Capital distribution of equity market and its statistical modeling

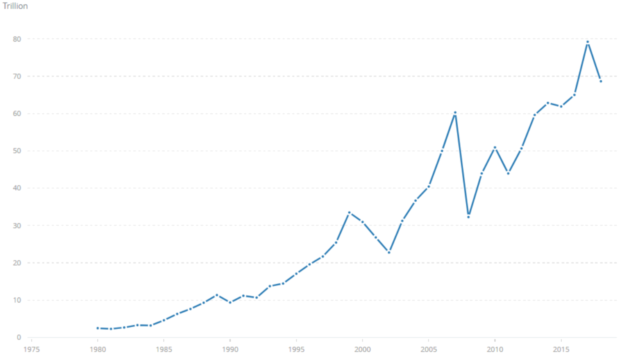
Ting-Kam Leonard Wong

University of Toronto

Ongoing project with Heng Kan and Colin Decker (U of T)

Equity market

As of 2018, the size of the world stock market (total market capitalization) was about US$69 trillion.



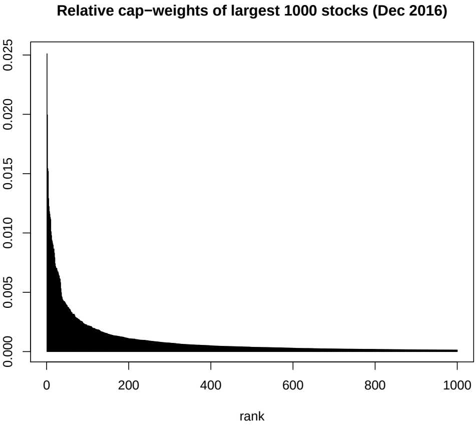
US market: about US$30 trillion in 2018.

Source: `https://data.worldbank.org/indicator/cm.mkt.lcap.cd`

Investing in the equity market

- Passive: Track a pre-defined benchmark (lower fee)
- Active: Aim to outperform by adopting various strategies

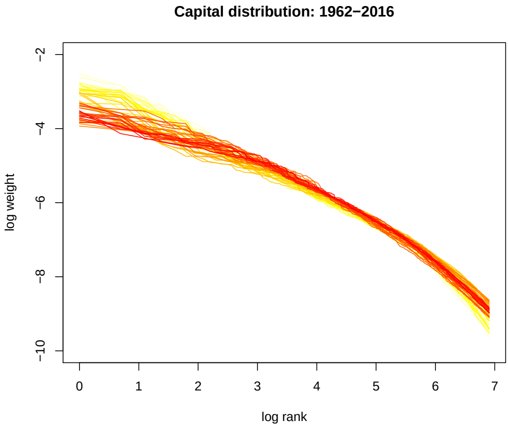
Benchmark: Usually a *cap-weighted* market index such as S&P500



Data: CRSP (Special thanks to Johannes Ruf and Desmond Xie)

Stability of capital distribution

Cap-weight of stock i: $\mu_i(t) = \mathrm{MC}_i(t) / \sum_j \mathrm{MC}_j(t)$.
Ranked weights: $\mu_{(1)}(t) \geq \mu_{(2)}(t) \geq \cdots \geq \mu_{(n)}(t)$.

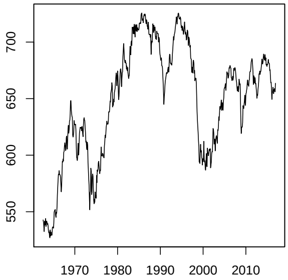


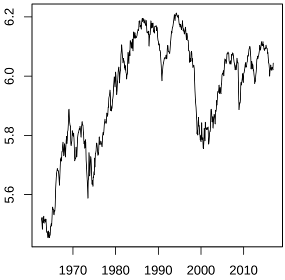
Market diversity (Fernholz 1997)

Diversity: a measure of the degree of concentration. Examples:

$$\Phi(\mu) = (\sum_j \mu_j^\lambda)^{1/\lambda} \quad (0 < \lambda < 1)$$

$$\Phi(\mu) = -\sum_j \mu_j \log \mu_j \quad \text{(Shannon entropy)}$$





Relevence of market diversity

Changes in market diversity explains a statistically and economically significant amount of variation in the relative returns of actively managed institutional large cap strategies. Data from Fernholz (2002) and Agapova, Greene and Ferguson (2011):

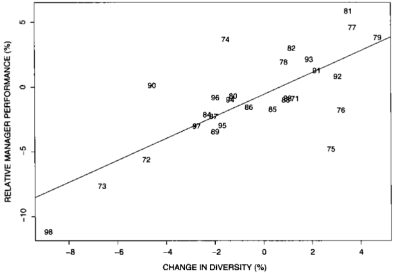
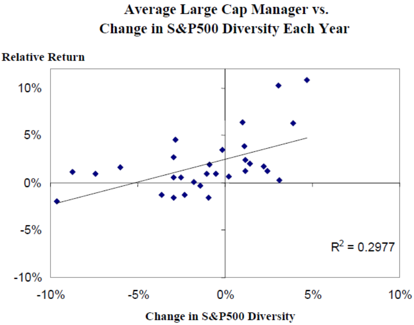


FIGURE 7.6. Manager performance relative to S&P 500 vs. change in $\mathbf{D}_p$. Manager performance data from Callan Associates: 1971–1998.



Theoretical explanation using SPT (Fernholz 2002)

Consider the *diversity-weighted portfolio*

$$\pi_i(t) = \frac{\mu_i(t)^\lambda}{\sum_{j=1}^n \mu_j(t)^\lambda}$$

which corresponds to the gradient of $\Phi(\mu) = (\sum_j \mu_j^\lambda)^{1/\lambda}$ which is *concave*:

$$\nabla\Phi(p) \cdot (q - p) \geq \Phi(q) - \Phi(p).$$

If $\varphi = \log \Phi$, then

$$\underbrace{\log\left(\nabla\varphi(p) \cdot \frac{q}{p}\right)}_{\text{relative log return}} \geq \underbrace{\varphi(q) - \varphi(p)}_{\text{change in diversity}}.$$

This is an example of Fernholz's *functionally generated portfolio*.

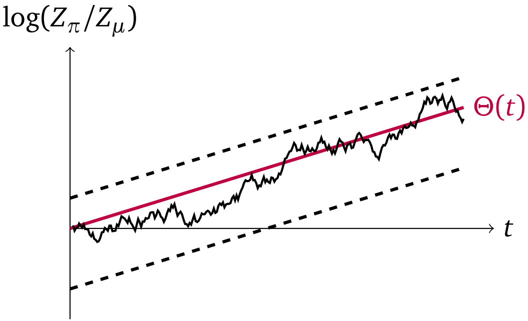
- Pal and W. (2015+): Optimal transport & information geometry.

Theoretical explanation using SPT

In a simplified Itô process model for the equity market, we have

$$\log \frac{Z_\pi(t)}{Z_\mu(t)} = \varphi(\mu(t)) - \varphi(\mu(0)) + \Theta(t),$$

where $\Theta(t)$ is an increasing process reflecting market volatility.



From this formula, the relative performance of the portfolio correlates with change in diversity.

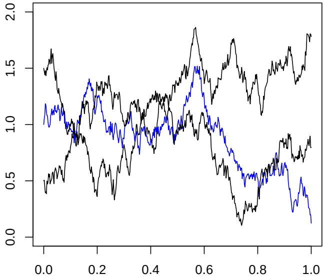
Modeling the capital distribution curve

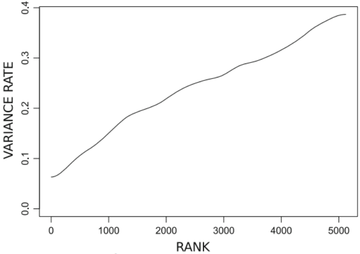
Mathematical approach

- Ranked-based models: an SDE system of *interacting* particles, each representing the market cap of a firm. A simple version:

$$d\,(\mathrm{Firm}_i(t)) = \gamma_{r_i(t)}dt + \sigma_{r_i(t)}dW_i(t),$$

where $r_i(t)$ is the *rank* of firm i at time t.





Smoothed values of σ_k^2, $k = 1, \ldots, 5{,}120$, for U.S. market: 1990–1999

Source (RHS): Fernholz, Ichiba and Karatzas (2013)

On the other hand Audrino Fernholz and Ferretti (2007) proposed and tested a model for forecasting market diversity.

Data

We are interested in dynamic modeling and forecasting, but as a first step we consider exploratory data analysis.

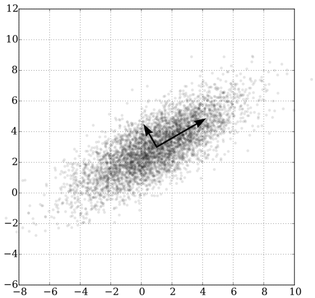
- Data source: CRSP (The Center for Research in Security Prices)
- Prepared and used by Johannes Ruf and Desmond Xie (2019)
- Daily data from 1962 to 2016.
- Market cap and daily returns for all US stocks.

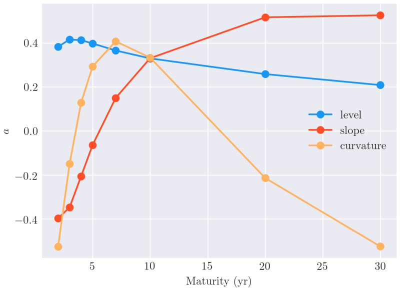
We focus on the largest 1000 stocks.

- Ranked-based
- Evolving universe, missing data.
- For simplicity we use monthly data in this study (reasonable time scale for large cap portfolio managers).
- Renormalize to get (relative) (ranked) market weights $\{\mu_{\leq}(t)\} \subset \Delta_{1000}$. May be regarded as a functional time series.

Principal component analysis

Performs dimension reduction by projecting the data linearly to a low dimensional subspace. A classic application in finance is *yield curve modeling*; the eigenvectors have useful interpretations.





Left: Wikipedia. Right: `https://quant.stackexchange.com`

Reminder of PCA

Suppose we observe data $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$. PCA aims to find a low-dimensional $\mathbf{y} = W\mathbf{x} \in \mathbb{R}^d$, $d < D$, that approximates $\mathbf{x}$:

$$\min_{U \in L(\mathbb{R}^d, \mathbb{R}^D), W \in L(\mathbb{R}^D, \mathbb{R}^d)} \sum_{i=1}^{N} \|\mathbf{x}_i - UW\mathbf{x}_i\|_2^2.$$

Here W and U are *linear* maps (i.e., matrices).

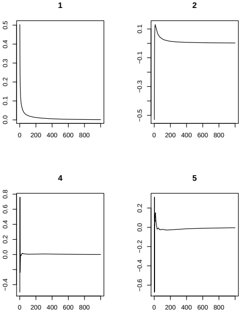
Theorem

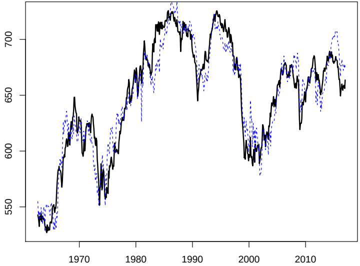
Let $A = \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^\top$, and let $\mathbf{u}_1, \ldots, \mathbf{u}_d$ be the eigenvectors corresponding to the largest d eigenvectors of A. Then the PCA problem is solved with

$$U = [\mathbf{u}_1, \ldots, \mathbf{u}_d], \quad W = U^\top.$$

Plain vanilla PCA does not work well

Since our data lies in the unit simplex, Euclidean PCA may (and does) not work well.

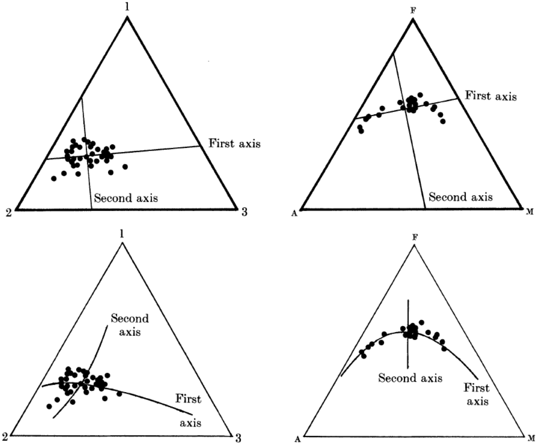




The blue series approximates the diversity using 5 eigenvectors. It does not track the short term fluctuations of the diversity.

Compositional data analysis, simplicial PCA

We will apply a version of PCA using the *Aitchison geometry*.



Source: Aitchison (1982)

Aitchison geometry

Consider the open unit simplex

$$\Delta_n := \{p = (p_1, \ldots, p_n) : p_1 + \cdots + p_n = 1\}.$$

Define the *closure* operator

$$C[\mathbf{x}] := \left(\frac{x_1}{x_1 + \cdots + x_n}, \ldots, \frac{x_n}{x_1 + \cdots + x_n} \right), \quad \mathbf{x} \in (0, \infty)^n.$$

Theorem

Define the operations

$$p \oplus q := C[p_1 q_1, \ldots, p_n q_n] \quad \textit{(perturbation)}$$
$$\lambda \otimes p := C[p_1^\lambda, \ldots, p_n^\lambda] \quad \textit{(powering)}$$

Then $(\Delta_n, \oplus, \otimes)$ *is an* $(n-1)$*-dimensional real vector space.*

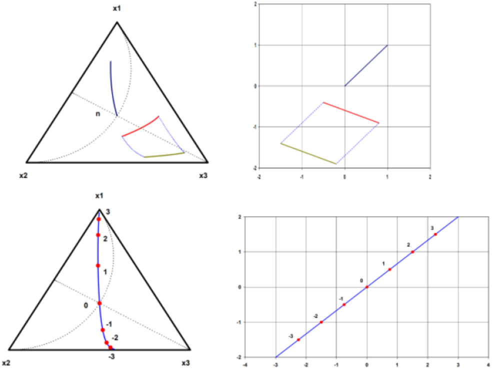
Aitchison geometry

Theorem (Simplex as a Hilbert space)

s Define

$$\langle p, q\rangle_A := \sum_{i=1}^{n} \log \frac{p_i}{g(p)} \log \frac{q_i}{g(q)},$$

where $g(x) = (x_1 \cdots x_n)^{1/n}$ *is the geometric mean. Then* $\langle \cdot, \cdot \rangle_A$ *is an inner product on* $(\Delta_n, \oplus, \otimes)$.



Isometric-log-ratio transform

Consider the orthonormal basis

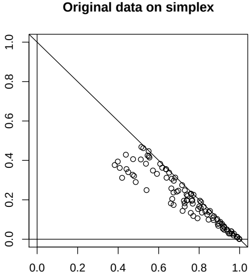
$$\mathbf{e}_i = C\left[\exp\left(\sqrt{\frac{1}{i(i+1)}},\ldots,\sqrt{\frac{1}{i(i+1)}},-\sqrt{\frac{i}{i(i+1)}},0,\ldots,0\right)\right],$$

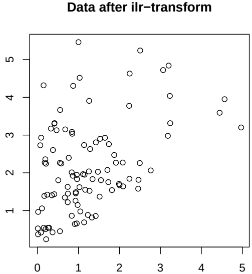
where $i = 1,\ldots,n-1$.

Definition (Egozcue et al (2003))

We define $\mathrm{ilr} : \Delta_n \to \mathbb{R}^{n-1}$ by

$$\mathbf{x} = \mathrm{ilr}(p) := (\langle p,\mathbf{e}_1\rangle_A,\ldots,\langle p,\mathbf{e}_{n-1}\rangle_A), \quad x_i = \sqrt{\frac{i}{i+1}}\log\frac{(p_1\cdots p_i)^{1/i}}{p_{i+1}}.$$



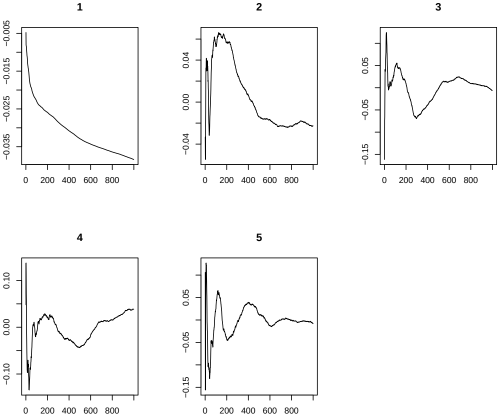


Simplicial PCA

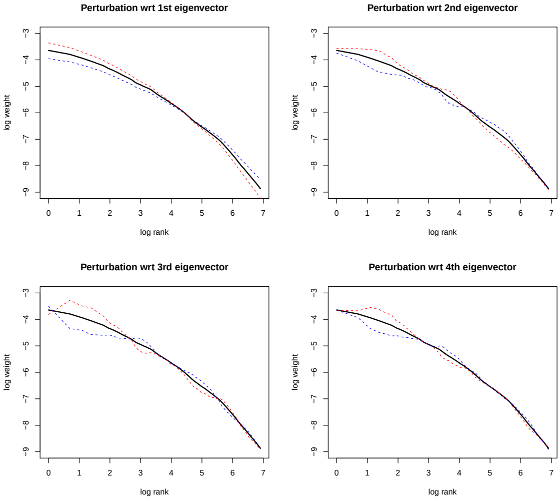
Aitchison (1982) uses instead the centered-log-ratio (clr) transform. Here we use the ilr-transform. Procedure:

original data → ilr → Euclidean PCA → inverse-ilr

Eigenvectors (in ilr-space) using the latest 20 years of data:



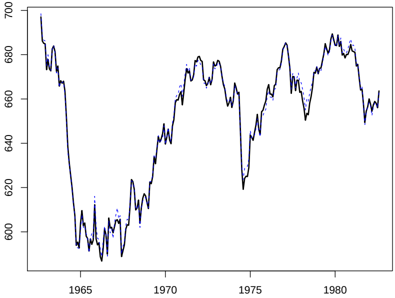
Interpretations of the eigenvectors

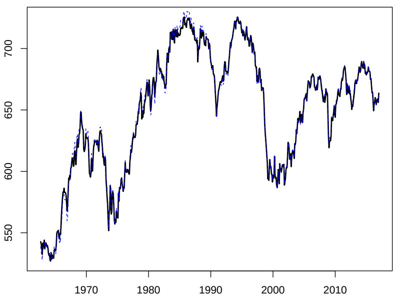


The shapes of the eigenvectors (except the first one and perhaps the 2nd) fluctuate over time. This may indicate some *structural changes* in the market despite stability of the capital distribution.

Approximation of market diversity

Again we use only the first 5 eigenvectors. The simplex PCA performs much better than the Euclidean PCA in capturing the market diversity.





Intuitively, the tracjectory of $\mu_{\geq}(\cdot)$ is close to a 5-dimensional submanifold of Δ_n. Right: Result using the entire dataset.

Future directions

- Further study of the geometry in relation to the properties of the data
- Financial interpretations
- Dynamic statistical models
- Forecasting
- Portfolio optimization
- Modeling of volatility using (information) geometric idea
- Combine with mathematical approaches in SPT