

Quantas Dimensões tem o Zoo? Redundância Informacional entre Características de Retorno no Mercado Acionário Brasileiro

Calebe Howang

Pedro Berta

Vitor Guardiano

Resumo

Este trabalho investiga, para o universo de ações do índice IBRA entre 2010 e 2024, quantas dimensões de risco e retorno genuinamente distintas um conjunto amplo de características candidatas, o chamado *factor zoo*, efetivamente carrega. A partir de dados da Economatica para 322 ativos, foram construídas 45 características candidatas não redundantes sob transformação de postos, organizadas em oito famílias (valor, momentum, tamanho, liquidez, lucratividade, investimento/crescimento, alavancagem e risco), padronizadas por uma transformação de postos *cross-sectional* e sintetizadas em uma matriz de correlação média representativa do período. Sobre essa matriz, aplicaram-se duas técnicas complementares de aprendizado não supervisionado: Análise de Componentes Principais (PCA) e *clustering* hierárquico com ligação por média, com base em uma métrica de distância derivada da correlação absoluta, para estimar a dimensionalidade latente do conjunto e selecionar um subconjunto de características representativas, mutuamente pouco redundantes. Os resultados indicam que 12 componentes principais explicam 81,96% da variância conjunta das características, e que 11 *clusters* emergem do critério de corte adotado no *clustering* hierárquico. Oito dos onze *clusters* (73%) carregam predominantemente em um único componente principal, evidência cruzada de que as duas técnicas leem a mesma estrutura de comunalidade, ainda que por critérios matemáticos distintos. Três características (o retorno de 1 mês, o *book-to-market* e o EBIT/capitalização) destacam-se como *singletons* no dendrograma. O conjunto de 11 características representativas selecionadas situa-se na mesma ordem de grandeza das estimativas obtidas para outros mercados por metodologias distintas, ainda que comparações diretas exijam cautela, dada a diferença de construtos medidos. Os resultados são interpretados como evidência de redundância informacional entre características, não como confirmação de que as dimensões identificadas correspondem a fatores de risco efetivamente precificados pelo mercado brasileiro, e estão condicionados a escolhas metodológicas explicitadas ao longo do trabalho, entre elas a hipótese de estabilidade temporal da estrutura de correlação estimada.

1 Introdução

Fatores de risco podem ser definidos, na linha de Berkin e Swedroe (2016), como características, ou conjuntos de características, compartilhadas por um amplo grupo de ativos que,

simultaneamente, ajudam a explicar seus retornos e sustentam prêmios sistematicamente superiores aos do mercado como um todo. Historicamente, o desenvolvimento desse arcabouço remonta ao modelo de precificação de ativos de capital (CAPM) de Sharpe (1964), no qual o único fator relevante era o prêmio de risco de mercado, ou seja, o retorno adicional, acima da taxa livre de risco, exigido por investidores para deter ativos arriscados. As décadas seguintes viram a incorporação sucessiva de novas dimensões de risco e retorno, sendo tamanho e valor incorporados no modelo de três fatores de Fama e French (1993), momento no modelo de quatro fatores de Carhart (1997), e lucratividade e investimento no modelo de cinco fatores de Fama e French (2015), em um processo contínuo de expansão do conjunto de variáveis consideradas relevantes para explicar o retorno esperado das ações.

Essa expansão, no entanto, ultrapassou rapidamente a capacidade da literatura de distinguir fatores genuinamente novos de versões redundantes de fatores já conhecidos. Em discurso presidencial à American Finance Association, Cochrane (2011) cunhou o termo *factor zoo* para descrever essa passagem do fator de mercado do CAPM como um único fator explicativo para um excesso de variáveis concorrentes propostas pela pesquisa acadêmica, um diagnóstico que a literatura seguinte aprofundou e, em parte, contestou.

Este trabalho parte dessa segunda leitura do problema. Testar a significância estatística de fatores isoladamente é uma abordagem sujeita aos vieses de mineração e múltiplos testes já documentados pela literatura. Assim, investigamos a estrutura de correlação entre um conjunto amplo de características candidatas por meio de técnicas de aprendizado não supervisionado. A lógica subjacente parte de uma hipótese de trabalho, não de um fato estabelecido: se duas ou mais características carregam essencialmente a mesma informação sobre risco e retorno, é razoável esperar que essa redundância se manifeste, ao menos em parte, em alta correlação entre elas ao longo do tempo. A relação inversa não é garantida, na qual características com baixa correlação direta podem, ainda assim, capturar uma mesma dimensão de risco subjacente por vias distintas, e duas características correlacionadas podem discordar exatamente na pequena parcela em que reside o prêmio de risco, mas essa hipótese é o que torna operacionalizável a pergunta de pesquisa, e pode ser identificada sistematicamente por Análise de Componentes Principais (PCA), que resume a variância conjunta das características em um número menor de componentes ortogonais, e por *clustering* hierárquico, que agrupa as variáveis segundo sua similaridade.

A aplicação dessa estratégia ao mercado brasileiro enfrenta desafios distintos dos observados em mercados desenvolvidos, para os quais a maior parte da literatura de redução de dimensionalidade de fatores foi originalmente desenvolvida. No cenário doméstico, o universo de ações líquidas é mais restrito, o histórico de dados confiáveis é mais curto, e a cobertura de determinadas características, sobretudo as que dependem de dados contábeis trimestrais, é mais heterogênea entre as empresas listadas. Por esse motivo, restringimos a amostra ao período de 2010 a 2024 e ao universo de ações do índice IBRA (Índice Brasil Amplo, o índice de ações mais abrangente da B3, com maior número de empresas do que o Ibovespa), e optamos por sintetizar a estrutura de correlação de todo o período em uma única matriz representativa, construída pela média das correlações mensais, em vez de estimar e comparar matrizes distintas por subjanelas temporais. Essa escolha simplifica a aplicação das técnicas de redução de dimensionalidade e evita fragmentar a amostra em janelas curtas demais para estimar correlações de forma confiável, ao custo de assumir que a estrutura de redundância entre as características é razoavelmente estável ao longo do período.

A pergunta de pesquisa que orienta este trabalho é, portanto: quantos fatores genuinamente independentes existem no universo de ações do IBRA entre 2010 e 2024, medidos por Análise de Componentes Principais e *Clustering* Hierárquico sobre a estrutura de correlação média das características candidatas, e quais variáveis capturam essas dimensões com menor redundância?

2 Referencial teórico: o *factor zoo* e a busca por dimensionalidade latente

2.1 O debate do *factor zoo*

O diagnóstico crítico do *factor zoo* se aprofundou na década seguinte ao discurso de Cochrane (2011). McLean e Pontiff (2016) mostraram que o poder preditivo de 97 variáveis preditoras cai, em média, 26% fora da amostra original e 58% após a publicação acadêmica. Ainda assim, o número de fatores publicados continuou a crescer: Harvey e Liu (2019) catalogaram 382 fatores publicados até janeiro de 2019, evidenciando a magnitude do problema. Em contraponto a essa leitura cética, Jensen, Kelly e Pedersen (2023) mostraram que a maior parte dos fatores catalogados é replicável e pode ser reorganizada em um número bem menor de temas latentes, deslocando o problema de "quantos fatores são falsos" para "quantas dimensões genuinamente distintas existem por trás deles". Essa reorganização, porém, não decorre de um algoritmo puramente estatístico: os autores agrupam os 153 fatores em 13 temas combinando a afinidade econômica entre eles (a interpretação teórica do que cada fator supostamente mede) com a correlação empírica de seus retornos, em um processo que mistura julgamento humano e evidência quantitativa. É exatamente esse componente humano que a próxima subseção discute, ao situar este trabalho na tradição metodológica que busca substituí-lo por um critério inteiramente estatístico.

2.2 Redução de dimensionalidade em modelos fatoriais

Parte da literatura já tentou responder a essa pergunta por vias distintas da simples elevação do padrão de significância. Green, Hand e Zhang (2017) testaram simultaneamente 94 características em regressões de Fama-MacBeth, o procedimento padrão de duas etapas que estima uma regressão *cross-sectional* (comparando todos os ativos em um mesmo instante) a cada período e resume os coeficientes pela média no tempo ajustadas para *data-snooping* (o viés de encontrar relações "significativas" apenas por testar um número muito grande de hipóteses, a mesma lógica da mineração de dados mencionada na Introdução) e encontraram que pouco mais de uma dezena carrega informação genuinamente independente sobre retornos. Kozak, Nagel e Santosh (2018, 2020) mostraram, por Análise de Componentes Principais sobre os retornos dos fatores, que um pequeno número de componentes ortogonais é suficiente para precificar o *cross-section* de ativos; Lettau e Pelger (2020), com uma generalização da PCA que penaliza o erro de precificação (*risk-premium* PCA), chegam a um resultado semelhante com apenas cinco fatores latentes. Jensen, Kelly e Pedersen (2023), por sua vez, resolveram o problema por agrupamento: reorganizaram 153 fatores em 13 temas com base na afinidade

econômica e na correlação entre eles. Este trabalho busca unir as duas linhagens metodológicas, aplicando-as de forma conjunta à estrutura de correlação das características no mercado brasileiro. A diferença em relação à abordagem de Jensen, Kelly e Pedersen (2023) está no critério de agrupamento: em vez de reunir os fatores por afinidade econômica, como fizeram esses autores, este trabalho utiliza um algoritmo de *clustering* hierárquico aplicado diretamente às correlações observadas entre as características.

3 Universo de ativos e construção das características

3.1 Fonte dos dados, universo investível e variáveis brutas

Os dados vêm da plataforma Economatica, para o universo de ações que compõem ou compuseram o índice IBRA entre 1^o de janeiro de 2010 e 31 de dezembro de 2024. O IBRA foi escolhido por ser mais amplo que o Ibovespa, o que reduz o risco de a estrutura de correlação ser dominada por poucos nomes líquidos, e o recorte a partir de 2010 busca equilibrar cobertura contábil e comparabilidade, evitando agravar o *survivorship bias* típico de janelas mais longas no mercado brasileiro.

A extração reúne 322 *tickers* em um único arquivo, cada um mapeado, por posição de coluna, a um dicionário canônico de vinte e três variáveis brutas (preço, volume, capitalização, número de ações, principais itens contábeis como patrimônio líquido, lucro líquido, receita, ativos, passivos, CPV, CAPEX e EBITDA, além de múltiplos e razões já calculados pelo fornecedor). Checagens mínimas de qualidade resultam em 321 *tickers* efetivamente carregados. As razões e múltiplos prontos do fornecedor não entram diretamente como características, e sim são recalculados a partir dos campos brutos, para manter o controle metodológico sobre cada definição, com exceção de P/E, EV/EBIT e EV/EBITDA, usados apenas para inferir o EBIT implícito.

3.2 Tratamento do painel: calendário, atividade, *look-ahead* e variáveis derivadas

O painel diário alinha cada ativo a um calendário comum de negociação, com regras voltadas a três riscos de viés, sobrevivência, informação contábil antecipada e múltiplos de preço tratados como dado contábil.

Para evitar viés de sobrevivência, a contagem de lacunas de preço de cada ativo começa no seu primeiro preço válido, não no início da janela, de modo que empresas listadas depois de 2010 não sejam tratadas como inativas antes mesmo de existir. A inativação permanente só ocorre após sessenta dias úteis consecutivos sem preço, já dentro do período listado. Ao final, 130 *tickers* entraram na amostra depois do início da janela e 86 sofreram inativação permanente, resultando em um corte transversal médio de cerca de 215 empresas por mês.

Para evitar o uso antecipado de informação contábil (*look-ahead*), as variáveis fundamentais recebem defasagem de sessenta dias úteis antes de serem propagadas, compatível com os prazos regulatórios brasileiros. Os múltiplos prontos do fornecedor (P/E, EV/EBIT,

EV/EBITDA) não recebem essa defasagem, pois já combinam preço corrente com fundamento *trailing*, sendo usados apenas no próprio dia para inferir o EBIT implícito e depois descartados como fatores, o que evita transformar um múltiplo em um degrau artificial entre divulgações.

O *enterprise value* é construído como capitalização corrente mais passivos defasados, um *proxy* que preserva bem o ordenamento do EV do fornecedor mas fica sistematicamente acima do valor reportado, por não abater caixa (limitação retomada na seção 6.1). Na frequência mensal, usada na análise principal, cada ativo entra com seu último preço válido do próprio mês, não a última data do calendário comum, o que evita que feriados sem pregão para todos os *tickers* anulem o mês inteiro. O painel mensal final cobre 180 meses e cerca de 38,7 mil linhas.

3.3 Características candidatas construídas a partir da base tratada

Sobre o painel tratado, foram construídas 45 características candidatas, organizadas em oito famílias, valor (formas *yield*, como *book-to-market* e *earnings-to-price*), *momentum* (janelas de um a doze meses, com variantes *skip-month*), tamanho (log da capitalização), liquidez (volume, *turnover*, medida de Amihud), lucratividade (ROE, ROA, margens, *gross profitability*), investimento e crescimento (crescimento de ativos, receita, lucro e CAPEX), alavancagem (passivos sobre patrimônio e sobre ativos) e risco (volatilidade, assimetria e curtose).

O conjunto foi deliberadamente depurado, antes da análise, de três classes de redundância que inflariam artificialmente a contagem de dimensões do *zoo*, transformações monótonas de uma característica já presente, formas recíprocas de múltiplos cujo denominador pode trocar de sinal (mantendo-se apenas as formas *yield*, de denominador estruturalmente positivo), e variáveis quase constantes no *cross-section*, que apareceriam como dimensões isoladas apenas por um artefato de discretização. Essa depuração é o motivo pelo qual as seções 4 e 5 encontram 45, e não um número maior, de características (a seção 6.2 quantifica o efeito de não aplicar esses filtros).

3.4 Atribuição dos *scores*: transformação de Kozak, Nagel e Santosh

Antes da análise de correlação, cada característica é transformada, em cada data, na escala *cross-sectional* proposta por Kozak, Nagel e Santosh (2018, 2020), cada ativo recebe um posto relativo entre 0 e 1, centrado pela média do *cross-section* e normalizado para que os pesos longos e curtos somem um em módulo. O resultado é um sinal *long-short* de alavancagem unitária, comparável entre características originalmente em unidades muito diferentes.

Um detalhe deliberado é que, quando todos os ativos empatam no posto, o valor permanece ausente e não é preenchido com zero, o que preserva a correlação *pairwise-complete* e a seletividade dos filtros de cobertura seguintes (o efeito do preenchimento por zero é quantificado na seção 6.2). Após a transformação, retêm-se apenas características com cobertura mínima de trinta ativos válidos em pelo menos 80% das datas e com variação suficiente para não serem consideradas degeneradas. As 45 características satisfazem ambos os critérios, resultando no painel de *z-scores* mensal que alimenta a matriz de correlação da seção 4.

4 Estrutura metodológica: PCA e *clustering* hierárquico sobre a matriz de correlação média

4.1 Construção da matriz de correlação representativa

A unidade de análise não é o retorno dos portfólios de fatores, e sim a estrutura de redundância entre as próprias características $z(c)$ no *cross-section*. Em cada data t , calcula-se a correlação de Pearson entre pares de características sobre os ativos com observação conjunta válida. Exigem-se pelo menos 30 observações válidas por coluna para que a coluna entre no cálculo daquela data. Ausências e infinitos são excluídos; pares sem amostra suficiente não entram na agregação. O Pearson é a escolha natural porque a pergunta é linear: “essas duas características ordenam o *cross-section* de forma proporcionalmente semelhante?”, e porque a distância usada no *clustering* (seção 4.3) é uma transformação monótona de $|\rho|$.

A matriz representativa do período é a média ponderada, ao longo do tempo, das correlações mensais, com peso igual ao número de ativos que entram no cálculo do par na data t :

$$\bar{\rho}(c, c') = \frac{\sum_t (n(c, c') \times \rho(c, c') \text{ em cada data } t)}{\sum_t n(c, c') \text{ em cada data } t}$$

Essa ponderação reduz a influência de datas com *cross-section* fino, problema recorrente no Brasil, e produz uma única matriz 45×45 a partir das correlações mensais. Como a média ponderada de correlações por data não preserva semidefinição positiva, a matriz utilizada na PCA é projetada no cone PSD (autovalores negativos *clipados* e diagonal renormalizada para 1) antes da decomposição espectral: sem esse reparo, autovalores negativos seriam artefatos numéricos da agregação, não propriedade econômica das características.

A interpretação econômica dessa matriz é deliberadamente restrita. $\bar{\rho}(c, c')$ alto indica que as duas características ordenam o *cross-section* de forma semelhante ao longo do período, não que ambas capturam o mesmo prêmio de risco em regressões de precificação. Em outras palavras, o objeto é a redundância informacional das características, e não a dimensionalidade do SDF, o *stochastic discount factor* (fator de desconto estocástico). Essa distinção é crítica: resultados de PCA/*clustering* aqui falam de quantas dimensões distintas há na nuvem de *signals*, e não de quantos fatores precificam o mercado.

Empiricamente, a matriz mensal revela correlação média *cross*-característica próxima de zero (média com sinal $\approx 0,09$), o que é esperado em um conjunto heterogêneo de famílias; a média dos valores absolutos *off-diagonal* é $\approx 0,21$. Dentro de famílias, a redundância permanece elevada, o que o *heatmap* da Figura 1 torna visível como blocos de cor intensa. Como a imagem 45×45 é densa para leitura pontual, três pares ilustram o que a escala de cores está dizendo:

1. Alavancagem quase idêntica: *lev_liab_to_assets* e *lev_liab_to_equity* têm $\bar{\rho} \approx 0,99$ (bloco vermelho escuro). Passivo/ativos e passivo/patrimônio ordenam o *cross-section* de forma praticamente igual, daí o *cluster* 11 da seção 4.3.
2. Liquidez com sinal invertido: Amihud (*liq_amihud_21d*) e volume médio em 21 dias (*liq_avg_vol_21d*) têm $\bar{\rho} \approx -0,98$ (bloco azul escuro). Correlação

negativa forte ainda é redundância informacional: iliquidez Amihud sobe quando o volume cai. Por isso, a distância do *clustering* usa $|\rho|$, não ρ .

3. Pouco em comum entre famílias distintas: *book-to-market* e retorno de 1 mês têm $\bar{\rho} \approx -0,08$ (cor próxima do neutro/branco). São sinais que quase não se alinham no período, coerentes com o aparecimento como *singletons* em *clusters* separados.

Outros pares na mesma lógica: volume médio versus volume em reais ($\bar{\rho} \approx 0,97$) e *momentum* 12 meses versus 12-1 ($\bar{\rho} \approx 0,96$). Esse padrão já antecipa que o "zoo" de 45 variáveis não corresponde a 45 dimensões independentes.

4.2 Análise de componentes principais (PCA)

A Análise de Componentes Principais (PCA) é uma técnica de aprendizado não supervisionado que resume a variação conjunta de um grande número de variáveis correlacionadas em um número menor de dimensões, chamadas componentes principais (James et al., 2013, cap. 10.2). Cada componente é uma combinação linear das 45 características candidatas, construída para capturar a maior variância conjunta possível, sujeita à restrição de ser ortogonal (não correlacionada) aos componentes já extraídos (James et al., 2013, cap. 10.2.1). A PCA responde, portanto, à pergunta “*quantas* direções ortogonais bastam para descrever a nuvem?”, não “*quais* variáveis nomeadas representam cada direção”.

Os *loadings* (correlação entre a característica original e aquele componente) reportados na Tabela 1 não são os autovetores brutos, mas cada autovetor multiplicado pela raiz quadrada do autovalor correspondente, o que dá a correlação entre cada característica original e o componente (Hastie, Tibshirani e Friedman, 2009, cap. 14.5.1). Como o traço de uma matriz de correlação (após o reparo PSD) é igual ao número de variáveis, a soma de todos os autovalores é 45, e a proporção de variância explicada por um componente é simplesmente seu autovalor dividido por 45.

Não existe um critério único e objetivo para decidir quantos componentes reter e James et al. (2013, cap. 10.2.3) são explícitos quanto a isso. Um critério alternativo e amplamente usado para matrizes de correlação é a regra de Kaiser (reter componentes com autovalor > 1). Neste trabalho, optamos por reter o menor número de componentes que, em conjunto, expliquem pelo menos 80% da variância total, um patamar convencional na prática de análise fatorial, escolhido por priorizar a descrição da nuvem completa em vez de um limiar por componente isolado.

Aplicada à matriz de correlação mensal $\bar{\rho}$, a PCA indica que 12 componentes são suficientes para atingir esse limiar (81,96% da variância, ver Figura 2), uma redução de 73% no número de dimensões necessárias para descrever a estrutura de correlação das 45 características. A Tabela 1 resume os componentes retidos. A coluna “características dominantes” lista as variáveis com maior associação ao componente, retenidas quando o valor absoluto do *loading* é pelo menos 0,30 (e até oito nomes por PC, por legibilidade). O número entre parênteses é o próprio *loading*: correlação entre a característica original e aquele componente. Sinal positivo significa que a característica se move no mesmo sentido do PC; sinal negativo, no sentido oposto. Por exemplo, no PC1, margem líquida (+0,82) e ROA (+0,81) estão fortemente alinhadas à dimensão dominante de rentabilidade/tamanho; no PC2, *ebit_to_ev*

(−0,51) entra com sinal contrário ao bloco de *momentum*, o que não a torna “irrelevante”, e sim oposta àquela direção.

O primeiro componente, sozinho, explica 24,32% da variância total e agrupa características de rentabilidade (margem líquida, ROA, ROE, EBITDA/ativos, margem EBITDA), *earnings-to-price* e tamanho (logaritmo da capitalização), um resultado que sugere que, no mercado brasileiro, empresas maiores tendem também a ser mais rentáveis, formando uma dimensão dominante de qualidade/tamanho. O segundo componente (10,33%) reúne majoritariamente medidas de *momentum* de médio/longo prazo. Os componentes seguintes recombina, em proporções decrescentes, as famílias de liquidez, valor por vendas, alavancagem, investimento e risco já descritas na seção 3.3, sem que nenhuma família aloque para um único componente exclusivo, o que já antecipa a sobreposição entre famílias que o *clustering* hierárquico da subseção 4.3 explora de forma complementar.

Uma métrica que acompanha os *loadings* é a comunalidade de cada característica: a soma dos quadrados de seus *loadings* nos 12 componentes retidos. Comunalidade baixa significa que a variável é mal recuperada pelos componentes retidos, resíduo idiossincrático relativo ao limiar de 80%. As características com menor comunalidade neste conjunto são o crescimento do patrimônio líquido *yoy* (0,51), o crescimento de receita *yoy* (0,57), o crescimento de ativos *yoy* (0,59), a margem bruta (0,60) e o *dividend yield* (0,63). Nenhuma delas é um *singleton* no dendrograma da seção 4.3: isolamento no *clustering* e baixa comunalidade na PCA medem propriedades distintas e não precisam coincidir.

Por fim, vale registrar uma tensão entre os critérios de corte: pela regra de Kaiser, o 12º componente (autovalor 0,995) ficaria de fora do corte, por estar ligeiramente abaixo de 1, apenas o critério de variância acumulada de 80% o inclui. Essa sensibilidade do resultado à escolha do critério é retomada como limitação na seção 6.

4.3 Clustering hierárquico

O *clustering* hierárquico é uma técnica de agrupamento que, ao contrário do *k-means*, não exige que se especifique de antemão o número de grupos (James et al., 2013, cap. 10.3.2). Em sua versão aglomerativa, a empregada neste trabalho, o algoritmo parte de cada característica como seu próprio *cluster* e, a cada passo, funde os dois *clusters* mais próximos entre si, até que reste um único *cluster* contendo todas as 45 características. O resultado é o dendrograma (James et al., 2013, cap. 10.3.2; Hastie, Tibshirani e Friedman, 2009, cap. 14.3.12). Enquanto a PCA pergunta *quantas* dimensões ortogonais bastam, o *clustering* pergunta *quais* variáveis medem praticamente a mesma coisa, e devolve grupos de variáveis reais, não combinações lineares sintéticas.

Para transformar a matriz de correlação $\bar{\rho}$ (seção 4.1) em uma medida de distância utilizável pelo algoritmo, definimos

$$d(c, c') = \sqrt{2(1 - |\bar{\rho}(c, c')|)}$$

Características fortemente correlacionadas positivamente ou negativamente ficam próximas (distância perto de zero), enquanto características sem relação linear relevante ficam distantes (distância máxima de $\sqrt{2} \approx 1,414$). Usar o valor absoluto da correlação é essencial: duas características que são versões invertidas uma da outra carregam a mesma

informação apesar do sinal negativo e devem ser tratadas como redundantes, não como opostas. A forma funcional $\sqrt{2(1-\rho)}$, e não simplesmente $1-\rho$, é a distância euclidiana entre as duas características quando padronizadas (Hastie, Tibshirani e Friedman, 2009, cap. 14.3.2) e, diferente de $1-|\rho|$, satisfaz a desigualdade triangular mesmo após a aplicação do valor absoluto.

A distância entre *clusters* com mais de um membro é definida pelo método de ligação (*linkage*). Adotamos a ligação por média (*average linkage*), que toma a distância entre dois *clusters* como a média de todas as distâncias cruzadas entre os membros dos dois grupos (James et al., 2013, cap. 10.3.2; Hastie, Tibshirani e Friedman, 2009, cap. 14.3.12). Em relação a alternativas comuns: a ligação simples tende a encadear grupos alongados; a completa favorece *clusters* compactos e pode fragmentá-los demais; a média é um compromisso padrão quando o interesse é a similaridade média entre membros, coerente com a leitura de redundância informacional.

O dendrograma pode ser cortado em qualquer altura para produzir um número específico de *clusters*. Adotamos um corte equivalente a exigir correlação absoluta média de pelo menos 0,3 entre *clusters* para que permaneçam distintos, o que, na escala de distância $\sqrt{2(1-|\bar{\rho}|)}$, corresponde a uma altura de corte de $\sqrt{2(1-0,3)} \approx 1,183$, resultando em 11 *clusters* a partir das 45 características candidatas, como ilustra a Figura 3. O limiar $|\rho| = 0,3$ é uma convenção de “associação fraca a moderada”: abaixo disso, tratamos os blocos como dimensões distintas para fins de redução do *zoo*. Esse número está na mesma ordem de grandeza dos 12 componentes retidos pela PCA, mas não coincide exatamente com ele: a convergência entre os métodos não é a igualdade numérica dos cortes, e sim o alinhamento qualitativo entre *clusters* e componentes dominantes (seções 4.4 e 5.1). A Tabela 2 resume a composição de cada *cluster* e sua coesão interna. Coesão, aqui, é a média dos $|\bar{\rho}|$ entre todos os pares de membros do mesmo *cluster*: quanto mais perto de 1, mais as variáveis do grupo ordenam o *cross-section* de forma semelhante (redundância alta); quanto mais perto de 0, mais frouxo é o agrupamento. Em *clusters* com um único membro (*singletons*), não existem pares internos, a coesão é, por definição, indefinida e reportada como “n/a (singleton)”.

Três *clusters* contêm um único fator, isolado por sua distância média a qualquer outro grupo exceder o corte adotado: o retorno de 1 mês (reversão/momentum de curtíssimo prazo, separado do bloco de *momentum* de médio/longo prazo, coerente com a literatura que trata a reversão de 1 mês como fenômeno distinto do *momentum* 12-1), o *book-to-market* e o EBIT/capitalização. Esses *singletons* não coincidem com os de menor comunalidade na PCA, o que é esperado, porque comunalidade baixa e isolamento no dendrograma medem propriedades distintas. O crescimento do lucro líquido *yoy* agrega-se ao grande bloco de rentabilidade/valor operacional (C8), e não aparece isolado: sob a especificação deste trabalho, ele co-move com ROA, margens e *earnings-to-price*.

Entre os *clusters* mais coesos estão o par CAPEX/ativos e CAPEX/receita (0,835) e o par de alavancagem (0,946). No outro extremo, o par assimetria e curtose (0,312) mostra a coesão mínima ainda suficiente para permanecer agrupado no corte escolhido. O maior *cluster* (C8, 13 fatores, reunindo rentabilidade, *yield* operacional, risco e *earnings-to-price*) tem coesão de 0,475 e carrega de forma coerente no PC1, o agrupamento mais alinhado à dimensão dominante da PCA.

4.4 Seleção de representantes por *cluster* e elo PCA-*clustering*

O *clustering* hierárquico particiona as 45 características em 11 grupos, mas ao contrário da PCA, que sintetiza variáveis novas, não produz, por si só, nenhuma redução do conjunto de trabalho: apenas rotula quem é semelhante a quem. O representante de cada *cluster* fecha essa lacuna. Adotamos a lógica de *medoid* de Hastie, Tibshirani e Friedman (2009, cap. 14.3.10): o *medoid* é o membro real do grupo, não um centróide sintético, que maximiza a correlação absoluta média com os demais membros do próprio *cluster*. Em termos intuitivos, é a característica “mais no meio” do bloco, medida em similaridade. Diferentemente dos componentes principais, o representante é uma variável nomeada e economicamente interpretável.

Em *clusters* de exatamente dois membros, o critério estatístico empata por construção: a correlação média de cada um "com os demais" reduz-se à correlação do próprio par. Nesses três casos, CAPEX (C3), momentos de ordem superior (C5) e alavancagem (C11), o desempate não é deixado à ordem do dendrograma; aplica-se um critério econômico explícito, documentado abaixo. Nos demais *clusters* ($n \neq 2$), o *medoid* estatístico decide sozinho. A Tabela 3 apresenta o representante final de cada *cluster*, o componente principal que o *cluster* carrega predominantemente, e um rótulo econômico sucinto.

Nos três empates de par, a escolha econômica é a seguinte. No *cluster* 3, prefere-se CAPEX/ativos a CAPEX/receita porque a intensidade de investimento escalada pelo estoque de capital é a métrica alinhada ao fator de investimento de Fama e French (2015) e à literatura de crescimento de ativos (Cooper, Gulen e Schill, 2008); CAPEX/receita mistura intensidade de *capex* com a ciclicidade da receita e, portanto, é menos limpa como *proxy* de investimento. No *cluster* 5, prefere-se a assimetria à curtose: o momento de terceira ordem é o canal mais diretamente discutido na literatura de efeito loteria e retornos esperados (Bali, Cakici e Whitelaw, 2011), já citada neste trabalho; a curtose permanece no *cluster* como redundante sob o corte adotado, mas não como rótulo da dimensão. No *cluster* 11, prefere-se passivo/ativo a passivo/patrimônio porque a alavancagem sobre o ativo total é limitada ao intervalo unitário e coincide com a noção clássica de *book leverage* como fração do balanço; o quociente sobre o patrimônio líquido explode quando o *equity* contábil se aproxima de zero (situação frequente em *distress* e em partes da amostra brasileira) e, portanto, é menos estável como representante *cross-sectional*.

Oito dos onze *clusters* (73%) mapeiam predominantemente para um único componente principal. Os três que "dividem" carga (tamanho/liquidez, crescimento e valor por vendas/margens) são precisamente os que cruzam famílias econômicas, um resultado esperado porque os PCs são ortogonais por construção e os *clusters* não. Essa tabela cruzada é a evidência operacional da tese metodológica do trabalho: a PCA responde *quantas* dimensões independentes parecem existir; o *clustering* responde *quais* variáveis medem praticamente a mesma coisa; a sobreposição entre um grupo do dendrograma e um mesmo componente é o sinal de que ambas as lentes descrevem a mesma dimensão econômica.

O conjunto de 11 representantes listados acima é a resposta operacional da pergunta de pesquisa formulada na Introdução: substitui, na prática, as 45 características candidatas por 11 variáveis reais e não sintéticas, mutuamente pouco redundantes segundo os dois critérios empregados neste trabalho.

5 Resultados

5.1 Consistência entre PCA e *clustering* hierárquico

A Análise de Componentes Principais e o *clustering* hierárquico partem de critérios matemáticos completamente distintos, a primeira olha para a matriz de correlação como um todo e o segundo compara as características duas a duas, mas chegam a respostas compatíveis sobre a dimensionalidade do *zoo* de 45 características candidatas: 12 componentes bastam para explicar $\sim 82\%$ da variância conjunta (seção 4.2), e 11 clusters resultam do corte do dendrograma em $|\bar{\rho}| \geq 0,3$ (seção 4.3). A coincidência não é numérica exata, e não precisa ser, uma vez que o que importa para a pergunta de pesquisa é que 73% dos clusters carregam em um único PC dominante (Tabela 3), evidenciando que as duas lentes descrevem a mesma estrutura de comunalidade.

Os três *singletons* (retorno de 1 mês, *book-to-market* e EBIT/capitalização) devem ser lidos com a cautela própria de qualquer isolamento sob um corte específico de $|\rho| = 0,3$: são dimensões distintas *relativamente a esse limiar*, não afirmações de que nenhuma outra característica do *zoo* compartilhe informação parcial com elas. Ainda assim, sua interpretação econômica é direta: a reversão de curtíssimo prazo é tratada pela literatura como fenômeno separado do *momentum* de médio prazo; o *book-to-market* permanece o *proxy* clássico de valor contábil; e o EBIT/capitalização captura um *yield* operacional sobre o preço que, neste *cross-section*, não se funde ao bloco de rentabilidade do PC1 no mesmo grau que o ROA ou o E/P.

Vale registrar que essa consistência entre PCA e *clustering* não constitui duas fontes independentes de evidência no sentido estatístico do termo: ambos operam sobre a mesma matriz $\bar{\rho}$ (seção 4.1). Ainda assim, não era garantido que as duas lentes apontassem para a mesma estrutura de blocos; a taxa de 73% de alinhamento *cluster*–PC é o que credencia a interpretação conjunta.

5.2 O *zoo* brasileiro em perspectiva internacional

O resultado de cerca de 11–12 dimensões genuinamente distintas encontrado para o mercado brasileiro situa-se na mesma ordem de grandeza da estimativa de Jensen, Kelly e Pedersen (2023), que reorganizam 153 fatores em 13 temas latentes a partir de uma amostra de 93 países.

A comparação com Green, Hand e Zhang (2017) permanece matizada. Os autores encontram 12 características com informação independente sobre retornos nos Estados Unidos ao longo de 1980–2014 como um todo, mas esse número cai para apenas 2 a partir de 2003, período que se sobrepõe à janela analisada neste trabalho (2010–2024). A distância entre ~ 12 (Brasil, redundância de sinais) e 2 (EUA pós-2003, significância de retorno) não deve ser lida como evidência de que o mercado brasileiro é estruturalmente menos redundante do que o americano: os dois números medem construtos diferentes. Além disso, a própria queda de 12 para 2 nos Estados Unidos ilustra exatamente o tipo de instabilidade temporal que a seção 6 reconhece como um risco não testado neste trabalho.

5.3 Síntese: dimensionalidade e representantes do *zoo* brasileiro

A resposta à pergunta de pesquisa que orienta este trabalho é, portanto, dupla e convergente em substância: o universo de 45 características candidatas ao estilo do *factor zoo*, construído para o índice IBRA entre 2010 e 2024, carrega cerca de 12 dimensões informacionais distintas na lente da PCA ($\sim 82\%$ da variância) e 11 blocos de redundância na lente do *clustering*, não 45. A Tabela 3 apresenta a lista operacional dessas dimensões: 11 características reais, *medoid* quando $n \neq 2$, e desempate econômico nos três pares, que substituem o conjunto original preservando a coesão intra-*cluster* reportada na Tabela 2.

6 Discussão crítica e limitações

6.1 Limitações substantivas do desenho

Uma primeira limitação diz respeito à hipótese de estabilidade temporal assumida na construção da matriz de correlação representativa (seção 4.1). Ao sintetizar toda a estrutura de correlação do período 2010–2024 em uma única matriz referente à média ponderada das correlações mensais, o trabalho pressupõe implicitamente que a redundância entre as características não varia de forma relevante ao longo do tempo. Essa hipótese nunca foi testada diretamente: não foram estimadas matrizes por subperíodos, nem foram comparados os *clusters* ou componentes que resultariam de janelas temporais distintas. Se a estrutura de redundância tiver mudado, em razão de crises de liquidez, mudanças regulatórias ou ciclos de mercado, as dimensões reportadas na seção 5 podem refletir uma média que não descreve bem nenhum subperíodo específico.

Uma segunda limitação é de interpretação, não de execução: os resultados deste trabalho medem redundância informacional entre características, não a dimensionalidade do modelo de precificação de ativos (o SDF) do mercado brasileiro, distinção já registrada na seção 4.1. Duas características podem ordenar o *cross-section* de forma semelhante ao longo do tempo, e por isso aparecer no mesmo *cluster* ou carregar peso no mesmo componente, sem que ambas sejam precificadas, ou sem que nenhuma das duas seja. Confirmar se as dimensões identificadas correspondem a fatores de risco efetivamente precificados exigiria uma etapa adicional de testes de precificação, por exemplo, regressões de Fama-MacBeth ou testes GRS (Gibbons-Ross-Shanken), que não foi realizada aqui.

Uma terceira limitação decorre da sensibilidade dos resultados às escolhas metodológicas feitas ao longo do trabalho, todas razoáveis, mas não únicas. Na PCA (seção 4.2), a regra de Kaiser (autovalor > 1) excluiria o 12^o componente, retido apenas pelo critério de variância acumulada de 80%, gerando um resultado de 11 componentes, numericamente alinhado ao número de *clusters*, mas igualmente defensável sob outro critério. No *clustering* hierárquico (seção 4.3), o corte adotado (equivalente a $|\rho| \geq 0,3$) é uma escolha específica; um corte ligeiramente diferente produziria um número diferente de *clusters*.

Uma quarta limitação está nos próprios dados e nas premissas operacionais do painel. O *lag* contábil de 60 dias úteis (seção 3.2) reduz, mas não elimina, o risco de usar informação contábil antes de sua divulgação pública. A inativação permanente após o primeiro *gap* longo de 60 pregões, embora inicie apenas após a listagem, ainda exclui ou reduz severamente ativos com negociação esporádica. O *enterprise value* utilizado é um *proxy* deliberado, capitaliza-

ção de mercado corrente mais passivos defasados, sem abatimento de caixa, porque o *export* não traz colunas separadas de caixa nem de dívida financeira líquida. Esse *proxy* preserva bem o ordenamento *cross-sectional* do EV do fornecedor, mas o nível fica sistematicamente acima do EV reportado (exatamente pelo não abatimento de caixa); razões que usam EV no denominador (EBITDA/EV, EBIT/EV, *sales-to-EV*) herdaram esse viés de nível, embora o *ranking* Kozak mitigue parte do efeito. Uma reconstrução alternativa via $(EV/EBITDA)_t \times EBITDA_{lag}$, que embutiria o caixa já descontado pelo fornecedor no múltiplo, foi considerada e descartada para a especificação reportada: alterava a partição de *clusters* e o alinhamento *cluster*-PC sem melhorar a interpretação econômica da pergunta central, de modo que se manteve o *proxy* transparente e auditável da seção 3.2. O EBIT implícito só é definido quando $EV/EBIT > 0$, de modo que os fatores baseados em EBIT correm sobre uma subamostra pró-lucrativa. Bancos entram com cobertura plena em ROE e *book-to-market*, mas se autoexcluem de razões que dependem de passivo desagregado ou de EBIT/receita no *export* utilizado.

Por fim, o conjunto de 45 características candidatas, embora organizado para refletir as famílias clássicas da literatura de fatores (seção 3.3) e depurado de duplicatas monótonas e recíprocas com sinal quebrado, é uma amostra curada, não uma réplica dos 382 fatores catalogados por Harvey e Liu (2019). A conclusão de que o *zoo* se reduz a cerca de 11–12 dimensões vale, estritamente, para esse conjunto; um conjunto de partida mais amplo ou diferente poderia revelar uma estrutura de redundância distinta.

6.2 Lições de especificações alternativas e artefatos evitados

A versão final da metodologia resultou de um processo de teste e descarte de alternativas, cada uma delas produzindo uma contagem de dimensionalidade diferente da reportada, o que confirma que essas escolhas de tratamento dos dados são necessárias, e não arbitrárias.

Seis problemas de construção do painel foram identificados e corrigidos ao longo do processo, entre eles o tratamento do gap pré-IPO, o tratamento de múltiplos de preço como fundamentos defasados, o critério de *snapshot* mensal, o preenchimento de ausências, a presença de duplicatas monótonas e recíprocas, e um fator degenerado ligado aos dias negociados. Três desses problemas merecem destaque por seu potencial de gerar viés significativo nos resultados finais.

O primeiro é o viés de sobrevivência por gap pré-IPO, se a contagem de lacunas de preço começasse no início do calendário unificado, em vez de no primeiro preço válido de cada ativo, cerca de 130 tickers listados após 2010 seriam tratados como inativos, empobrecendo a amostra e distorcendo a estrutura de correlação em direção a um universo artificialmente mais sobrevivente.

O segundo é o preenchimento de ausências por zero, que dilui artificialmente as correlações entre características com cobertura distinta. O efeito é mensurável, a correlação teórica de $-1,00$ entre formas recíprocas, como E/P e P/E , caía para cerca de $-0,65$ sob esse tratamento, além de neutralizar os filtros de cobertura.

O terceiro é a manutenção de duplicatas monótonas e formas recíprocas com denominador de sinal variável, que infla artificialmente o número de características e pode quebrar a monotonicidade do posto, criando a falsa impressão de que uma variável é uma dimensão idiossincrática isolada, quando na verdade o resultado é apenas um artefato do sinal do

denominador.

Combinados, esses problemas levavam a uma contagem de aproximadamente 53 características, 14 componentes principais e 14 a 15 *clusters*, com *singletons* economicamente pouco interpretáveis. Corrigidos, o resultado converge para os números reportados na seção 5, 45 características, 12 componentes, 11 *clusters* e 73% de alinhamento *cluster*-PC, o que ilustra o risco de estimar a dimensionalidade do *zoo* sem antes auditar cuidadosamente o painel de dados.

7 Conclusão

Este trabalho investigou, para o universo de ações do índice IBRA entre 2010 e 2024, quantas dimensões de risco e retorno genuinamente distintas um conjunto de 45 características candidatas ao estilo do *factor zoo* efetivamente carrega. A resposta, obtida por duas técnicas de aprendizado não supervisionado operando sobre critérios matemáticos distintos (Análise de Componentes Principais e *clustering* hierárquico), converge em substância para um número muito menor do que 45: 12 componentes explicam $\sim 82\%$ da variância conjunta, 11 *clusters* particionam o dendrograma, e 73% desses *clusters* carregam em um único componente dominante. Onze características reais, medoid estatístico com desempate econômico nos três *clusters* de par, substituem o conjunto original preservando a coesão intra-cluster observada.

A contribuição deste trabalho está em aplicar, de forma sistemática e reproduzível, uma abordagem de redução de dimensionalidade, na linha metodológica de Kozak, Nagel e Santosh (2018, 2020) e de Jensen, Kelly e Pedersen (2023), à estrutura de correlação de um conjunto amplo de características no mercado acionário brasileiro, combinando PCA e *clustering* hierárquico de forma complementar, e explicitamente cruzada, em vez de recorrer a apenas uma das duas técnicas. Os resultados, no entanto, são condicionados às escolhas metodológicas documentadas ao longo do trabalho (a hipótese de estabilidade temporal da matriz de correlação, os critérios de corte adotados na PCA e no *clustering*, e o escopo do próprio conjunto de 45 características) e não estabelecem, por si só, que as dimensões identificadas correspondem a fatores de risco efetivamente precificados pelo mercado.

Testar essa precificação diretamente e verificar a estabilidade da estrutura de redundância encontrada em subperíodos distintos são extensões naturais deste trabalho. Por ora, fica registrado que o *zoo* de fatores, ao menos no recorte de características e período aqui analisado, é bem menor do que o número de suas espécies catalogadas sugere.

Referências

- BALI, Turan G.; ÇAKICI, Nusret; WHITEHEAD, Robert F. Maxing out: stocks as lotteries and the cross-section of expected returns. *Journal of Financial Economics*, v. 99, n. 2, p. 427-446, 2011.
- BANZ, Rolf W. The relationship between return and market value of common stocks. *Journal of Financial Economics*, v. 9, n. 1, p. 3-18, 1981.
- BERKIN, Andrew L.; SWEDROE, Larry E. *Your complete guide to factor-based investing: the way smart money invests today*. Buckingham: BAM Alliance Press, 2016.
- CARHART, Mark M. On persistence in mutual fund performance. *The Journal of Finance*, v. 52, n. 1, p. 57-82, 1997.
- COCHRANE, John H. Presidential address: discount rates. *The Journal of Finance*, v. 66, n. 4, p. 1047-1108, 2011.
- COOPER, Michael J.; GÜLEN, Huseyin; SCHILL, Michael J. Asset growth and the cross-section of stock returns. *The Journal of Finance*, v. 63, n. 4, p. 1609-1651, 2008.
- FAMA, Eugene F.; FRENCH, Kenneth R. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, v. 33, n. 1, p. 3-56, 1993.
- FAMA, Eugene F.; FRENCH, Kenneth R. A five-factor asset pricing model. *Journal of Financial Economics*, v. 116, n. 1, p. 1-22, 2015.
- GREEN, Jeremiah; HAND, John R. M.; ZHANG, Xiaoquan Frank. The characteristics that provide independent information about average U.S. monthly stock returns. *The Review of Financial Studies*, v. 30, n. 12, p. 4389-4436, 2017.
- HARVEY, Campbell R.; LIU, Yan. A census of the factor zoo. Working Paper, Duke University, 2019.
- HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The elements of statistical learning: data mining, inference, and prediction*. 2. ed. New York: Springer, 2009.
- JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. *An introduction to statistical learning: with applications in R*. New York: Springer, 2013.
- JENSEN, Theis I.; KELLY, Bryan T.; PEDERSEN, Lasse H. Is there a replication crisis in finance? *The Journal of Finance*, v. 78, n. 5, p. 2465-2518, 2023.
- KOZAK, Serhiy; NAGEL, Stefan; SANTOSH, Shrihari. Interpreting factor models. *The Journal of Finance*, v. 73, n. 3, p. 1183-1223, 2018.
- KOZAK, Serhiy; NAGEL, Stefan; SANTOSH, Shrihari. Shrinking the cross-section. *Journal of Financial Economics*, v. 135, n. 2, p. 271-292, 2020.

- LETTAU, Martin; PELGER, Markus. Factors that fit the time series and cross-section of stock returns. *The Review of Financial Studies*, v. 33, n. 5, p. 2274-2325, 2020.
- MCLEAN, R. David; PONTIFF, Jeffrey. Does academic research destroy stock return predictability? *The Journal of Finance*, v. 71, n. 1, p. 5-32, 2016.
- SHARPE, William F. Capital asset prices: a theory of market equilibrium under conditions of risk. *The Journal of Finance*, v. 19, n. 3, p. 425-442, 1964.

Apêndice

Tabela 1 — Componentes principais retidos (matriz de correlação mensal)

Componente	Autovalor	Var. explicada	Var. acumulada	Características dominantes (loading)
PC1	10,944	24,32%	24,32%	net_margin (+0,82), ROA (+0,81), ebitda_to_assets (+0,78), ROE (+0,75), ebit_to_assets (+0,73), ebitda_margin (+0,71), earnings_to_price (+0,65), size_log_mcap (+0,63)
PC2	4,649	10,33%	34,65%	mom_ret_6m (+0,72), mom_ret_6_1 (+0,67), mom_ret_12m (+0,66), mom_ret_12_1 (+0,63), mom_vol_adj_12_1 (+0,62), mom_ret_3m (+0,62), mom_ret_3_1 (+0,57), ebit_to_ev (−0,51)
PC3	4,227	9,39%	44,04%	sales_to_ev (−0,62), asset_turnover (−0,59), liq_dollar_vol (+0,58), liq_avg_vol_21d (+0,57), sales_to_price (−0,54), amihud (−0,54), liq_turnover (+0,45), liq_vol_per_share (+0,44)
PC4	3,629	8,06%	52,11%	asset_turnover (−0,62), sales_to_ev (−0,59), liq_dollar_vol (−0,57), liq_avg_vol_21d (−0,56), amihud (+0,52), liq_vol_per_share (−0,50), sales_to_price (−0,49), liq_turnover (−0,47)
PC5	2,397	5,33%	57,44%	liab_to_assets (+0,66), liab_to_equity (+0,57), book_to_market (−0,46), ebitda_growth_yoy (+0,41), sales_to_ev (−0,37), ebit_to_mcap (+0,36), ebit_margin (+0,33), dividend_yield (−0,31)
PC6	2,255	5,01%	62,45%	ebit_to_mcap (+0,54), gross_profitability (−0,38), book_to_market (+0,38), sales_growth_yoy (−0,37), liab_to_equity (+0,35), ebitda_to_ev (+0,33), equity_growth_yoy (−0,32), earnings_growth_yoy (−0,31)

Componente	Autovalor	Var. explicada	Var. acumulada	Características dominantes (loading)
PC7	2,045	4,54%	66,99%	capex_to_assets (+0,79), capex_to_revenue (+0,70), liq_turnover (−0,36), asset_growth_yoy (+0,36)
PC8	1,689	3,75%	70,75%	book_to_market (−0,59), ebitda_growth_yoy (−0,39), sales_growth_yoy (−0,38), gross_profitability (+0,31)
PC9	1,606	3,57%	74,31%	vol_60d (−0,46), liq_turnover (−0,39), downside_vol_60d (−0,36), gross_profitability (−0,30), kurt_126d (−0,30)
PC10	1,329	2,95%	77,27%	mom_ret_1m (−0,63), mom_ret_3m (−0,44), mom_ret_12_1 (+0,36), mom_vol_adj_12_1 (+0,35)
PC11	1,117	2,48%	79,75%	kurt_126d (+0,39), mom_ret_1m (−0,38), skew_126d (+0,34), earnings_to_price (+0,32)
PC12	0,995	2,21%	81,96%	skew_126d (+0,52), kurt_126d (+0,47), mom_ret_3_1 (−0,32)

Tabela 2 — Clusters hierárquicos (average linkage, corte equivalente a $|\bar{\rho}| \geq 0,3$)

Cluster	Nº fatores	Coesão (média abs. ρ)	Características
1	1	n/a (singleton)	mom_ret_1m
2	7	0,646	mom_ret_3_1, mom_ret_3m, mom_ret_12m, mom_ret_12_1, mom_vol_adj_12_1, mom_ret_6_1, mom_ret_6m
3	2	0,835	inv_capex_to_assets, inv_capex_to_revenue
4	6	0,662	size_log_mcap, liq_dollar_vol, liq_amihud_21d, liq_avg_vol_21d, liq_turnover, liq_vol_per_share
5	2	0,312	oth_kurt_126d, oth_skew_126d
6	4	0,394	inv_asset_growth_yoy, inv_equity_growth_yoy, inv_ebitda_growth_yoy, inv_sales_growth_yoy
7	6	0,545	val_sales_to_price, prof_asset_turnover, val_sales_to_ev, prof_gross_margin, prof_ebit_margin, prof_ebitda_margin
8	13	0,475	prof_gross_profitability, risk_downside_vol_60d, risk_vol_60d, val_dividend_yield, prof_ebit_to_assets, prof_ebitda_to_assets, val_ebit_to_ev, val_ebitda_to_ev, inv_earnings_growth_yoy, val_earnings_to_price, prof_roe, prof_net_margin, prof_roa
9	1	n/a (singleton)	val_book_to_market
10	1	n/a (singleton)	val_ebit_to_mcap
11	2	0,946	lev_liab_to_assets, lev_liab_to_equity

Tabela 3 — Representantes por cluster e alinhamento ao PC dominante

Cluster	Representante	Coesão	PC(s) dominante(s)	Coerente (1 PC)?	Dimensão econômica
1	z_mom_ret_1m	n/a (singleton)	PC10	Sim	Momentum, reversão de curtíssimo prazo
2	z_mom_ret_12m	0,646	PC2	Sim	Momentum de médio e longo prazo
3	z_inv_capex_to_assets	0,835	PC7	Sim	Intensidade de investimento (CAPEX)
4	z_liq_dollar_vol	0,662	PC1, PC3, PC4	Não	Tamanho e liquidez de mercado
5	z_oth_skew_126d	0,312	PC12	Sim	Assimetria, preferência por loteria
6	z_inv_asset_growth_yoy	0,394	PC1, PC8	Não	Crescimento de ativos, vendas, EBITDA
7	z_val_sales_to_ev	0,545	PC1, PC3, PC4	Não	Valor por vendas e margens
8	z_prof_roa	0,475	PC1	Sim	Rentabilidade, yield operacional e risco
9	z_val_book_to_market	n/a (singleton)	PC8	Sim	Valor contábil (book-to-market)
10	z_val_ebit_to_mcap	n/a (singleton)	PC6	Sim	Valor operacional sobre capitalização
11	z_lev_liab_to_assets	0,946	PC5	Sim	Alavancagem financeira

Figura 1 — *Heatmap* de correlação de Pearson entre as 45 características candidatas, matriz mensal representativa

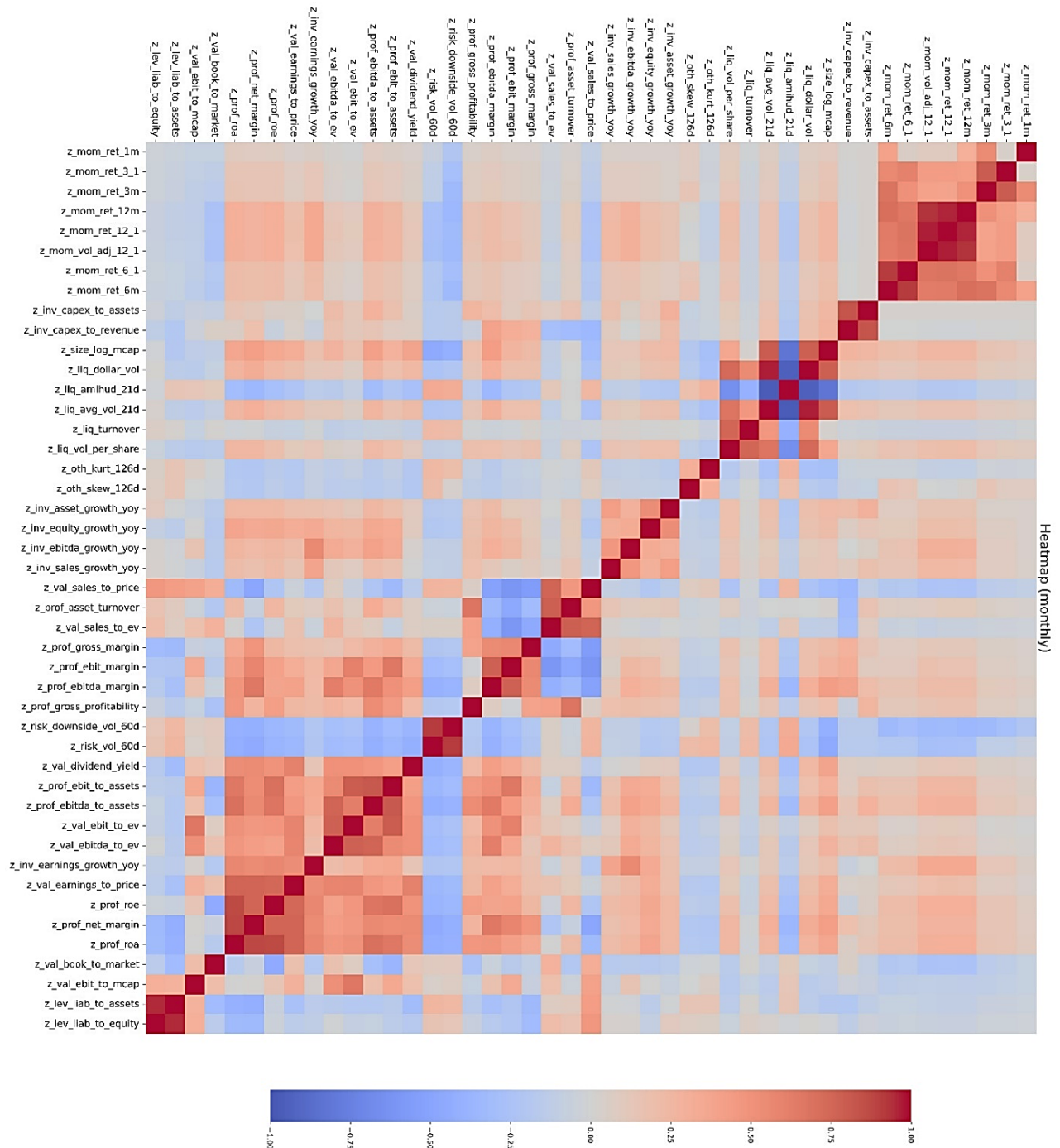
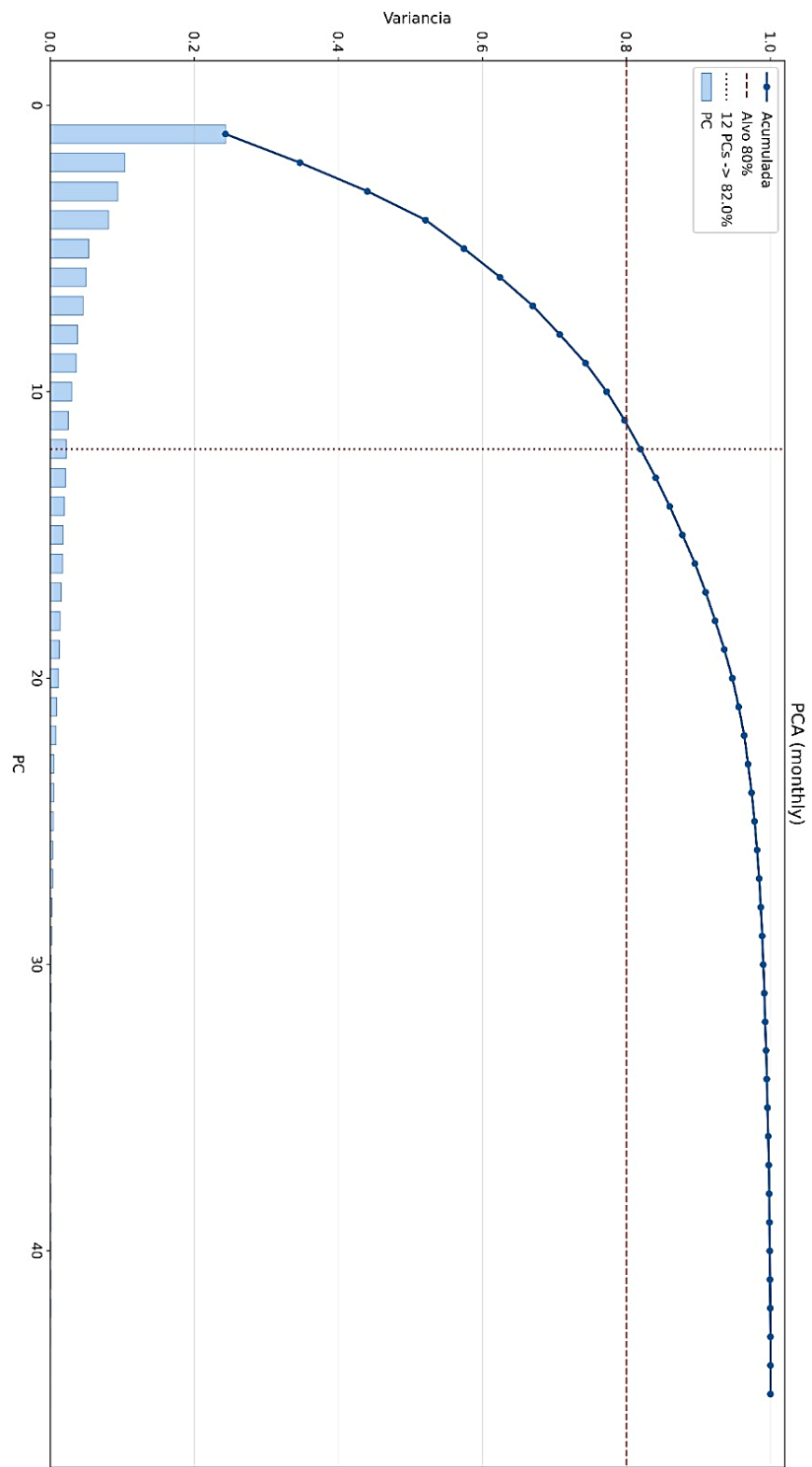
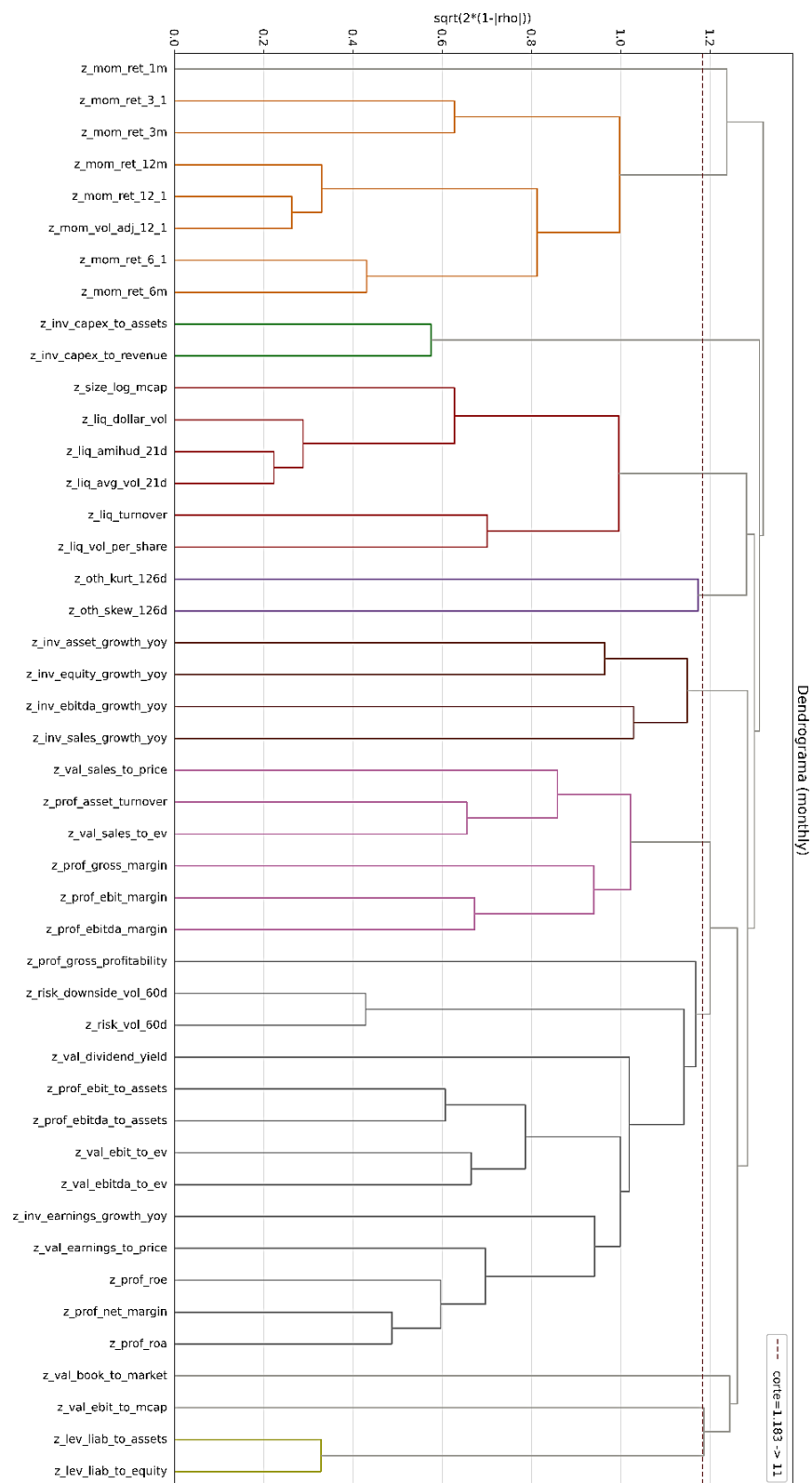


Figura 2 — Variância acumulada pelos componentes principais



Fonte: data/panel_output/correlation_v4/variancia_acumulada_monthly.png.

Figura 3 — Dendrograma da clusterização hierárquica (*average linkage*)



Fonte: data/panel_output/correlation_v4/dendrograma_fatores_monthly.png.