Decompose Anything!

Grant Holtes

July 21, 2024

Introduction

Decomposition is providing an answer to the question of *why* a certain value is what it is, and being able to communicate that in a way that is easily understood and accepted by a wide range of audiences. This is a task that generally leans towards the art side of the art-science spectrum, and this note aims to provide a brief overview of the techniques that I have found useful.

The note starts by covering simple additive decomposition, where the whole is defined by the sum of its parts. It then covers decomposition by multiplication, where a whole is the product of its parts, with applications to ratios and compounded return values. It then takes a detour to discuss multi-asset return attribution and the issues that arise in this specific use case. This is contrasted with the statistical technique of factor modelling and its approach to attributing returns. Finally models are discussed, which includes some model-agnostic techniques to attribute a model's output to its inputs and some that aim to measure a model's sensitivity to its inputs.

Sum Decomposition

Sum decomposition is arguably the simplest and most intuitive method. Take a quantity of interest and split it into parts such that the whole, S is the sum of the parts, c_i.

$$S = \sum_{i=1}^{N} c_i \tag{1}$$

For a simple example, consider a national retail chain: They could choose to decompose their dollar sales values by state of sale, which would be a valid decomposition. They could also choose to decompose sales by the category of item sold, which is equally valid. You would choose between these based on the question you wanted to answer or the narrative you wanted to create.

In other applications the validity of the decomposition is not guaranteed and you may be left with a residual. Consider a decomposition of equity returns over a year. This can be done a number of ways, but one is to decompose returns into:

1. Inflation
2. Income (dividend yield / buy backs)

Given all three of Return, Inflation and Income are known values, you will almost certainly end up with some residual term here. Given the context we can label this calculated residual as something like capital growth, as in the below chart. Roger G Ibbotson, 2001 provide a great graphic that shows how the return on equity from 1926-2000 can be decomposed in wide range different ways, all of which use this additive methodology.

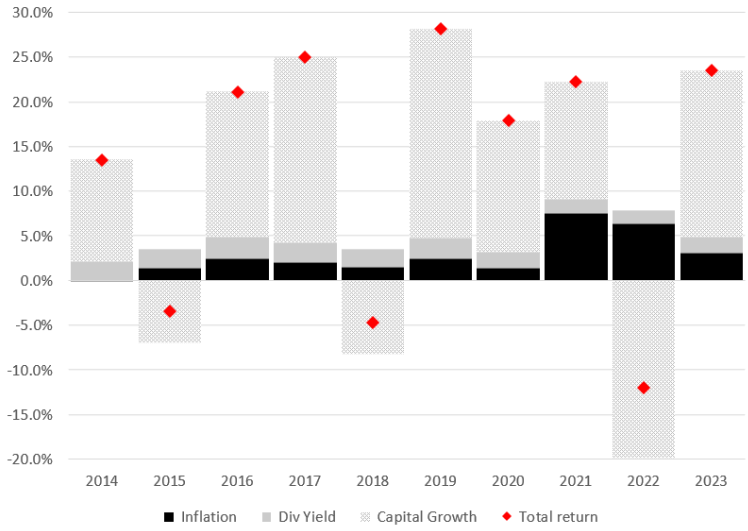

Figure 1: Simple equity returns decomposition for the S&P 500. Data from Robert Shiller's CAPE model.

Hierarchical Sum Decomposition

To further improve communication it may be beneficial to have some hierarchy to the decomposition. To use the retail store example, you could first divide sales by region, then for each region divide by product category.

Ratio Decomposition

The only thing better than a ratio is more ratios, so its great that you can decompose a ratio into the product of a ratios. Lets say that we care about the ratio of two variables, A and B. We can "expand" $\frac{A}{B}$ into a chain of ratios:

$$\frac{A}{B} = \frac{A}{x}\frac{x}{y}\frac{y}{z}\frac{z}{B} \tag{2}$$

If we choose the other variables x, y, z carefully, this can be a great technique to explore how the components of the ratio of interest have changed over time.

Lets consider GDP per capita[1] as as example - Here our ratio is $\frac{GDP}{Pop}$. For example lets ignore the

[1]Real GDP per capita is used here

roles of capital and investment and investigate how changes in labour have impacted our GDP per capita. Here we propose that there are 3 key ratios we can decompose GDP per capita into: GDP per hour worked, a proxy for productivity; Hours worked per employed person; and Employment rate.

$$\frac{GDP_t}{Pop_t} = \frac{GDP_t}{Hours_t}\frac{Hours_t}{Employed_t}\frac{Employed_t}{Pop_t} \tag{3}$$

The results of this decomposition can be seen in the graphs below.

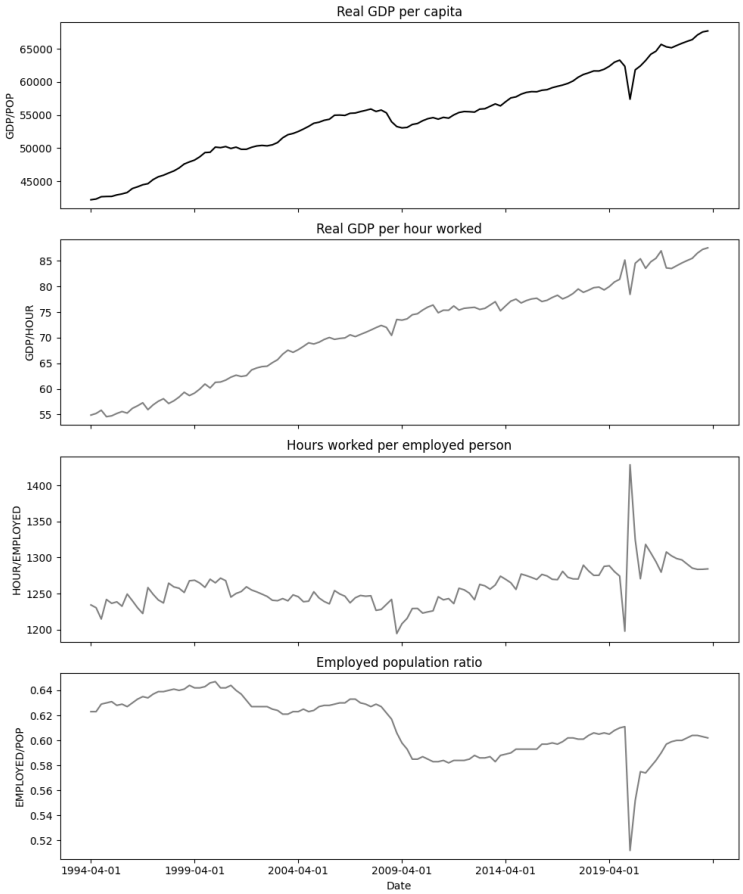

Figure 2: Decomposition of real GDP per capita

This method is most helpful with time series data, with each ratio calculated for each observation period. In this case, the change in each ratio is also a valid decomposition. The percentage growth in $\frac{GDP}{Pop}$ is equal to the compounding of each of the decomposition ratios.

Date	GDP/POP	GDP/HOUR	HOUR/EMPLOYED	EMPLOYED/POP
2015-01-01	58152	76.78	1277	0.593
2019-10-01	63283	81.41	1274	0.610
Growth	**8.82%**	6.03%	-0.22%	2.87%
2019-10-01	63283.72	81.41	1274	0.61
2024-01-01	67695.29	87.56	1284	0.602
Growth	**6.97%**	7.55%	0.78%	-1.31%

Table 1: Decomposition of GDP per capita growth

For example, in the table we can see that GDP per capita growth was greater in 2015-19 (8.82%) than in 2019-24 (6.97%). In 2015-19, 6% of this 8.82% growth was from our productivity measure and 2.9% was from an increase in employment rate. However, in 2019-24, we got 7.55% growth in productivity, which was undermined a 1.3% decrease in employment rate, giving us our 7% total GDP per capita growth.

You can see how this method can be used to string together a narrative on why changes have occurred in the ratio of interest. We are implying adding our decomposed growth rates when we should really be compounding them, but it gives a fair idea of where the changes have occurred.

There is a art in selecting the other variables to use in these decomposition's, and the method will obviously still work even if you select completely unrelated variables. For this reason its probably best to have some theoretical basis for the decomposing ratios.

Decomposition of bond returns

A specific use case of this technique is decomposing bond returns into roll (return due to reduction in duration) and yield curve shifts (return due to changes in underlying rates at a constant duration).

Let the price of a bond with a duration D at time period t be given as $p(D,t)$. The one year return on a bond can then be calculated as:

$$\frac{p(D-1,t+1)}{p(D,t)} - 1 \tag{4}$$

This can decomposed into:

$$\frac{p(D-1,t)}{p(D,t)} \frac{p(D-1,t+1)}{p(D-1,t)} - 1 \tag{5}$$

The first term is the roll term component (holds time constant), while the second term is the yield curve shift component (holds duration constant).

Decomposing Compounded Returns

In the ratio decomposition method we came across the issue that we have a valid decomposition of a growth rate into compounding components, but we want to decompose them into additive components.

Consider the decomposition of a growth rate x into y, z in:

$$x = (1 + y)(1 + z) - 1 \tag{6}$$

If x, y, z are all small, the approximation will hold that $x \approx y + z$, but there will always be a small residual[2] which is disappointing. In exact additive decomposition this is solved by simply labelling the residual as an interaction effect, giving it a place as its own "component".

Of course we could convert it all to logs, with $ln(1 + x) = ln(1 + y) + ln(1 + z)$. This removes the residual, but now everything is in log terms, which doesn't lend itself to easy communication. What we really want is a way to convert y, z to some new values y', z' such that $x = y' + z'$. This would allow for easy communication of the decomposition, for example:

> Our revenue growth (x) is 5%. 4% of this is caused by increased USD sales (y) and 1% is caused by an appreciation of the Australian dollar (z). So 80% of our sales growth was caused by factors under our control.

One method is to linearly re-scale the components to force their sum to be equal to x:

$$y' = y\frac{x}{y + z} \tag{7}$$

This is equivalent to spreading out the residual between each component, with the allocation being proportional to the size of the component - a *weighted distribution of the interaction effect.*

Another approach would be to split the residual equally between each component, perhaps on the basis that every component is equally "responsible" for the existence of the interaction component. This *proportional distribution of the interaction effect* may be arguable in the 2-component case, but becomes less justifiable in the case of 3 or more variables. Consider the case of a decomposition into $x = (1 + a)(1 + b)(1 + c) - 1$. When approximated as $x = a + b + c$ it has the residual of $abc + ab + ac + bc$. If c happens to zero it makes little sense to attribute any of the remaining residual (ab) to it when calculating c'.

Returns Attribution

Multi-asset portfolio returns attribution is a specific decomposition task that is worthy of separate mention. The task is usually to explain why the return on a multi-asset portfolio differs from the return on a benchmark or other portfolio. The methods described below are only a subset of those outlined in Bacon, 2019.

Its worth noting that there are other returns attributions techniques that answer asset-class specific questions. The same techniques here can be applied within equity or real asset classes, where the asset-class segments are replaced by some other category, such as industry or geography. Bottom up or stock level attribution can be used to attribute returns to individual equities. Fixed income offers its own challenges, where decompositions may look to attribute returns to decisions on duration and carry.

[2]The residual is yz in this case, simply found by expanding the original definition

Single Period Arithmetic Attribution

This method is described by Brinson and Fachler, 1985 and uses the following notation:

The portfolio weight on asset class i: w_i

The benchmark weight on asset class i: W_i

The portfolio return on asset class i: r_i

The benchmark return on asset class i: b_i

The overall benchmark return: b

In this model the return is decomposed into 3 components:

Asset Allocation: Differences in return due to differences in asset class weights. For each asset class this is given as $(w_i – W_i)(b_i – b)$, with the total contribution being given by the sum over all asset classes.

Asset Selection: Differences in return due to differences in asset selection within each asset class. For each asset class this is given as $W_i(r_i – b_i)$, with the total contribution being given by the sum over all asset classes.

Interaction: Differences in return due to the cross product of asset class and asset selection effects. For each asset class this is given as $(r_i – b_i)(W_i - w_i)$, with the total contribution being given by the sum over all asset classes. Some prefer to combine this with the asset selection effect, which is justified by assuming that asset class weights are chosen first, then the assets themselves.

The sum of these effects give the total difference in returns between the portfolio and the benchmark, $r - b$. This decomposition is often shown graphically as a series of areas, as below:

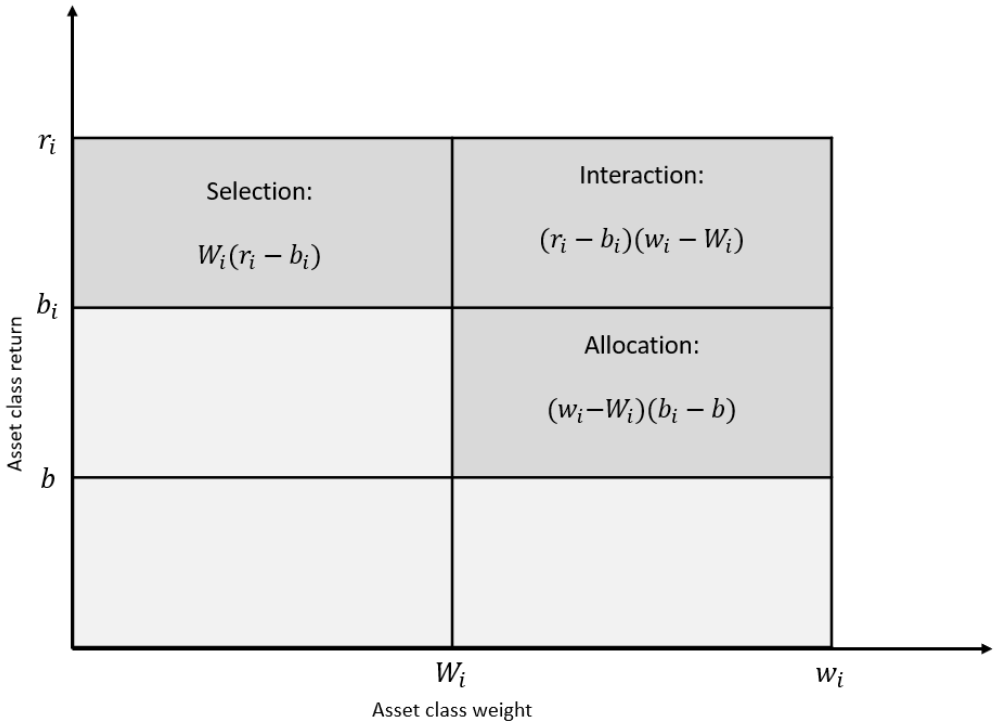

Figure 3: Brinson and Fachler, 1985 decomposition

This method has the same rather unsatisfactory residual was discussed in the previous section on decomposing single period compounded returns, we have just renamed it as *interaction*. It is also unclear how to apply this method across multiple time periods.

Single-and-Multi-Period Geometric Returns Attribution

One way to avoid the issues described before is using geometric returns and the ratio decomposition method. The difference in geometric returns between the portfolio and the benchmark for a given period is given by:

$$R = \frac{1+r}{1+b} - 1 \tag{8}$$

We can decompose this into the asset allocation and asset selection effects as before by using the ratio technique:

$$R = \frac{1+r}{1+b} - 1 = \frac{1+b_s}{1+b}\frac{1+r}{1+b_s} - 1 \tag{9}$$

Where b_s is the return on a portfolio with the asset class weights as in the portfolio and the asset class returns of the benchmark: $b_s = \sum_{i=0}^{N} w_i b_i$. This represents the return due to asset allocation only.

It follows that A gives the return from asset allocation and S gives the return from stock selection.

$$A = \frac{1 + b_s}{1 + b} - 1 \tag{10}$$

$$S = \frac{1 + r}{1 + b_s} - 1 \tag{11}$$

All these terms compound across time periods, so this methodology easily handles multi-period use cases, as is shown below for R.

$$R_t = \frac{1 + r_t}{1 + b_t} - 1 \tag{12}$$

$$R = \prod_{t=1}^{T} (\frac{1 + r_t}{1 + b_t}) - 1 \tag{13}$$

The same expansion is valid for A_t, S_t, with the components compounding to the whole for the whole multi-period set $R = (1+A)(1+S)-1$ and each period individually $R_t = (1+A_t)(1+S_t)-1$.

As with the single period arithmetic attribution before, it would be beneficial to decompose A, S further, showing the contribution to each of A_t, S_t from each asset class. Bain, 1996, Burnie et al., 1998 and Bacon, 2002 suggest the following expressions to capture these asset class components:

$$A_{i,t} = (w_i - W_i)(\frac{1 + b_{i,t}}{1 + b_t} - 1) \tag{14}$$

$$S_{i,t} = w_i \frac{r_{i,t} + b_{i,t}}{1 + b_{s,t}} \tag{15}$$

Multi-Period True Geometric Returns Attribution

The issue with the asset class decomposition's outlined above is that the return within each time period is expressed additively, with $b = \sum_{i=1}^{I} W_i b_i$. This means that the below expression is true:

$$R = (1 + \sum_{i=1}^{I} A_i)(1 + \sum_{i=1}^{I} S_i) - 1 \tag{16}$$

However, this means that we cannot compound these decompositions as per the equation below, which would be the desired use of these decompositions:

$$R + 1 \neq \prod_{t=1}^{T} \prod_{i=1}^{I} (1 + A_{i,t})(1 + S_{i,t}) \tag{17}$$

There is a slight correction term or residual that is required to make the above statement hold true. Forgy, 2002 propose a number of different derivations for this correction term, of varying levels of complexity[3].

[3]Vary from very to extremely complex

The good news is this correction term is usually very small, even for large returns relative to the benchmark. As in the section on *decomposing single-period compounded returns*, we can perhaps get an adequate answer by simply distributing the residual across the components in a weighted manner. This is expressed a three step process to avoid finding a closed form solution for the residual, which would be a large and messy equation.

Step 1: Calculate components Calculate $A_{i,t}, S_{i,t}$ as per Equation 14 and Equation 15.

Step 2: Calculate the correction term For each time period, calculate the correction term required. Lets call these $\epsilon_{A,t}, \epsilon_{S,t}$. These represent the residual geometric return for each period that is not capture in the product between each of the asset class level allocation and selection components.

These are defined by $A_t + 1 = \epsilon_{A,t} \prod_{i=1}^{I}(1 + A_{i,t})$ and $S_t + 1 = \epsilon_{S,t} \prod_{i=1}^{I}(1 + S_{i,t})$, which simply rearranges to:

$$\epsilon_{A,t} = \frac{1 + A_t}{\prod_{i=1}^{I}(1 + A_{i,t})} \tag{18}$$

$$\epsilon_{A,t} = \frac{1 + S_t}{\prod_{i=1}^{I}(1 + S_{i,t})} \tag{19}$$

Step 3: Distribute residual Spread the correction term over each of the asset class components for each time period. Here the weight on the asset class is used to distribute the residual, but alternate weighting schemes would work equally well.

$$\tilde{A}_{i,t} = A_{i,t} \sqrt[w_i * t]{\epsilon_{A,t}} \tag{20}$$

$$\tilde{S}_{i,t} = S_{i,t} \sqrt[w_i * t]{\epsilon_{S,t}} \tag{21}$$

The desired property now holds:

$$R = \prod_{t=1}^{T} \prod_{i=1}^{I} (1 + \tilde{A}_{i,t})(1 + \tilde{S}_{i,t}) - 1 \tag{22}$$

This equation can be re-arranged to give a range of results:

1. Allocation and selection return contributions from a specific asset class for a specific time period are given by $\tilde{A}_{i,t}, \tilde{S}_{i,t}$ respectively.
2. Total return contributions from a specific asset class for a given time period is given by $(1 + \tilde{A}_{i,t})(1 + \tilde{S}_{i,t}) - 1$.
3. Cumulative selection return contribution from a specific asset class over all periods (or some subset of periods) is given by $\prod_{t=1}^{T}(1 + \tilde{S}_{i,t}) - 1$. The same method works for allocation and total return contribution.
4. We can still get the original allocation and selection effects, aggregated over all asset classes. For example, the allocation contribution for a specific period is given by $A_t = \prod_{i=1}^{I}(1+A_{i,t})-1$ and the allocation effect over all time periods would be $A_t = \prod_{t=1}^{T} \prod_{i=1}^{I}(1 + A_{i,t}) - 1$

Despite all these efforts, you are still left with a geometric decomposition, which has the issues discussed in the *decomposing compounded returns* section. As discussed in that section, if communication in additive terms is required, some additional re-scaling of the results will be required. If returns are relatively small, this scaling factor will not be large.

Factor Models

Factor models provide a simple linear statistical framework for decomposing a variable into potential explanatory factors, plus an idiosyncratic or unexplained component.

Single factor models

A *single factor model* refers to a statistical model that explains variations in a dependent variable X using only one explanatory variable or factor F. This model is typically expressed as:

$$X = \alpha + \beta F + \epsilon \tag{23}$$

where:

- X is the dependent variable (e.g., asset returns),
- F is the explanatory factor (e.g., market returns),
- α is the intercept (constant term),
- β is the coefficient representing the relationship between X and F,
- ϵ is the error term representing unexplained variations in X.

This model assumes that the variation in X can be predominantly explained by variations in the single explanatory factor F.

Multi-factor models

A multi-factor model is simply a multivariate extension of the single factor model:

$$X = \alpha + \sum_{i=1}^{N} \beta_i F_i + \epsilon \tag{24}$$

Decomposition with Factor Models

Once the factor loadings, β, have been estimated (usually using linear regression), the variable of interest can be decomposed into:

- $\beta_i F_i$ - The component attributable to factor i.
- ϵ - The unexplained or idiosyncratic component.

As the factor model is linear, this is an additive decomposition. As with the ratio decomposition, the choice of factors is crucial in getting a satisfactory result. Generally the factors should have some theoretical causal relationship with the variable of interest[4].

An example is given below, where market, currency and commodity price factors are used to supplement a market factor in decomposing returns on BHP, a resource extraction firm.

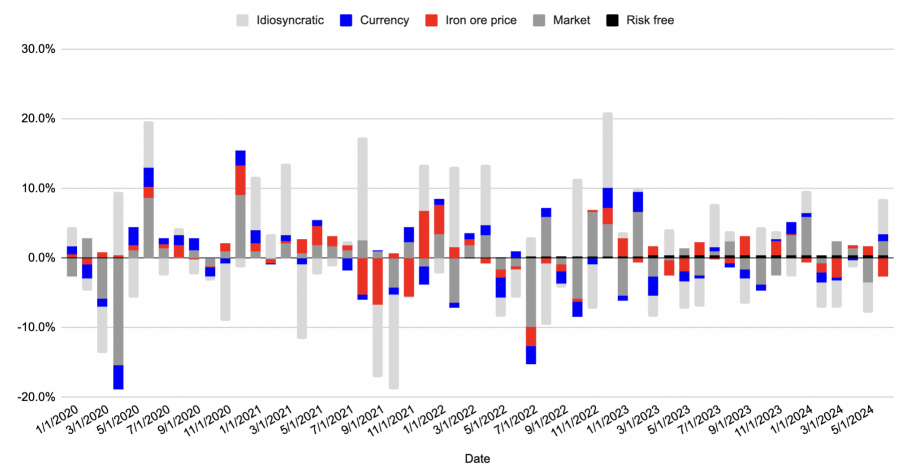

Figure 4: Factor decomposition of BHP's price return using a market factor, a current factor and a iron price factor. This is an overly simplistic example, as there is multicollinearity issues between currency and resource price returns in Australia

Excess Returns

When working with return data in factor models, it is convention to use excess returns for both the dependant variable and the factors, with the advantages of:

- Controlling for differing interest rates over time.
- Isolating the component of both asset and factor returns that are over-and-above those that are presumed to be available in the risk free rate.

Note that if the factor is constructed as a long-short portfolio with zero net exposure, it is not necessary to subtract the risk free rate.

How to Build a Factor Model

Determine Factors Factors should have an economic rationale and can be macroeconomic variables or portfolio returns. Examples include Chen, Roll, and Ross's macro factors and Fama and French's portfolio factors. It is easier statistically if factors are returns.

[4]Note that this may not be the set of factors with the highest explanatory power or R^2, as simply optimising for this may result in spuriously correlated variables being used.

Collect Data Gather data for both factors and test assets. It is recommended that you remove predictable components from macroeconomic factors. For factors that are based on portfolio returns, decide on rebalancing frequency (e.g., annually like Fama and French or monthly like Carhart).

Estimate Regressions There are two methods: Time-Series Regression or Cross-Sectional Regressions. Here we describe Time-Series Regression which is used when all factors are returns.

1. Calculate excess returns as required
2. Estimate regression slopes / factor loadings (β_i).
3. Factor premia (λ) can be estimated as the mean of factor returns.
4. Test if all $\alpha = 0$ using F-tests and check if factors are strongly related to expected returns. Sense check that the factor loadings have the expected signs.

Common Factor Model Use Cases

Capital Asset Pricing Mode

The Capital Asset Pricing Model (CAPM) is a single factor model and a foundational theory in finance that describes the relationship between the expected return $E[R_i]$ of an asset i and its systematic risk, measured by beta β_i, relative to the overall market. It is expressed by the formula:

$$E[R_i] = R_f + \beta_i(E[R_m] - R_f) \tag{25}$$

where:

- R_f is the risk-free rate,
- $E[R_m]$ is the expected return of the market portfolio.

This can be estimated using regression on:

$$R_{i,t} - R_{f,t} = \alpha + \beta(R_{m,t} - R_{f,t}) + \epsilon \tag{26}$$

Fama-French Four Factor Model

The Fama-French four factor model, developed by Eugene Fama and Kenneth French, extends the traditional Capital Asset Pricing Model (CAPM) by incorporating additional factors that aim to better explain stock returns. The model decomposes the total return of a stock into four distinct factors:

1. **Market Risk (Market Factor):** This factor captures the systematic risk associated with overall market movements, represented by the return of a broad market index such as the S&P 500.
2. **Size Factor (SMB - Small Minus Big):** This factor decomposes returns based on the market capitalization of stocks.

3. **Value Factor (HML - High Minus Low):** The value factor decomposes returns based on the book-to-market ratio of stocks. It reflects differences in returns between value and growth stocks.

4. **Profitability Factor (RMW - Robust Minus Weak):** Added later, this factor decomposes returns based on the profitability of companies.

Factor Covariance Matrix

As described by Jianqing Fan and Mincheva, 2011 and Dmitri Mossessian, 2014, factor models are a common and effective tool at estimating covariance matrices of asset returns. This is especially true where a large number of assets are included, which causes dimensionality issues with the usual sample covariance estimator.

I have covered the basics of how this method works in my note on the matter (Holtes, 2024), but the basic steps are:

1. Select assets, X, factors, F, and compute excess returns on both as required.
2. Compute factor loadings, β_i, for each asset, as in the usual multi-factor model.
3. Compute the covariance matrix between the factors, $\hat{cov}(F)$. This can use the sample covariance estimator or more exotic methods, as in Dmitri Mossessian, 2014.
4. Compute the asset covariance matrix as $\hat{cov}(X) = B^T \hat{cov}(F) B$.
5. Overwrite the asset variances to account for idiosyncratic volatility with $diag(\hat{cov}(X) = \hat{var}(X)$

Model Output Attribution

Another common attribution task is to understand why a model has provided a specific output. This is used to answer a range of questions:

- How important is each feature in determining a model's output, on average over a dataset, the *feature importance* question.
- Why is a specific output different to the average output over a dataset, one type of attribution.
- Given a two sets of inputs (a old and new), how can the change in a model's output be attributed to the change in its inputs?

This note will focus on the last question. To put some notation around this problem, lets describe the vector of model inputs as x and the model as $f(x)$. The problem then can be described as attributing the change in model output, $f(x_1) - f(x_2)$ to each input variable as the input is changed from x_1 to x_2.

Incremental Evaluation

One approach to this problem is to change x_1 one variable at a time until it is equal to x_2, with difference in model output with each change being the margin contribution to the overall change that can be attributed to input variable that was changed.

```
x_temp1 = xtemp2 = x1
contribition = empty array of length of x
for variable_index in x2:
    x_temp1[variable_index] = x2[variable_index]
    contribition[variable_index] = f(x_temp1) - f(x_temp2)
    x_temp2 = x_temp1
```

This method provides a simple additive decomposition of the change in model output, but it has the obvious issue that it is sensitive to the order in which the variables are changed, with different results being possible if the order of the variables is changed, which is an undesirable property.

Repeated Incremental Evaluation

A simple solution to this problem is to evaluate the attribution for every possible order of the input variables, then take the average of the attributions to each variable. This way any interaction effects between individuals that are sensitive to order should be spread between the variables in question.

SHAP (SHapley Additive exPlanations) Values

SHAP values formalise the repeated incremental evaluation method with a foundation in game theory, with their applications discussed by Lundberg and Lee, 2017 and Lipovetsky and Conklin, 2001. The high level method for calculating SHAP values is:

1. Calculate the model output at the baseline inputs, x_1. Generally the average model output is used as the baseline, but here we want a specific output of $f(x_1)$ to act as the baseline.
2. Determine every possible subset of the inputs.
3. For each subset, create a model input, x, where the variables in the subset are set to their values in x_2 while the remaining variables are left at their baseline values.
4. Once the results for every subset are known, calculate the SHAP value for each variable as the weighted marginal contribution to the new result. That is, what is the difference between the model's output for a given subset, S with and without the variable, i:

$$f(S \cup i) - f(S) \tag{27}$$

   The weights used are determined by the size of the subset, S, and the total number of variables, N. The SHAP value for a given variable is then determined as:

$$\phi_i = \sum_{S \subseteq N \setminus i} \frac{|S|!(|N| - |S| - 1)!}{|N|!}(f(S \cup i) - f(S)) \tag{28}$$

Given that in the repeated incremental evaluation method every subset is also considered, the main difference in SHAP values is the weighting scheme, where smaller subsets are given higher weights because each feature's contribution is more significant when fewer other features are present.

Computational Complexity and Monte Carlo methods

Iterating over every subset or order of variables is massively expensive, with $2^n - 1$ possible subsets to consider. A popular approach to approximating these attributions is to randomly sample subsets (or orders) rather than exhaustively evaluating every option.[5]

Model Sensitivity Analysis

Sensitivity analysis is closely related to model output attribution, but is a less strict process. Rather than exactly decomposing an output, sensitivity aims to provide intuition on how large a change in output is likely given a change in an input.

Simple Sensitivity Analysis

To estimate the sensitivity of a model $f(x)$ to each input x_i we can approximate the gradient of the model's output with respect to each input using the simple method below.

1. Calculate $f(x)$ for the given input x.
2. Vary each input x_i by a small amount ϵ.
3. Compute $f(x')$ where $x'_i = x_i + \epsilon$ and $x'_j = x_j$ for $j \neq i$.
4. Estimate sensitivity as $\text{Sensitivity}_i \approx \frac{f(x') - f(x)}{\epsilon}$.
5. Repeat for each input x_i to determine sensitivity to changes in x.

This method provides a numerical measure of how much $f(x)$ changes with one unit variations in each input x_i. This is a local gradient estimate only - it is only valid very near the input value used.

LIME (Local Interpretable Model-agnostic Explanations)

LIME helps us understand why a model made a particular prediction by approximating the model locally around the instance being explained. This provides estimates for the sensitivity of the model to each input, at the current input values.

1. **Select an Instance**: Choose the specific instance (data point) you want to explain.
2. **Perturbation**: Create many new data points by slightly altering the chosen instance. This involves changing features to see how these changes affect the model's prediction.
3. **Prediction**: Use the original complex model to predict the outcomes for these new, slightly altered data points.
4. **Weighted Linear Model**: Fit a simple, interpretable model (like a linear regression) to these new data points. The new model is weighted so that points more similar to the original instance have more influence on the explanation.

[5] There are also more efficient model-specific methods in the cases of tree based algorithms.

5. **Interpretation**: Use the coefficients of this simple model to explain the prediction. These coefficients show the importance of each feature in making the prediction.

This will provide a very similar output to the simple sensitivity analysis, but may be more robust due to its use of multiple perturbations for each variable.

Note:

Views expressed are the author's, and may differ from those of JANA investments. This material does not constitute investment advice and should not be relied upon as such. Investors should seek independent advice before making investment decisions. Past performance cannot guarantee future results. The charts and tables are shown for illustrative purposes only.

References

Bacon, C. (2002). Excess returns—arithmetic or geometric? *Journal of Performance Measurement*.

Bacon, C. (2019). Performance attribution history and progress. *CFA INSTITUTE RESEARCH FOUNDATION*.

Bain, W. (1996). *Investment performance measurement*. Woodhead Publishing.

Brinson, G. P., & Fachler, N. (1985). Measuring non-us equity portfolio performance. *Journal of Portfolio Management*.

Burnie, S., Knowles, J., & Teder, T. (1998). Arithmetic and geometric attribution. *Journal of Performance Measurement*.

Dmitri Mossessian, V. V. (2014). Robust estimation of risk factor model covariance matrix. *FactSet Research Systems Inc*.

Forgy, E. (2002). Geometric issue and sector selection for performance attribution. *Available at SSRN: https://ssrn.com/abstract=1420229 or http://dx.doi.org/10.2139/ssrn.1420229*.

Holtes, G. (2024). Practical guide to factor covariance matrix construction [see https://www.grantholtes.com].

Jianqing Fan, Y. L., & Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *Annals of Statistics*.

Lipovetsky, S., & Conklin, M. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*.

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *NeurIPS*.

Roger G Ibbotson, P. C. (2001). The supply of stock market returns.