

エグゼクティブサマリー

生成 AI を用いた仮想市場データシナリオの生成

動機 (Motivation): ヒストリカルデータのみでは、あらゆるシナリオを網羅できない可能性があります。合成的でありながら現実的なデータの活用により、より堅牢なリスク管理が可能となります。

目標 (Goal): TONA 3M 先物の動きを対象に、新規かつ現実的な合成市場シナリオを創出すること。

近年、生成 AI および AI は、さまざまな産業において革新的な変化をもたらしている。生成 AI とは、データ内のパターンを学習し、統計的特性が類似する新たなデータを生成する技術である。自然界や社会、さらには金融データにも、特定の要因がどのように相互作用し、変化するかを示すパターンが存在する。金融の文脈では、これは市場におけるさまざまな金融商品の共動きとして表れる。AI や機械学習アルゴリズムを設計し、これらのパターンを学習させることで、それに従うデータを生成することが可能である（図 1）。技術的な言葉では、データの確率密度関数（PDF）を学習し、そこからサンプリングを行うと表現される。この PDF の学習と、そこからのサンプリングは、いずれも容易ではない課題である。

ニューラルネットワークを用いたモデルは、現実のデータに存在する複雑なパターンや非線形性を捉える柔軟性を備えているため、非常に強力である。非線形性とは、構成要素の効果が総和が、全体としての効果と異なることを意味する。個々の効果は加算的ではない。数学的には、関数 f が $f(x+y) \neq f(x) + f(y)$ である場合、非線形であると定義される。構成要素間の非線形な相互作用は、単純な線形相互作用とは全く異なる相乗効果を生み出す可能性がある。現実世界のデータの多くは非線形であり、適切にモデル化するためには非線形手法が必要である。

本ホワイトペーパーでは、1 年以内に満期を迎える最初の 4 つの TONA 3M 先物契約の共動きをモデル化することに焦点を当てている。ニューラルネットワークを基盤とした生成モデル「変分オートエンコーダ（VAE）」を用いている。変分オートエンコーダは、複雑かつ高次元のデータを非線形に変換し、低次元の潜在空間に圧縮する。各データサンプルは、VAE の潜在空間において条件付きガウス分布にマッピングされる。この潜在空間はデータ空間よりも次元が低く、データサンプル同士が近接することで、滑らかで連続的な低次元表現が得られる。現実的な新しいサンプルは、任意の 2 つの実データサンプル間を補間するか、潜在空間の空いている領域からサンプリングすることで生成できる。TONA3M を VAE でモデル化した後、この VAE の潜在空間をサンプリングすることでシナリオを生成する。また、類似の手法を用いて古典的な主成分分析（PCA）を活用し、新たなシナリオを作成し比較している（図 2）。

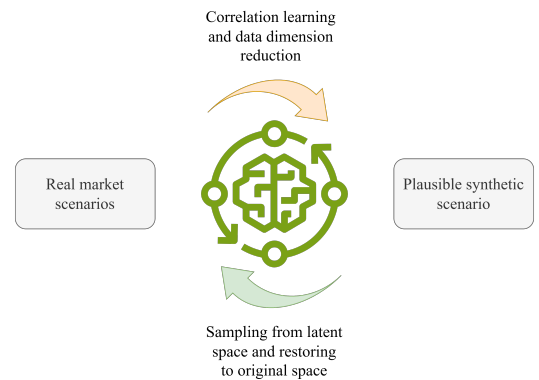


図 1: 市場シナリオにおける生成 AI

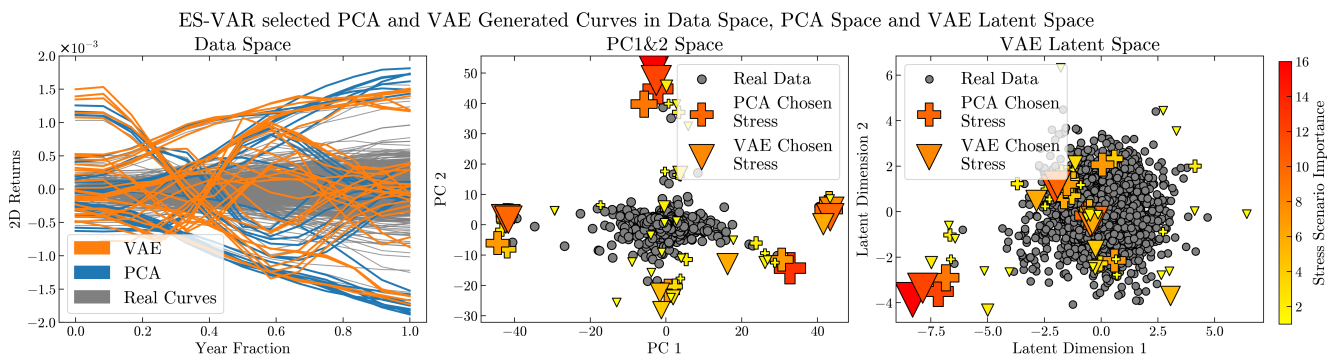


図 2: ES-VaR は、潜在空間における最遠点サンプリングを用いて、合成された VAE および PCA 生成曲線を選択した。すべての曲線は、中央と右側の図でそれぞれ PCA 潜在空間と VAE 潜在空間において表現されている。ヒートマップは、ES-VaR 計算におけるポートフォリオ固有のストレスシナリオ選択頻度を示している。

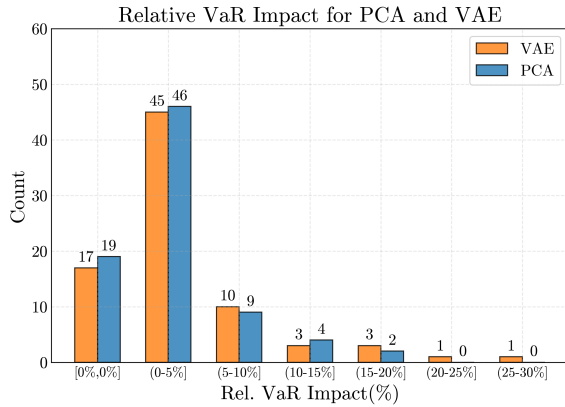
ES-VaR（期待ショートフォール）における合成シナリオの影響を検証するため、TONA 3M 先物の異なる組み合わせ

せを用いて仮想的なポートフォリオを生成し、簡単なテストを構築する。計算対象が4つの特定の銘柄に限定されているため、ポートフォリオ P_N は、それぞれの銘柄の満期に応じた数量のベクトルとして定義できる。

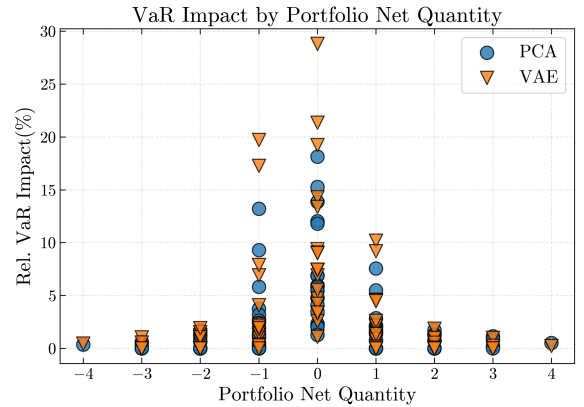
$$P_N = (I_{N1}, I_{N2}, I_{N3}, I_{N4}) \quad (1)$$

ここで、 I_{N1} 、 I_{N2} 、 I_{N3} 、 I_{N4} は、それぞれ1番目から4番目までの満期に近い銘柄のネットポジション数量を表す。本稿では、すべての銘柄について -1 、 0 、 1 のポジションを設定し、空の $(0, 0, 0, 0)$ ポートフォリオを除外した結果、 $3^4 - 1 = 80$ 通りのポートフォリオ組み合わせを作成した。

VAE および PCA のいずれも、図3に示されているように、VaR 影響が0%から30%の範囲に収まる新たなシナリオを見出すことができた。さらに、両手法の計算において、過去のシナリオを含む学習データセットも加えて比較を行った結果、VaR 影響は0%から10%の範囲に収まった。

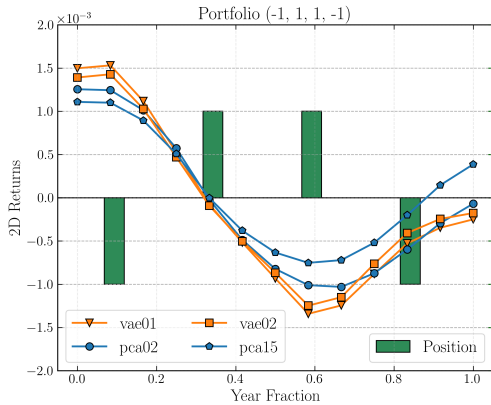


(a) 全ポートフォリオにおける相対的 VaR インパクトの概要

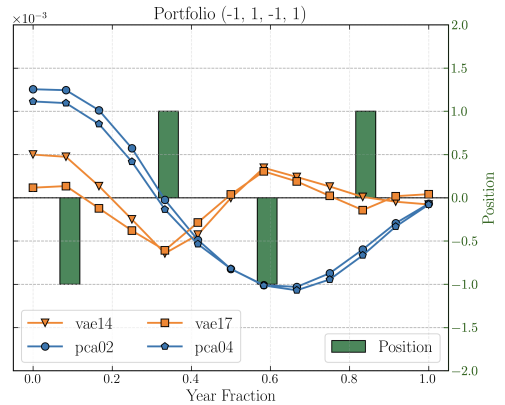


(b) ポートフォリオのネットポジション別の相対的 VaR インパクト

図3: 図3aは、VAE および PCA の合成シナリオを仮想ストレスシナリオとして追加した後のES-VaRの相対変化を示している。図3bでは、VAE および PCA による計算におけるポートフォリオのネット数量と VaR 影響の比較が示されている。



(a) 最大 VaR インパクトを持つポートフォリオ $(-1, 1, 1, -1)$



(b) 異なるシナリオ形状を持つポートフォリオ $(-1, 1, -1, 1)$

図4: ネット数量が0の2つのポートフォリオにおける、ES-VaRが選択した仮想シナリオが示されている。ポートフォリオ $(-1, 1, 1, -1)$ は、VAE および PCA のいずれの生成シナリオにおいても、すべてのポートフォリオ中で最大の VaR 影響を示し、それぞれ28.8%および18.1%であった。図4bは、ES-VaRが選択したVAE および PCA 生成シナリオが異なる形状を持つ事例を示している。このポートフォリオにおける VaR 影響は、VAE 生成シナリオが5.5%、PCA 生成シナリオが1.3%であった。

VAE と PCA の結果を比較したところ、VAE はより極端な高値や安値を示す (図4a) とともに、より多様な曲線形状のバリエーションを持つ (図4b) ことから、VaR 影響が大きくなる傾向が見られた (図3a)。これらの特性は、リスクカバレッジを向上させるための極端なシナリオ探索において望ましいものである。VAE は、データ分布の局所的な共分散構造を考慮して新しいデータサンプルを生成できるため、PCA が捉えられないデータ分布の微細なニュアンスを反映し、より多様な合成データを得ることができる。一方、PCA は大域的な共分散構造しか見ておらず、局所的な特性を捉えられない。選択されたシナリオをさらに詳しく見ると、図5に示されているように、一部は学習データに似た特徴や形状

を共有していることが確認できる。これは手法の特性上当然のことであり、学習データと完全に異なるシナリオを生成するものではないことは留意すべきである。ただし、どちらの手法も元のシナリオ以上の損失をもたらす新しいシナリオを生成できており、リスクカバレッジやリスク理解の向上に寄与すると言える。とはいえ、VaR に新たな合成シナリオを加える際には、計算結果に過度な影響を与えないよう慎重に行う必要がある。また、本稿の範囲では、単一の計算日を用いて仮想的なポートフォリオの小規模なセットをテストしている点にも言及すべきである。

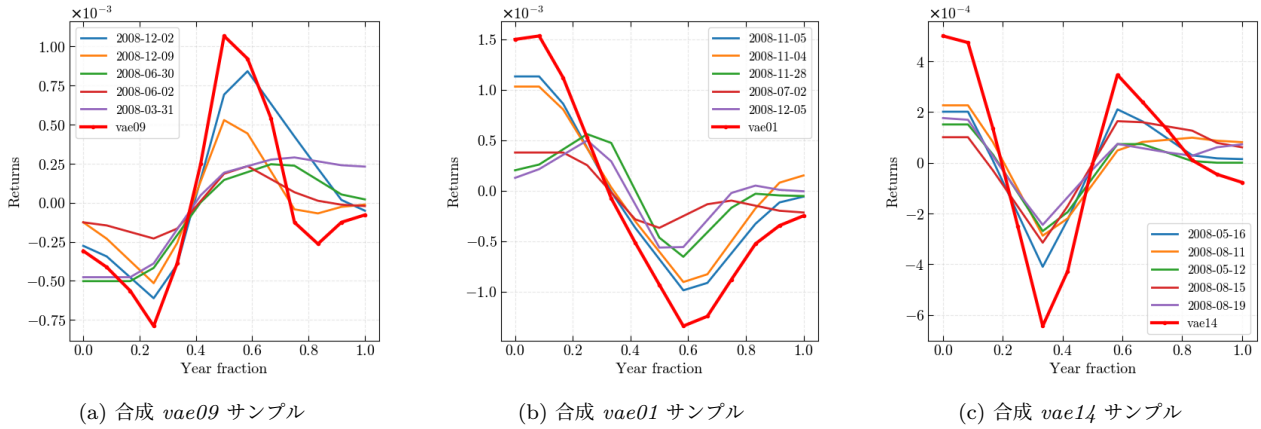


図 5: 選択された合成 VAE シナリオ（太い赤線）と最も近い訓練データの比較（細い色付き線）

当初の目的は、極端でありながらも現実的なシナリオをサンプリングする手法を見出すことであった。抽象的な金融データにおいて、現実的かどうかを定義することは困難である。金融データの特長上、専門家でない限り、データを目視で確認しただけで現実的かどうかを判断するのは難しい。そのため、統計的指標に基づいて、生成データが実際のデータと類似した分布を持っているかどうかを確認するしかない。生成データは、実データの分布の輪郭に沿い、曲線も類似した形状を持つことが確認できる。さらに、多くのシナリオが ES-VaR によって選択されており、これらのシナリオがポートフォリオの価値に大きな変化をもたらすほど極端であることを示している。

ただし、どの技術にも注意すべき点や欠点は存在する。生成 AI は複雑なモデルやアーキテクチャに基づいており、その背後には複雑な数学が存在する。これは、次の二つの意味でブラックボックスである。

1. アーキテクチャが非常に複雑であるため、意思決定や予測の過程を説明することが難しい点。
2. それらを支配する数学は、一般に理解しやすいものではない点。

これにより、AI モデルは非常に強力である一方、その仕組みを理解することが難しくなり、利用者が限界を十分に把握できなくなる可能性がある。さらに、導入のボトルネックの一つは、十分なデータが存在しないことである。AI モデルは、学習のために大量の高品質なデータを必要とする。加えて、ニューラルネットワークを訓練する際に、最適なアーキテクチャや各種ハイパーパラメータ、さらには学習期間を決定するための万能なアルゴリズムは存在しない。多くの場合、試行錯誤、直感、問題特有の知見を組み合わせに進められている。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ことを申し添えておきたい。

PCA の線形性は、その仕組みを説明しやすく、実装や利用が容易であることを意味している。ただし、PCA 変換を用いてデータを生成する際には、慎重なサンプリングが求められ、これは一般に容易な作業ではない。主成分を手動で変化させることで合成データを生成することは可能であるが、専門家による慎重な監督が必要となる。一方、VAE は、複雑かつ高次元のデータ分布を単純で連続的な潜在分布に変換し、生成モデルとしての理論的な枠組みを備えている。より単純で低次元の潜在空間は、サンプリングが容易になり、次元が低いために計算処理も扱いやすくなる。

結論として、PCA と VAE に基づく生成的アプローチはそれぞれ独自の利点を持ち、理想的には併用することで合成データを生成できる。例えば、市場と強く相関していることが多い第一および第二主成分を除去し、残りの複雑な依存関係を VAE を用いてモデル化することができる。さらに、第一および第二主成分をもとに残差主成分の分布を予測する条件付きモデルを構築することも可能である。これらの手法を組み合わせることで、VAE やその他のニューラルネットワークを基盤とする生成モデルの説明性を高めることができる。