Private information distributions in securities markets

Kerry Back, Kevin Crotty, and Tao Li

Abstract

We propose to estimate a structural model of endogenous informed trading that is a hybrid of the PIN and Kyle models. Back, Crotty, and Li (RFS 2018) estimate a version of the model under the simplifying assumption of a binary distribution of private information. We propose to estimate the model using more general private information distributions, which requires optimizing likelihood functions containing numeric integration. The resulting estimates can improve our understanding of the nature of information asymmetry in the cross-section of stocks.

1. Model of Informed Trading

Back, Crotty, and Li (2018, hereafter BCL) estimate a hybrid model of information asymmetry by maximum likelihood. Private information events occur with probability α and the private signal is negative with probability p_L. Noise trading is normally distributed with volatility σ. A public information process V is assumed to be a geometric Brownian motion on each day with a constant volatility δ. In the generalized version of the model, BCL show that, for a general distribution G of a private signal, intraday equilibrium pricing function p at time t for cumulative order flow y is given by:

$$p(t,y) = \int_{-\infty}^{y_L} G^{-1}\left(\frac{F(z)}{\alpha}\right) f(z \mid t,y)\,\mathrm{d}z + \int_{y_H}^{\infty} G^{-1}\left(\frac{F(z)-1+\alpha}{\alpha}\right) f(z \mid t,y)\,\mathrm{d}z\,. \tag{1}$$

where y_L and y_H are order flow thresholds that depend on model parameters and F is the distribution of noise trading. Due to computational considerations, BCL assume that private information follows a binary distribution.[1]

BCL estimate the model using the joint distribution of intraday returns and order flows. As is standard, they assume that each day is a separate realization of the model and that parameters are constant within each year for each stock. The data consist of cumulative order flows and intraday returns, sampled at regular intervals of length Δ. The log-likelihood

[1] In the binary case where the private information is either high H or low L and noise trading is normally distributed, the pricing function p is given by

$$p(t,y) = L \cdot \mathrm{N}\left(\frac{y_L - y}{\sigma\sqrt{1-t}}\right) + H \cdot \mathrm{N}\left(\frac{y - y_H}{\sigma\sqrt{1-t}}\right)\,. \tag{2}$$

function $\mathcal{L}$ for an observation period of n days satisfies

$$-\mathcal{L} = n(k+1)\log\sigma + \frac{1}{2\sigma^2\Delta}\sum_{i=1}^{n} Y_i'\Sigma^{-1}Y_i + n(k+1)\log\delta + \frac{1}{2\delta^2\Delta}\sum_{i=1}^{n} U_i'\Sigma^{-1}U_i + \frac{n\delta^2}{8} + \sum_{i=1}^{n}\left(\sum_{j=1}^{k} U_{ij} + \frac{3}{2}U_{i,k+1}\right), \quad (3)$$

where k is the number of intraday observations. Y_i is the vector of cumulative order flows for day i. U_i is the vector $(U_{i1}, \ldots, U_{i,k+1})'$ of log pricing differences

$$U_{ij} = \log\left(\frac{P_{ij}}{P_{i0}} - p(t_j, Y_{ij})\right) \quad (4)$$

between the observed return and the model's pricing function. BCL assume a binary signal to ease the computational burden involved in maximizing this likelihood. We propose to estimate the model for individual stock-years over the last 25 years using general distributional forms G for private information (e.g., lognormal or triangular distributions).

2. Need for XSEDE resources

The underlying data for the project are the NYSE Trade and Quote database. The typical daily quote dataset can range in size from 20-50 GBs, so processing the data requires substantial computing resources. The data can be collapsed to hourly observations of cumulative intraday returns and order-flows, which is the input data for the optimization.

The maximization of the likelihood for the general model requires optimizing functional forms that require numeric integration. The estimation is done in Python. This is computationally intensive and slow relative to the binary signal model, but can provide a fuller picture of the nature of private information.

Optimization of the likelihoods can be run in parallel. If we consider four different functional forms for the private information distribution, estimating the model for the approximately 120,000 firm-years in a panel of NYSE and NASDAQ stocks for the last 25 years would require optimizing approximately 4.8 million likelihood functions.[2] This is not feasible without substantial computing power like that provided by XSEDE.

References

Back, Kerry, Crotty, Kevin, and Tao Li, 2017. "Identifying Information Asymmetry in Securities Markets." *Review of Financial Studies*, forthcoming.

[2]This assumes starting each firm-year optimization from 10 random initial parameter values.