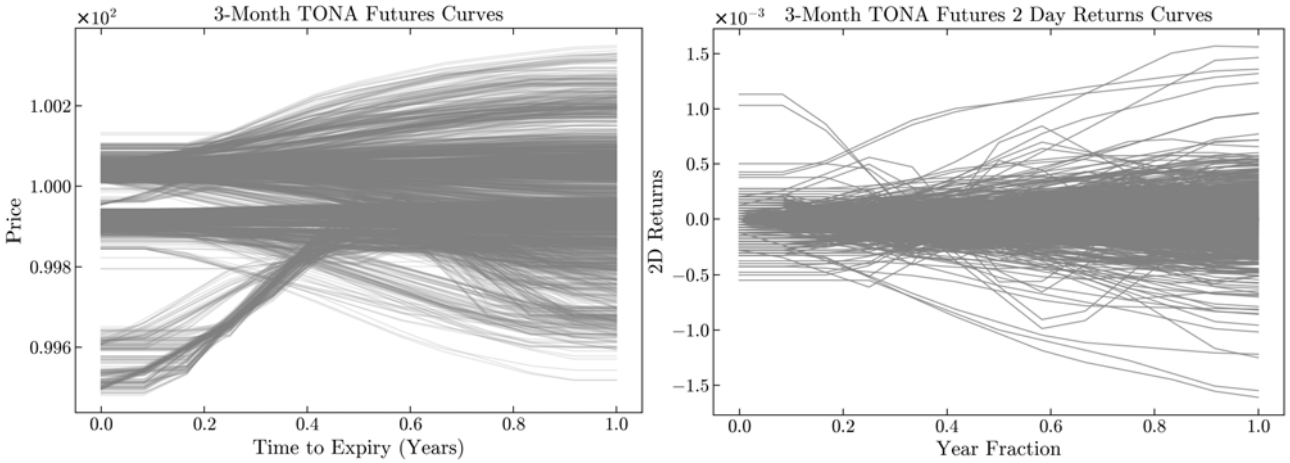
構造的で連続的な潜在空間を持つ変分オートエンコーダのアーキテクチャは、拡張の観点から極めて有用である。VAE は GAN のような敵対的損失を伴って拡張され、高い生成品質に沿った明示的な潜在空間を可能にした [17] 。VAE はまた、大幅なスピードアップを可能にする拡散モデルと組み合わされている [18] 。同様に、フローベースのモデルからのアイデアは、VAE 潜在空間とも組み合わされている [19] 。

3　金融向け合成データのための生成 AI

本節では、VAE と PCA の両手法を用いて合成 TONA 3M 先物カーブを作成するための方法論およびその結果を詳述する。

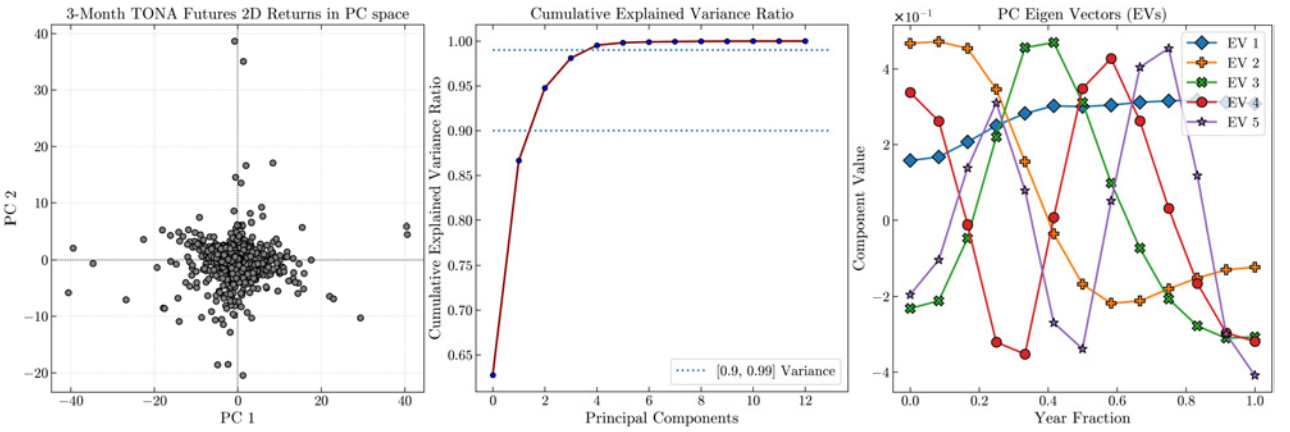
3.1　TONA 3M 先物カーブ

セクション 1.4 において導入された MPOR に整合する市場変動を捉えるため、図 3.1 に示すように、TONA 3M 先物の生のカーブを用いて 2 日間の対数リターン（2D）を算出した。本稿では、各リターン・カーブを独立した標本として扱い、データの自己相関性は考慮しないものとする。1 年以内の満期に対して計算されたリターン・カーブは、以下の通りである。



(a) 先物曲線　　(b) 2 日間 Log Return 曲線

図 3.1: 図 3.1a には、過去の TONA 3M 先物カーブを示す。これに対応する 2 日間の対数リターン（2D log return）曲線は、2 営業日の MPOR における先物カーブの変動を捉えたものであり、図 3.1b に示す。これらの過去リターンシナリオは、生成モデルの学習データとして用いられる

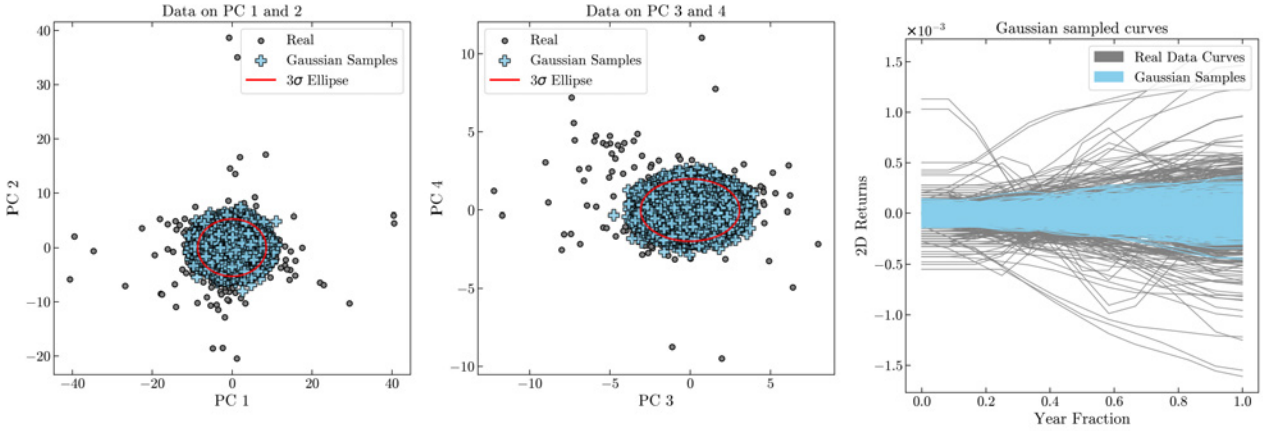


(a) 第 1・第 2 主成分への射影　　(b) 寄与率　　(c) 固有ベクトル

図 3.2: リターンデータの主成分予測 (図 3.2a) 。分散は固有ベクトルによって説明される寄与率 (図 3.2b) と最初の数個の固有ベクトル (図 3.2c) によって説明される

3.2 PCA を生成モデルとして使用する

合成データを生成する最も簡単で単純な方法は、まず 図 3.2 のように主成分空間における変換されたデータへのガウス分布へのデータ変換である。次に図 3.3 のように多変量ガウスをこの主成分空間のデータに当てはめ、多変量ガウスをサンプリングし、合成主成分をデータ空間に逆変換することができる。主成分分析は、テール分布を容易に取り入れることができないことがわかる。たとえ主成分が無相関であっても、図 3.4 に示すように、主成分は必ずしも独立しているとは限らない。

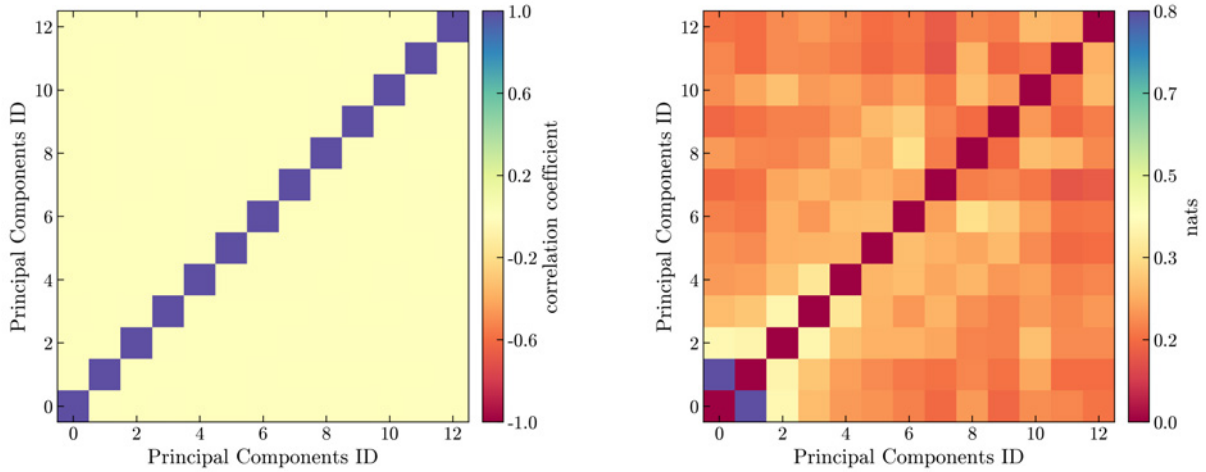


(a) 第 1・第 2 主成分へのガウスフィッティング

(b) 第 3・第 4 主成分へのガウスフィッティング

(c) 曲線としてのガウスサンプル

図 3.3: ガウス分布のフィッティングを介した主成分潜在空間でのサンプリング (図 3.3a, 図 3.3b) 。ガウス分布はファットテールの動きを捉えることができない。サンプリングされた点は、曲線空間に戻され、ガウスサンプルが、データ内のコア運動のみを捕捉することを明確に見ることができる (図 3.3c)



(a) 主成分の相関行列

(b) 主成分の相互情報量行列

図 3.4: 主成分の相関関係は設計によってゼロである (図 3.4a) 。PCA 変換によって線形依存性を取り除いた後でも、主成分間の相互情報はゼロではなく、これは主成分における非線形依存性の存在を示し、その結果、実際のデータ特徴も同様に示すことが分かる (図 3.4b)

非対角項における相互情報量の平均は 0.234 nats であり (図 3.4b)、2 つの変数がガウス分布であると仮定すると、相関係数は 0.61 となる。2 つのガウス確率変数の相互情報量と相関係数を結ぶ計算式は:

$$|r| = \sqrt{1 - e^{-2I}} \tag{1}$$

ここで、r は相関係数であり、I は相互情報量である。主成分分析による変換で線形相関を除いたデータにしても、線形相関の 0.61 に相当する強さの非線形の依存性が残る (式 1) 。この理由は、相関係数が有界である一方、相互情報量が一般に正であり、上限を有さないためである。完全な相関は、無限の相互情報量を有する。主成分分析はデータにおける非線形関係を捕捉することができないので、各種テナーの共運動を正確に捕捉し、ファットテール効果を捕捉するための非線形方法が必要である。

3.3 VAE を生成モデルとして使用する

リターン・カーブに含まれる非線形な依存関係をモデル化するために、VAE を用いてそれらの同時分布を学習させる。トレーニング済みの VAE モデルが得られれば、その潜在空間をサンプリングおよび探索することで、新たな曲線を生成することが可能となる。VAE の潜在空間は、強い正則化により、図 3.5 に示すように、はるかに良好な構造を有している。

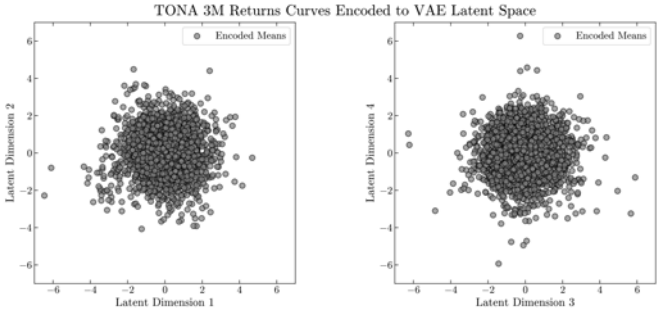


図 3.5: VAE の最初の 4 つの潜在空間の次元

VAE 潜在空間は滑らかであるため、潜在空間内の近くの点は、デコーダによって類似のデータ点に変換される。その滑らかさおよび連続性により、任意の実際の曲線間でデータサンプルをサンプリングすることが可能になる。データスペースまたは VAE 潜在スペースの両方において、2 つの異なる曲線間を補間することができることが分かる。

データ空間補間軌跡は、図 3.6 中段プロットの主成分射影で青線のように直線的であるが、オレンジ線で示す VAE 空間での線形補間は、主成分空間に射影した場合に曲線的であり、一般的にはデータ分布の等高線に従う。したがって、中間体補間点も、単純な補間によって生成されるよりもはるかに現実的である。

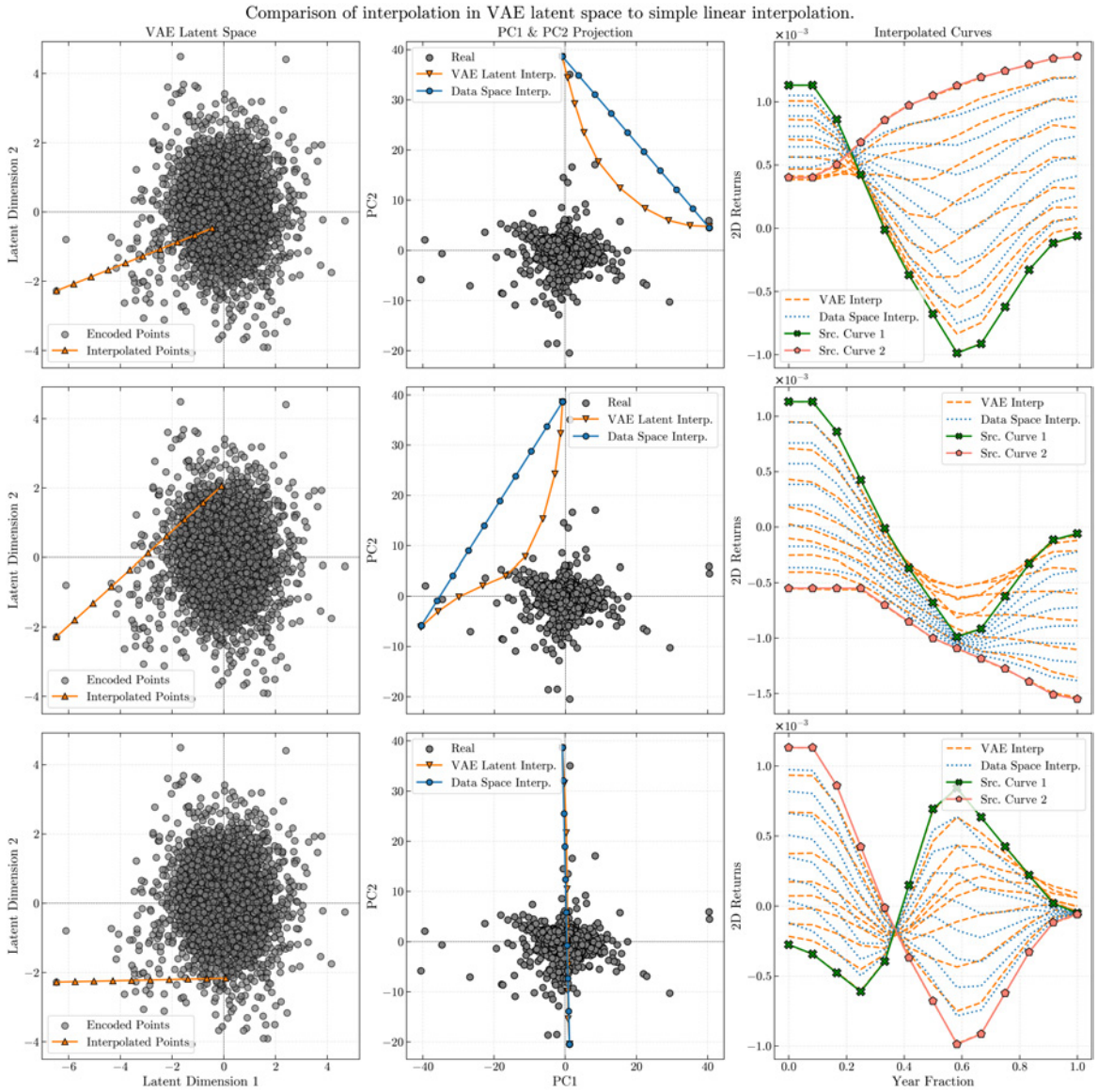


図 3.6: VAE 潜在空間とデータ空間の補間。VAE より現実的なカーブを作成し、データ分布の輪郭に従う

3.4 階層型 VAE を使用したデータ分布のサンプリング

VAE の潜在空間からサンプリングするために、複数の戦略を採用することができる。最初に、VAE スペースを一般的にサンプリングする方法と、いくつかの基本的な問題とそれらの解決法について議論する。第 1 の手順として、VAE の潜在空間にガウス分布を当てはめ、次いで、ガウスベクトルをランダムにサンプリングし、それらをデータ空間に復号することができる。

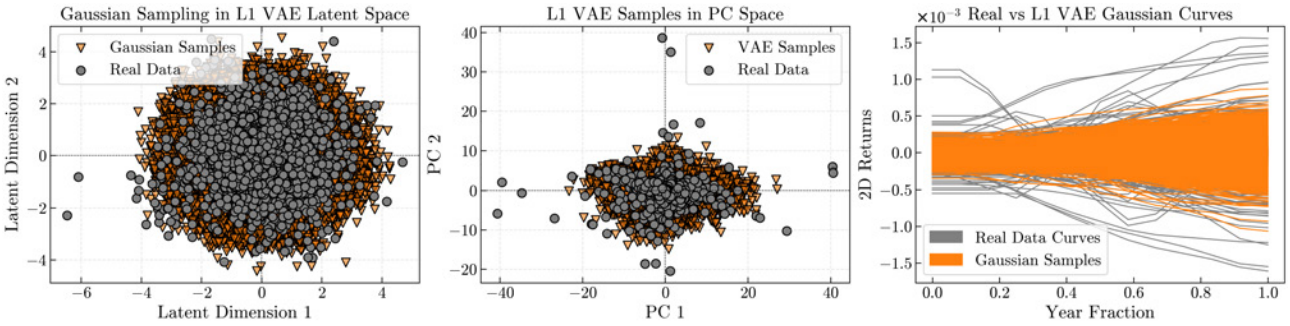


図 3.7: 訓練された VAE 潜在空間はガウス分布ではない。これは、もとのデータ空間よりはるかに良好に振る舞うが、依然としてガウス分布ではない

しかしながら、潜在空間に示されているように (図 3.7) 、標準正規化に対する固有の VAE ベースの正則化にもかかわらず、集約された潜在空間は、ガウス分布に近い必要はない。したがって、潜在空間にガウス分布をフィッティングすることは、テールを無視し、極端に不十分なサンプリングをもたらす。図 3.8 のマハラノビス距離プロット Q-Q プロットは、潜在空間が依然として非常に非ガウス的であることを示している。n 次元正規分布データのマハラノビス距離は自由度 n のカイ二乗分布に従う。初めの VAE による潜在空間に別の VAE をフィッティングさせる。潜在空間と同じ次元の別の VAE をフィッティングすることによって、この問題を回避することができる。第 2 レベルの VAE の潜在空間は、より良好に振る舞うが、依然として、容易なサンプリングを可能にするのに十分なガウス分布ではない。第 3 レベルの VAE は、図 3.8 に示すような非常にガウス型の潜在空間を導くものに当てはめ、ガウス型分布を当てはめてレベル 3 の VAE の潜在空間をサンプリングし、それからレベル 2 のサンプルを得て、それを図 3.9 に示すようにレベル 1 のサンプルに変換する。レベル 3 の潜在空間では、サンプリングは、図 3.9 の中央の列および右端の列に示されるように、外れ値を含む分布全体をうまく包含することが分かる。

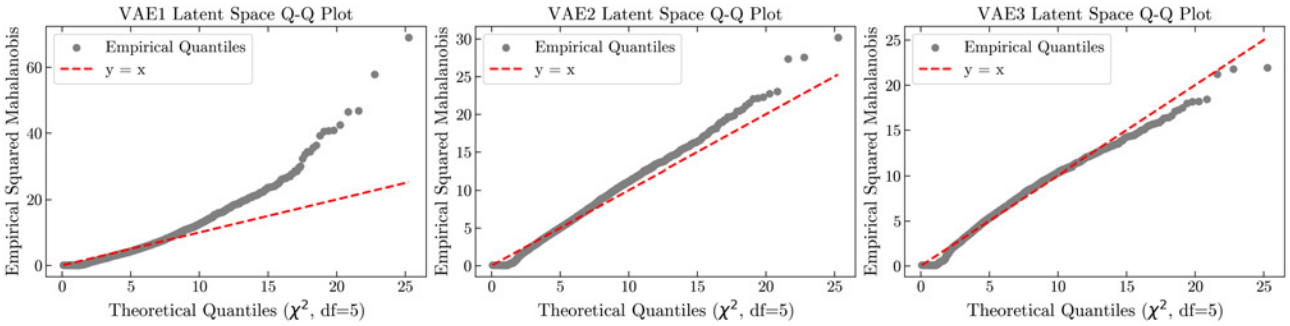


図 3.8: マハラノビス距離の Q-Q プロットにより、レベル 2 VAE は依然としてガウス分布ではなく、ガウスサンプリングは機能しない。一方、レベル 3 VAE はガウス分布に非常に近く、ガウスサンプリングが機能する

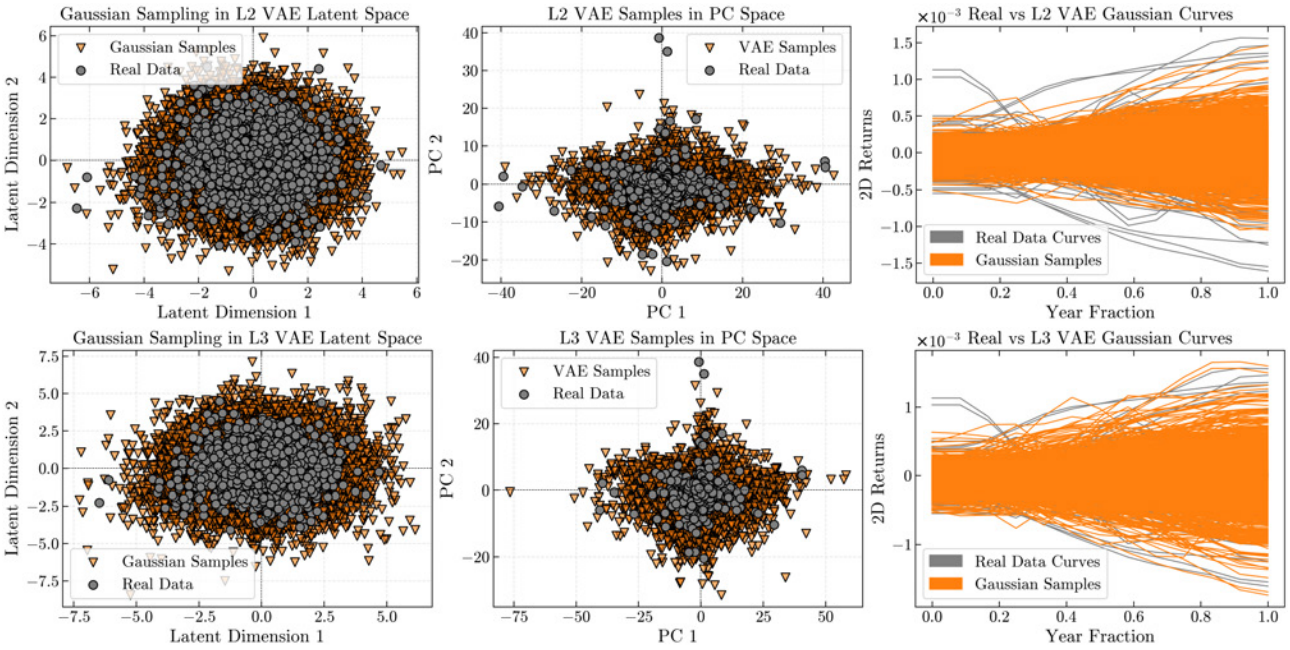


図 3.9: 上段は、レベル 2 の VAE のサンプルを示す。下段は、レベル 3 の VAE の潜在空間と、主成分空間の対応するサンプルを示す。階層的 VAE は、潜在空間を、ガウス分布によってフィッティングすることによって効果的にサンプリングすることができるガウス分布にますますマッチするように変換する。デコードされた曲線は右側に示されている

3.5 個々の曲線周辺のサンプリング

TONA 3M 先物のリターン曲線を 2 つ示している。1 つは主成分空間にあり、もう 1 つは VAE 潜在空間である。もし極端な点を見つけ、それを主成分空間あるいは VAE 潜在空間のガウス分布ノイズを介して摂動させることができれば、それらのヒストリカルシナリオの異なるバージョンを生成することができる。この場合、レベル 1 の VAE のみを使用する。PCA 潜在空間では、シナリオを選択し、その周りを主成分と同じ共分散行列で標本化することができる。

図 3.10 に示すように、主成分と同じ共分散を持つガウス分布に従って特定のシナリオの周りの主成分空間でサンプリングする場合、平均値と傾きを記述する最初の 2 つの固有ベクトルが最も強い影響を与える。これにより、サンプリングされたシナリオは、ほとんどがソースシナリオと同様であるが、平行移動および小さな傾斜変化を伴う。しかし、VAE 潜在空間でサンプリングしながら、潜在空間共分散行列に従ってサンプリングした。これにより、非線形にサンプリングし、所与の曲線の周りのより広い範囲の曲線をサンプリングすることができる。グローバル共分散行列を使用する代わりに、VAE ではデータをサンプリングするためにローカル共分散を使用することを可能にし、グローバル構造のみを捕捉する第 1 主成分および第 2 主成分に全ての比重をおくのではなく、必要な場合には異なるデータ特性により多くの重みを与えている。

図 3.10 の第 3 列は、データ点が主に第 1 主成分または第 2 主成分によってあらわされる場合、VAE ガウス分布摂動では、それぞれ第 1 主成分または第 2 主成分に沿ってより大きな摂動をもたらすことを見て取ることができる。一方の主成分空間では、主に第 2 主成分によってあらわされるサンプルでさえ、第 1 主成分が主成分共分散行列のガウス分布摂動での主たるものとなっている。

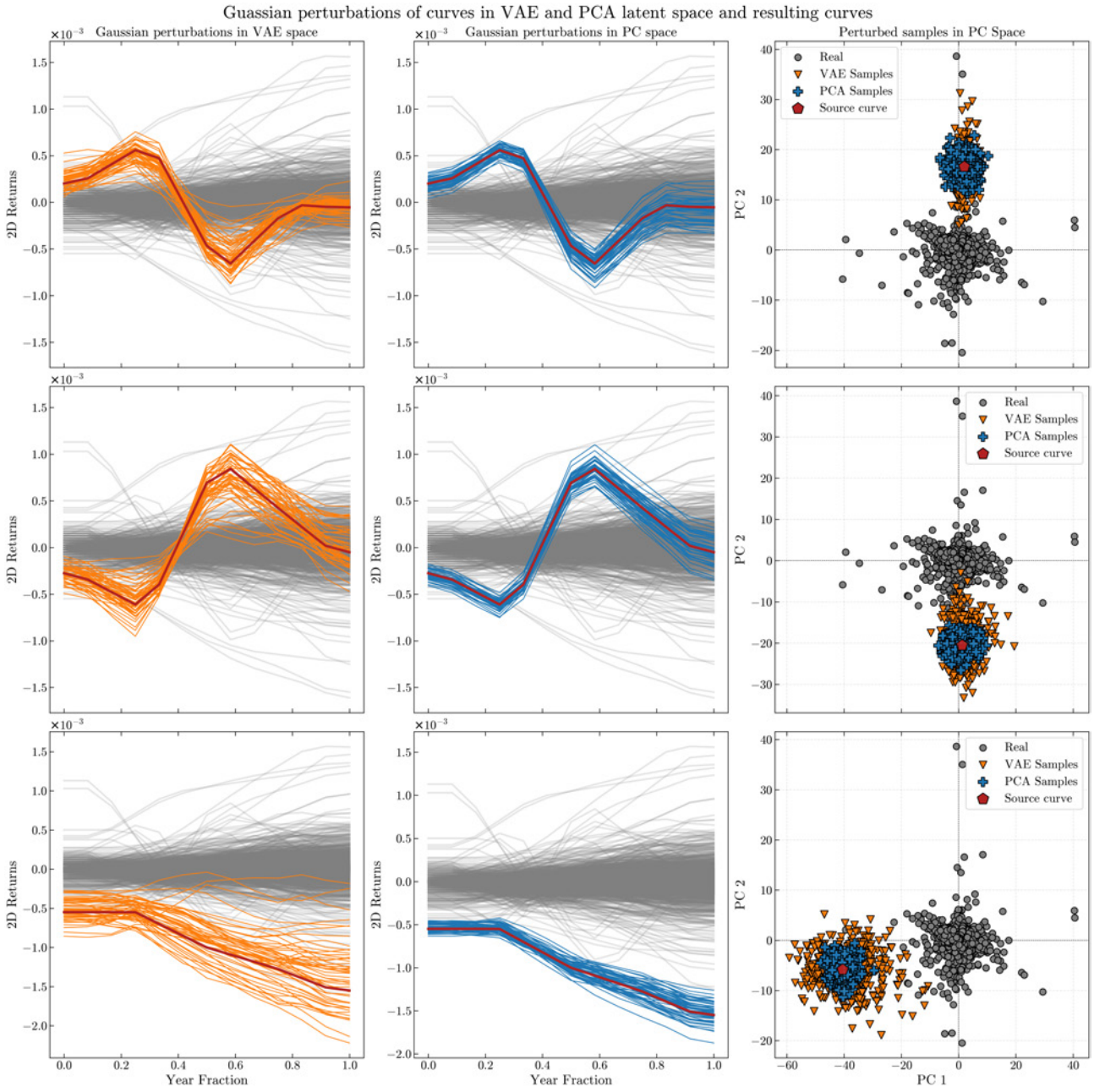


図 3.10: VAE 空間対主成分空間のガウス分布摂動の例。すべての摂動サンプルは、比較のために主成分空間に示される

3.6 極値のサンプリング

極端なシナリオをサンプリングするために、分布のテール部分を見てそこからサンプルをとって見てみよう。手順は以下の通りである:

1. そのデータを、VAE と PCA それぞれの潜在次元における各象限に分割する
2. 各象限について、各象限のなかで正規分布の上位 20 パーセンタイルを選択し、外れ値を選択する
3. 新しいサンプルを生成するために、様々なレベルのガウスノイズを有する様々な反復においてこれらの外れ値を突然変異させる
4. すべての突然変異点について、実際のデータの中で最も近い隣接点を探す
5. 3 の反復ごとに、実際のデータポイントに対して最も近い隣接距離を持つ突然変異ポイントを選択する

この手順は、VAE 潜在空間および PCA 潜在空間でも実施することができる。

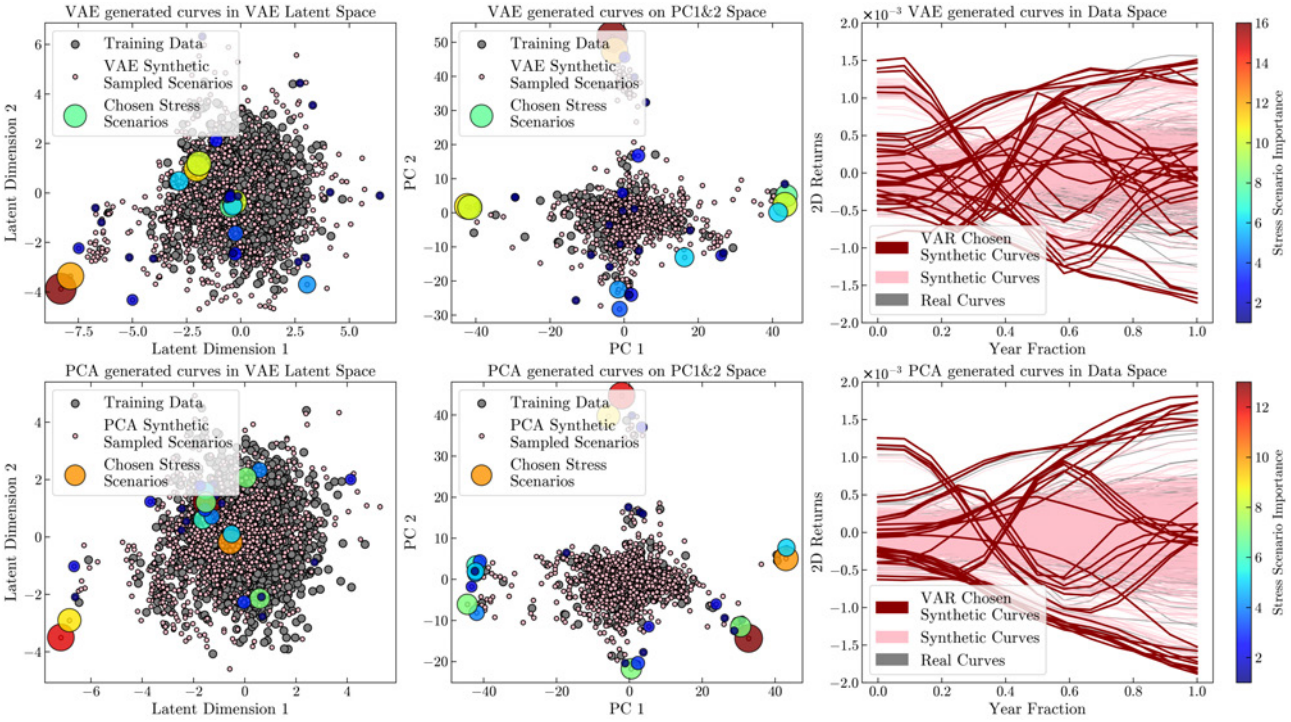


図 3.11: 上段は、VAE 潜在空間における最も遠いポイントのサンプリングによって生成されたカーブを示している。下段には、主成分空間で最も遠いポイントサンプリングを行うことで同じデータが表示される。右端は、セクション 4 で述べるさまざまなポートフォリオの VaR 計算で選択されたカーブも示している。主成分空間サンプリングの場合、PC1 および 2 の極値のみがほとんど選択されるが、VAE 生成曲線の場合、選択された曲線は、青色のシナリオの上部中央の図によって示されるように、必ずしも PC1、PC2 の極値にある必要はないことがわかる。カラー・バーは「ストレス・シナリオの重要性」を示しており、次章で詳しく述べるように、VaR の計算において特定のシナリオがストレス・シナリオとして選択されたポートフォリオの数を数えている

4 VaR 計算における仮想シナリオ

VAE と PCA の合成シナリオを実際にどのように用いるかを示し、リスク・カバレッジをどのように広めていくかをさらに探求するために、ここでは VaR の計算にそれらを統合することで見ていく。セクション 3.1 と同様に、TONA 3M 先物に分析の範囲を限定する。また、分析の範囲をさらに限定するために、満期に近い 4 つの限月に焦点を当てる。

TONA 3M 先物の異なる組み合わせを用いて仮想ポートフォリオを生成することによって、簡単なテストを構築することができる。計算の範囲は、4 つの特定の商品に限定されるので、ポートフォリオ P_N を、各商品の限月ごとのネット・ポジションのベクトルとしてそれぞれ定義することができる:

$$P_N = (I_{N1}, I_{N2}, I_{N3}, I_{N4}) \tag{2}$$

ここで、I_{N1}, I_{N2}, I_{N3}, I_{N4} は、それぞれ、第 1、第 2、第 3、および第 4 の最も満期に近い商品のネット・ポジション量である。さらに、ネット・ポジション量をネット・ロングを正とネット・ショートを負と定義する。一例として、20250217 に最も満期の近い 4 つの TONA 3M 先物の満期は、それぞれ 202503、202506、202509、および 202512 である。次に、2 つのロング 202503 ポジションと 1 つのショート 202509 ポジションを持つ仮想のサンプル・ポートフォリオを、次のように定義することができる。$P_X = (2, 0, 0, -1)$ 。本稿では、すべて限月において-1、0、1 のネット・ポジションの組み合わせに対する仮想ポートフォリオを作成し、空のポートフォリオを除外すると 80 $(3^4 - 1 = 80)$ のポートフォリオの組み合わせが得られる。

$$\begin{aligned} P_1 &= (-1, -1, -1, -1) \\ P_2 &= (-1, -1, -1, \ \ 0) \\ P_3 &= (-1, -1, \ \ 0, -1) \\ &\cdots \\ P_{80} &= (\ \ 1, \ \ 1, \ \ 1, \ \ 1) \end{aligned} \tag{3}$$

また、サンプル計算日として 20250217 を使用し、JSCC のホームページに掲載されているものと同じ国債先物清算資格での計算パラメータを使って計算する [20] 。比較のために、(1) JSCC の公式ストレスシナリオ、(2) JSCC の公式ストレスシナリオ + VAE の合成シナリオ、(3) JSCC の公式ストレスシナリオ + PCA の合成シナリオの 3 種類のストレスシナリオを用いて VaR を計算する。この結果は、我々のヒストリカル VaR 計算テストのための以下の計算入力要約に基づいている:

表 1: VaR の計算範囲と分析パラメータ

パラメータ	詳細
計算日	20250217
日中証拠金の計算方法	VaR (期待ショートフォール)
期待ショートフォールストレス損失数	2
期待ショートフォール信頼区間	97.5%
ヒストリカル・シナリオ	JSCC のウェブサイトで公表された公式シナリオ (1250 シナリオ)。
ストレスシナリオ	(1) 公式 JSCC シナリオ。 (2) 公式 JSCC シナリオ + PCA 合成シナリオ。 (3) 公式 JSCC シナリオ + VAE 合成シナリオ。
対象ポートフォリオ	TONA 3M 先物の 202503、202506、202509 および 202512 限月のすべての-1、0、1 の正味量の組み合わせは、 (0、0、0、0 を除くと) 合計 80 の組み合わせ

4.1 VaR 算出結果

仮想 VAE シナリオと PCA シナリオをストレスシナリオの集合に追加することによる $\frac{\mathrm{VaR}_{\mathrm{Syntethic}}}{\mathrm{VaR}_{\mathrm{Base}}} - 1$ で表される相対的なインパクトが 図 4.1a 見られる。予想されるように、ストレスシナリオを追加すると、以前よりも高いポートフォリオ・ロスを持つシナリオを見つけるための追加的な機会が与えられるため、VaR の数は常に同等以上になる。要約すると、一部のポートフォリオでは VaR への影響がなく、ほとんどのポートフォリオでは VaR への影響は 5% 未満であり、少数のポートフォリオではそれよりもさらに大きい影響があった。また、VAE シナリオは、PCA と比較して、わずかに高い VaR 影響をもたらす傾向があることもわかる。

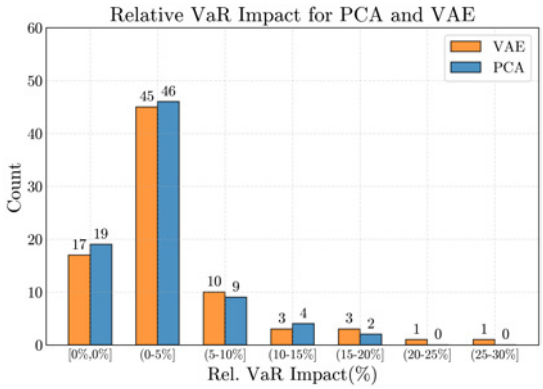
VAE と PCA の VaR の影響が最も大きい上位 5 ポートフォリオを表 2 、表 3 に示す。すべてのポジションのネット建玉 ($\sum_{i=1}^{4} I_{\mathrm{Ni}}$) (下表の Net Qty) が 0 に近いポートフォリオが、VAE と PCA の両方に対して最大の VaR 影響を持っているように見えることを示している。これは、合成シナリオが当初のシナリオでは持たなかった新しいカレンダースプレッドシナリオを含んでいることを示している。この傾向は、図 4.1b にも見られる。VaR の影響をポートフォリオのネット建玉の関数として表している。

表 2: VAE で生成されたシナリオの VaR の最大影響ポートフォリオ上位 5 位。シナリオ ID はセクション 4.2で定義され、ポートフォリオ選択頻度に基づいている。シナリオは本文中で、*vae01, vae02* または *pca01, pca02* などとして参照している

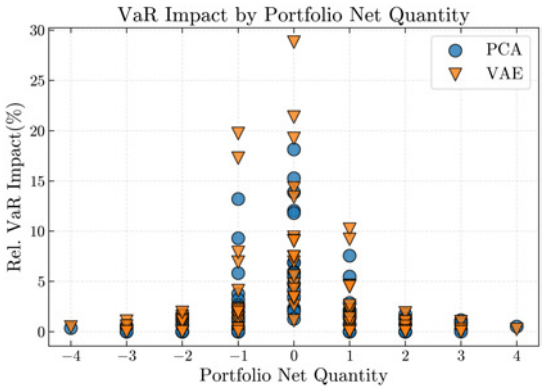
I_{N1}	I_{N2}	I_{N3}	I_{N4}	**Net Qty.**	**VAE vs. Original Diff.**	**PCA vs. Original Diff.**	**VAE vs. PCA Diff.**	**VAE Selected Scenarios (*vae–*)**	**PCA Selected Scenarios (*pca–*)**
-1	1	1	-1	0	28.8%	18.1%	9.0%	01, 02	02, 15
-1	0	1	0	0	21.3%	15.3%	5.3%	01, 02	02, 04
-1	-1	1	0	-1	19.7%	13.2%	5.8%	01, 02	02, 04
-1	1	0	0	0	19.2%	13.9%	4.7%	01, 02	02, 04
-1	0	1	-1	-1	17.3%	9.3%	7.3%	01, 02	02, 15

表 3: PCA で生成されたシナリオの VaR の最大影響ポートフォリオ上位 5 位。シナリオ ID はセクション 4.2で定義され、ポートフォリオ選択頻度に基づいている

I_{N1}	I_{N2}	I_{N3}	I_{N4}	**Net Qty.**	**VAE vs. Original Diff.**	**PCA vs. Original Diff.**	**VAE vs. PCA Diff.**	**VAE Selected Scenarios**	**PCA Selected Scenarios**
-1	1	1	-1	0	28.8%	18.1%	9.0%	01, 02	02, 15
-1	0	1	0	0	21.3%	15.3%	5.3%	01, 02	02, 04
-1	1	0	0	0	19.2%	13.9%	4.7%	01, 02	02, 04
-1	-1	1	0	-1	19.7%	13.2%	5.8%	01, 02	02, 04
0	1	-1	0	0	14.3%	12.0%	2.0%	12, 25	14, 05



(a) 仮想ストレスシナリオとして、VAE と PCA の合成シナリオを加えた後の相対 VaR の影響



(b) VAE と PCA の両方の計算におけるポートフォリオの Net Quantity と VaR の影響の比較

図 4.1: VAE/PCA の合成シナリオを仮想的なストレスシナリオとして追加した後の VaR の相対的な変化は、図 4.1a に示されている。VAE および PCA 計算におけるポートフォリオ純数量と VaR インパクトの比較は、図 4.1b に示されている

4.2 VaR の選択シナリオ

次に、合成 PCA シナリオと VAE シナリオの設定から ES-VaR によって選択されたシナリオの種類を分析する。2 つの方法によって生成された約 500 のシナリオのうち、31 の VAE シナリオおよび 29 の PCA シナリオがそれぞれ選択された。図 4.2 に示すように、いくつかのシナリオも複数のポートフォリオから選択されている。さらに議論を簡単にするために、同じ図に基づいて仮想シナリオ ID を導入する。図中の *vae01* シナリオは 16 のポートフォリオによって選択され、*pca01* は 13 のポートフォリオによって選択された。全体として、VAE シナリオは 116 回選択され、PCA シナリオは 114 回選択された。この数は、選択されたストレスシナリオの理論上の最大数と比較すると比較的高い。この最大値は、表 1 に記載される計算構成から $\text{portfolios} \times \text{stress losses} = 80 \times 2 = 160$ ストレスシナリオであることを前提としている。

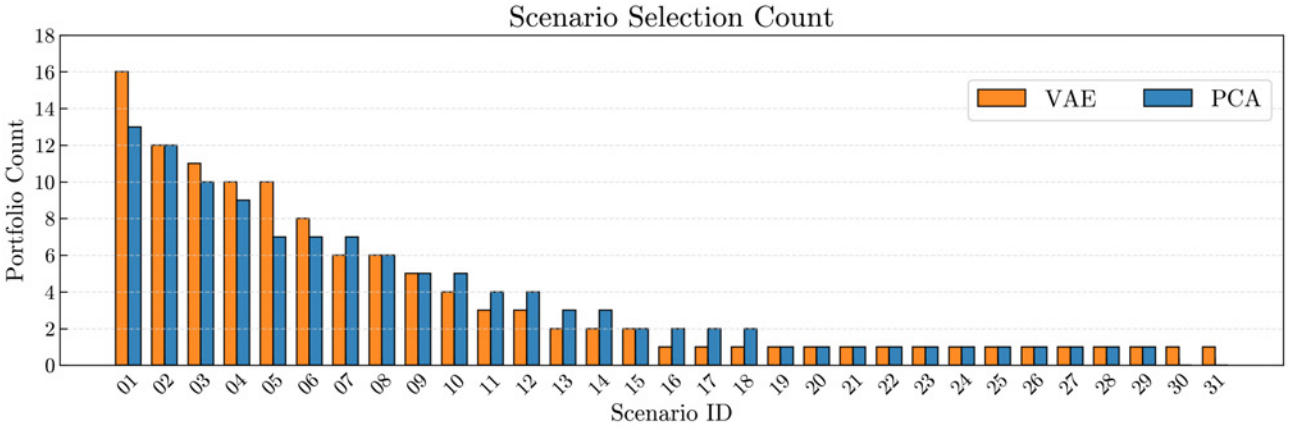
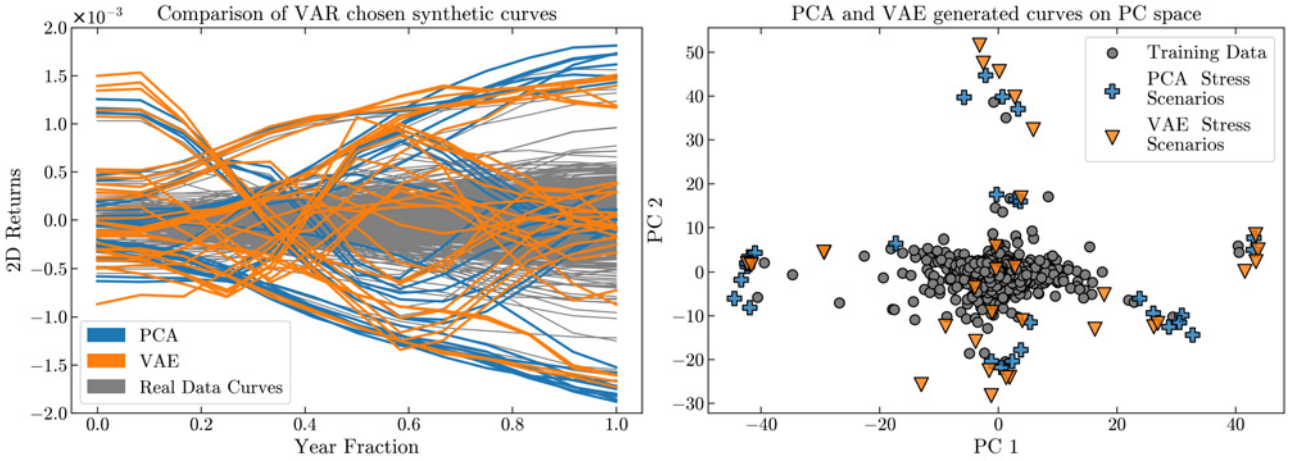


図 4.2: シナリオ別シナリオ選択回数 *vae01* シナリオは 16 のポートフォリオによって選択され、*pca01* シナリオは 13 のポートフォリオによって選択された

図 4.3a は、さらに、期待ショートフォールの計算においてポートフォリオによって選択されたすべての仮想 VAE カーブと PCA カーブの概要を示している。一般に、PCA シナリオはより滑らかである傾向があり、一方、VAE シナリオはより多くの曲線形状変動を有する。同じシナリオは、図 4.3b でも見ることができ、ここでは、それらを第 1 および第 2 の PCA 構成要素上に投影する。注目すべきことは、ES 選択された PCA シナリオは、第 1 および第 2 の PCA 構成要素分布のテールに大部分が見出され、その分布はデータにおける主な変化部分になっている。逆に、VAE シナリオのいくつかは、分布の中央部分に見出され、VAE サンプリングは PCA の第 1 主成分や第 2 主成分では説明できなかったより多くの微妙な変化も含むストレスシナリオも導入していることを示す。



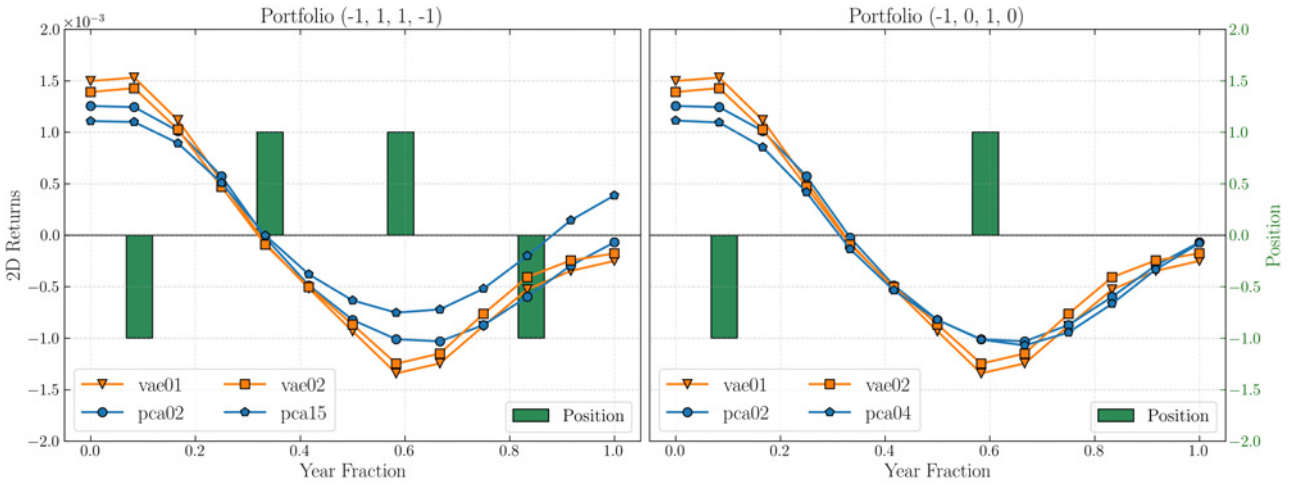
(a) 仮想ポートフォリオ計算のために ES が選択した VAE シナリオと PCA シナリオ

(b) 選択された VAE および PCA 生成曲線の第 1・第 2 主成分空間への射影

図 4.3: データ空間および PCA 空間での仮想ポートフォリオ計算において、ES により選択された VAE および PCA シナリオ

4.3 ポートフォリオ固有の分析

特定のシナリオが選択された理由をよりよく理解するためには、それを特定のポートフォリオの内容と比較する必要がある。表 2 、表 3 では、VAE と PCA の 2 つの最大の VaR 影響は、(-1、1、1、-1) ポートフォリオと (-1、0、1、0) ポートフォリオであることがわかった。それぞれのポートフォリオとシナリオを図 4.4a と図 4.4b に示す。



(a) 最大の VaR インパクトの (-1、1、1、-1) ポートフォリオ

(b) 2 番目に大きな VaR インパクトの (-1、0、1、0) ポートフォリオ

図 4.4: 図 4.4a は (-1、1、1、-1) ポートフォリオ、図 4.4b は (-1、0、1、0) ポートフォリオにおける VAE および PCA シナリオ選択を示す。(-1、1、1、-1) ポートフォリオは、PCA と VAE の両方について、すべてのポートフォリオの VaR に最も大きな影響を与えた。(-1、0、1、0) ポートフォリオ (図 4.4b) は、PCA と VAE の両方に対して、すべてのポートフォリオの 2 番目に最大 VaR 影響を持っていた。ポジションと逆向きのシナリオリターンは損失となる

予想されるように、最も高いポートフォリオ・ロスをもたらすリターン・シナリオは、ポジション量として反対の符号をもつ。(-1、1、1、-1) ポートフォリオは、ショート・ロング・ロング・ショート・ポジションの組み合わせを持っているため、*vae01*、*vae02*、*pca02*、*pca15* などのアップ・ダウン・アップ・シナリオは、特に 1 番目と 3 番目のテナーで大きな損失を引き起こしている。図 4.4a では、VAE シナリオと PCA シナリオの形状は似ているが、VAE シナリオのリターンは、第 1、第 3 テナーでそれぞれ PCA より高いリターンと低いリターンになっていることがわかる。これは、表 2, 表 3 で観測された VaR のインパクトの違いを説明している。同様のことが図 4.4b でも見て取れる。ショートロングのポジションの組み合わせでは VAE と PCA の両方においてアップダウンシナリオが選定される結果となっている。

図 4.5 に示すように、(-1、1、-1、1) ポートフォリオでは、VAE と PCA のシナリオカーブの形状が異なることが分かった例があるが、このポートフォリオでは、VAE の VaR 影響は 5.5%、PCA の VaR 影響は 1.3% であり、VAE と PCA の差は 6 番目に大きい。同図に見られるように、PCA シナリオの結果、第 1 限月、第 4 限月の損失は第 3 限月の利益で大幅に相殺された。これは、ほとんどのテナーで損失をもたらした VAE シナリオと比較することができる。

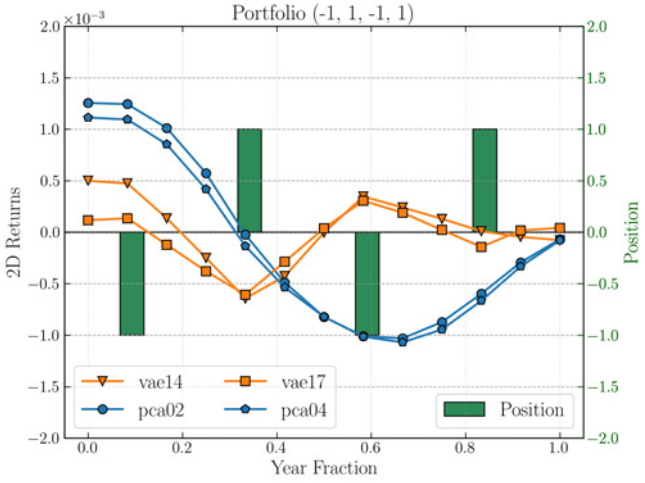


図 4.5: (-1、1、-1、1) ポートフォリオの VAE シナリオと PCA シナリオ。ポジションと逆向きのシナリオリターンは損失となる

4.4 合成シナリオとヒストリカルシナリオ比較

前節で議論した VaR の結果は、仮想ストレス・シナリオのプールに新しい合成シナリオを加えることによって得られたものである。理論的には、これらのシナリオは、実際のヒストリカルトレーニング・データ・シナリオと同一となるものであり、かつ、これらのシナリオがヒストリカル VaR ウィンドウ外であり、かつ、ヒストリカルストレス日として選択されていない場合には、VaR 影響を依然としてもたらす可能性がある。これは、

1. JSCC オリジナルシナリオ + 訓練データシナリオ
2. JSCC オリジナルシナリオ + 訓練データシナリオ + VAE 合成シナリオ
3. JSCC オリジナルシナリオ + 訓練データシナリオ + PCA 合成シナリオ

を使用した VaR の影響を比較する追加比較を行う動機となる。他のすべての計算パラメータは、表 1 で定義されているものと同じである。

図 4.6 に見られるように、VaR の影響は、これまでの比較ほど大きくはない。なぜなら、我々の参考計算は、以前よりも高いベースナンバーを生み出すからである。それにもかかわらず、VAE および PCA のいずれも、それぞれ 59 のポートフォリオが VaR インパクトがゼロでない状態にある。このことは、VAE と PCA の両方のサンプリング方法が、過去のトレーニングデータよりも高いポートフォリオ損失をもたらす新しいシナリオの組み合わせを生み出すことを示している。VAE と PCA の VaR の影響が最大上位 5 つのポートフォリオをそれぞれ表 4 、表 5 に示す。さらに、図 4.7 では、5 番目に大きな VAE での影響 (1、-1、-1、1) を持つポートフォリオについて、選択されたシナリオを見ることができる。

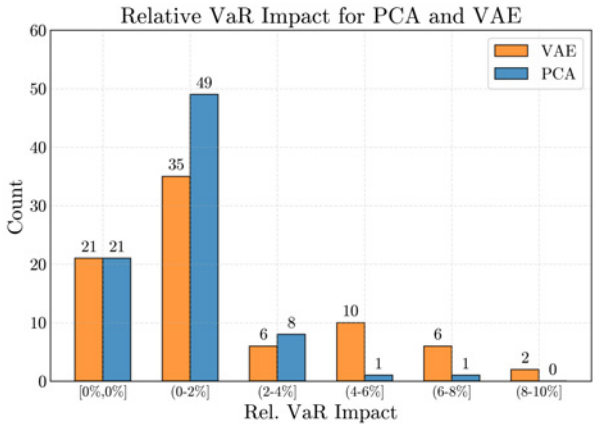


図 4.6: 元のストレス・シナリオとトレーニング・データ・シナリオに加えて、仮想ストレス・シナリオとして VAE と PCA の合成シナリオを追加した後の相対 VaR の影響

表 4: 元のストレスシナリオおよびトレーニングデータシナリオに VAE によって生成された合成シナリオを追加した場合の VaR の影響ポートフォリオ上位 5 位 N/A は合成シナリオ以外が選ばれたことを示す

I_{N1}	I_{N2}	I_{N3}	I_{N4}	**Net Qty.**	**VAE vs. Original Diff.**	**PCA vs. Original Diff.**	**VAE vs. PCA Diff.**	**VAE Selected Scenarios (*vae*–)**	**PCA Selected Scenarios (*pca*–)**
0	1	-1	0	0	8.6%	6.4%	2.0%	12, 25	14, 05
-1	1	0	0	0	8.4%	3.6%	4.7%	01, 02	02, 04
-1	1	1	-1	0	7.9%	0.0%	7.9%	01, 02	N/A, N/A
-1	1	1	-1	0	7.1%	1.7%	5.3%	01, 02	02, 04
1	-1	-1	1	0	6.7%	2.5%	4.1%	09, 26	18,05

表 5: 元のストレスシナリオおよびトレーニングデータシナリオに PCA によって生成された合成シナリオを追加した場合の VaR の影響ポートフォリオ上位 5 位 N/A は合成シナリオ以外が選ばれたことを示す

I_{N1}	I_{N2}	I_{N3}	I_{N4}	**Net Qty.**	**VAE vs. Original Diff.**	**PCA vs. Original Diff.**	**VAE vs. PCA Diff.**	**VAE Selected Scenarios**	**PCA Selected Scenarios**
0	1	-1	0	0	8.6%	6.4%	2.0%	12, 25	14, 05
1	1	-1	0	1	6.6%	4.1%	2.4%	10, 12	05, 14
-1	0	0	1	0	4.6%	3.6%	0.9%	01, N/A	N/A, 02
-1	1	0	0	0	8.4%	3.5%	4.7%	01, 02	02, 04
-1	-1	0	1	-1	4.2%	3.2%	1.0%	01, 02	02, N/A

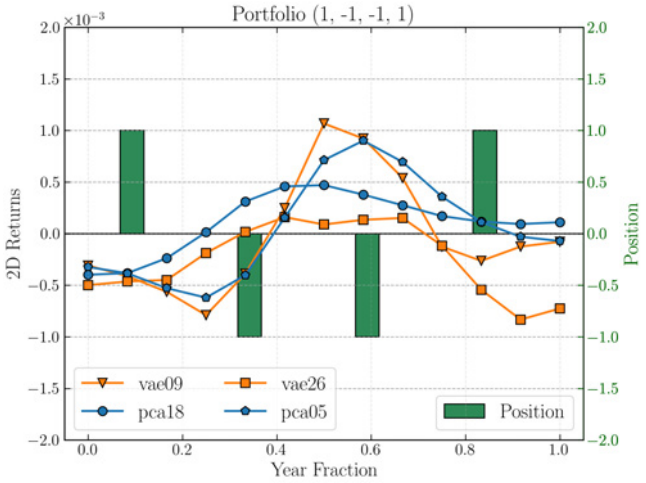
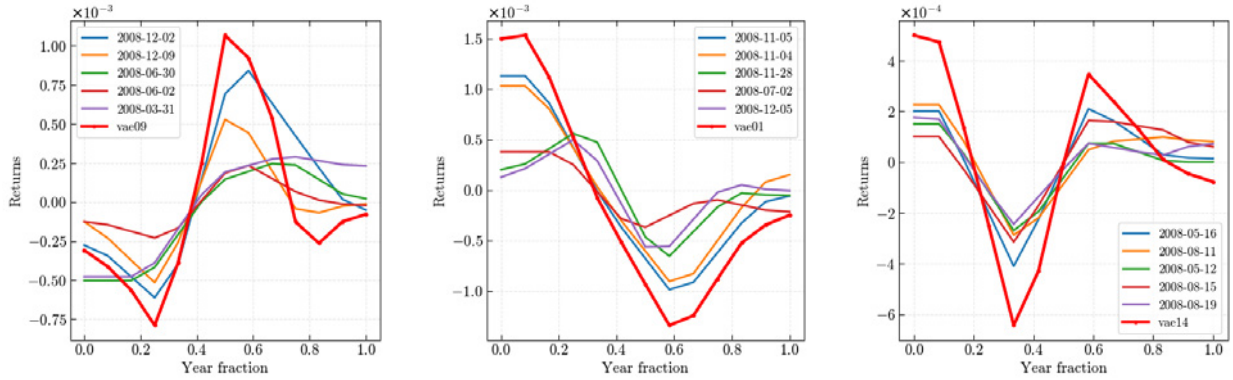


図 4.7: (1、-1、-1、1) ポートフォリオの VAE と PCA シナリオを、それぞれ 6.7% と 2.5% の VaR の影響をもって選択した。VAE は、PCA よりも 4.1% 高い VaR 結果をもたらす

図 4.8a では、選択された VAE 合成シナリオを、潜在空間における最も近い周囲の実データシナリオと比較している。図に見られるように、合成シナリオは、高いほう低いほうにそれぞれ拡大し、その結果、それは、その実データの対応物よりも選択されることになる。前に論じた *vae01* と *vae14* のシナリオでは、図 4.8b と図 4.8c に同じタイプの比較を見ることができる。ここでは、同様の挙動を観察することができる。3 つの図すべてに見られるように、選択されたすべての合成シナリオは、2008 年のヒストリカルシナリオに近いが、特定のポートフォリオでは損失がさら

に大きくなるような変動を伴っていた。



(a) 合成 *vae09* サンプル　(b) 合成 *vae01* サンプル　(c) 合成 *vae14* サンプル

図 4.8: 選択された合成シナリオ（太い赤線）と、最も近いトレーニングデータの近傍（細い色付きの線）の比較

4.5 VaR インパクト・サマリー

本章では、VAE シナリオと PCA シナリオの導入が、一連の仮想ポートフォリオの期待ショートフォール計算にどのような影響を及ぼすかを検討した。両手法とも、VaR が 0% から 30% の範囲で影響を与える新しいシナリオを見つけることができたことを示した。さらに、特定のシナリオが選択された理由をポートフォリオ・レベルで議論し、選択された PCA シナリオと VAE シナリオの間でいくつかの特性を比較した。選択された VAE のシナリオは PCA での第 1 主成分や第 2 主成分では説明できなかったより多くの微妙な変化を補足できることが分かった。要約すると、VAE 合成シナリオは、より極端な高低と組み合わせをもつより多くの曲線形状変動シナリオを有するため、PCA よりも高い VaR 影響をもたらした。

5 考察

実世界データは、しばしば非常に複雑であり、非線形関係を有する。通常、正規分布やその他の標準分布に従うことはない。より単純なモデルを用いて現実世界のデータをモデル化しようとすると、現実的でない仮定を課すことになる。金融でのデータは、典型的には、ガウス分布ではなく非線形である。相関関係は、状況に応じて強まったり弱まったりする。生成 AI の方法では、古典的な統計的手方法をこえて、データに単純化された仮定を課すことなく、データから直接的に学習する。ニューラルネットワークの柔軟性のために、それらは変数間の複雑な依存性を捕捉することができる。それらは、なめらかなデータ分布を学習することができ、合成ではあるが現実的な見た目をもつデータを生成するのに役立つ。

本稿では、TONA 3M 先物の 2 日間のリターン曲線を見て、PCA と VAE の両方を使用した合成データの生成を検討した。多変量ガウス分布を用いて PCA 空間をサンプリングすることは、ガウス分布が実データに存在するテール確率を過小評価すると予想されるデータにおいてテールを捕捉することに失敗したケースを確認できた。また、PCA では汎用で使えるサンプリング方法論がない。特定のデータをサンプリングし、主成分共分散行列に従ったガウス分布摂動を追加すると、一般に平行移動の水準と傾きにとらえられる第 1 主成分と第 2 主成分が優勢になる。その共分散が成り立たない領域におけるグローバル共分散を用いて既存の極端なサンプルを摂動することによって新しいサンプルを生成することは、貧弱で実態に合わないデータサンプルをもたらす。

VAE はデータに非線形変換を行い、確率的な潜在空間を作成する。これらは設計による確率モデルであるため、生成モデリングに使用できる。VAE は、潜在空間に課せられたガウス分布にもとづいて、潜在空間におけるサンプリング

のための一般的な方式を提供する。潜在空間上のガウス分布は、ガウシアンに近づく方向に正則化する (ただし、必ずしもそうではない)。VAE 潜在空間におけるサンプリングとデコーディングバックにより、特徴間の非線形依存性も捕捉できる。第 1 主成分と第 2 主成分が優勢な主成分分析ベースのサンプリングとは異なり、VAE 潜在空間においてガウス分布摂動を行う場合、これらはデータ空間における非ガウス分布摂動ととらえられる。VAE では、データ内の他のモードやファクターが、第 1 主成分や第 2 主成分のみにほとんどのウェイトを設定するのではなく、必要に応じてより多くのウェイトを設定できる。これは、図 3.10 の右端の列でも同様に見ることができる。

VAE の潜在空間は、現実的な補間に特に有用である。図 3.6 に示すように、VAE 潜在空間におけるシナリオ間を線形補間する場合、補間軌跡はデータ空間線形補間とは異なり、データ分布の輪郭に密接に従う。さらに、潜在空間の多変量ガウス分布サンプルのみを用いて、データ分布の全部からサンプル化するために階層的 VAE を用いた。

また、セクション 3.6 で述べた戦略を用いて、データ分布における極端な事象をサンプリングする。我々は、VAE および主成分潜在表現の両方に対して同じ方法論を使用し、主成分ベースの極端な点サンプリングでは、低位の主成分次元によってあらわされる極端なシナリオを捕捉することができないことも見てきた。また、第 1 および第 2 の主成分が支配的でない領域であるにも関わらず、支配的になってしまっている不自然な状況になってしまっている。最も高いリターンが小さいが、特定のポートフォリオに大きな損失をもたらすような方法で分配されるシナリオを考えることができる。

続いて、セクション 4 では、PCA と VAE を導入することによって生成される仮想ストレスシナリオが、一連の仮想サンプル・ポートフォリオの期待ショートフォール計算にどのような影響を与えるかを検討した。生成されたシナリオを元のヒストリカルシナリオとストレスシナリオのセットに追加すると、PCA と VAE の両方に対する VaR の影響が 0%–30% の範囲であることがわかった。

セクション 4.4 では、同様の比較を行ったが、両方の計算にヒストリカルシナリオのトレーニングデータセットを含めた結果、VaR は 0%~10% の範囲で影響を受けた。どちらの方法も、当初のシナリオよりも高い損失をもたらす新しいシナリオを生み出すことができたので、リスク・カバレッジとリスク理解の向上に役立つと主張することができる。VaR に新規合成シナリオを追加する際には、計算結果への影響が大きすぎたり、望まぬ影響をおこすことを避けるために、慎重に行うべきである。VAE と PCA の結果を比較すると、VAE は、より多くの曲線形状のばらつきと、より極端な変動のために、より高い VaR 影響を有することがわかった。これらはどちらも、リスク・カバレッジを改善できる極端なシナリオを探す際に望ましい特性である。選択されたシナリオをさらに見ると、オリジナルのトレーニングデータのいくつかと同様の特性および形状を共有していることが分かる。これは設計によるものであるが、言及する価値がある。トレーニングデータとはまったく異なるシナリオを生成するために、導入されたサンプル方法論を期待すべきではない。また、本稿の範囲内で、われわれは、単一の計算日を用いて、少数の仮想ポートフォリオのみを検証した点にも言及する価値がある。われわれが使用している金融商品の満期までの時間とポートフォリオのポジション構成が最も高い損失を生み出すシナリオのタイプを決定する。追加のポートフォリオの組み合わせおよびより多くの計算日をテストすることは有益であろうが、本論文の対象外としている。当初の目標は、極端ではあるがもっともらしいシナリオをサンプリングする方法を見つけることであった。抽象的な金融データの場合、妥当性を定義することは困難である。写真、声、音楽などのデータの場合、技術の進化は、偽物から本物を識別し、不可解なものから妥当なものを識別する優れた能力を我々に提供している。しかし、金融データの場合の課題は、高度に訓練された専門家ではない人が見るだけでは、データが現実的であるか非現実的であるか、判断することが難しい。統計的尺度を使用して、生成されたデータが実際のデータと同様のデータ分布に従うかどうかを確認することしかできない。もしそれが同様の分布に従うなら、それは現実的であり、妥当であると仮定することができる。図 3.6 と図 3.9 では、我々の補間と合成サンプルがそれぞれ分布の輪郭に従い、データ分布の形状を保存している。さらに、これらのシナリオの多くは、期待ショートフォールによって、ポートフォリオの価値に重大な変更をもたらすのに十分なほど極端であることを示して選ばれていた。手短に言えば、生成 AI 法の使用、特に合成市場データの生成は、以下の点で我々を助けとなる

1. 複数のリスクファクター間の複雑な依存関係の共同モデル化
2. コピュラフィッティング、アドホックサンプリング技術のようなアドホック方法の必要性を取り除くか、または減らすこと

しかし、どのような技術でも、いくつかの欠点があり、注意すべき点がある。生成 AI は、複雑なモデルとアーキテクチャに基づいており、その背後に複雑な数学がある。それらは、2 つの意味でブラックボックスである:

1. それらのアーキテクチャ上の複雑さのために、意思決定と予測は説明が難しい
2. それらを支配する数学は容易には利用できない

これは、AI を強力にするが、同時に理解することを困難にし、ユーザが制限を完全に理解することを困難にし得る。もう 1 つのボトルネックは、データの欠如である。AI モデルは、訓練されるべき大量の高品質データを必要とする。さらに、最適なアーキテクチャや、ニューラルネットワークの学習に関わる様々なハイパーパラメータ、さらには学習時間を決定するための汎用的なアルゴリズムは存在しない。通常、それらの選定は、実験、直感、そして問題固有の知見を組み合わせて行われる。実際のところ、AI モデルの訓練には多くの試行錯誤が伴うのが実情である。

5.1 最後に

主成分分析の線形性は、実装および利用を非常に説明しやすく容易にする。しかしながら、主成分変換を使用してデータを生成することは、注意深いサンプリングを必要とし、これは、通常簡単ではない。さらに、主成分空間のまばらさや連続性の欠如は、生成モデルとしてのその使用を妨げる。より小さい次元では、これは専門家にとって大きな問題ではないかもしれないが、データの次元数が増加すると、我々の直感が誤りやすくなり、計算の複雑さが増す。要するに、主成分を手動で変化させることによる合成データ生成は、データを生成することを可能にするが、注意深い人間の専門家の監督を必要とする。一方、VAE は複雑な高次元データ分布を単純で連続な潜在分布に変換し、生成モデルとしての能力を保証する背後に健全な理論的枠組みを持つ。よりシンプルで低次元の潜在空間は、より良いサンプリングを可能にし、低次元性のため、取り扱いが容易である。両方のアプローチは、それらおのおのの利点を有し、理想的には、合成データを生成するために一緒に使用することができる。例えば、市場との相関が高いデータの第 1 主成分と第 2 主成分を除去し、VAE を用いてデータの残りの複雑な依存関係をモデル化することができる。また、第 1 主成分と第 2 主成分に基づいて残差主成分分布を予測するための条件付モデルを構築することもできる。さらに、両方のアプローチを一緒に使用することは、VAE または他のニューラルネットワークベースの生成モデルアプローチに対するより説明能力を提供する。

5.2 今後の方向性

本白書では、JSCC で合成金融データを生成するために生成 AI モデルを用いることの実現可能性を検討した。この研究のさらなる方向性は、データの時間依存を考慮したり、複数のリスクファクターをモデリングし、敵対的生成ネットワーク (GAN)、敵対的オートエンコーダまたは拡散モデルのようなより高度なモデルを使用することである [11], [14], [15], [17]。リスク管理の観点からも、時系列での予測モデル化は特に興味深い。それとは別に、生成 AI モデルは、データ内の異常なパターンにフラグを付けることができる異常検出にも使用することができる。

6　謝辞

本ワーキングー・ペーパーの執筆にあたり、多くの方々にご支援いただきました。はじめに、株式会社日本証券クリアリング機構の小沼泰之社長、三輪光雄執行役員、田村康彦執行役員には、清算機関のリスク分析における AI 活用という大変貴重な分野での執筆の機会と貴重なフィードバックをいただきました。厚く御礼申し上げます。テーマ選定や分析アプローチについては、取引所取引清算部の濱﨑圭一氏と藤本祐太氏にご協力をいただきました。深く感謝いたします。

7 参考文献

[1] A. Rehlon and N. Dan, “Central counterparties: What are they, why do they matter and how does the Bank supervise them?” Bank of England, Jun. 2013.

[2] “3-Month TONA Futures,” *Japan Exchange Group*. https://www.jpx.co.jp/english/derivatives/products/interest-rate/3m-tona-futures/01.html.

[3] L. Ouyang *et al.*, “Training language models to follow instructions with human feedback.” arXiv, Mar. 2022. doi: 10.48550/arXiv.2203.02155.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models.” arXiv, Apr. 2022. doi: 10.48550/arXiv.2112.10752.

[5] A. Ramesh *et al.*, “Zero-Shot Text-to-Image Generation.” arXiv, Feb. 2021. doi: 10.48550/arXiv.2102.12092.

[6] W. S. McCulloch and W. Pitts, “A logical calculus of the ideas immanent in nervous activity,” *The bulletin of mathematical biophysics*, vol. 5, pp. 115–133, 1943.

[7] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain.” *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[8] K. Hornik, M. Stinchcombe, and H. White, “Multilayer feedforward networks are universal approximators,” *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[9] K. Pearson, “Principal components analysis,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 6, no. 2, p. 559, 1901.

[10] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes.” arXiv, 2013. doi: 10.48550/arXiv.1312.6114.

[11] I. J. Goodfellow *et al.*, “Generative adversarial nets,” in *Advances in neural information processing systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2014.

[12] L. Dinh, J. Sohl-Dickstein, and S. Bengio, “Density estimation using Real NVP.” arXiv, Feb. 2017. doi: 10.48550/arXiv.1605.08803.

[13] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep Unsupervised Learning using Nonequilibrium Thermodynamics.” arXiv, Nov. 2015. doi: 10.48550/arXiv.1503.03585.

[14] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 6840–6851.

[15] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-Based Generative Modeling through Stochastic Differential Equations.” arXiv, Feb. 2021. doi: 10.48550/arXiv.2011.13456.

[16] “Midjourney,” *Midjourney*. https://www.midjourney.com/website.

[17] A. Plumerault, H. L. Borgne, and C. Hudelot, “AVAE: Adversarial variational auto encoder,” *CoRR*, vol. abs/2012.11551, 2020, Available: https://arxiv.org/abs/2012.11551

[18] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models.” arXiv, Apr. 2022. doi: 10.48550/arXiv.2112.10752.

[19] D. J. Rezende and S. Mohamed, “Variational Inference with Normalizing Flows,” *arXiv.org*. https://arxiv.org/abs/1505.05770v6, May 2015.

[20] Japan Exchange Group, “Archive of Weekday Files.” https://jscc-h.jpx.co.jp/jscc/listed-derivatives/weekday/VaRParameter_20250217_1600.csv.