

Comparación entre el CAPM y el análisis de componentes principales *Sparse* como herramientas para la indexación

Comparison between CAPM and Principal Components Analysis *Sparse* as an indexing tool

Nelson Aldana-Martínez*

* Magíster en Inteligencia Analítica para la Toma de Decisiones. Docente-Investigador, Observatorio de Economía y Operaciones Numéricas (ODEON), Universidad Externado de Colombia, Bogotá (Colombia). [nelson.aldana1@uexternado.edu.co], [ORCID: 0000-0003-4232-4073].

Artículo recibido: 27 de mayo de 2022

Aceptado: 10 de junio de 2022

Para citar este artículo:

Aldana-Martínez, N. (2022). Comparación entre el CAPM y el análisis de componentes principales *Sparse* como herramientas para la indexación. *Odeon*, 21, 25-54.

DOI: <https://doi.org/10.18601/17941113.n21.03>.

Resumen

En un mundo que cada vez es más propenso a la inversión pasiva, es importante plantear un escenario en el que no se tenga que comprar todo el índice bursátil para este propósito. A través del cálculo de las betas móviles respecto al índice de los activos estables se busca hallar portafolios con menos acciones que todo el índice, en este caso el Standard and Poor's 500. En contraste, haciendo uso de la técnica de reducción de dimensionalidad “análisis de componentes principales *Sparse*”, se busca encontrar conjuntos de activos que también cumplan este propósito. Para esto se ponen a prueba tres experimentos: 1) optimizar con la data de entrenamiento para encontrar los pesos de cada uno; 2) ponderar de igual manera y 3) ponderar de acuerdo con el cálculo del índice, con bolsas de 200, 100, 50 y 25 activos elegibles.

Palabras clave: análisis de datos; portafolios de inversión; indexación; inversión pasiva.

Clasificación JEL: C13, C32, C61, G11.

Abstract

Globally is now more passive investing than ever. Bearing on mind that it is important to set up a scenario where an investor does not have to buy the entire index for this purpose. By searching stable assets respecting to their index, throw the rolling beta calculation, to find the fewer stocks to reach this purpose. On the other hand, use the dimensionality reduction technique, Sparse principal components analysis, helped to find a set of assets that also serve to this. To look for them, three experiments were tested, the first one optimizing with the training data to find the weights, weighting in the same way and weighting according to the calculation of the index, with baskets of 200, 100, 50 and 25 assets.

Key words: Data analysis; investment portfolios; indexing; passive investment.

JEL classification: C13, C32, C61, G11.

Introducción

En los últimos 30 años ha habido un incremento en las estrategias pasivas de inversión. Anadu *et al.* (2018) y Appel *et al.* (2016) hablan de un cambio del 5 % en 1995 al 37 % en 2018. En parte, esto se explica por una mayor aversión al riesgo generada desde la crisis de 2008; desde ese momento y hasta 2018 ha habido un aumento en al menos 5 trillones de dólares en este tipo de estrategias.

Por su parte, Stambaugh (2014) deja claro que la inversión pasiva se ha convertido en una alternativa a la inversión activa y, además, se ha observado cómo los fondos mutuos en general han bajado sus comisiones y, paulatinamente, se ha hecho una transición hacia fondos activos atados a índices, dado el atractivo de este tipo de inversiones. Malikov (2018) también destaca que el ascenso de la inversión pasiva se debe, igualmente, al declive del costo de la administración que esta conlleva.

Esto ha dado lugar a un estudio riguroso de los efectos de este fenómeno y las ventajas y desventajas que puede presentar. Por una parte, hace que los precios sean más eficientes (Lee, 2020; French, 2008; Appel *et al.*, 2016), lo que ha llevado a corroborar la hipótesis de mercado eficiente¹ que planteó Fama en 1970 y 1991. Esta eficiencia se traduce en una mayor transferencia de información privada al mercado, haciendo que se refleje en un cambio de valor de los activos disponibles para inversión.

Cabe recordar que “el mercado puede ser eficiente si un número suficiente de inversionistas tiene acceso a la información disponible” (Fama, 1970); además, en su revisión de 1991, Fama llega a la conclusión de que los administradores pasivos llevan a que haya una mayor eficiencia de mercado, dada la participación que adquieren.

Según Lee (2020), “los precios de los fondos indexados se convierten en más eficientes al decrecer los costos de transacción de las estrategias pasivas, de manera que más inversionistas escogen estrategias desinformados”, por lo cual reflejan de una forma óptima el valor de las acciones y dejan entrever un mejor flujo de información entre todos los agentes del mercado.

Esto también se explica dado que un mayor flujo de liquidez inyectado por el crecimiento de la inversión pasiva deja que los administradores activos puedan tener facilidad para la ejecución de sus operaciones y, por tanto, para transferir la información a los precios, lo que deja como resultado un mercado cada vez más eficiente, en el cual se puede decir que los precios son el reflejo del valor de las compañías, dada la información disponible.

Además, uno de los beneficios de la creciente migración a los fondos pasivos de inversión es que al ser menores los inversionistas activos, “los inversionistas pasivos aprovechan de manera óptima los esfuerzos de adquisición de

1 “El mercado en el que el precio refleja la información disponible, es llamado eficiente” (Fama, 1970).

información de los comerciantes activos en un agregado de mercado” (Malikov, 2018). Aunque este mismo autor también concluye que el mercado solamente reflejará de manera total la información con un rezago de 2 años, lo que deja entrever la necesidad de la existencia de inversionistas activos que ayuden a que esta información se vea reflejada en el precio.

Farboodi *et al.* (2018) señalan que el crecimiento de las firmas ha hecho que se refleje de una manera óptima la información que estas arrojan al mercado. Es decir, cuando una empresa es más grande, la información que está disponible para el mercado es más visible; lo anterior, unido a que en la era de la información se ha visto el fenómeno de que solo las compañías que entran a un índice bursátil que requiere una capitalización bursátil mayor son las que reflejan esto.

Esto ha ayudado a resolver una disyuntiva que se podría crear entre los diferentes inversionistas, donde no valdría la pena pagar una comisión más elevada para poder obtener un alfa respecto al mercado. Sin embargo, en un mercado con muchos valores, los administradores activos se decantan por la maximización de la ratio de información mediante la explotación de la información privada y las oportunidades de inversión, que llevan a un mercado eficiente.

De hecho, French (2008) resalta que los inversionistas activos son necesarios dado que hacen una mejor distribución de recursos en la sociedad, debido a su afán de buscar los retornos y la rotación de capital que esto produce. De hecho, el juego del *trading* activo, luego de costos, comisiones y gastos, es un juego de suma cero para las dos partes de la transacción. Lo que gana uno de los inversionistas, puede ser la pérdida de otro.

Además, los múltiples estudios de French (2008), Malkiel (2012) y Stambaugh (2014) demostraron que el agregado de los administradores activos de portafolio ha llegado a tener un menor desempeño que si se hubiera invertido en el índice. “Los inversionistas harían mucho mejor comprando y manteniendo un fondo indexado que intentando comprar y vender valores individuales o fondos mutuos administrados activamente [...] los grandes costes de transacción restan sustancialmente a los rendimientos de la inversión” (Malkiel, 2012).

“Bajo supuestos razonables, el típico inversionista podría incrementar su retorno medio anual en 67 puntos básicos en el periodo 1980-2006 si cambiara a una estrategia pasiva de inversión” (French, 2008), dando lugar a la motivación de encontrar una forma pasiva de inversión que sea consistente y buena para el inversionista. Ahora bien, se podría entonces pensar en invertir en un Exchanged Traded Fund que siga el mercado, sin embargo, existe una de las típicas preguntas que se han planteado a través del tiempo en finanzas: ¿qué es el mercado?

Desde que se definió el modelo del Capital Asset Pricing Model (CAPM) por parte de Sharpe (1964), Litner (1965) y Treynor (1961), se ha dado un debate acerca de cuál es el mercado sobre el que se debe comparar el activo para obtener su retorno esperado. Gârleanu y Pedersen (2021) cuestionan cuál es el portafolio que debe escoger un inversionista pasivo como *benchmark*, dado que los índices bursátiles, aunque tengan una gran cantidad de activos, no necesariamente son el todo.

Aunque cabe resaltar que el Standard and Poor's 500 (SPX500) –que es el índice bursátil que se suele seguir cuando se habla del mercado, por la tradición de sus componentes y la madurez empresarial que este tiene– se ha vuelto uno de los más eficientes, dada la transferencia de información que le ha producido el fenómeno de la inversión pasiva.

Esto ha llevado a que, según Stambaugh (2014), los inversionistas activos pasen de tener un 2,2 % de *tracking error* en 1980 a un 1,2 % en 2006, lo que acerca cada vez más la inversión activa a la pseudo-indexación. Asumir un *tracking error* de 0 % es el equivalente a tener un $\beta_{im} = 1$, o dejar solo el riesgo sistemático, que no es diversificable y corresponde a la totalidad del mercado.

Roll (1977), en su artículo “A critique on the asset pricing theory's test”, llega a varias conclusiones, entre ellas, que existe una relación de si y solo si entre la linealidad del retorno-beta y la eficiencia de media-varianza del portafolio de mercado. Sin embargo, la conclusión que es más relevante para el objetivo de este artículo es que si existe un portafolio compuesto por activos riesgosos, las medias muestrales de los retornos serán funciones lineales de la muestra de betas.

Es justamente esto lo que lleva a pensar que se puede llegar a una indexación o inversión pasiva de un índice bursátil, dado que teóricamente es posible con un $\beta = 1$, y con el estudio de estos se puede llegar a tener un *tracking error* de 0. En el presente artículo se busca estudiar si el uso del CAPM puede llevar a tener una inversión indexada al SPX500, aproximándose a obtener un $\beta = 1$, con la menor cantidad de activos posibles. Adicionalmente, se compara este método de indexación con un análisis de componentes principales con *Sparse* (PCA *Sparse*), como técnica de reducción de dimensionalidad para obtener el menor *tracking error* posible con la menor cantidad de activos.

El SPX500 tiene aproximadamente el 47,64 % del producto interno bruto (PIB) mundial en capitalización de mercado de sus 504 componentes, siendo uno de los índices que más se acercan al mercado. Sin embargo, como lo demostró Roll, no existe un índice bursátil que tenga el mercado como un todo, por lo que se puede hacer uso del SPX500 en la aproximación de un portafolio eficiente de

media varianza, y, por tanto, al calcular el β de un activo contra él, puede asumirse que se está midiendo qué parte del riesgo sistémico se ha dejado por fuera.

Antes de entrar en los detalles técnicos de los experimentos propuestos es importante recalcar que el SPX500 se calcula de la siguiente manera:

$$SPX500 = \sum_{i=1}^n \frac{P_i * Q_i}{divisor} \quad (1)$$

Donde:

P_i : precio de la acción i en el momento de cálculo.

Q_i : Cantidad de acciones i usadas para el cálculo del índice.

$divisor$: Nivel necesario para obtener el nivel del índice en el momento t .

Esto se traduce en la definición de que es un índice bursátil calculado por la capitalización de mercado, ajustada por el flotante de acuerdo con la metodología publicada por la casa que lo calcula.

1. Capital Asset Pricing Model (CAPM)

El CAPM asume que la economía tiene un número k de inversionistas que buscan maximizar su utilidad:

$$U_k = E_k - \frac{\sigma_k^2}{\tau_k} \quad (2)$$

Donde:

U_k : utilidad del inversionista k .

E_k : retorno esperado del inversionista k .

σ_k^2 : varianza del portafolio.

τ_k : tolerancia al riesgo del inversionista k .

Dado que el CAPM se basa en el trabajo de Markowitz, el riesgo de este depende del componente riesgoso de las acciones y de las correlaciones que tienen con el

resto. Por tal motivo, se puede decir que el riesgo del portafolio está integrado por esos dos componentes. Los inversionistas deben invertir todo lo que poseen en el portafolio, por lo tanto:

$$\sum_i X_{ik} = 1 \quad (3)$$

Dado que los mercados se equilibran y que todos los activos están en los portafolios de los k inversionistas existentes en la economía, el portafolio ponderado por capital tendrá una condición de optimalidad de primer orden dada por:

$$E_i - \frac{2}{\tau_m} \sigma_{im} = \lambda_{fm} \quad (4)$$

Donde:

E_i : retorno esperado de la acción i .

τ_m : tolerancia al riesgo ponderada por capital o la tolerancia al riesgo de la sociedad.

σ_{im} : covarianza del activo i con el portafolio de mercado (que incluye todos los activos disponibles).

λ_{fm} : utilidad marginal social de la riqueza.

Despejando el retorno esperado de (4) se puede demostrar que este se puede expresar en función de la covarianza del portafolio o de la correlación (ρ_{im}) de este:

$$E_i = \lambda_{fm} - \frac{2\rho_{im}\sigma_m\sigma_i}{\tau_m} \quad (5)$$

O expresado en términos de β , dado que $\beta = \frac{\rho_{im}\sigma_i}{\sigma_m}$:

$$E_i = \lambda_{fm} - \frac{2\sigma_m^2}{\tau_m} \beta_{im} \quad (6)$$

Dejando claro que en equilibrio existe una relación lineal entre los retornos esperados de las acciones y las correlaciones con el portafolio de mercado, dado que (6) aplica para todos los portafolios, incluso para el de mercado consigo mismo, en caso en el que $\beta = 1$.

Finalmente, el CAPM implica que el portafolio de mercado es eficiente, todos los portafolios eficientes son equivalentes a invertir en el portafolio de mercado y existe una relación lineal entre el retorno esperado y el β .

Dadas estas características, y en línea con el propósito de encontrar un portafolio que consiga seguir el SPX500, se buscó hacer un portafolio con la menor cantidad de activos posibles que sean equivalente al SPX500. Es decir, se seleccionó un portafolio con $n \leq 504$, en el que el *tracking error* sea mínimo y el $\beta_{pi} = 1$. Definiendo *tracking error* como la desviación estándar de la diferencia entre el retorno del portafolio y el retorno del índice:

$$TE = \sqrt{\frac{\sum_{t=1}^n \left((R_p - R_i)_t - \overline{(R_p - R_i)} \right)^2}{n}} = \sigma_{R_p - R_i}$$

Donde:

R_p : rentabilidad del portafolio.

R_i : rentabilidad del índice.

$\overline{(R_p - R_i)}$ la media de la diferencia de las rentabilidades.

n : número de periodos evaluados.

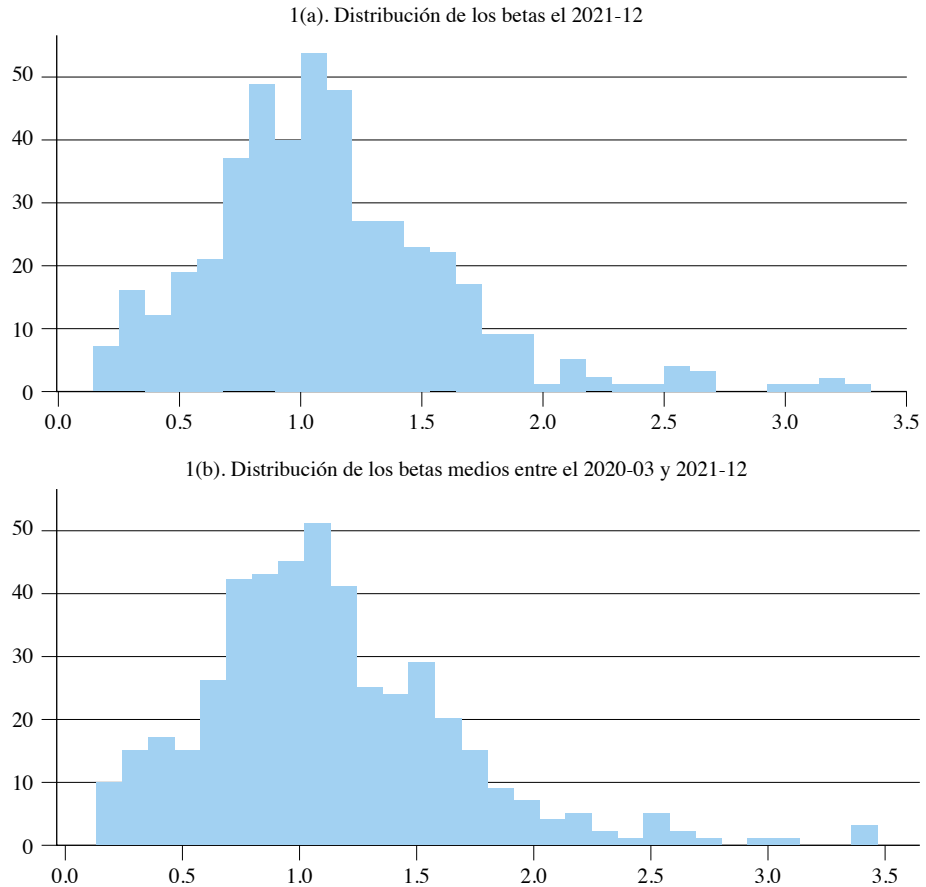
Dicho esto, se procedió con el análisis de los datos para el cálculo de los β de cada una de las acciones que componen el índice. Es importante recalcar que el procedimiento de esta sección tiene varios conjuntos de datos, los primeros son los que se usaron para calcular los coeficientes β_i , que van desde marzo de 2015 hasta diciembre de 2021 con cierres mensuales; la data de entrenamiento del modelo, que fueron los precios diarios del año 2021 y, finalmente, los precios del primer trimestre de 2022, que sirvieron como prueba del modelo.

Para el primer conjunto de datos se tomaron los valores desde enero de 2015 hasta diciembre de 2021. Descartando aquellas acciones que por su historia no tienen observaciones completas para hacer el estudio, se obtienen 459, perdiendo

el 9 % del índice. Una vez efectuada la limpieza de datos, se procedió a calcular el β_{im} móvil de 60 meses, desde marzo de 2020 hasta diciembre de 2021, dejando como resultado 22 medidas para cada uno de los activos sujetos de estudio. Como se puede observar en la figura 1, tanto el promedio del coeficiente en la ventana de tiempo, como el valor en el último momento tienen un rango similar de 0-3,5.

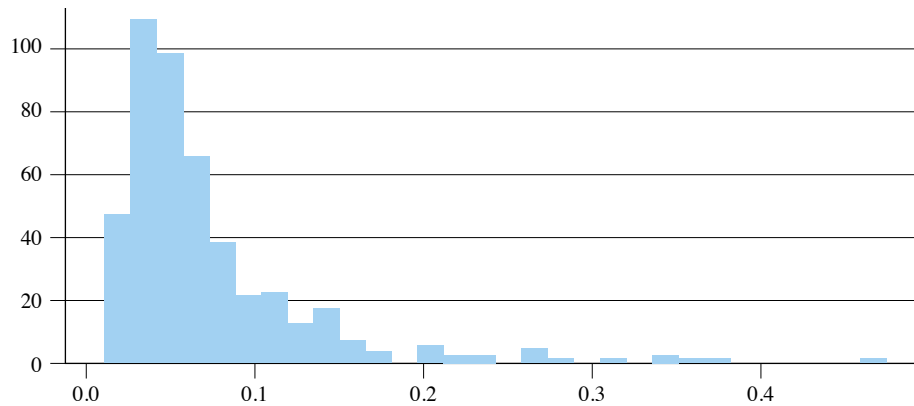
Ahora bien, este último fenómeno también se observa en la distribución del coeficiente de variación de la beta móvil que se muestra en la figura 2, donde el 50 % de los datos se ubicaron por debajo de 0,1750, mostrando así una variación baja a través del periodo de estudio.

Figura 1: Distribución de las betas en 2021-12 y distribución de la media de los betas móviles entre 2020-03 y 2021-12



Fuente: elaboración propia.

Figura 2: Distribución del coeficiente de variación de las betas móviles entre 2020-03 y 2021-12



Fuente: elaboración propia.

Al observarse una relativa estabilidad de los betas en la figura 2, donde los coeficientes de variación tienen una asimetría a la derecha, se plantean dos formas de abordar el problema: la primera es mediante una optimización del *tracking error*, con los activos con menores coeficientes de variación histórica de los betas móviles (200, 100, 50 y 25) y hallar el portafolio que más se acerque al índice con el menor número de activos; dado que entre menor es el *tracking error* es más cercano a la indexación, se espera que el $\beta_{mp} = 1$; la segunda es hacer dos ponderaciones con estas mismas acciones, la primera con $w_i = 1/n$ y la

segunda con $w_i = \frac{marketcap_i}{\sum marketcap}$, esto dada la metodología de cálculo del índice.

Para esto se tomaron como data de entrenamiento los precios del año 2021, a fin de hallar el portafolio que en este periodo de tiempo permita la indexación y se probará con la data del primer trimestre de 2021.

1.1. Optimización del *tracking error*

Con los 200, 100, 50 y 25 activos seleccionados de acuerdo con los coeficientes de variación más bajos de las betas móviles, se procedió a plantear el siguiente problema de optimización:

$$\min_{TE} \sqrt{\frac{\sum_{t=1}^n \left((R_p - R_i)_t - \overline{(R_p - R_i)} \right)^2}{n}}$$

s.a. $Q_j \in \mathbb{Z}^+$

$P_{j1} * Q_j = \text{USD } 500,000$

Donde:

R_p : rentabilidad del portafolio.

R_i : rentabilidad del índice.

$\overline{(R_p - R_i)}$ la media de la diferencia de las rentabilidades.

Q_j : cantidad de acciones del activo j.

P_{j1} : precio del activo j en el momento 1.

n : número de periodos evaluados.

De esta manera se obtiene el número de acciones que se tienen que comprar de los activos que hacen que el *tracking error* sea mínimo, en este caso se obtiene, con una posición inicial. Sin embargo, teniendo en cuenta que el modelo se va a testear con los datos del trimestre inmediatamente siguiente a aquel con el que se entrenó, se hace una modificación a las restricciones, de la siguiente manera:

$$\min_{TE} \sqrt{\frac{\sum_{i=1}^n \left((R_p - R_i)_t - \overline{(R_p - R_i)} \right)^2}{n}}$$

s.a. $Q_i \in \mathbb{Z}^+$

$P_{in} * Q_i = \text{USD } 500,000$

Donde:

R_p : rentabilidad del portafolio.

R_i : rentabilidad del índice.

$\overline{(R_p - R_i)}$ la media de la diferencia de las rentabilidades.

Q_j : cantidad de acciones del activo j.

P_{jn} : precio del activo j en el momento n .

n : número de periodos evaluados.

Una vez calculados los modelos se obtuvieron los resultados de la tabla 1 con la data de entrenamiento; teniendo en cuenta esto, se calcularon los datos de los dos escenarios planteados en las ecuaciones 8 y 9:

Tabla 1: Resultados de las optimizaciones de los portafolios en la data de entrenamiento

	Modelo ecuación 8				Modelo ecuación 9			
Corte coef. var.	0,0483	0,0352	0,0280	0,0232	0,0483	0,0352	0,0280	0,0232
Número de Activos	200	100	50	25	200	100	50	25
Activos Usados	91	53	30	16	85	51	31	16
Beta	1,0015	0,9897	1,0047	1,0297	1,0031	0,9939	0,9918	1,0207
<i>Tracking error</i> ²	1,38 %	3,60 %	3,95 %	6,20 %	1,37 %	3,09 %	3,96 %	5,87 %

Fuente: elaboración propia.

Sin embargo, como predecir el pasado es una forma fácil de saber qué hacer, se puso como data de prueba al primer trimestre de 2022; invirtiendo en la cantidad de acciones que se hallaron en el punto anterior, se obtuvieron los resultados que se muestran en la tabla 2. Se observa que no necesariamente hay una consistencia del *tracking error* y la beta con respecto a los datos de entrenamiento en todos los portafolios, dada la forma de cálculo donde se hallaron para un momento en específico y no con una metodología consistente. Sin embargo, es de apreciar que las medidas se mantuvieron por debajo del 6,2 % en todos los nodos probados y la beta se ubicó en un rango entre 0,98 y 1,03, dejando entrever una alta correlación entre los portafolios obtenidos y el índice.

2 Para efectos de todas las comparaciones, el *tracking error* siempre se encontrará anualizado.

Tabla 2: Resultados de las optimizaciones de los portafolios en la data de prueba

	Modelo ecuación 8				Modelo ecuación 9			
Corte coef. var.	0,0483	0,0352	0,0280	0,0232	0,0483	0,0352	0,0280	0,0232
Número de Activos	200	100	50	25	200	100	50	25
Activos Usados	91	53	30	16	85	51	31	16
Beta	0,9670	0,9810	0,9279	0,9601	0,9670	0,9837	0,9141	0,9444
Tracking error	3,57 %	5,20 %	5,18 %	6,70 %	3,53 %	4,93 %	5,36 %	6,57 %

Fuente: elaboración propia.

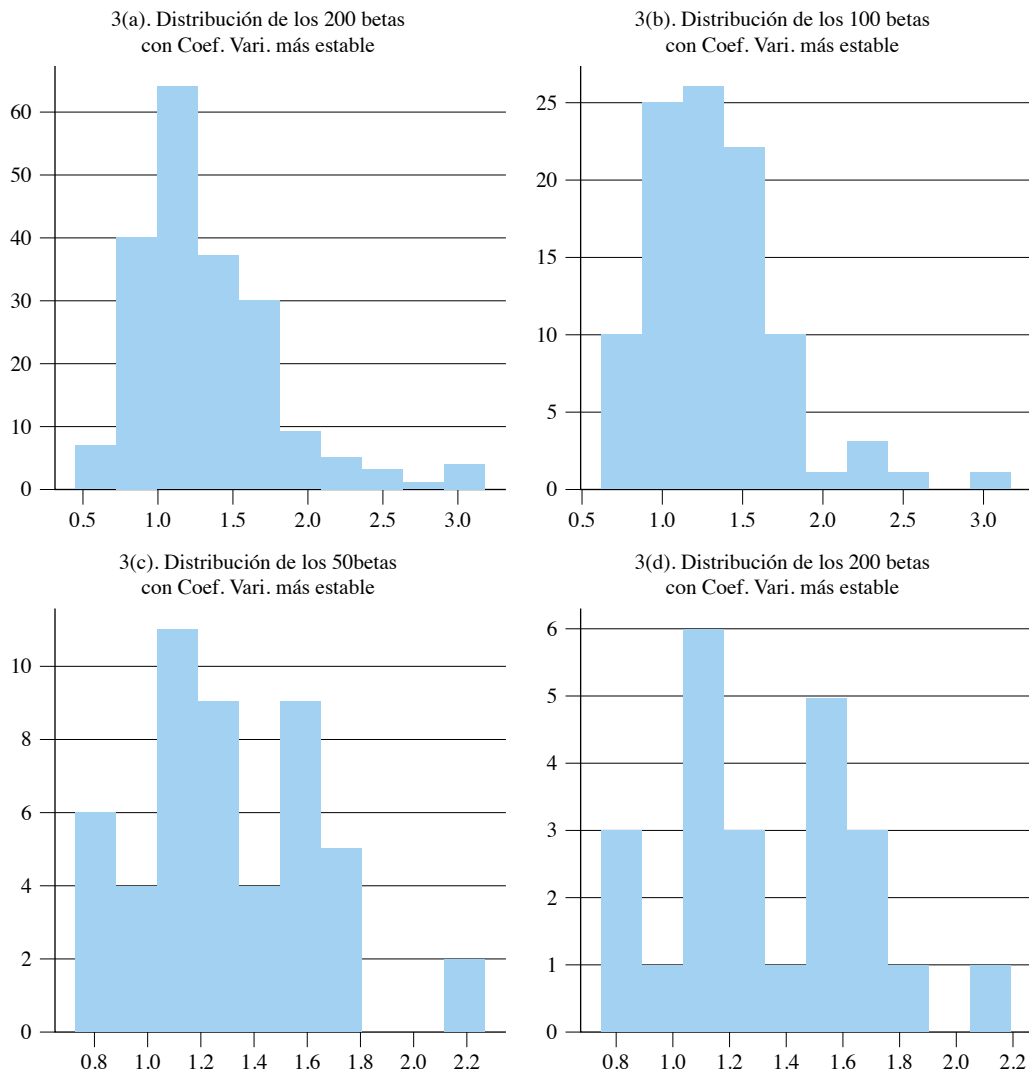
Dado que se obtuvieron diferencias entre los resultados de entrenamiento y prueba, vale la pena evaluar alternativas. En la siguiente sección se equiponderó cada activo $Q_j = \frac{500000}{j} \div P_{j1}$ y se hará una ponderación de acuerdo con su capitalización de mercado ajustada por el flotante libre, como lo establece la metodología del índice.

1.2. Ponderación de acuerdo con pesos equivalentes y

$$w_i = \frac{\text{marketcap}_i}{\sum \text{marketcap}}$$

Debido a que se hizo una selección de activos de acuerdo con un criterio de estabilidad de las betas, y teniendo en cuenta que, como se observa en la figura 3, las betas más estables están ubicadas en torno al 1 y se compensan entre sí, se procedió a equiponderar los activos y a calcular cuál fue su desempeño para, de esta manera, saber si es una estrategia pasiva exitosa o, por el contrario, se debe usar otra.

Figura 3: Histogramas de las betas calculadas en diciembre de 2021



Fuente: elaboración propia.

Como se observa en la tabla 3, los resultados de invertir de USD 500.000 en los diferentes activos que conforman el SPX500, teniendo como criterio de selección el coeficiente de variación de sus betas y equiponderando la inversión en cada uno de ellos, no mejoró con respecto a optimizar los portafolios, con criterio de éxito para tener un *tracking error* mínimo. Es de resaltar que para ambos casos se usaron los datos del primer trimestre de 2022.

Tabla 3: Resultados de invertir la misma cantidad de dinero en cada uno de los activos

Corte coef. var.	0,0483	0,0352	0,0280	0,0232
Número de activos	200	100	50	25
Beta	0,8918	0,9270	0,9131	0,8916
Tracking error	6,92 %	7,49 %	8,77 %	10,36 %

Fuente: elaboración propia.

Ahora bien, el índice tiene una metodología de cálculo diferente, lo que podría llevar a plantearse una ponderación por capitalización de mercado, ajustado por flotante en el momento de hacer la inversión. En la tabla 4 se muestra el resultado de llevar a cabo este experimento. Sin embargo, al llevar a cabo la evaluación de prueba del modelo se observa que tampoco hubo un incremento respecto a los resultados obtenidos en el modelo de la optimización, aunque se resalta que, al ponderar los activos de esta manera, sí existe una mejora respecto a darles el mismo peso a todos.

Tabla 4: Resultados de invertir en cada uno de los activos de acuerdo con su capitalización de mercado

Corte coef. var.	0,0483	0,0352	0,0280	0,0232
Número de activos	200	100	50	25
Beta	1,0159	1,0011	1,0346	0,9125
Tracking error	4,06 %	7,12 %	8,32 %	11,70 %

Fuente: elaboración propia.

Luego de hacer los tres experimentos para obtener una indexación en el primer trimestre del año 2022, se puede observar que, aunque se pudo llegar a coeficientes de *tracking error* inferiores al 3,6 % en la data de prueba y coeficientes β_{mp} en torno al 1, no es suficiente para poder decir que se está indexado de manera total³. Por esta razón, se estudió con una técnica de reducción de

3 Una estrategia indexada es aquella cuyo *tracking error* sea menor que 2 %.

dimensionalidad, que en trabajos pasados demostró ser la mejor selección de las acciones para lograr una indexación (Aldana, 2021).

2. Análisis de componentes principales *Sparse*

Las técnicas de reducción de dimensionalidad han sido usadas normalmente para encontrar nuevas variables que representen el mismo conjunto de datos. Es decir, el análisis de componentes principales se usa para hallar componentes (n) que son combinaciones lineales del set de variables originales (m) y explican con un menor número ($n < m$) una gran variabilidad de la data.

Es una técnica de aprendizaje de máquina no supervisado, con la que se pretende encontrar relaciones dentro del conjunto de datos, dado que las variables están interrelacionadas entre sí, derivando en un nuevo conjunto de componentes que no están correlacionadas, y que de manera ordenada explican la variabilidad desde el primer elemento.

Sin embargo, dado que los componentes principales son combinaciones lineales de las variables originales, que tienen un coeficiente diferente de 0 generalmente, Zou *et al.* (2006) introdujeron el Sparse Principal Component Analysis que, como su nombre lo indica, busca escasos miembros dentro de cada uno de estos. Este modelo está construido con base en el análisis de componentes principales, como un tipo de problema de optimización de regresión, con una penalización cuadrática.

Para hacer esto, Zou *et al.* (2006) usaron una red elástica y una penalización *lasso* para lograr la obtención de los componentes principales *Sparse*, con una mayor cantidad de coeficientes 0 en las combinaciones. Se propone el computo del modelo de acuerdo con estos y de igual forma en la que se calcula la librería Sci-kit Learn (Pedregosa *et al.*, 2011) de la siguiente manera:

$$(U^*, V^*) = \arg \min_{U, V} \frac{1}{2} \|X - UV\|_{Fr}^2 + \alpha \|V\|_{1,1}$$

$$s.a. \|U_k\|_2 \leq 1 \quad \forall 0 \leq k < n_{componentes}$$

Donde: $\|X - UV\|_{Fr}^2$ es la norma frobiana y $\|V\|_{1,1}$ representa la norma de la matriz de entradas, que es la suma de los valores absolutos de todas las entradas en la matriz.

Dado que con este método lo que se busca es hacer que los componentes principales estén representados por menos variables de las que se usan en el conjunto de datos, se pueden identificar de una mejor manera cuáles son las que más aportan a cada uno. Es de recordar que el objetivo buscado es encontrar la manera de replicar el SPX500 con la menor cantidad de activos posibles. Por esto se procedió a encontrar 200, 100, 50 y 25 componentes.

Por lo tanto, con el conjunto de datos de entrenamiento que corresponde a los retornos logarítmicos diarios del año 2021, se procedió a hacer un análisis de componentes principales *Sparse* con 200, 100, 50 y 25 componentes. Así, se buscan los activos que, en cada uno de ellos, aporten la mayor cantidad de variabilidad, haciendo de esta manera la selección de acciones que se usó para la evaluación de cada uno de los portafolios que buscan seguir el índice.

En la primera parte, con los datos de entrenamiento, se buscó optimizar el *tracking error*, luego se equiponderaron los activos, dado que cada componente representa una nueva variable no correlacionada con el resto y, finalmente, se probó con una ponderación por capitalización de mercado, ajustada por flotante.

2.1. Optimización del *tracking error*

Buscando el portafolio que minimizara el *tracking error*, se usó la data de entrenamiento para hallar el número de acciones en las que se debía invertir en un portafolio de USD 500.000, de igual manera a como se planteó en las ecuaciones 8 y 9, y a través de la curva se obtuvo un mejor *tracking error* respecto a lo que se observó en el experimento de la sección 2.1. Esto se debe a que la metodología usada levanta la linealidad que se asume en todo el modelo del CAPM.

Tabla 5: Resultados de las optimizaciones de los portafolios en la data de entrenamiento

Número de componentes	Modelo ecuación 8				Modelo ecuación 9			
	200	100	50	25	200	100	50	25
Activos usados	119	66	45	25	130	68	45	25
Beta	0,9853	0,9480	0,9502	0,9437	0,9841	0,9447	0,9546	0,9492
Tracking error	1,23 %	2,85 %	2,99 %	4,22 %	1,26 %	2,63 %	2,83 %	4,24 %

Fuente: elaboración propia.

Ahora bien, aunque los resultados obtenidos con la data de entrenamiento son mejores, no ocurre lo mismo con los datos de prueba, donde se pueden observar

resultados mixtos, dado que para 200, el *tracking error* luego de optimizar es peor. Sin embargo, para el resto de los casos se mantiene la tendencia a obtener mejores resultados, lo que da a entender un mejor desempeño del modelo.

Tabla 6: Resultados de las optimizaciones de los portafolios en la data de prueba

Número de componentes	Modelo ecuación 8				Modelo ecuación 9			
	200	100	50	25	200	100	50	25
Activos usados	119	66	45	25	130	68	45	25
Beta	1,0036	0,9669	0,9416	0,9057	0,9966	0,9184	0,9592	0,9070
Tracking error	11,71 %	4,23 %	5,26 %	5,70 %	11,12 %	4,46 %	4,88 %	5,70 %

Fuente: elaboración propia.

Sin embargo, como los experimentos se deben repetir con los nuevos activos obtenidos de la reducción de dimensionalidad, se plantearon dos más, uno con un portafolio equiponderado $Q_j = \frac{500000}{j} \div P_{j1}$ y otro ponderando por la capitalización bursátil de las compañías ajustadas por el flotante.

2.2. Ponderación de acuerdo con pesos equivalentes y

$$w_i = \frac{\text{marketcap}_i}{\sum \text{marketcap}}$$

Dada la metodología usada para la selección de activos, vale la pena equiponderar los mismos en el portafolio, teniendo en cuenta que esta busca nuevas variables que no estén correlacionadas entre sí, lo que hace que el efecto correlación de los activos, se aminore. Comparando este experimento con el que se obtiene del CAPM, se puede ver una disminución en el *tracking error* de todos los portafolios, lo que deja ver que, a priori sí, se puede tener una mejor indexación con esta metodología que con la anterior.

Tabla 7: Resultados de invertir la misma cantidad de dinero en cada uno de los activos

Número de componentes	200	100	50	25
Beta	0,8476	0,8280	0,8518	0,8583
Tracking error	6,05 %	6,35 %	6,82 %	6,71 %

Fuente: elaboración propia.

Ahora bien, vale la pena ponderar de la misma manera como se hace en el índice, para tener un proxy de mercado. Aunque los resultados no son óptimos, sí se obtienen mejores que usando la selección de acciones con el CAPM para los sets de 100, 50 y 25.

Tabla 8: Resultados de invertir en cada activo de acuerdo con su capitalización de mercado

Número de componentes	200	100	50	25
Beta	0,9705	0,8539	0,9648	0,8626
Tracking error	19,98 %	6,41 %	6,93 %	8,23 %

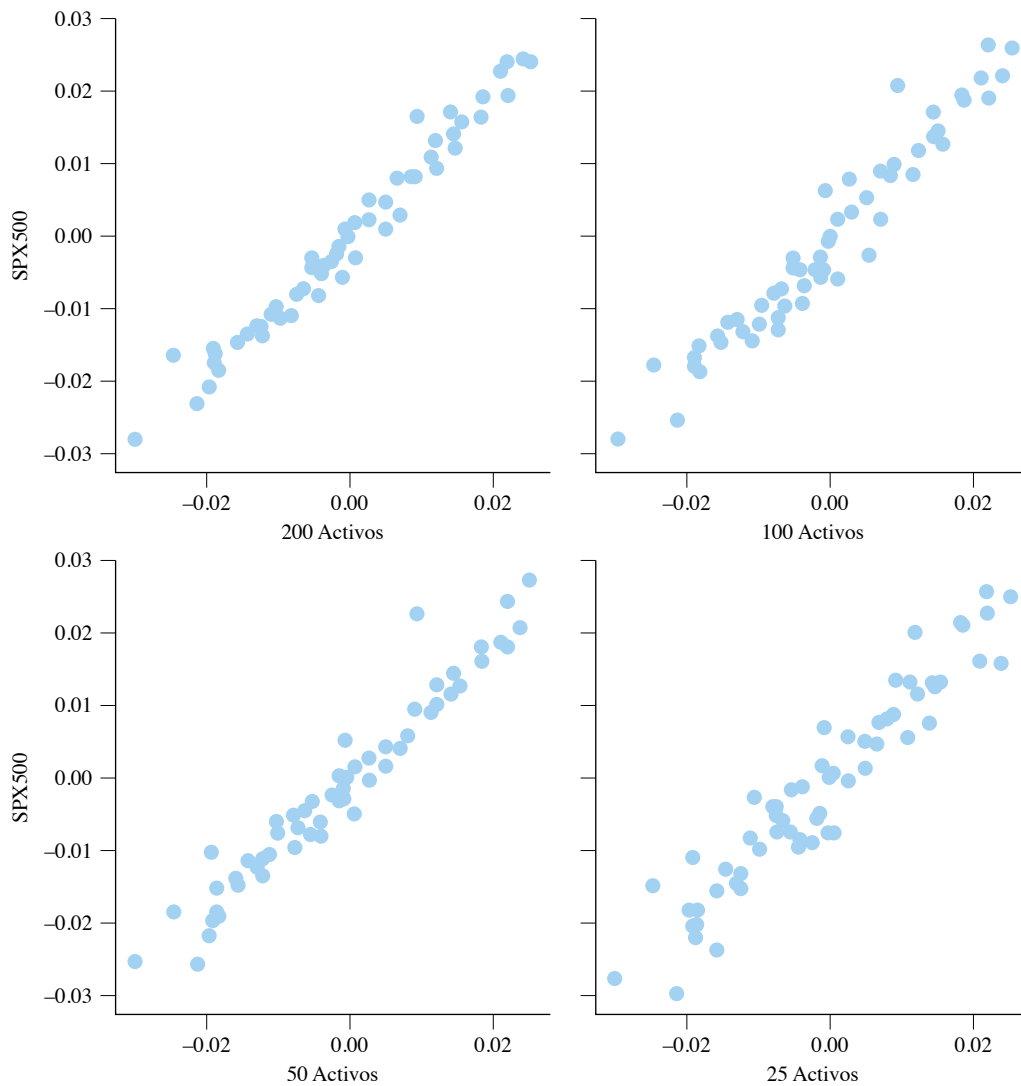
Fuente: elaboración propia.

3. Conclusiones

Una vez concluidos los experimentos en cada una de las metodologías, se observa que, aunque la aproximación de forma lineal no es la mejor para el mercado financiero, puede seguir siendo provechosa dados los resultados que se dieron con modelos como el CAPM. Además de la reducción de costo computacional y la fácil interpretación.

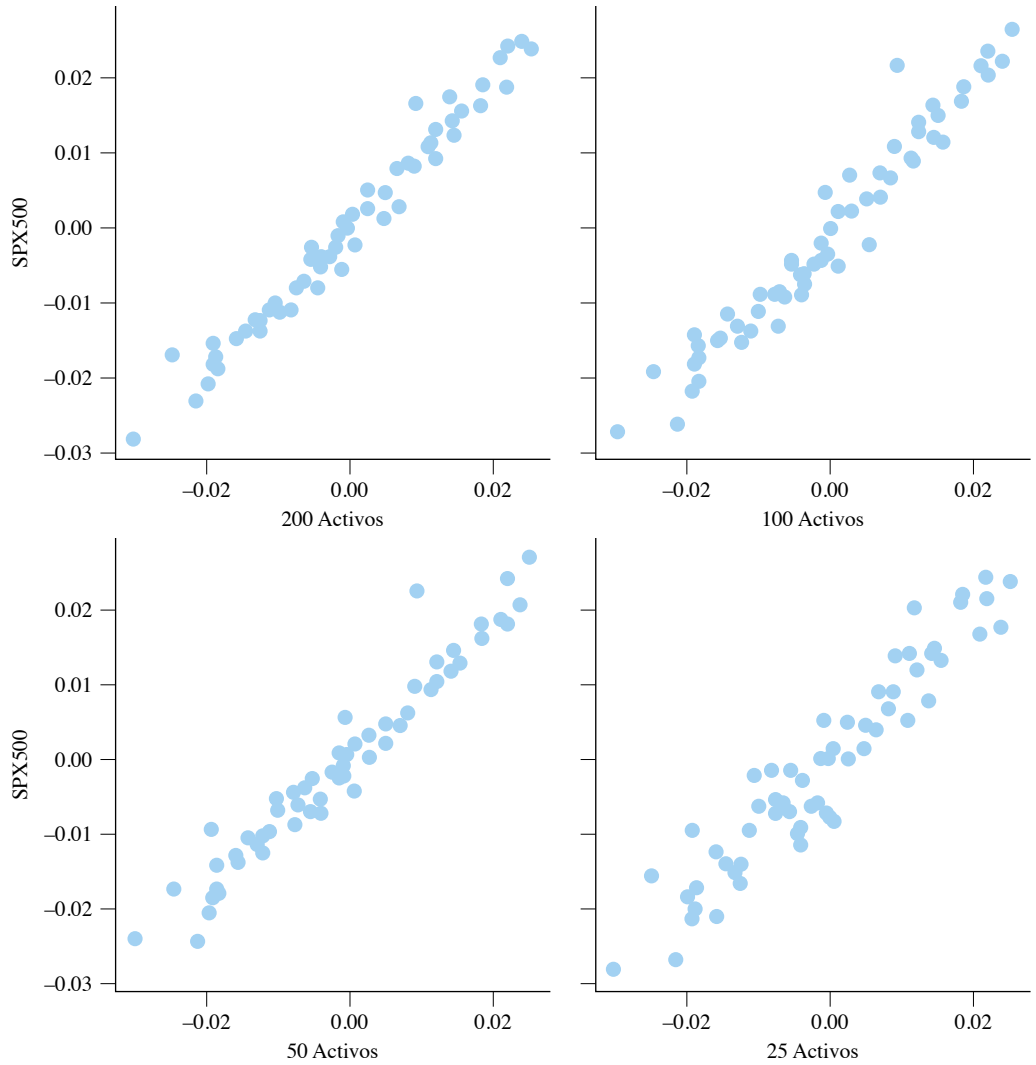
En cuanto a la información del mercado, se puede observar cómo se empieza a reflejar su incorporación a los precios en cada una de las metodologías, debido a que se dan indicadores *tracking error* consistentes en la mayoría de los casos. Además, como se puede observar en las figuras 4 a 11, existe una clara correlación entre los retornos del índice y los diferentes portafolios hallados.

Figura 4: Gráficos de dispersión de los portafolios seleccionados con el CAPM y el problema de optimización 8 y el SPX500 para la data de prueba



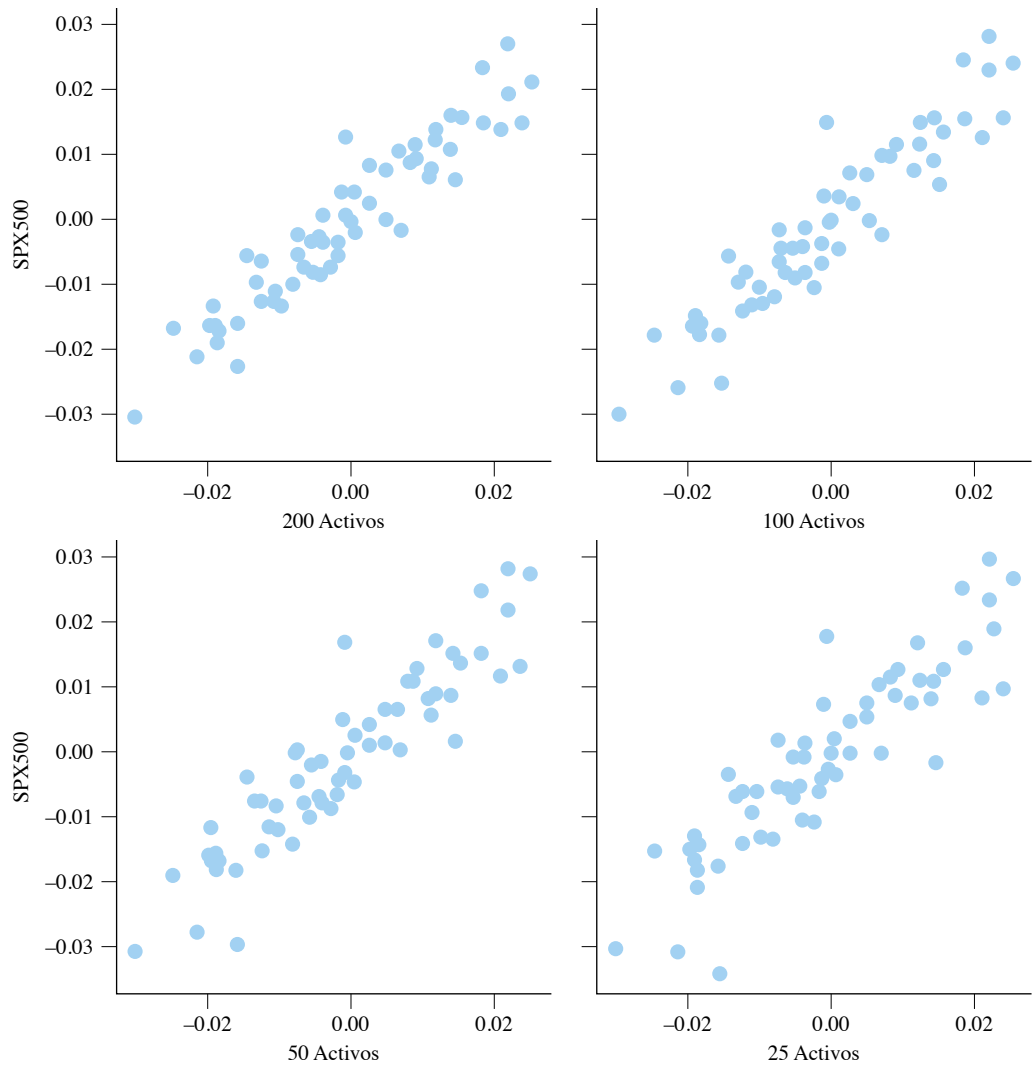
Fuente: elaboración propia.

Figura 5: Gráficos de dispersión de los portafolios seleccionados con el CAPM y el problema de optimización 9 y el SPX500 para la data de prueba



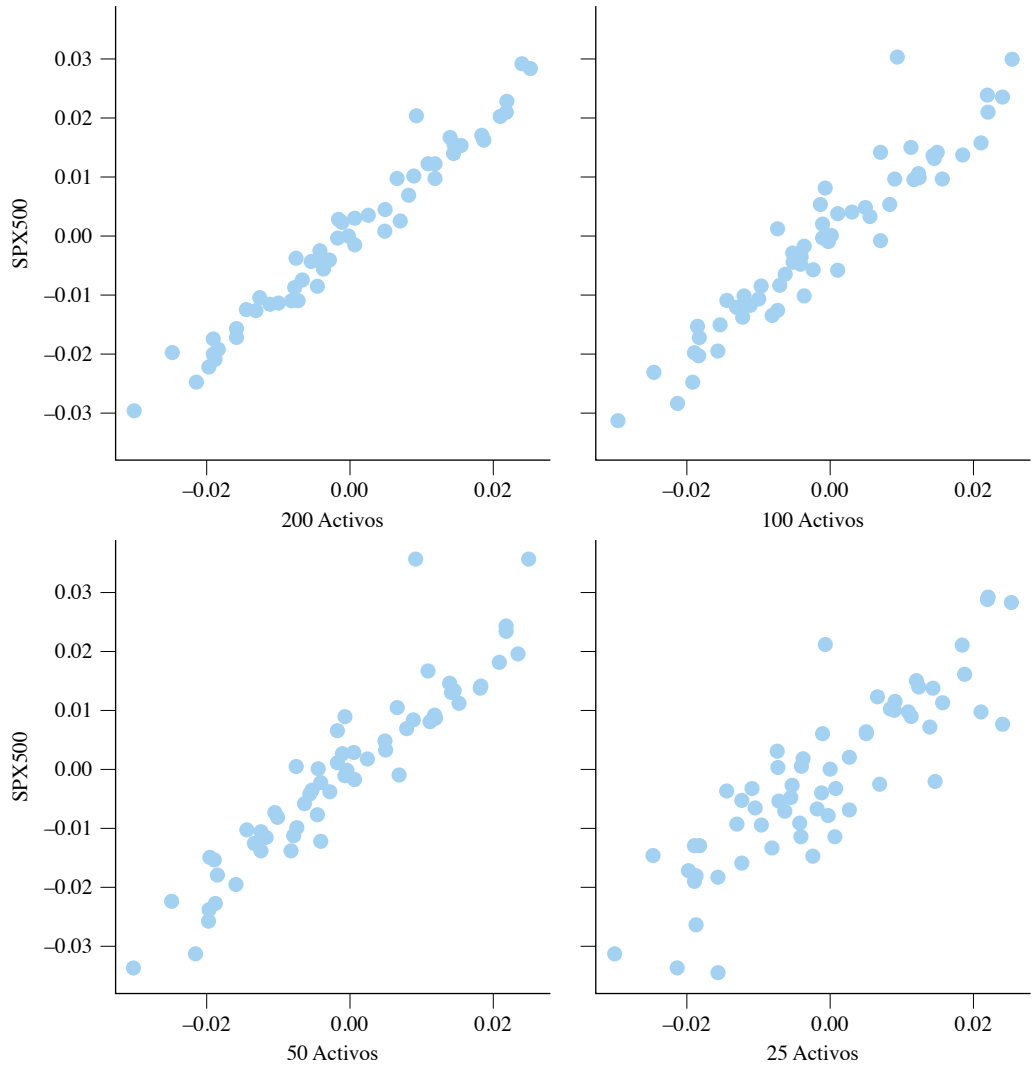
Fuente: elaboración propia.

Figura 6: Gráficos de dispersión de los portafolios seleccionados con el CAPM equiponderando los activos y el SPX500 para la data de prueba



Fuente: elaboración propia.

Figura 7: Gráficos de dispersión de los portafolios seleccionados con el CAPM y ponderándolos por capitalización bursátil y el SPX500 para la data de prueba

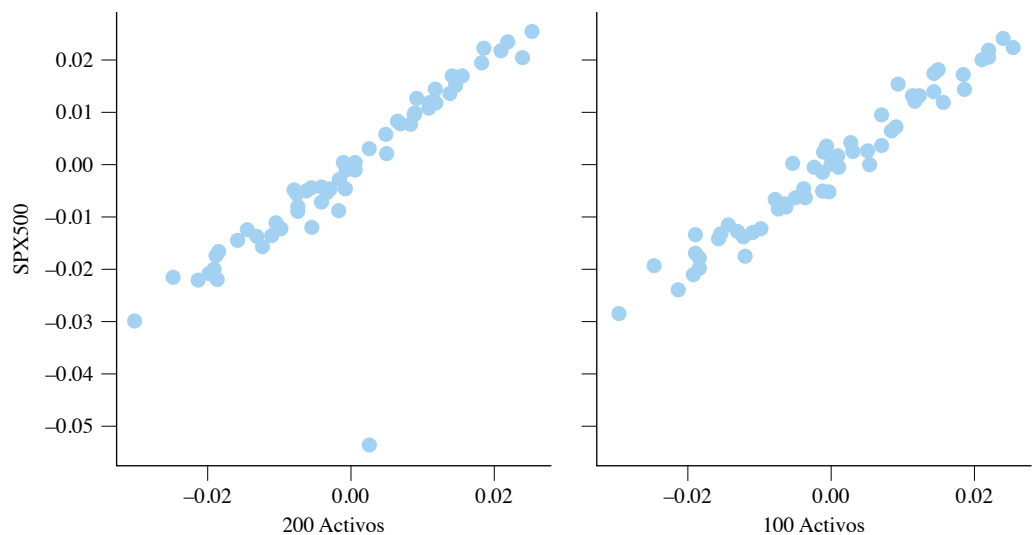


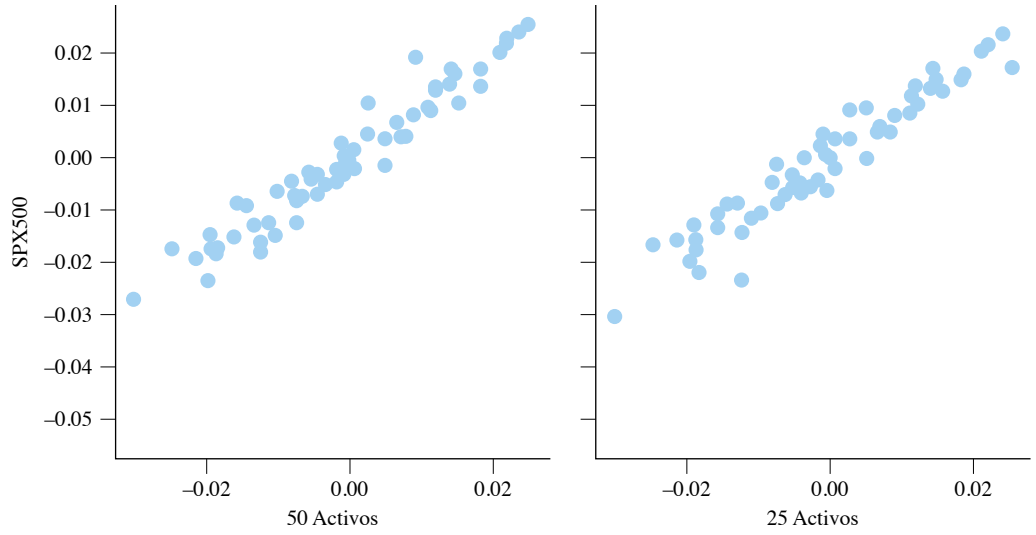
Fuente: elaboración propia.

Cabe rescatar, además, que dentro de los portafolios hallados por el CAPM, la marcada relación lineal explica los resultados con los beta que se ubicaron entre 0,89 y 1,04 para todas las carteras de inversión. Esto lleva a una primera conclusión de los experimentos llevados a cabo con esta metodología, aunque si bien es cierto que existe error dentro de los mercados financieros por su aproximación a una caminata aleatoria, se puede ver cómo el obtener un portafolio dentro de una submuestra con media-varianza eficiente lleva a resultados consistentes con el modelo.

Igualmente, en el análisis de componentes principales *Sparse* se puede observar una consistencia a través de los diferentes portafolios, con un dato atípico para la selección con 200 activos en las optimizaciones, dadas las acciones usadas para el experimento, que además llevó a otros datos atípicos causados por un par de componentes del portafolio (Google, Netflix y Amazon), que se notó menos en los otros experimentos dado que tenían una ponderación obtenida por equivalencias y no por las correlaciones (en el caso de la optimización) ni por la capitalización de mercado.

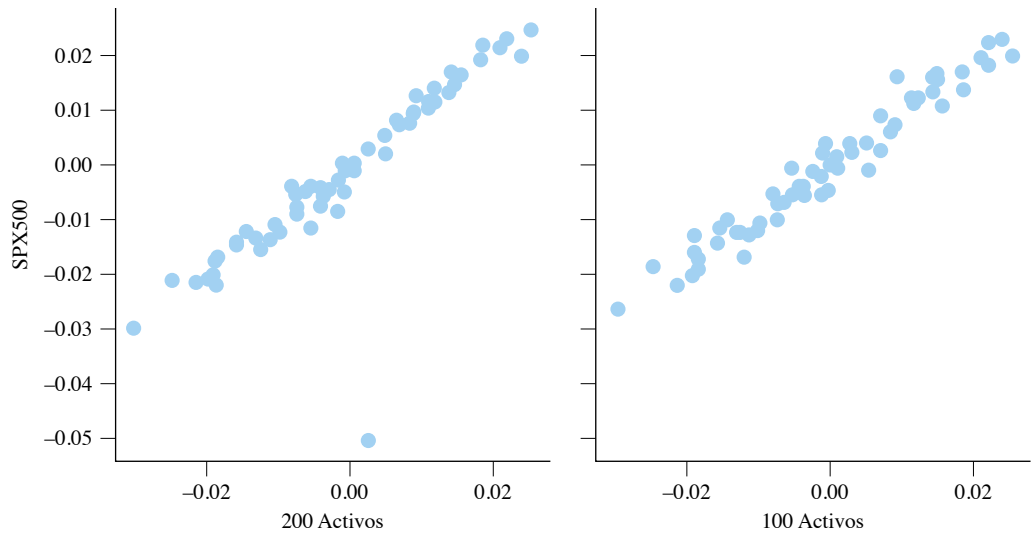
Figura 8: Gráficos de dispersión de los portafolios seleccionados con *SparsePCA* y el problema de optimización 8 y el SPX500 para la data de prueba

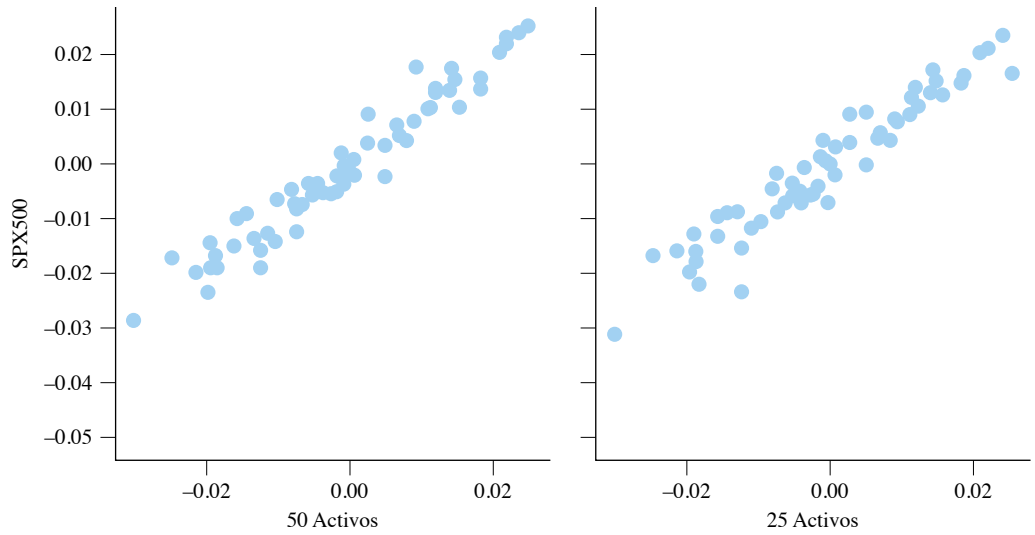




Fuente: elaboración propia.

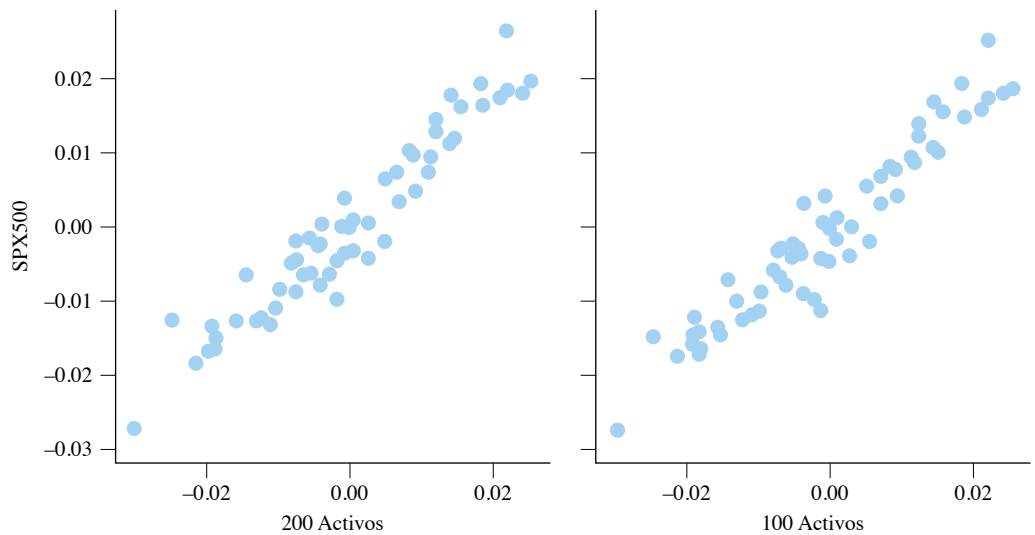
Figura 9: Gráficos de dispersión de los portafolios seleccionados con *SparsePCA* y el problema de optimización 9 y el SPX500 para la data de prueba

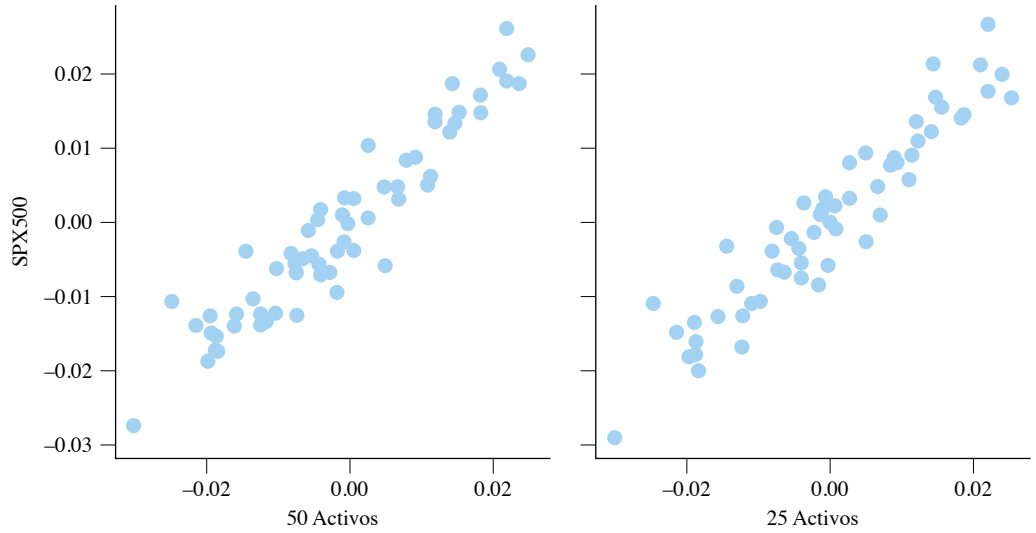




Fuente: elaboración propia.

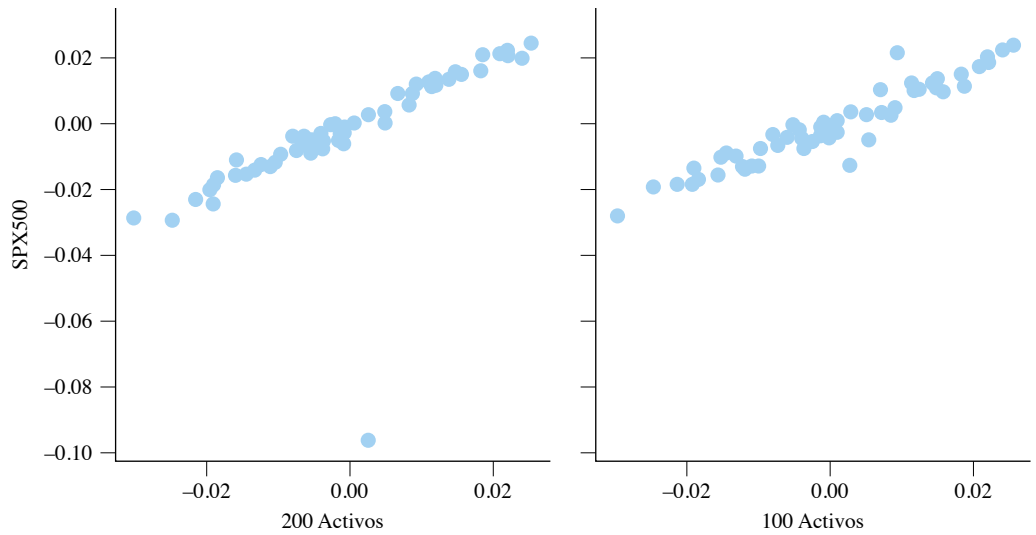
Figura 10: Gráficos de dispersión de los portafolios seleccionados con *SparsePCA* y equiponderando los activos y el SPX500 para la data de prueba

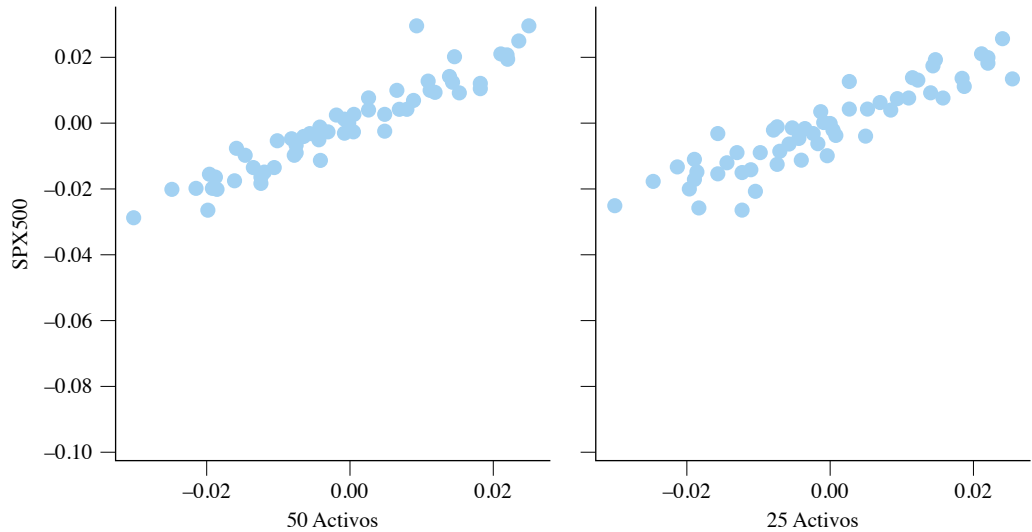




Fuente: elaboración propia.

Figura 11: Gráficos de dispersión de los portafolios seleccionados con *SparsePCA* y equiponderando los activos y el SPX500 para la data de prueba





Fuente: elaboración propia.

Finalmente, luego de los múltiples experimentos realizados, se puede llegar a la segunda conclusión: la transmisión de información al mercado es rápida y eso expone por qué con algunos pocos activos se puede explicar la gran mayoría de un índice que tiene 504 integrantes. Adicionalmente, se puede considerar cada uno de los experimentos llevados a cabo como una forma de buscar una distribución de portafolio pseudo indexada para una estrategia activa. Es decir, tener alguna de estas como la parte pasiva y con el resto del capital buscarle un alfa al mercado.

Cabe resaltar que, además, se debe tener en cuenta que esto abre otro tipo de estudios que pueden partir de aquí, como el hecho de investigar con *main-fold learning* el comportamiento de los activos que componen el SPX500, de manera que se levante aún más el supuesto de linealidad que hay detrás de los modelos usados.

Al partir del hecho de que la primera sección tiene base teórica en el CAPM, que es un modelo ampliamente aceptado en el mercado y que ha sido muy estudiado, y que el análisis de datos tal como se presentó es una parte incipiente (al menos para lo que respecta a la indexación), se puede concluir que el segundo modelo es mejor para este propósito, dados los resultados. Sin embargo, se debe seguir indagando con técnicas que sean menos lineales y, además, cada tres meses se deben reevaluar los modelos dada la periodicidad de rebalanceo del índice.

Referencias

- Aldana-Martínez, N. (2021). *Comparison of dimensionality reduction techniques of spx500: An approach to indexing* [Tesis de maestría, Universidad de los Andes]. <http://hdl.handle.net/1992/55768>
- Anadu, K., Kruttli, M., McCabe, P., Osambela, E. y Shin, C. H. (2018). The shift from active to passive investing: potential risks to financial stability? *Finance and Economics Discussion Series (2018-060r1)*
- Appel, I. R., Gormley, T. A. y Keim, D. B. (2016). Passive investors, not passive owners. *Journal of Financial Economics*, 121(1), 111-141.
- Bach, F. Mairal, J., Ponce, J. y Sapiro, G. (2009). *Online Dictionary Learning for Sparse Coding*. Proceedings of the 26th International Conference on Machine Learning.
- Cremers, M., Ferreira, M. A., Matos, P. y Starks, L. (2016). Indexing and active fund management: International evidence. *Journal of Financial Economics*, 120(3), 539-560.
- Farboodi, M., Matray, A., y Veldkamp, L. (2018). Where Has All the Big Data Gone? En *SSRN Electronic Journal*. Elsevier BV. <https://doi.org/10.2139/ssrn.3164360>
- Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical Work. *The Journal of Finance*, 25(2), 383.
- Fama, E. F. (1991). Efficient capital markets: II. *The Journal of Finance*, 46(5), 1575-1617.
- French, K. R. (2008). Presidential Address: The cost of active investing. *The Journal of Finance*, 63(4), 1537-1573.
- Gârleanu, N. y Pedersen, L. H. (2021). Active and passive investing: Understanding Samuelson's Dictum. *The Review of Asset Pricing Studies*, 12, 389-446.
- Guerra-Urzola, R., van Deun, K., Vera, J. C. y Sijtsma, K. (2021). A guide for Sparse PCA: Model comparison and applications. *Psychometrika*, 86(4), 893-919.
- Jenatton, R., Obozinski, G. y Bach, F. (2009). *Structured Sparse Principal Component Analysis*. <http://arxiv.org/abs/0909.1440>

- Lee, J. (2020). Passive Investing and Price Efficiency. *SSRN Electronic Journal*.
- Lintner, J. (1965). The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. *The Review of Economics and Statistics*, 47(1), 13.
- Malikov, G. (2018). Information, participation, and passive investing. *SSRN Electronic Journal*.
- Malkiel, B. (2012). *A random walk down Wall Street* (10 ed.). WW Norton.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... *et al.* (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Roll, R. (1977). A critique of the asset pricing theory's tests Part I: On past and potential testability of the theory. *Journal of Financial Economics*, 4(2), 129-176.
- Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3), 425-442.
- Stambaugh, R. F. (2014). Presidential address: Investment noise and trends. *Journal of Finance*, 69(4), 1415-1453.
- Treynor, J. L. (1961). Market Value, Time, and Risk. En *SSRN Electronic Journal*. Elsevier BV. <https://doi.org/10.2139/ssrn.2600356>
- Zou, H., Hastie, T. y Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2), 265-286.