УДК 519.25 DOI: 10.14529/cmse250201

УЛУЧШЕНИЕ ОЦЕНКИ СТОИМОСТНОЙ МЕРЫ РИСКА МНОГОМЕРНЫХ ПОРТФЕЛЕЙ С ПОМОЩЬЮ СМЕСИ ВЕРОЯТНОСТНЫХ АНАЛИЗАТОРОВ ГЛАВНЫХ КОМПОНЕНТ

 Н.В. Волков

Московский физико-технический институт

(141701 Московская область, Долгопрудный, Институтский пер., д. 9)

E-mail: nikita.v.volkov@phystech.edu

Поступила в редакцию: 23.03.2025

В работе предлагается новый подход для оценки стоимостной меры риска (VaR) многомерных портфелей, основанный на смеси вероятностных анализаторов главных компонент (mPPCA) и информационном критерии Акаике. Проверяется эффективность рассматриваемого подхода на основе исторических данных с учетом различного количества компонент смесей в методе mPPCA. Исследование проводится на 100 сильно и 100 слабо диверсифицированных портфелях акций индекса S&P 500 за период 2009–2023 гг., используя скользящие окна размером 350 торговых дней. Вероятностный метод главных компонент (PPCA) позволяет моделировать сложные зависимости между активами и учитывать «тяжелые» хвосты распределений. Благодаря этому метод mPPCA превосходит классический метод главных компонент (PCA) в точности оценки. Помимо этого, за счет понижения размерности модель оказывается вычислительно существенно легче и стабильнее, чем смесь гауссовских распределений (GMM). В работе показывается зависимость волатильности и «тяжести» хвостов распределений лог-приростов стоимости портфеля как от оптимального количества компонент в методе mPPCA, так и от минимального достаточного количества основных компонент в методах PCA и PPCA для объяснения 80 % дисперсии в данных. Новый подход с оптимизацией количества компонент методом mPPCA показывает более высокие результаты, чем подходы с методами GMM, PCA и PPCA, особенно на слабо диверсифицированных портфелях. В работе описаны подходы по оптимизации обучения метода mPPCA и проведена обширная оценка эффективности на основе исторических данных (бэктестинг). Использование JIT-компиляции, «теплого старта» обучения метода mPPCA на каждом новом положении окна и трехступенчатый алгоритм поиска меры VaR позволяют существенно ускорить эксперименты по сравнению с обычной реализацией.

Ключевые слова: Value at Risk, VaR, PCA, PPCA, mPPCA, бэктестинг, понижение размерности.

ОБРАЗЕЦ ЦИТИРОВАНИЯ

Волков Н.В. Улучшение оценки стоимостной меры риска многомерных портфелей с помощью смеси вероятностных анализаторов главных компонент // Вестник ЮУрГУ. Серия: Вычислительная математика и информатика. 2025. Т. 14, № 2. С. 5–25. DOI: 10.14529/cmse250201.

Введение

Каждой компании и каждому банку необходимы методы измерения и контроля своих финансовых рисков для различных будущих сценариев, чтобы поддерживать целевой рейтинг. В качестве классической меры риска используется стоимостная мера риска (VaR, Value At Risk). Если в портфеле инвестора или компании есть только один актив, то оценка параметров одномерного распределения и вычисление меры VaR являются стандартной задачей [1].

Оптимизация финансового портфеля при высоких размерностях значительно сложнее. Это связано с необходимостью оценивать многомерное совместное распределение сильно скоррелированных активов [2–4]. Как правило, для решения этой проблемы предлагается использовать методы понижения размерности, в частности метод главных компонент (PCA, Principal Component Analysis), введенный в работе [5]. Применение метода PCA к оценке финансовых рисков описано во многих работах (см., например, [6–8]). В работе [9] авторы с применением метода PCA выделили некоррелированные компоненты, которые являются линейными комбинациями исходных активов и объясняют достаточную долю дисперсии портфеля, а затем, основываясь на них, упростили построение границы эффективных портфелей в задаче Марковитца, введенной в [10]. На основе полученного подхода в работе [11] авторы применили метод выделения главных компонент для определения диверсифицированности портфеля. В [12, 13] данный подход обобщен так, что максимально диверсифицированные портфели должны включать в себя активы в практически одинаковых долях, минимизируя при этом дисперсию доходности портфеля. В [8] метод PCA и ядерный метод главных компонент (Kernel PCA) применяются для оптимизации долгосрочного инвестирования на основе диверсификации Марковица.

Также для оптимизации многомерных портфелей и оценки их риска часто применяется смесь гауссовских распределений (GMM, Gaussian Mixture Models), которая позволяет моделировать «тяжелые» хвосты (высокую вероятность по сравнению с гауссовским распределением сильных отклонений случайной величины от своего среднего) за счет использования нескольких гауссовских компонент [14, 15]. Однако для портфелей большой размерности практическое применение метода GMM затруднено, так как требуется оценка полной матрицы ковариаций для каждой смеси, что существенно увеличивает вычислительную сложность, а также может привести к неустойчивости оценок. Чтобы преодолеть эти проблемы, в ряде исследований предлагается использовать регуляризацию и штрафные подходы для стабилизации процесса оценивания и уменьшения эффективной размерности задачи. В частности, в работе [16] предложен подход на основе максимизации правдоподобия со штрафом и с разреживанием (sparsity) обратных ковариационных матриц, что уменьшает количество параметров и стабилизирует оценки. Также для повышения точности оценивания ковариационных матриц компонент в работе [17] предложена аналитическая регуляризация типа Ледуа—Вольфа (Ledoit–Wolf shrinkage), встроенная в алгоритм ожидания-максимизации (EM, Expectation-Maximization,[18]).

В данной работе рассматривается альтернативный подход, сочетающий в себе преимущества методов PCA и GMM в приложении к оценке риска многомерных портфелей, основанный на смеси вероятностных анализаторов главных компонент (mPPCA, Mixture of Probabilistic Principal Component Analyzers), введенной в работах [19, 20]. Предложенный метод активно исследуется, например, в приложении к мониторингу мультимодальных промышленных процессов и поиску в них неисправностей [21–25], при этом он крайне мало рассмотрен в литературе в применении к задачам финансов [26, 27] и вовсе не рассматривался применительно к задаче оценки финансового риска. В отличие от PCA, этот метод предполагает вероятностную модель данных и оптимизирует не долю объясненной дисперсии, а функцию правдоподобия. Таким образом, благодаря вероятностной модели, отдельной оценке шума и совмещению нескольких PPCA, становится возможным точнее учитывать «тяжелые» хвосты распределений и лучше оценивать меру VaR. При этом метод PCA, напротив, крайне неустойчив к выбросам, которые часто встречаются в финансовых данных. Одновременно метод mPPCA обладает преимуществом смеси GMM в использовании смеси нормальных распределений, но при этом за счет понижения размерности не страдает проблемой оценки полноразмерных матриц ковариаций.

Кроме того, разработан и протестирован гибридный подход для оценки меры $\mathrm{VaR}_{5\%}$ с адаптивным выбором нужного количества компонент в методе mPPCA с использованием информационного критерия Акаике (AIC, Akaike Information Criterion) [28]. Отметим, что при оптимизации портфеля крайне важно быстро оценивать риск без повторного обучения модели для каждого случая. Именно этот принцип лежит в основе предложенного метода.

Для корректной оценки качества метода расчета меры VaR необходимо проведение оценки эффективности на основе исторических данных (бэктестинга) [15]. В работе проводится масштабный бэктестинг результатов оценки меры $\mathrm{VaR}_{5\%}$ с помощью метода mPPCA на различных 100 сильно и 100 слабо диверсифицированных портфелях из акций индекса S&P 500 за период 2005–2023 гг. для mPPCA с разным количеством смесей и адаптивным подбором смесей на основе критерия AIC. Обучение алгоритмов проходит на скользящих окнах размером в 350 торговых дней [29, 30], после чего делается оценка VaR на следующий за окном день. Оценка меры VaR для каждого портфеля бэктестируется на прохождение классического Купик-теста [31]. Рассмотрение разных видов портфелей важно, так как все они могут возникать при решении задачи оптимизации портфеля и критично, чтобы алгоритм оценки риска корректно на них работал.

В работе проводится анализ зависимости изменчивости цены (волатильности) и «тяжести» хвостов распределений лог-приростов следующего дня от оптимального количества компонент смеси в методе mPPCA и от минимально достаточного количества основных компонент в методах PPCA и PCA (для объяснения 80 % дисперсии данных). Отдельно предлагается алгоритм расчета доли объясненной дисперсии в случае метода PPCA, его теоретическое обоснование и применение на реальных данных.

В работе также предложены следующие подходы к оптимизации обучения смеси метода PPCA и проведения обширного бэктестинга: компиляция «точно в нужное время» (JIT, Just In Time) Python-кода за счет использования библиотеки Numba [32] и адаптации всего кода под эту библиотеку; присвоение стартовой точке при обучении метода mPPCA в новом положении окна значения оптимума в предыдущем положении окна; трехэтапный поиск численного решения уравнения при финальной оценке VaR.

В разделе 1 представлены основные понятия и подходы, используемые в работе, обосновывается расчет доли объясненной дисперсии в случае метода PPCA. Раздел 2 содержит описание алгоритмов и процесса их обучения. Результаты экспериментов рассмотрены в разделе 3. В заключении приводятся резюме работы, краткая сводка полученных результатов и итоговые выводы.

1. Теоретическая часть

1.1. Probabilistic PCA и смесь PPCA

В отличие от классического метода PCA, вероятностный метод главных компонент (PPCA, Probabilistic PCA) [19] предполагает вероятностную модель данных, которая обеспечивает бóльшую устойчивость к выбросам и пропущенным данным. Пусть имеется выборка из n объектов

$$X = \{x_i\}_{i=1}^{n}, \quad x_i \in \mathbb{R}^d. \tag{1}$$

Предполагается существование латентного пространства размерности $q < d$, и связь наблюдений x_i с латентными переменными $y_i \in \mathbb{R}^q$ задается линейной моделью с гауссовским шумом:

$$x_i = \mu + W y_i + \varepsilon_i, \quad y_i \sim \mathcal{N}(0, I_q), \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_d), \tag{2}$$

где $\mu \in \mathbb{R}^d$ — вектор средних, W — матрица вложения размерности $d \times q$, а $\sigma^2 I_d$ — ковариационная матрица шума. Тогда безусловное распределение x_i имеет вид

$$x_i \sim \mathcal{N}(\mu, C), \quad C = WW^\top + \sigma^2 I_d, \tag{3}$$

где C — ковариационная матрица, а апостериорное распределение латентных переменных при наблюдении x_i задается выражением

$$y_i \mid x_i \sim \mathcal{N}\left(M^{-1} W^\top (x_i - \mu),\ \sigma^2 M^{-1}\right), \quad M = W^\top W + \sigma^2 I_q. \tag{4}$$

Оценка параметров (W, μ, σ^2) производится максимизацией логарифма правдоподобия

$$\mathcal{L}(W, \mu, \sigma^2) = -\frac{nd}{2} \log(2\pi) - \frac{n}{2} \log\det(C) - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^\top C^{-1} (x_i - \mu) \tag{5}$$

с использованием EM-алгоритма (см. [19]).

Для моделирования более сложных распределений метод PPCA естественно обобщается на случай смеси из K компонент [20] — mPPCA с K компонентами. Каждая компонента имеет собственный набор параметров (μ_k, W_k, σ_k^2), и модель смеси записывается в виде

$$p(x) = \sum_{k=1}^{K} \pi_k\, \mathcal{N}(x \mid \mu_k, W_k W_k^\top + \sigma_k^2 I_d), \tag{6}$$

где $\{\pi_k\}_{k=1}^{K}$ — веса смеси, удовлетворяющие условиям $\pi_k \geq 0$, $\sum_{k=1}^{K} \pi_k = 1$.

Параметры модели метода mPPCA оцениваются максимизацией логарифма правдоподобия смеси

$$\mathcal{L}_{\text{mPPCA}} = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k\, \mathcal{N}(x_i \mid \mu_k, W_k W_k^\top + \sigma_k^2 I_d) \right]. \tag{7}$$

Таким образом, метод mPPCA сочетает преимущества вероятностного подхода и снижения размерности, эффективно описывая многомерные распределения с «тяжелыми» хвостами и сложной структурой, что особенно важно в задачах оценки финансовых рисков.

1.2. Определение доли объясненной дисперсии для PPCA

В классическом методе PCA доля дисперсии, объясненной первыми q главными компонентами, определяется как отношение суммы первых q собственных значений к сумме всех собственных значений ковариационной матрицы данных:

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_q}{\lambda_1 + \lambda_2 + \cdots + \lambda_d}. \tag{8}$$

Однако метод PPCA оперирует вероятностной моделью, в которой, в том числе, дополнительно оценивается и шум с дисперсией σ^2, поэтому подход для определения доли дисперсии требует модификации. Ниже предлагается и теоретически обосновывается подход для расчета доли объясненной дисперсии в случае метода PPCA.

Утверждение 1. Пусть в модели PPCA размерность наблюдаемого пространства равна d, а размерность латентного пространства — $q < d$. Согласно [19] оценка максимального правдоподобия матрицы вложений W имеет вид

$$W = U_q(\Lambda_q - \sigma^2 I_q)^{1/2} R, \tag{9}$$

где

- $U_q \in \mathbb{R}^{d\times q}$ — матрица, столбцами которой являются первые q собственных векторов выборочной ковариационной матрицы;
- $\Lambda_q = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ — диагональная матрица, содержащая первые q собственных значений ковариационной матрицы данных;
- σ^2 — оценка максимального правдоподобия дисперсии шума, вычисляемая как среднее оставшихся $(d - q)$ собственных значений;
- $R \in \mathbb{R}^{q\times q}$ — произвольная ортогональная матрица ($RR^\top = I_q$).

В таком случае аналог доли объясненной дисперсии для модели PPCA определяется следующим образом:

$$\mathrm{coverage}(q) = \frac{\mathrm{trace}(WW^\top)}{\mathrm{trace}(WW^\top) + d\sigma^2} = \frac{\sum_{i=1}^q (\lambda_i - \sigma^2)}{\sum_{i=1}^q (\lambda_i - \sigma^2) + d\sigma^2}. \tag{10}$$

Доказательство. Из оригинальной работы по методу PPCA [19] следует, что оценка максимального правдоподобия для матрицы вложений W представима в виде (9). Рассмотрим выражение для $WW^\top$:

$$WW^\top = U_q(\Lambda_q - \sigma^2 I_q)^{1/2} R \left(U_q(\Lambda_q - \sigma^2 I_q)^{1/2} R\right)^\top. \tag{11}$$

Учитывая ортогональность матрицы R и матрицы собственных векторов U_q, имеем

$$WW^\top = U_q(\Lambda_q - \sigma^2 I_q) U_q^\top. \tag{12}$$

Следовательно, след матрицы $WW^\top$ выражается через собственные значения как

$$\mathrm{trace}(WW^\top) = \mathrm{trace}\big(U_q(\Lambda_q - \sigma^2 I_q) U_q^\top\big) = \sum_{i=1}^q (\lambda_i - \sigma^2). \tag{13}$$

Полная дисперсия в методе PPCA состоит из дисперсии, объясняемой латентными переменными, и дисперсии шума. Так как дисперсия шума равна σ^2 для каждой из d компонент, суммарный вклад шума равен $d\sigma^2$. Таким образом, итоговое выражение для аналога доли объясненной дисперсии имеет вид

$$\mathrm{coverage}(q) = \frac{\sum_{i=1}^q (\lambda_i - \sigma^2)}{\sum_{i=1}^q (\lambda_i - \sigma^2) + d\sigma^2}. \tag{14}$$

□

Полученная формула (10) позволяет аналитически рассчитывать аналог доли объясненной дисперсии для метода PPCA, обеспечивая корректное сопоставление с классическим методом PCA и удобство при выборе оптимальной латентной размерности.

1.3. Информационный критерий Акаике для mPPCA

Для автоматического выбора оптимального числа компонент смеси mPPCA в каждом скользящем окне используется информационный критерий Акаике [28], что позволяет балансировать между качеством аппроксимации данных и сложностью модели.

Обозначим набор параметров смеси mPPCA с K компонентами и размерностью латентного пространства q следующим образом:

$$\theta \;=\; \{\pi_k,\, \mu_k,\, W_k,\, \sigma_k^2\}_{k=1}^{K}, \tag{15}$$

где π_k — веса компонент смеси, μ_k — векторы средних компонент, W_k — матрицы вложений, а σ_k^2 — дисперсии шума.

Для модели mPPCA с K компонентами, размерностью данных d и размерностью латентного пространства q число p параметров вычисляется по формуле

$$p \;=\; \underbrace{Kdq}_{\text{матрицы } W_k} \;+\; \underbrace{Kd}_{\text{векторы средних } \mu_k} \;+\; \underbrace{K}_{\text{дисперсии } \sigma_k^2} \;+\; \underbrace{(K-1)}_{\text{веса смеси } \{\pi_k\}}\;, \tag{16}$$

где

- Kdq — количество параметров матриц W_k (каждая размером $d \times q$);
- Kd — число параметров средних векторов μ_k (каждый вектор размерности d);
- K — число дисперсий σ_k^2 (одна на каждую компоненту смеси);
- $(K-1)$ — число свободных параметров для весов смеси $\{\pi_k\}$, так как выполняется условие нормировки $\sum_{k=1}^{K} \pi_k = 1$.

Таким образом, полная формула для числа параметров модели записывается следующим образом:

$$p \;=\; K(dq + d + 1) + (K-1). \tag{17}$$

Формула для критерия AIC смеси mPPCA задается выражением

$$\mathrm{AIC}_{\mathrm{mPPCA}} \;=\; 2\,p \;-\; 2\,\mathcal{L}_{\mathrm{mPPCA}}\big(\hat{\theta}\big), \tag{18}$$

где $\hat{\theta}$ — оценка параметров, полученная после обучения модели на окне, а $\mathcal{L}_{\mathrm{mPPCA}}\big(\hat{\theta}\big)$ — значение логарифма правдоподобия mPPCA при параметрах $\hat{\theta}$.

1.4. Оценка меры VaR

Для вычисления меры VaR заданного уровня α для конкретного портфеля с весами активов w полученные многомерные распределения проецируются на одномерное распределение доходностей портфеля r_p следующим образом:

$$r_p \sim \sum_{k=1}^{K} \pi_k \mathcal{N}\left(w^T \mu_k,\, w^T C_k w\right). \tag{19}$$

Затем численно решается нелинейное уравнение, определяющее квантиль уровня α:

$$F_{r_p}(\mathrm{VaR}_\alpha) = \alpha, \tag{20}$$

где $F_{r_p}(\cdot)$ — функция распределения смеси нормальных распределений доходностей портфеля r_p. Решение уравнения проводится в три этапа:

- определяются границы поиска в виде минимального и максимального квантилей среди всех смесей;
- границы поиска уточняются с помощью метода бисекции;
- производится уточнение корня методом Ньютона—Рафсона (история появления метода [33]).

Таким образом, предложенный подход на основе метода mPPCA учитывает вероятностную структуру многомерных данных и эффективно вычисляет меру VaR, оставаясь более устойчивым и менее вычислительно затратным, чем классические подходы PCA и GMM.

1.5. Бэктестинг меры VaR

Пусть $\{X_t\}_{t=1}^n$ обозначает ряд доходностей, а $\mathrm{VaR}_\alpha(X_t)$ — оценка меры VaR на уровне α. Для каждого дня определим индикатор исключения:

$$I_t \;=\; \begin{cases} 1, & \text{если } X_t < -\,\mathrm{VaR}_\alpha(X_t), \\ 0 & \text{иначе.} \end{cases} \tag{21}$$

Если модель корректна, то количество исключений $n_1 = \sum_{t=1}^n I_t$ можно рассматривать как реализацию биномиальной случайной величины с параметром α. Для проверки гипотезы о корректности модели в работе используется тест Купика [31], основанный на асимптотическом поведении отношения правдоподобия.

Тест Купика [31] оценивает, совпадает ли общее количество нарушений с ожидаемым. Пусть

$$n_1 \;=\; \sum_{t=1}^{n} I_t, \quad n_0 \;=\; n - n_1, \quad \hat{p} \;=\; \frac{n_1}{n}. \tag{22}$$

Тогда статистика теста безусловного покрытия ($\mathrm{LR_{uc}}$) определяется как

$$\mathrm{LR_{uc}} \;=\; -\,2\,\ln\!\left[\frac{(1-\alpha)^{\,n_0}\,\alpha^{\,n_1}}{(1-\hat{p})^{\,n_0}\,\hat{p}^{\,n_1}}\right]. \tag{23}$$

При справедливости нулевой гипотезы «$p = \alpha$» (т. е. корректном безусловном покрытии) эта статистика асимптотически имеет распределение χ^2 с 1 степенью свободы. Если значение $\mathrm{LR_{uc}}$ велико и его p-value меньше 5%, то гипотеза о совпадении средней доли нарушений с α отвергается.

2. Описание алгоритмов

Код основных методов выложен в открытом доступе в репозитории [34].

2.1. Реализация алгоритма mPPCA и процедура обучения

Пусть на текущем скользящем окне размером Δ торговых дней имеются наблюдения доходностей активов в виде матрицы $X \in \mathbb{R}^{\Delta\times d}$, элементы которой обозначим через $x_{i,j}$, где i — индекс наблюдения (день), j — индекс актива, Δ — длина окна, d — количество активов.

Z-нормализация данных. Перед обучением модели данные нормализуются по каждому признаку (активу):

$$x_{i,j}^{\mathrm{norm}} = \frac{x_{i,j} - \bar{x}_j}{s_j}, \quad \text{где} \quad \bar{x}_j = \frac{1}{\Delta}\sum_{i=1}^{\Delta} x_{i,j}, \quad s_j = \sqrt{\frac{1}{\Delta}\sum_{i=1}^{\Delta}(x_{i,j} - \bar{x}_j)^2}. \tag{24}$$

Инициализация параметров модели. Перед началом скольжения окна параметры модели инициализируются следующим образом.

- Нормализованные данные разделяются на K кластеров методом K-means (см. [35, 36]).
- К каждому из полученных кластеров применяется метод PCA, из результатов которого получают начальные приближения параметров смеси mPPCA:
  - векторы средних μ_k;
  - матрицы вложений W_k;
  - дисперсии шума σ_k^2;
  - веса компонент смеси π_k (на основании количества точек в каждом из кластеров).

Обучение на основе применения EM-алгоритма. На каждом временном окне параметры метода mPPCA оцениваются с использованием EM-алгоритма [18], включающего E-шаг (Expectation step) и M-шаг (Maximization step), которые повторяются, пока разница между новыми оценками параметров и предыдущими не станет незначительной.

E-шаг. Вычисляются вероятности принадлежности (responsibilities) каждого наблюдения x_i каждой компоненте смеси:

$$r_{i,k} = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, W_k W_k^\top + \sigma_k^2 I_d)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_i \mid \mu_j, W_j W_j^\top + \sigma_j^2 I_d)}, \tag{25}$$

где $r_{i,k}$ — вероятность принадлежности наблюдения x_i компоненте k.

M-шаг. На основе полученных на E-шаге вероятностей $r_{i,k}$ параметры смеси обновляются следующим образом.

- Веса компонент смеси:
$$\pi_k \leftarrow \frac{1}{\Delta} \sum_{i=1}^{\Delta} r_{i,k}. \tag{26}$$
- Векторы средних компонент:
$$\mu_k \leftarrow \frac{\sum_{i=1}^{\Delta} r_{i,k} x_i}{\sum_{i=1}^{\Delta} r_{i,k}}. \tag{27}$$
- Выборочные ковариационные матрицы компонент:
$$S_k = \frac{1}{\Delta \pi_k} \sum_{i=1}^{\Delta} r_{i,k} (x_i - \mu_k)(x_i - \mu_k)^\top. \tag{28}$$
- Матрицы вложений:
$$W_k \leftarrow S_k W_k \left(\sigma_k^2 I_q + M_k^{-1} W_k^\top S_k W_k \right)^{-1}, \tag{29}$$
  где матрица M_k определяется выражением
$$M_k = W_k^\top W_k + \sigma_k^2 I_q. \tag{30}$$
- Дисперсии шума:
$$\sigma_k^2 \leftarrow \frac{1}{d} \mathrm{trace} \left[S_k - S_k W_k M_k^{-1} W_k^\top \right]. \tag{31}$$

Важно отметить, что в ходе этих операций приходится обращать матрицы не исходной размерности $d \times d$, а только матрицы размерности $q \times q$, где q — латентная размерность. Для повышения численной устойчивости при обращении матриц в любом случае используется регуляризация.

Обучение метода mPPCA на новом положении окна производится с так называемым «теплым стартом», при котором в качестве начальных приближений параметров модели берутся оценки, полученные на предыдущем положении окна. Это позволяет значительно сократить число итераций до достижения сходимости EM-алгоритма.

Оптимальное количество компонент смеси на каждом окне выбирается на основе минимизации информационного критерия Акаике, что позволяет автоматически балансировать между сложностью модели и точностью аппроксимации данных.

Алгоритм выбора числа компонент K по критерию AIC

1. Задаем диапазон значений K от $K_{\min}$ до $K_{\max}$ (в работе от 1 до 4).
2. Для каждого значения $K = K_{\min}, K_{\min}+1, \ldots, K_{\max}$:
   - оцениваем параметры модели $\hat{\theta}$ с помощью EM-алгоритма;
   - вычисляем соответствующее значение $\mathrm{AIC}_{\mathrm{mPPCA}}(K)$.
3. Оптимальным числом компонент K^* считается то, при котором достигается минимальное значение критерия AIC:

$$K^* = \arg\min_K \big[\mathrm{AIC}_{\mathrm{mPPCA}}(K)\big]. \tag{32}$$

Таким образом, в каждом скользящем окне автоматически определяется оптимальное число компонент смеси mPPCA, исходя из компромисса между точностью аппроксимации данных и сложностью модели.

Описанная реализация алгоритма mPPCA с пошаговой нормализацией данных, «теплым стартом» и автоматическим выбором оптимального числа компонент по критерию AIC значительно повышает эффективность оценки VaR многомерных финансовых портфелей.

2.2. Численное решение для меры VaR

В реализации вычисления VaR_α для смеси одномерных нормальных распределений используется следующая схема.

1. Начальные границы поиска $\big[L, U\big]$ берутся на основе α-квантилей для каждой гауссовской компоненты смеси:

$$L = \min_k F^{-1}\big(\alpha;\, \mu_k,\, \sigma_k\big), \quad U = \max_k F^{-1}\big(\alpha;\, \mu_k,\, \sigma_k\big), \tag{33}$$

   где F — функция распределения для нормального распределения $\mathcal{N}(\mu_k, \sigma_k^2)$. Благодаря этому отрезок $[L, U]$ наверняка содержит искомый квантиль смеси.
2. На отрезке $[L, U]$ выполняется несколько шагов классического метода бисекции, чтобы быстро получить аппроксимацию корня уравнения $F_{\mathrm{mix}}(x) = \alpha$ с точностью порядка 10^{-2} по аргументу.
3. Полученный результат из бисекции выступает начальными данными для метода Ньютона—Рафсона, с помощью которого выполняется поиск решения с точностью уже порядка 10^{-6}. На каждом шаге для текущего приближения x вычисляется

$$x_{\mathrm{new}} = x_{\mathrm{old}} - \frac{F_{\mathrm{mix}}(x_{\mathrm{old}}) - \alpha}{f_{\mathrm{mix}}(x_{\mathrm{old}})}, \tag{34}$$

   где $f_{\mathrm{mix}}(\cdot)$ — смесь плотностей.

Симуляция на десяти миллионах различных нормальных смесей показывает ускорение подсчета квантилей с нужным уровнем точности более чем в три раза по сравнению с каждым из методов отдельно. Например, метод бисекции занимает 161 секунду, а трехшаговый — 45 секунд. Все эксперименты здесь и далее выполнялись в бесплатном Google Colab на центральном процессоре (CPU, Central Processing Unit), частота 2.2–2.8 ГГц, без использования графического процессора (GPU, Graphics Processing Unit).

2.3. JIT-компиляция и отказ от внешних библиотек

Для повышения скорости вычислений используется `@jit(nopython=True)` (или `@njit`) из библиотеки **Numba**, что дает перевод вычислительно интенсивных участков в машинный код. Это позволяет:

- избегать медленных Python-циклов в критических местах (все операции выполняются над NumPy-массивами или в простых `for`-циклах, которые Numba оптимизирует);
- минимизировать вызовы «тяжелых» библиотек (например, `scipy.stats`). Вместо этого функции CDF, PDF, PPF нормального распределения реализованы вручную, а вся линейная алгебра (например, умножение матриц) делается через NumPy-функции, совместимые с Numba;
- Данный подход вместе с использованием «теплого» старта дает ускорением в более чем 100 раз. К прмеру обучение mPPCA со всеми оптимизациями с двумя смесями на 650 точках на 221 активе занимает 72 секунды, а без оптимизаций — более 4 часов.

3. Вычислительные эксперименты

3.1. Результаты бэктестинга

Был проведен масштабный бэктестинг результатов оценки меры $\mathrm{VaR}_{5\%}$ с помощью метода mPPCA на различных 100 сильно и 100 слабо диверсифицированных портфелях на акциях из индекса S&P 500 за период 2009–2023 гг. (тестирование; обучение начинается ранее). Под слабо диверсифицированными портфелями понимаются портфели, в которых 90 % весов сосредоточены в 1–5 ведущих акциях, а под сильно диверсифицированными — портфели, где веса случайным образом распределены между всеми акциями. Рассматривались методы PCA, mPPCA и GMM с разным количеством компонент. Оценки меры VaR проверялись на прохождение классического Купик-теста. В табл. 1 приведены доли портфелей, на которых тест пройден успешно.

Таблица 1. Сравнение результатов бэктестирования оценок $\mathrm{VaR}_{5\%}$

Метод	**Слабо диверс.**	**Сильно диверс.**
PCA	0.00	1.00
GMM 1	0.37	0.00
GMM 2	0.46	0.00
GMM 3	0.51	0.00
mPPCA 1	0.78	1.00
mPPCA 2	0.83	1.00
mPPCA 3	0.86	1.00
mPPCA 4	0.88	1.00
mPPCA AIC	0.89	1.00

Из полученных результатов следует, что для слабо диверсифицированных портфелей классический метод PCA не справляется с задачей, демонстрируя нулевую долю успешных прохождений теста Купика. Метод GMM существенно улучшает результаты по сравнению с методом PCA, однако даже при использовании трех компонент смеси доля успешно пройденных тестов составляет лишь 51 %. Однако при этом метод GMM крайне плохо показывает себя на диверсифицированных портфелях.

Наилучшие результаты достигаются при использовании метода mPPCA: уже при одной компоненте доля успешных прохождений значительно превышает результаты методов GMM и PCA, достигая 78 %. Дальнейшее увеличение числа компонент смеси до четырех последовательно повышает качество оценки, доходя до 88 %. Адаптивный выбор количества компонент по критерию AIC дополнительно улучшает результаты (до 89 %). Стоит отметить, что дальнейшее увеличение числа компонент смеси для метода mPPCA свыше четырех уже не приносит улучшения.

Таким образом, метод mPPCA демонстрирует явные преимущества перед методами PCA и GMM в оценке меры VaR, особенно для слабо диверсифицированных портфелей, совмещая эффективность понижения размерности и гибкость вероятностной модели смеси.

3.2. Влияние оптимального по критерию AIC числа компонент в mPPCA на лог-приросты следующего дня

Был также проведен анализ зависимости абсолютных значений лог-приростов следующего дня от оптимального по критерию AIC количества компонент в методе mPPCA (далее «группа»). Соответствующая ящичная диаграмма (box plot) изображена на рис. 1. Такие показатели, как среднее, стандартное отклонение, ассиметрия и коэффициент эксцесса лог-приростов для каждой из групп приведены в табл. 2. Здесь и далее важно отметить, что речь идет о значениях лог-приростов каждого из активов на следующий день за текущим окном, на котором обучается модель и происходит оценка оптимального количества смесей в методе mPPCA. Из таблицы видно существенное изменение стандартных отклонений и коэффициентов асимметрии в зависимости от количества компонент в смеси. В группе с двумя смесями самые маленькие стандартное отклонение и коэффициент асимметрии, в группе с тремя смесями эти показатели средние, а в группе с четырьмя — самые высокие.

Таким образом, использование метода mPPCA дает возможность выявления скрытых рыночных режимов и учета потенциального уровня волатильности и экстремальных значений для оценки риска на следующий день.

Таблица 2. Статистики лог-приростов следующего дня в зависимости от оптимального количества смесей по критерию AIC

Группа	**Среднее**	**СКО**	**Асимметрия**	**Эксцесс**
2	0.0007	0.015	0.027	13.461
3	0.0008	0.020	0.170	20.516
4	0.0005	0.027	0.534	17.942

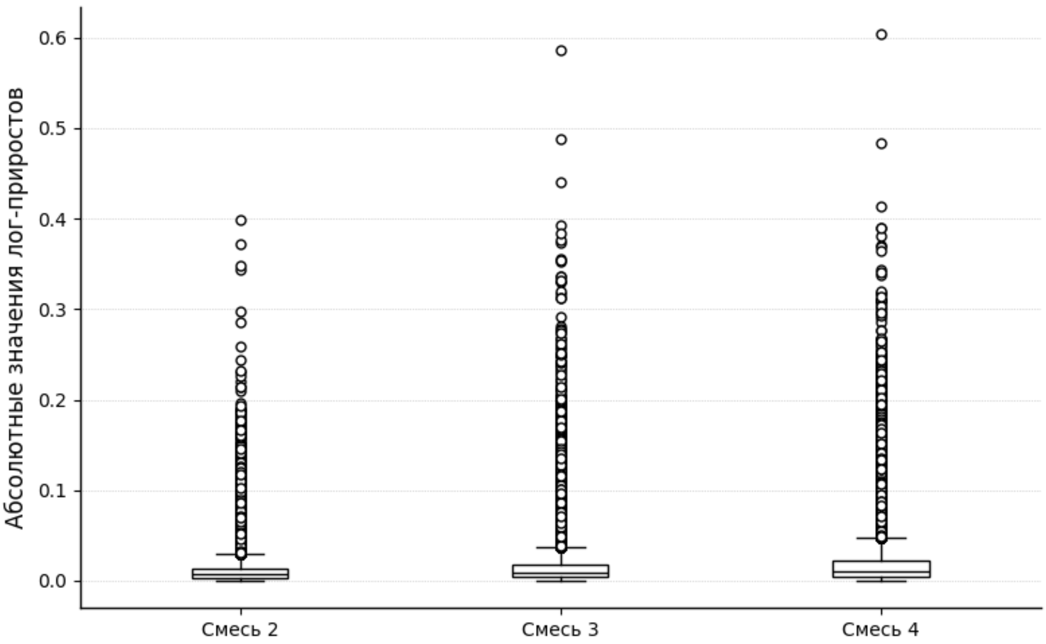

Рис. 1. Ящичная диаграмма (box plot) абсолютных значений лог-приростов и оптимального количества смесей в mPPCA по критерию AIC

3.3. Влияние минимально достаточного числа главных компонент в методах PCA и PPCA на абсолютные лог-приросты следующего дня

Была также рассмотрена зависимость абсолютных лог-приростов следующего дня от минимального достаточного количества основных компонент в методах PCA и PPCA для объяснения α процентов дисперсии данных. В качестве α бралось 80 % из эмпирических соображений. На рис. 2 изображены графики минимального достаточного количества основных компонент в методах PCA и PPCA, из которого видно, что количества требуемых компонент у методов PCA и PPCA крайне близки.

На рис. 3 приведены ящичные диаграммы абсолютных лог-приростов следующего дня, сгруппированных по числу необходимых главных компонент в методах PCA и PPCA, достаточных для объяснения 80 % дисперсии данных. Из графиков видно, что в группе $(0, 40]$ наблюдаются наиболее крупные выбросы и более высокая медиана абсолютных значений, что указывает на повышенную волатильность.

В табл. 3 приведены параметры распределений абсолютных значений лог-приростов для тех же групп. Сравнение показывает, что в группе $(0, 40]$ абсолютные лог-приросты характеризуются максимальным стандартным отклонением, наиболее высокими значениями асимметрии и эксцесса. Это согласуется с большим разбросом абсолютных лог-приростов на рис. 3. В группах с большим количеством необходимых компонент (более 40) волатильность и «толщина» хвостов распределения заметно снижаются, что отражается и в уменьшении выбросов на диаграмме абсолютных значений, и в относительно меньших значениях стандартного отклонения, асимметрии и эксцесса в таблице.

Дополнительно было проведено исследование связи между минимальным достаточным количеством главных компонент и средней корреляцией активов внутри скользящего окна.

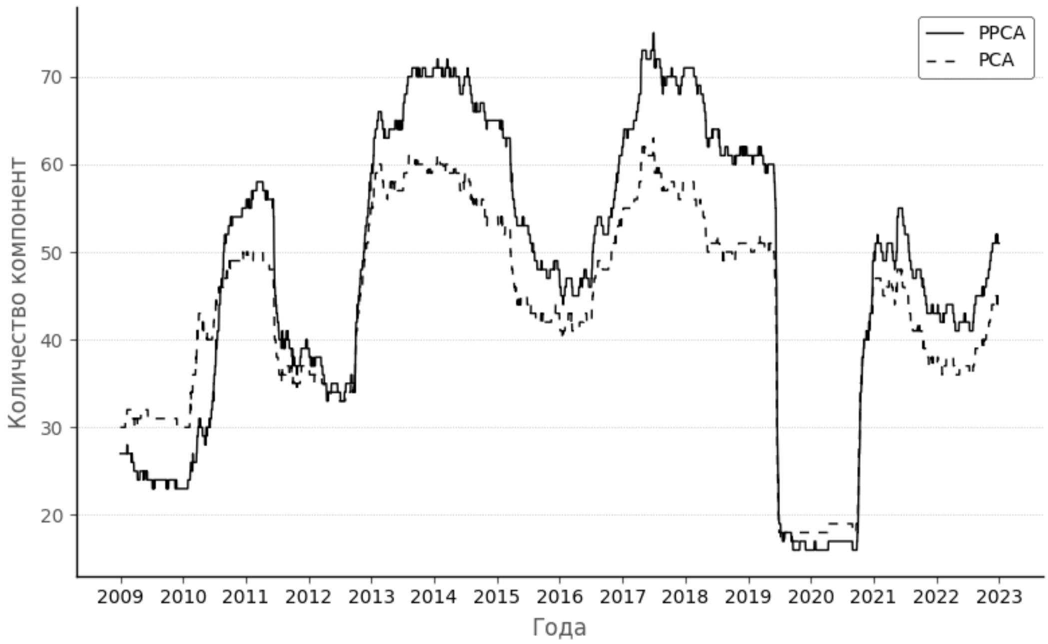

Рис. 2. Графики минимального достаточного количества основных компонент для объяснения 80 % дисперсии в методах PCA и PPCA

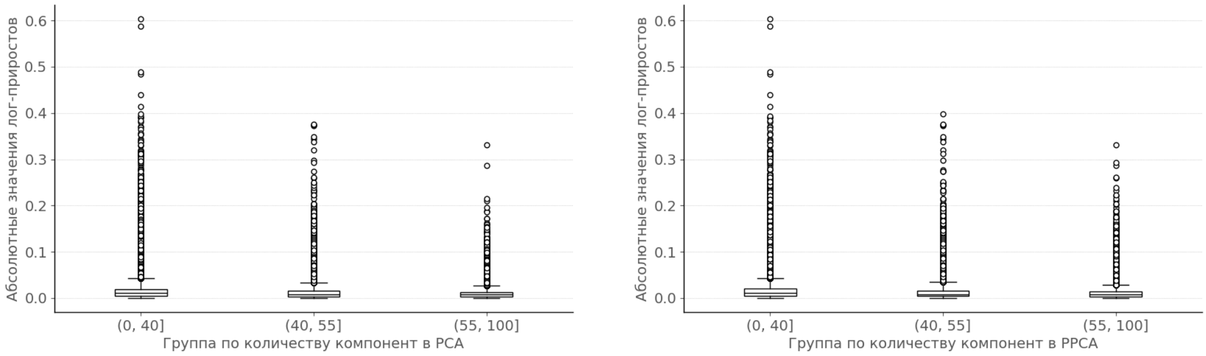

Рис. 3. Ящичные диаграммы (box plot) абсолютных значений лог-приростов и минимального достаточного количества основных компонент для методов PCA и PPCA

Таблица 3. Параметры распределений абсолютных значений лог-приростов на следующий день в зависимости от группы для методов PCA и PPCA

Группа	**Среднее**		**СКО**		**Асимметрия**		**Эксцесс**	
	PCA	PPCA	PCA	PPCA	PCA	PPCA	PCA	PPCA
(0, 40]	0.015	0.016	0.018	0.019	4.656	4.613	46.877	45.665
(40, 55]	0.011	0.012	0.012	0.013	3.689	3.914	34.699	38.115
55 и более	0.010	0.010	0.010	0.011	3.562	3.533	34.619	31.096

На каждом положении окна вычислялось среднее значение всех попарных корреляций активов, после чего оно сопоставлялось с минимально необходимым количеством компонент.

Полученный коэффициент корреляции между этими двумя величинами составил −78 % и −79 % для методов PCA и PPCA соответственно. Это демонстрирует тесную обратную зависимость между рыночной корреляцией и числом необходимых компонент. Из рис. 4 видно, что периоды высоких средних корреляций совпадают с уменьшением количества главных компонент. Это наблюдение подтверждается и ящичной диаграммой на рис. 5: чем меньше необходимое количество компонент, тем выше средний уровень корреляций активов. С экономической точки зрения это естественный результат — во время кризиса корреляция активов значительно увеличивается, а значит, уменьшается количество основных компонент, с помощью которых можно достаточно точно описать общее движение портфеля. Однако в случае отсутствия кризиса все активы ведут себя по-разному, корреляции уменьшаются и, соответственно, требуется большее количество компонент для описания портфеля.

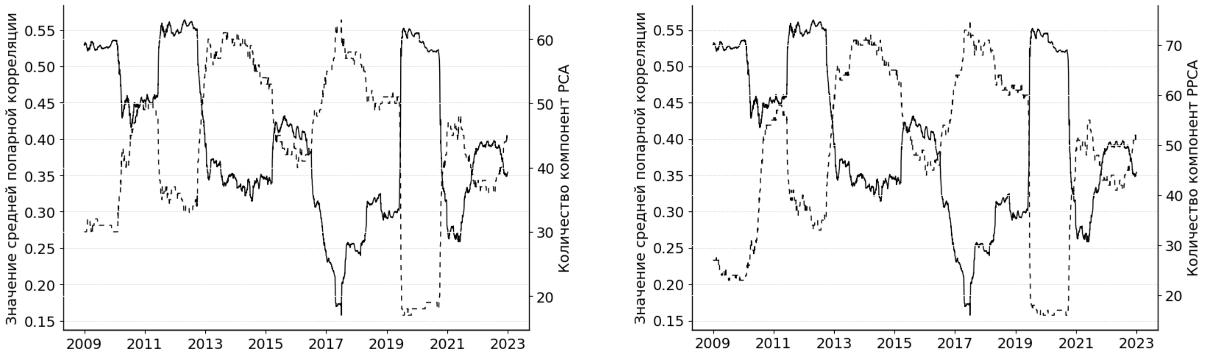

Рис. 4. Средняя попарная корреляция активов на окне и минимальное достаточное количество компонент для объяснения нужной доли дисперсии в PCA и PPCA

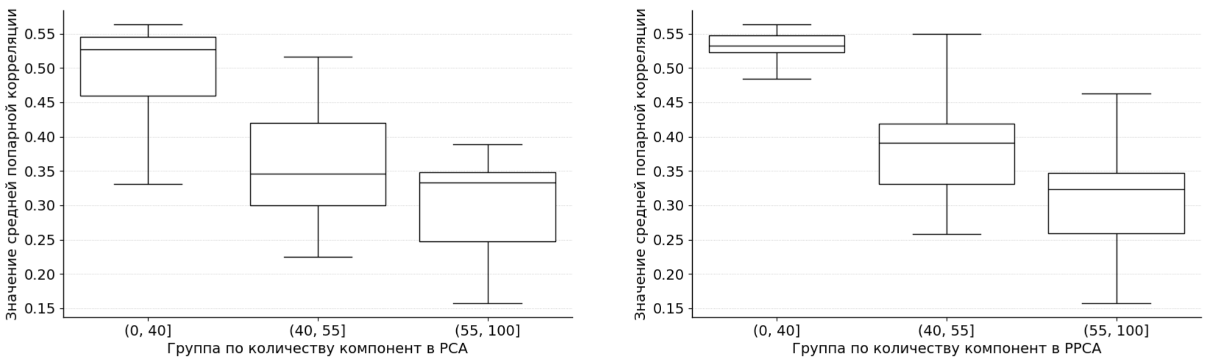

Рис. 5. Ящичные диаграммы (box plot) средней попарной корреляции активов внутри групп по количеству минимально достаточных компонент в PCA и PPCA

Заключение

В работе предложен подход к оценке меры $\mathrm{VaR}_{5\%}$ для многомерных портфелей на основе смеси PPCA с адаптивным выбором оптимального количества компонент смеси по критерию AIC. Данный подход позволяет учитывать нелинейные зависимости и «тяжелые» хвосты распределений доходностей активов, что существенно улучшает точность оценки меры VaR в сравнении с традиционными методами PCA и GMM, особенно для слабо диверсифицированных портфелей. Масштабный бэктестинг на основе теста Купика, проведенный на 100 сильно и 100 слабо диверсифицированных портфелях из акций индекса S&P 500 за период 2009–2023 гг. на скользящих окнах размером в 350 торговых дней, подтвердил преимущество предложенного метода, где доля успешно пройденных тестов увеличилась до

89 % по сравнению с методом mPPCA при фиксированном числе компонент смеси, классических методов PCA и GMM.

В ходе работы выявлено, что оптимальное по критерию AIC количество компонент смеси в mPPCA демонстрирует прямую связь с характеристиками оцененного распределения лог-приростов следующего дня. Увеличение числа компонент соответствует росту волатильности и степени «тяжести» хвостов распределения. Предложен подход для расчета объясненной дисперсии в случае метода PPCA и показано, что минимальное достаточное число главных компонент (для объяснения 80 % дисперсии) в методах PCA и PPCA обратно коррелирует (–78 % и –79 %) со средней попарной корреляцией активов и также служит индикатором повышения волатильности и риска следующего торгового дня. Таким образом, предложенные статистики могут быть использованы как индикаторы стрессовых состояний финансовых рынков.

Важным результатом работы стало достижение высокой вычислительной эффективности алгоритма. Использование JIT-компиляции (Numba), «теплого старта» при перемещении окна и трехшагового алгоритма нахождения VaR позволило существенно ускорить обучение и бэктестинг предложенного подхода по сравнению с прямой реализацией.

Литература

1. Duffie D., Pan J. An overview of value at risk // Journal of Derivatives. 1997. Vol. 4, no. 3. P. 7–49. DOI: 10.3905/jod.1997.407971.
2. Ledoit O., Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection // Journal of Empirical Finance. 2003. Vol. 10, no. 5. P. 603–621. DOI: 10.1016/S0927-5398(03)00007-0.
3. Fan J., Liao Y., Liu H. An overview of the estimation of large covariance and precision matrices // The Econometrics Journal. 2016. Mar. Vol. 19, no. 1. P. C1–C32. DOI: 10.1111/ectj.12061.
4. Pafka S., Kondor I. Noisy covariance matrices and portfolio optimization II // Physica A: Statistical Mechanics and its Applications. 2003. Vol. 319. P. 487–494. DOI: 10.1016/S0378-4371(02)01499-1.
5. Pearson K. LIII. On lines and planes of closest fit to systems of points in space // The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901. Vol. 2, no. 11. P. 559–572. DOI: 10.1080/14786440109462720.
6. Kelly B., Malamud S., Pedersen L.H. Principal Portfolios // The Journal of Finance. 2023. Vol. 78, no. 1. P. 347–387. DOI: 10.1111/jofi.13199.
7. Mavungu M. Computation of financial risk using principal component analysis // Algorithmic Finance. 2023. Vol. 10, no. 1-2. P. 1–20. DOI: 10.3233/af-220339.
8. Kulikov A.V., Polozov D.S., Volkov N.V. Long-term investment optimization based on Markowitz diversification // Business Informatics. 2024. Vol. 18, no. 3. P. 56–69. DOI: 10.17323/2587-814X.2024.3.56.69.
9. Partovi M.H., Caputo M. Principal Portfolios: Recasting the Efficient Frontier // Economics Bulletin. 2004. Jan. Vol. 7. P. 1–10.
10. Markowitz H. Portfolio Selection // The Journal of Finance. 1952. Vol. 7, no. 1. P. 77–91. DOI: 10.2307/2975974.

11. Meucci A. Managing diversification // Risk. 2009. P. 74–79.
12. Lohre H., Neugebauer U., Zimmer C. Diversified Risk Parity Strategies for Equity Portfolio Selection // The Journal of Investing. 2012. Aug. Vol. 21. P. 111–128. DOI: `10.3905/joi.2012.21.3.111`.
13. Lohre H., Opfer H., Ország G. Diversifying risk parity // The Journal of Risk. 2014. June. Vol. 16. P. 53–79. DOI: `10.21314/JOR.2014.284`.
14. Luxenberg E., Boyd S. Portfolio construction with Gaussian mixture returns and exponential utility via convex optimization // Optimization and Engineering. 2024. Vol. 25, no. 1. P. 555–574. DOI: `10.1007/s11081-023-09814-y`.
15. Indrė Morkūnaitė D.C., Leipus R. Evaluation of Value-at-Risk (VaR) using the Gaussian Mixture Models // Research in Statistics. 2024. Vol. 2, no. 1. P. 2346075. DOI: `10.1080/27684520.2024.2346075`.
16. Ruan L., Yuan M., Zou H. Regularized Parameter Estimation in High-Dimensional Gaussian Mixture Models // Neural Computation. 2011. Vol. 23. P. 1605–1622. DOI: `10.1162/NECO_a_00128`.
17. Halbe Z., Bortman M., Aladjem M. Regularized Mixture Density Estimation with an Analytical Setting of Shrinkage Intensities // IEEE Transactions on Neural Networks and Learning Systems. 2013. Vol. 24. P. 460–470. DOI: `10.1109/TNNLS.2012.2234477`.
18. Dempster A.P., Laird N.M., Rubin D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society. Series B (Methodological). 1977. Vol. 39, no. 1. P. 1–38. DOI: `10.1111/j.2517-6161.1977.tb01600.x`.
19. Tipping M.E., Bishop C.M. Probabilistic Principal Component Analysis // Journal of the Royal Statistical Society Series B. 1999. Vol. 61, no. 3. P. 611–622. DOI: `10.1111/1467-9868.00196`.
20. Tipping M.E., Bishop C.M. Mixtures of Probabilistic Principal Component Analyzers // Neural Computation. 1999. Feb. Vol. 11, no. 2. P. 443–482. DOI: `10.1162/089976699300016728`.
21. Lyu Y., Zhou L., Cong Y., *et al.* Multirate mixture probability principal component analysis for process monitoring in multimode processes // IEEE Transactions on Automation Science and Engineering. 2023. Vol. 21, no. 2. P. 2027–2038. DOI: `10.1109/TASE.2023.3253285`.
22. Tra V., Amayri M., Bouguila N. Unsupervised fault detection for building air handling unit systems using deep variational mixture of principal component analyzers // IEEE Transactions on Automation Science and Engineering. 2024. Vol. 21, no. 4. P. 6787–6803. DOI: `10.1109/tase.2023.3331347`.
23. Zhang J., Chen M., Hong X. Nonlinear process monitoring using a mixture of probabilistic PCA with clusterings // Neurocomputing. 2021. Vol. 458. P. 319–326. DOI: `10.1016/j.neucom.2021.06.039`.
24. Zhao B., Xiao X., Zhang W., *et al.* Self-Paced Probabilistic Principal Component Analysis for Data with Outliers // ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. P. 3737–3741. DOI: `10.1109/icassp40776.2020.9054487`.

25. Chao Han S.L., House L. Covariance-Guided Mixture Probabilistic Principal Component Analysis (C-MPPCA) // Journal of Computational and Graphical Statistics. 2015. Vol. 24, no. 1. P. 66–83. DOI: 10.1080/10618600.2014.891460.

26. Tzagkarakis G., Caicedo-Llano J., Dionysopoulos T. Exploiting market integration for pure alpha investments via probabilistic principal factors analysis // Journal of Mathematical Finance. 2013. Vol. 3, no. 1. P. 192–200. DOI: 10.4236/jmf.2013.31A018.

27. Begušić S., Kostanjčar Z. Cluster-Specific Latent Factor Estimation in High-Dimensional Financial Time Series // IEEE Access. 2020. Vol. 8. P. 164365–164379. DOI: 10.1109/ACCESS.2020.3021898.

28. Akaike H. A new look at the statistical model identification // IEEE Transactions on Automatic Control. 1974. Dec. Vol. 19, no. 6. P. 716–723. DOI: 10.1109/TAC.1974.1100705.

29. Brutti Righi M., Sergio Ceretta P. On the existence of an optimal estimation window for risk measures // Applied Finance Letters. 2015. Nov. Vol. 4, no. 1-2. P. 28–32. DOI: 10.24135/afl.v4i1and2.30.

30. Волков Н.В. Оценка VaR портфелей с применением методов понижения размерности PCA и PPCA // Труды МФТИ. 2025. Т. 17, № 1. С. 127–141.

31. Kupiec P.H. Techniques for verifying the accuracy of risk measurement models // Journal of Derivatives. 1995. Vol. 3, no. 2. P. 73–84. DOI: 10.3905/jod.1995.407942.

32. Lam S.K., Pitrou A., Seibert S. Numba: A LLVM-based Python JIT Compiler // Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, LLVM 2015, Austin, TX, USA, November 15, 2015. ACM, 2015. P. 1–6. DOI: 10.1145/2833157.2833162.

33. Ypma T.J. Historical development of the Newton–Raphson method // SIAM Review. 1995. Vol. 37, no. 4. P. 531–551. DOI: https://doi.org/10.1137/1037125.

34. Volkov N. MPPCA for VaR Estimation: Source Code Repository. 2025. URL: https://gitlab.com/n.volkovsky/mppca-for-var-estimation (дата обращения: 20.03.2025).

35. MacQueen J. Some methods for classification and analysis of multivariate observations // Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, Statistics. Vol. 5. University of California Press. 1967. P. 281–298.

36. Lloyd S. Least squares quantization in PCM // IEEE Transactions on Information Theory. 1982. Vol. 28, no. 2. P. 129–137. DOI: 10.1109/TIT.1982.1056489.

Волков Никита Васильевич, кафедра дискретной математики, Московский физико-технический институт (национальный исследовательский университет) (Долгопрудный, Российская Федерация)

DOI: 10.14529/cmse250201

IMPROVING VALUE-AT-RISK ESTIMATION FOR MULTIVARIATE PORTFOLIOS USING A MIXTURE OF PROBABILISTIC PRINCIPAL COMPONENT ANALYZERS

© 2025 N.V. Volkov

Moscow Institute of Physics and Technology
(Institutskiy Pereulok 9, Dolgoprudny, Moscow Region, 141701 Russia)
E-mail: nikita.v.volkov@phystech.edu
Received: 23.03.2025

This paper proposes a novel approach for estimating the Value-at-Risk (VaR) of multidimensional portfolios, based on a mixture of probabilistic principal component analyzers (mPPCA) and the Akaike information criterion. The effectiveness of the proposed approach is evaluated on historical data, accounting for various numbers of mixture components in the mPPCA method. The study is performed on 100 highly diversified and 100 weakly diversified stock portfolios of the S&P 500 index over the 2009–2023 period, using rolling windows of 350 trading days. Probabilistic principal component analysis (PPCA) can model complex dependencies among assets while capturing the "heavy" tails of return distributions. As a result, the mPPCA method surpasses conventional principal component analysis (PCA) in the accuracy of VaR estimation. In addition, by reducing dimensionality, the model is computationally much more efficient and stable than a mixture of Gaussian distributions (GMM). The paper demonstrates how portfolio return volatility and tail heaviness depend both on the optimal number of components in mPPCA and on the minimal sufficient number of principal components in PCA and PPCA needed to explain 80 % of the variance in the data. The new approach, which optimizes the number of components in mPPCA, consistently achieves higher performance than GMM, PCA, or PPCA, especially for less diversified portfolios. The paper describes methods for optimizing mPPCA training and provides an extensive historical performance evaluation (backtesting). By employing just-in-time compilation, a "warm start" for mPPCA with each new window position, and a three-step algorithm for VaR estimation, the experiments are significantly accelerated compared with the standard implementation.

Keywords: Value at Risk, VaR, PCA, PPCA, mPPCA, backtesting, dimensionality reduction.

FOR CITATION

Vokov N.V. Improving Value-at-Risk Estimation for Multivariate Portfolios Using a Mixture of Probabilistic Principal Component Analyzers. Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering. 2025. Vol. 14, no. 2. P. 5–25. (in Russian) DOI: 10.14529/cmse250201.

References

1. Duffie D., Pan J. An overview of value at risk. Journal of Derivatives. 1997. Vol. 4, no. 3. P. 7–49. DOI: 10.3905/jod.1997.407971.
2. Ledoit O., Wolf M. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance. 2003. Vol. 10, no. 5. P. 603–621. DOI: 10.1016/S0927-5398(03)00007-0.

3. Fan J., Liao Y., Liu H. An overview of the estimation of large covariance and precision matrices. The Econometrics Journal. 2016. Mar. Vol. 19, no. 1. P. C1–C32. DOI: 10.1111/ectj.12061.
4. Pafka S., Kondor I. Noisy covariance matrices and portfolio optimization II. Physica A: Statistical Mechanics and its Applications. 2003. Vol. 319. P. 487–494. DOI: 10.1016/S0378-4371(02)01499-1.
5. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. 1901. Vol. 2, no. 11. P. 559–572. DOI: 10.1080/14786440109462720.
6. Kelly B., Malamud S., Pedersen L.H. Principal Portfolios. The Journal of Finance. 2023. Vol. 78, no. 1. P. 347–387. DOI: 10.1111/jofi.13199.
7. Mavungu M. Computation of financial risk using principal component analysis. Algorithmic Finance. 2023. Vol. 10, no. 1-2. P. 1–20. DOI: 10.3233/af-220339.
8. Kulikov A.V., Polozov D.S., Volkov N.V. Long-term investment optimization based on Markowitz diversification. Business Informatics. 2024. Vol. 18, no. 3. P. 56–69. DOI: 10.17323/2587-814X.2024.3.56.69.
9. Partovi M.H., Caputo M. Principal Portfolios: Recasting the Efficient Frontier. Economics Bulletin. 2004. Jan. Vol. 7. P. 1–10.
10. Markowitz H. Portfolio Selection. The Journal of Finance. 1952. Vol. 7, no. 1. P. 77–91. DOI: 10.2307/2975974.
11. Meucci A. Managing diversification. Risk. 2009. P. 74–79.
12. Lohre H., Neugebauer U., Zimmer C. Diversified Risk Parity Strategies for Equity Portfolio Selection. The Journal of Investing. 2012. Aug. Vol. 21. P. 111–128. DOI: 10.3905/joi.2012.21.3.111.
13. Lohre H., Opfer H., Ország G. Diversifying risk parity. The Journal of Risk. 2014. June. Vol. 16. P. 53–79. DOI: 10.21314/JOR.2014.284.
14. Luxenberg E., Boyd S. Portfolio construction with Gaussian mixture returns and exponential utility via convex optimization. Optimization and Engineering. 2024. Vol. 25, no. 1. P. 555–574. DOI: 10.1007/s11081-023-09814-y.
15. Indrė Morkūnaitė D.C., Leipus R. Evaluation of Value-at-Risk (VaR) using the Gaussian Mixture Models. Research in Statistics. 2024. Vol. 2, no. 1. P. 2346075. DOI: 10.1080/27684520.2024.2346075.
16. Ruan L., Yuan M., Zou H. Regularized Parameter Estimation in High-Dimensional Gaussian Mixture Models. Neural Computation. 2011. Vol. 23. P. 1605–1622. DOI: 10.1162/NECO_a_00128.
17. Halbe Z., Bortman M., Aladjem M. Regularized Mixture Density Estimation with an Analytical Setting of Shrinkage Intensities. IEEE Transactions on Neural Networks and Learning Systems. 2013. Vol. 24. P. 460–470. DOI: 10.1109/TNNLS.2012.2234477.
18. Moon T. The expectation-maximization algorithm. IEEE Signal Processing Magazine. 1996. Vol. 13, no. 6. P. 47–60. DOI: 10.1109/79.543975.

19. Tipping M.E., Bishop C.M. Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society Series B. 1999. Vol. 61, no. 3. P. 611–622. DOI: 10.1111/1467-9868.00196.

20. Tipping M.E., Bishop C.M. Mixtures of Probabilistic Principal Component Analyzers. Neural Computation. 1999. Feb. Vol. 11, no. 2. P. 443–482. DOI: 10.1162/089976699300016728.

21. Lyu Y., Zhou L., Cong Y., *et al.* Multirate mixture probability principal component analysis for process monitoring in multimode processes. IEEE Transactions on Automation Science and Engineering. 2023. Vol. 21, no. 2. P. 2027–2038. DOI: 10.1109/TASE.2023.3253285.

22. Tra V., Amayri M., Bouguila N. Unsupervised fault detection for building air handling unit systems using deep variational mixture of principal component analyzers. IEEE Transactions on Automation Science and Engineering. 2024. Vol. 21, no. 4. P. 6787–6803. DOI: 10.1109/tase.2023.3331347.

23. Zhang J., Chen M., Hong X. Nonlinear process monitoring using a mixture of probabilistic PCA with clusterings. Neurocomputing. 2021. Vol. 458. P. 319–326. DOI: 10.1016/j.neucom.2021.06.039.

24. Zhao B., Xiao X., Zhang W., *et al.* Self-Paced Probabilistic Principal Component Analysis for Data with Outliers. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. P. 3737–3741. DOI: 10.1109/icassp40776.2020.9054487.

25. Chao Han S.L., House L. Covariance-Guided Mixture Probabilistic Principal Component Analysis (C-MPPCA). Journal of Computational and Graphical Statistics. 2015. Vol. 24, no. 1. P. 66–83. DOI: 10.1080/10618600.2014.891460.

26. Tzagkarakis G., Caicedo-Llano J., Dionysopoulos T. Exploiting market integration for pure alpha investments via probabilistic principal factors analysis. Journal of Mathematical Finance. 2013. Vol. 3, no. 1. P. 192–200. DOI: 10.4236/jmf.2013.31A018.

27. Begušić S., Kostanjčar Z. Cluster-Specific Latent Factor Estimation in High-Dimensional Financial Time Series. IEEE Access. 2020. Vol. 8. P. 164365–164379. DOI: 10.1109/ACCESS.2020.3021898.

28. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974. Dec. Vol. 19, no. 6. P. 716–723. DOI: 10.1109/TAC.1974.1100705.

29. Brutti Righi M., Sergio Ceretta P. On the existence of an optimal estimation window for risk measures. Applied Finance Letters. 2015. Nov. Vol. 4, no. 1-2. P. 28–32. DOI: 10.24135/afl.v4i1and2.30.

30. Volkov N.V. Estimating VaR of diversified portfolios using PCA and PPCA dimensionality reduction methods. Proceedings of MIPT. 2025. Vol. 17, no. 1. P. 127–141. (in Russian).

31. Kupiec P.H. Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives. 1995. Vol. 3, no. 2. P. 73–84. DOI: 10.3905/jod.1995.407942.

32. Lam S.K., Pitrou A., Seibert S. Numba: A LLVM-based Python JIT Compiler. Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC, LLVM 2015, Austin, TX, USA, November 15, 2015. ACM, 2015. P. 1–6. DOI: 10.1145/2833157.2833162.

33. Ypma T.J. Historical development of the Newton–Raphson method. SIAM Review. 1995. Vol. 37, no. 4. P. 531–551. DOI: https://doi.org/10.1137/1037125.

34. Volkov N. MPPCA for VaR Estimation: Source Code Repository. 2025. URL: https://gitlab.com/n.volkovsky/mppca-for-var-estimation (accessed: 20.03.2025).

35. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, Statistics. Vol. 5. University of California Press. 1967. P. 281–298.

36. Lloyd S. Least squares quantization in PCM. IEEE Transactions on Information Theory. 1982. Vol. 28, no. 2. P. 129–137. DOI: 10.1109/TIT.1982.1056489.