НИУ ВШЭ ФКН x ФЭН

Экономика и анализ данных

Москва 2025

Приложение метода главных компонент к формированию инвестиционного портфеля.

Аляутдинов Карим БЭАД245

Мотивировка

- Современные финансовые портфели включают десятки и сотни активов.
- Из-за высокой величины портфеля сложно оценить его риск.
- Классические методы оценки риска (например, классический VaR) теряют точность при коррелированных активах.

Почему важно оценивать риск портфеля?

- Оценка риска — фундамент эффективного портфельного управления.
- От качества модели риска зависит надёжность стратегий и точность прогнозов потерь.
- После кризисов 2008 и 2020 годов внимание исследователей и инвесторов вновь сосредоточилось на факторных моделях, способных глубже описывать рыночную структуру и взаимосвязи между активами.

Что такое VaR?

Value-at-Risk (VaR) — это статистическая мера, оценивающая максимально возможный убыток портфеля за заданный период времени при определённом уровне доверия.

$$VaR_\alpha = -q_\alpha(r_t)$$

$q_\alpha(r_t)$ — квантиль распределения доходностей портфеля на уровне доверия α

Пример:
1-дневный VaR на уровне 95% = 2% означает, что с вероятностью 95% потери портфеля за день не превысят 2%.

VaR через ковариацию:

$$VaR_\alpha = z_\alpha \sqrt{w^T \Sigma w}$$

ω — веса портфеля

Σ — ковариационная матрица доходностей

PCA (метод главных компонент)

Основная идея данного метода понижения размерности заключается в поиске q-мерной гиперплоскости в d-мерном пространстве, проекция исходных данных на которую обладает максимальной дисперсией

- ***PCA ищет главные компоненты, вдоль которых дисперсия данных максимальна.***
- ***Позволяет уменьшить размерность данных без сильной потери информации.***

PCA + VaR

- ***Вместо работы с исходными активами используем главные компоненты (PC).***
- ***Ковариация главных компонент отражает основной рыночный риск.***
- ***Это позволяет более устойчиво и точно оценивать VaR портфеля, уменьшая влияние шумных корреляций между отдельными активами.***

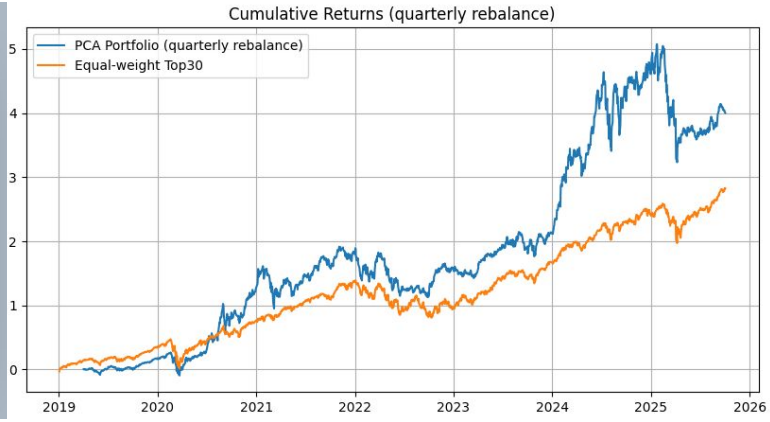
Замечание: по сути каждая компонента - линейная комбинация активов, которая показывает наибольшее направление дисперсии

PCA + VaR

Для применения метода PCA необходимо определить минимальное количество компонент, достаточное для восстановления распределения исходного портфеля с приемлемым качеством, и произвести оценку VaR 5%, используя только выбранные с помощью PCA компоненты. В качестве размера окна для оценки обучения PCA и оценки VaR подобрано оптимальное значение в 350 дней, которое балансирует между исторической значимостью и актуальностью данных.

Пример кода PCA на python

```python
def cov_and_pca(returns, n_components=None):
    R = returns - returns.mean()
    lw = LedoitWolf().fit(R.values)
    cov = lw.covariance_
    if n_components is None:
        n_components = min(10, R.shape[1])
    pca = PCA(n_components=n_components)
    pca.fit(R.values)
    components = pd.DataFrame(pca.components_.T, index=returns.columns, columns=[f'PC{i+1}' for i in
range(pca.components_.shape[0])])
    return cov, components, pca.explained_variance_ratio_
```



[Ссылка на полный код]

Статистика Андерсона-Дарлинга

Пусть $X = \{x_i\}_{i=1}^{n}$ и $Y = \{y_i\}_{j=1}^{m}$ — две выборки, в каждой из которых элементы отсортированы по возрастанию. Положим $Z = \{z_i\}_{i=1}^{k}$, $k = m + n$, — объединение X и Y, упорядоченное по возрастанию элементов.

Определение 2. Статистикой Андерсона – Дарлинга называется величина

$$\mathrm{AD} = \frac{1}{nm} \sum_{i=1}^{k-1} \frac{(kc_i - mi)^2}{i(k-i)},$$

где $c_i = |\{x : x \in X : x \leq z_i\}|, i = 1, \ldots, k.$

Гипотеза о совпадении распределений, из которых сгенерированы две выборки, отвергается, если значение статистики Андерсона – Дарлинга превышает критическое значение, которое вычисляется исходя из уровня значимости и размеров выборок.

Минусы PCA

- ***PCA работает хорошо для диверсифицированных портфелей, но показывает проблемы на слабо диверсифицированных (на следующем слайде).***
- ***В слабо диверсифицированных портфелях веса концентрируются в небольшой группе активов → искажения в оценке VaR 5%.***
- ***PCA детерминирован и нацелен на максимизацию объяснённой дисперсии, не моделирует шум и не даёт вероятностной интерпретации компонент.***
- ***Не всегда точно выделяет ключевые факторы риска.***
- ***PPCA, напротив, учитывает стохастическую природу (по сути учитывает шум) данных и чаще правильно оценивает риск на слабо диверсифицированных портфелях.***

Сравнение PPCA и PCA

На диверсифицированных портфелях оба метода показывают хорошие результаты — проходят тесты на VaR.

На слабо диверсифицированных портфелях:

- *PCA: менее точная оценка VaR 5%, не учитывает стохастический шум, плохо моделирует хвосты распределений (сильно отклоняющиеся от среднего данные).*
- *PPCA: сохраняет высокую точность, корректно восстанавливает портфель, учитывая случайный компонент доходностей.*

Вывод: PPCA предпочтительнее для портфелей с высокой концентрацией активов и нестандартными распределениями доходностей.

Таблица 1

Слабо диверсифицированные портфели

$VaR_{5\%}$	Industrials		Financials		Health Care		Все акции		100 случайных	
	PCA	PPCA	PCA	PPCA	PCA	PPCA	PCA	PPCA	PCA	PPCA
Бин Тест	14 %	91 %	71 %	91 %	4 %	92 %	0 %	78 %	1 %	87 %
Д.и. 95%	14 %	91 %	73 %	91 %	5 %	92 %	0 %	78 %	1 %	87 %
Д.и. 99%	29 %	98 %	81 %	99 %	7 %	96 %	0 %	90 %	4 %	91 %

Таблица 2

Диверсифицированные портфели

$VaR_{5\%}$	Industrials		Financials		Health Care		Все акции		100 случайных	
	PCA	PPCA	PCA	PPCA	PCA	PPCA	PCA	PPCA	PCA	PPCA
Бин Тест	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Д.и. 95%	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Д.и. 99%	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %

PPCA (Probabilistic PCA)

Идея: PPCA расширяет обычный PCA, вводя **вероятностную модель** для данных. Каждый объект $x_i \in \mathbb{R}^d$ описывается как:

$$x_i - \bar{x} = W y_i + \epsilon$$

где:

- $y_i \sim \mathcal{N}(0, I_q)$ — латентные переменные размерности $q < d$,
- W — матрица весов $d \times q$,
- $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ — независимый гауссовский шум.

Оценка параметров: Параметры (W, σ) находятся через максимизацию логарифма правдоподобия:

$$\mathcal{L}(X; W, \sigma) = -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\det(C) - \frac{1}{2}\sum_{i=1}^{n} x_i^T C^{-1} x_i$$

Используется EM-алгоритм для итеративной оптимизации.

Итоги

- Было рассмотрено 2 метода сжатия компонент, PCA и PPCA
- Оба показали хорошие результаты на диверсифицированных портфелях
- PPCA отлично справился с задачей и на выборе из слабо диверсифицированных портфелях


HSE UNIVERSITY
EST 1992