Valuing Financial Data

Maryam Farboodi
MIT

Dhruv Singal
Columbia

Laura Veldkamp
Columbia

Venky Venkateswaran
NYU

Value of Financial Data to Investors

- Data is a valuable asset for investors. How valuable?

  What is an investor's willingness to pay? This is a demand, not an equilibrium transactions price.

- Data valuation is not easy
  - How much you can profit from data depends on who else knows that data, who knows similar data, and how aggressively they will trade on it .

    $\Rightarrow$ impossible data requirements: Everyone else's data sets, preferences, price impact, investment mandates ...

- Our contribution: sufficient statistics that bypass the need to know others' information sets and characteristics.

$\Rightarrow$ A tool to put a dollar value on a piece of data.
It depends only on returns and *your* characteristics

Findings

- Investor characteristics change the value of data by orders of magnitude
  - Wealth
  - Price impact of trades
  - Investment style (set of investable assets)
  - Existing information (in paper)
  - Trading frequency (in paper)

  Data is not a common value asset: same data valued very differently
- Demand elasticity of data tied to asset market elasticity.
- Related work puts one value on each piece of data

  Manela & Kadan ('20, '21), Glode, Green, Lowrey ('12), Kacperczyk & Sundaresan ('19), Davila & Parlatore ('20), Farboodi, Matray, Veldkamp, Venkateswaran ('21)

Measuring different data values for different traders is a first step to trace out data demand curve.

Outline

NREE model with rich heterogeneity and covariances

- 1 riskless asset and n risky assets pay a dividend stream, a vector AR(1) :

$$d_{t+1} = \mu + G(d_t - \mu) + y_{t+1}$$

  $y_{t+1} \sim \mathcal{N}(0, \Sigma_d)$, iid across time, correlated across assets.

- Stochastic demand (vector) $x_t \sim N(0, \Sigma_x)$
- N OLG investors i with heterogeneous preferences / investment sets $\mathbb{E}\left[U_i(c_{it+1}) \mid \mathcal{I}_{it}\right]$, with $U_i'' < 0$.
- A data point: noisy signal of dividend innovation.
  Others may know it. Data may be correlated with what others know:

$$s_{it} = y_{t+1} + \zeta_{it} z_{t+1} + \xi_{it}$$

  - $z_{t+1} \sim N(0, \Sigma_z)$: public noise
  - $\xi_{it} \sim N(0, K_{it}^{-1})$: private noise

- Information set: $\mathcal{I}_{it} = \{\mathcal{I}_{i,t-1}, s_{it}, d_t, p_t\}$

Equilibrium Solution

- **Equilibrium**
  - Investors learn from prices and data. Update beliefs with Bayes' law.
  - Choose portfolios q_{it} to max EU, accouting for price impact dp/dq_i.
  - Price p_t equates demand and supply:

$$p_t = A_t + B(d_t - \mu) + C_t y_{t+1} + D_t x_{t+1} + F_t z_{t+1}$$

- A second-order approximation to expected utility yields mean-variance.

$$\mathbb{E}\left[U(c_{it+1}) \mid \mathcal{I}_{it}\right] \approx \rho_i \mathbb{E}\left[c_{it+1} \mid \mathcal{I}_{it}\right] - \frac{\rho_i^2}{2} \mathbb{V}\left[c_{it+1} \mid \mathcal{I}_{it}\right] + \text{const}_{it}$$

Absolute risk aversion ρ_i can depend on time-t wealth (not CARA)

Investor Expected Utility

- Profits $(\Pi_t) \rightarrow$ returns (R_t) for measurement

$$(p_{t+1} + d_{t+1}) - rp_t \rightarrow (p_{t+1} + d_{t+1}) \oslash p_t - r$$

- Substitute optimal portfolio q_{it}, equilibrium price, price information and take expectations over realization of random outcomes and signals

$$\tilde{U}(\mathcal{I}_{it}) \approx \underbrace{\mathbb{E}\left[R_t\right]' \hat{V}_{it}^{-1} \mathbb{E}\left[R_t\right]}_{(\text{Sharpe ratio})^2} + \text{Tr}\left[\underbrace{\left(\mathbb{V}[R_t] - \mathbb{V}[R_t \mid \mathcal{I}_{it}]\right)}_{\text{variance reduction}} \hat{V}_{it}^{-1}\right] + r\rho_i \bar{w}_{it}$$

  - R_t: returns for i's investable assets based on his investment style
  - $\hat{\hat{V}}_{it}$: conditional variance of this return, adj for price impact $\frac{dp}{dq_i}$

Sufficient Statistics to Value Data

- **Dollar value of data:** investor indifferent between
  having the data $\equiv$ no data+ additional riskless wealth

$$\$\text{Value of Data}_i = \frac{1}{r\rho_i}\left(\tilde{U}(\mathcal{I}_{it} + \text{data}) - \tilde{U}(\mathcal{I}_{it})\right)$$

- **The approach:**
  - Unconditional means and variances are easy. Challenge is $\mathbb{V}[R_{t+1} \mid \mathcal{I}_{it}]$.
  - Small insight: For linear normals, Bayes law and OLS coincide.
    $\mathbb{V}[R_{t+1} \mid \mathcal{I}_{it}]$ is the expected squared residual from OLS regression.
  - Estimate a return forecasting regression with data investor already knows.

    Estimate same regression, with historical values of data we are valuing.

    Use the average squared residuals from those two regressions as moments in equilibrium model to value data.

The Insight: Others' Info Disappeared!

- It must be wrong: We know that information others know is less valuable. How could this effect disappear?
  - It did not disappear. It matters through R_{t+1}
  - Data others know is in prices p_t. It does not forecast returns beyond prices
  - Conditioning on it on top of prices will not affect $\mathbb{V}[R_{t+1} \mid \mathcal{I}_{it}] \Rightarrow$ it won't increase utility

- It's obvious: Utility looks like in many REE models. What's new?
  - Mapping many models into these sufficient stats is new.
    Models with: heterogeneous investors, style constraints, data that is private, partially public or correlated with what others know ...
  - Return-based sufficient stats are a crucial recent step forward for NREE.

Whether data is public, private or correlated with what others know is crucial, but it matters through conditional variances

Outline

Estimation: Procedure

- Data to be valued X_t, and existing data Z_t

$$R_{t+1} = \gamma_2 Z_t + \varepsilon_t^Z$$
$$R_{t+1} = \beta_1 X_t + \beta_2 Z_t + \varepsilon_t^{XZ}$$

- Conditional variance without data we're valuing for a sample $1, \ldots, T$

$$\mathbb{V}[R_{t+1} \mid \mathcal{I}_{it}] \approx \widehat{\mathrm{Cov}}[\varepsilon_t^Z] = \frac{1}{T - |Z|} \sum_{t=1}^{T} \varepsilon_t^Z \varepsilon_t^{Z\prime}$$

- Conditional variance with data

$$\mathbb{V}[R_{t+1} \mid \mathcal{I}_{it} + \text{data}] \approx \widehat{\mathrm{Cov}}[\varepsilon_t^{XZ}] = \frac{1}{T - |Z| - |X|} \sum_{t=1}^{T} \varepsilon_t^{XZ} \varepsilon_t^{XZ\prime}$$

- Plug these in equilibrium expected utility to get data value.

Estimation: Data Sources

- A proof of concept: Value the same data for many different investors.
- The data we value:
  1. Institutional Brokers Estimate System (I/B/E/S) earnings forecasts
     a panel of 26,606 analysts for 15,780 firms from 1985-2015
  2. In the paper: GDP Growth, announced 1 year in advance
     Only you got to see their estimate one year early: Worth \$0.9m
     (for rich, unrestricted investor with price impact.)
     We can value data that does not yet exist.
- The data we use:
  1. stock prices: CRSP 1985-2015, annual
  2. Investors also know: Previous dividends and D/P
     Compustat, end of previous year (so they are known to investors)

Estimation: Challenges

- There is a ton of data

  $\rightarrow$ need to aggregate and reduce dimensionality

  Our approach:

  1. Group assets into portfolios,
  2. Use value-weighted mean of median forecasts of earnings growth for each portfolio.

- What else does the investor already know?

  Our approach:

  - Value data for an investor who only knows macro variables and prices.

- **Key:** methodology can be easily adapted to other approaches/assumptions

  proof of concept for the tool!

Outline

Different Values for the Same Data

How much are this year's IBES forecasts worth, to an investor who only knows D_t and D_t / P_t? A take-it-or-leave-it offer.

	Investment Style				
	Small	Large	Growth	Value	All
Perfect Competition					
Investor with $500,000 Wealth	0.00	$1.7k	$2.5k	$490	$3.5k
Investor with $250m Wealth	0.00	$566k	$844k	$164k	$1.2m
With Price Impact					
Investor with $500,000 Wealth	0.00	$1.6k	$2.5k	$410	$1.4k
Investor with $250m Wealth	0.00	$24k	$57k	$1.5k	$253k

Purple: Richer investors value data much more than poorer ones.

Yellow: Investment style matters enormously.

Red: Price impact reduces the value of data - a little or a lot.

The dispersion of valuations for the same data is immense!

Effects of Investor Heterogeneity

	Effect on Data Value	
More wealth	$\uparrow$	
Price impact	$\downarrow$	
Investment style	it matters	
Previously purchased data	$\downarrow$ modestly	in the paper
Trading horizon	modest effect that varies	in the paper

Directional effects intuitive but effects often compete.

Magnitudes would be tough to guess without our tool.

Data Demand Elasticity

Results teach us about data markets:
Inelastic asset demand creates more elastic data demand.
Why?

- Inelastic asset market demand (more price impact) lowers data value most for investors with the highest data values. (red cells)
  They're the ones that wanted to use the data to trade aggresively.

- Data values become **less heterogeneous**, b/c of asset price impact.
- If the price of data changes $\rightarrow$ big swing in data demand.
- That's high price elasticity of data demand.

◂ ex-post macro data

Conclusion

- Data is one of the most valuable assets in the modern economy

  We need tools to value it.

- Data has enormously variable private values.
  The same data is worth vastly different amounts to investors with different wealth and style, with and without price impacts, ...

- Next steps to understand data markets:
  - Estimate distributions of investor characteristics to produce a demand curve.
  - Understand the data supply side.

  Then we can do asset pricing theory ... for data!

  and data matters for firm decisions and thus, for corporate finance!

Extra Slides

Macro Data: *Ex-post GDP Growth*

	Portfolio Type					
	Small	Large	Growth	Value	SP500	All
Panel A: Perfect Competition						
Dollar Value (in $000, ann.) for:						
Investor with $500,000 Wealth	4.03	4.09	2.93	3.71	1.83	5.22
Investor with $250m Wealth	1367.65	1387.57	995.22	1260.04	620.80	1769.76
Panel B: With Price Impact						
Dollar Value (in $000, ann.) for:						
Investor with $500,000 Wealth	4.03	4.06	2.93	2.72	1.78	3.57
Investor with $250m Wealth	201.35	89.62	167.54	6.90	19.14	909.20

123 observations.

Main point: We can value non-traded data.

return

Quadratic Approximation of Utility

- A 2^{nd}-order Taylor expansion around $E[c]$: $U(c) = c - \frac{\rho_i}{2}(c - E[c])^2$
- Take an expectation: $E[U(c)] = E[c] - \frac{\rho_i}{2} V[c]$
- What if you approximate around $E[c]$ and have info set $\mathcal{I}$?

$$E[E[U(c)|\mathcal{I}]] = E[c] - \frac{\rho_i}{2} V[c|\mathcal{I}] - \frac{\rho_i}{2} E[(E[c|\mathcal{I}] - E[c])^2]$$

- The last term isn't real consumption risk. It's information risk.
  The investor is scared of learning something new.
  (No preference for early resolution of uncertainty).
- Baseline value of data: only prices investment risk.
- You can discount the value of data for information fear.
  Sufficient stats will to do that. Value of data falls, on average xx%.
- Another reason data values are heterogeneous:
  Some investors might as scared to learn as they are to invest.

Data Value with Skewed Variables

- Consider a normal variable x and a concave transformation g.
  If $g'' < 0$, then $g(x)$ has negative skewness.
- We can write utility as an indirect function of this skewed payoff: $\tilde{U}(g(x))$.
- A second-order approximation of this utility in x will now have a $-\rho = \tilde{U}''/\tilde{U}'$ term, plus an additional g'' term from the change-of variable function.
  $\rightarrow$ negative skewness adds to effective risk aversion!
- Risk-neutral pricing removes preference curvature and puts it in Q-probs.
  We take curvature out of a probability transformation $\rightarrow$ preferences.
  Both use a Radon-Nikodym derivative to pose an equivalent problem with a change of measure.
- Punchline: We can value data with negative skewness.
  Like valuing normal data, with extra risk aversion.

Our Price Impact Estimates vs. Price Elasticity Estimates

- What is relevant for data value: dp/dq_i for an individual investor.
- Koijen - Yogo measure: $d\log q/d\log p$.
  In a competitive market, $dp/dq = 0$.
  But portfolio choices still respond to price: $dq/dp < 0$.
  These are conceptually different objects.
- Gabaix - Koijen measure: dp/dq for aggregate market events.
  GK is a macro elasticity. When our investor trades on data, they are a micro shock.
  Should these be the same? A detailed discussion in GK explains why micro and macro elasticities differ.