Big Data and Firm Dynamics: Online Appendix

Maryam Farboodi, Roxana Mihet, Thomas Philippon, and Laura Veldkamp

1. Model Solution Details

There are two sources of uncertainty in firm i's problem at date t: the (random) optimal technique $\theta_{i,t}$, and the aggregate price P_t. Let $(\hat{\mu}_{i,t}, \Sigma_{i,t})$ denote the conditional mean and variance of firm i belief about $\theta_{i,t}$ given its information set at date t, $\mathcal{I}_{i,t}$.

In this section, we will first describe the firm belief updating process about its optimal technique. Next, we argue that in this environment, the firm's optimal production choice is deterministic, and thus the price is deterministic as well. Finally, we lay out the full set of equations that characterize the equilibrium of this economy with two groups of firms.

Belief updating. The information problem of firm i about its optimal technique $\theta_{i,t}$ can be expressed as a Kalman filtering system, with a 2-by-1 observation equation, $(\hat{\mu}_{i,t}, \Sigma_{i,t})$.

We start by describing the Kalman system, and show that the sequence of conditional variances is deterministic. Note that all the variables are firm specific, but since the information problem is solved firm-by-firm, for brevity we suppress the dependence on firm index i.

At time t, each firm observes two types of signals. First, date $t-1$ output provides a noisy signal about θ_{t-1}:

$$y_{t-1} = \theta_{t-1} + \epsilon_{a,t-1}, \tag{1}$$

where $\epsilon_{a,t} \sim \mathcal{N}(0, \sigma_a^2)$. We provide model detail on this step below. Second, the firm observes $n_t = z k_t^{\alpha}$ data points as a bi-product of its economic activity. The set of signals $\{s_{t,m}\}_{m\in[1:n_{i,t}]}$ are equivalent to an aggregate (average) signal $\bar{s}_t$ such that:

$$\bar{s}_t = \theta_t + \epsilon_{s,t}, \tag{2}$$

where $\epsilon_{s,t} \sim \mathcal{N}(0, \sigma_\epsilon^2/n_t)$. The state equation is

$$\theta_t - \bar{\theta} = \rho(\theta_{t-1} - \bar{\theta}) + \eta_t,$$

where $\eta_t \sim \mathcal{N}(0, \sigma_\theta^2)$.

At time, t, the firm takes as given:

$$\begin{aligned}\widehat{\mu}_{t-1} &= \mathbb{E}\big[\theta_t \mid s^{t-1}, y^{t-2}\big] \\ \Sigma_{t-1} &= Var\big[\theta_t \mid s^{t-1}, y^{t-2}\big]\end{aligned}$$

where $s^{t-1} = \{s_{t-1}, s_{t-2}, \dots\}$ and $y^{t-2} = \{y_{t-2}, y_{i,t-3}, \dots\}$ denote the histories of the observed variables, and $s_t = \{s_{t,m}\}_{m\in[1:n_{i,t}]}$.

We update the state variable sequentially, using the two signals. First, combine the priors with y_{t-1}:

$$\begin{aligned}\mathbb{E}\big[\theta_{t-1} \mid \mathcal{I}_{t-1}, y_{t-1}\big] &= \frac{\Sigma_{t-1}^{-1}\widehat{\mu}_{t-1} + \sigma_a^{-2} y_{t-1}}{\Sigma_{t-1}^{-1} + \sigma_a^{-2}} \\ Var\big[\theta_{t-1} \mid \mathcal{I}_{t-1}, y_{t-1}\big] &= \big[\Sigma_{t-1}^{-1} + \sigma_a^{-2}\big]^{-1} \\ \mathbb{E}\big[\theta_t \mid \mathcal{I}_{t-1}, y_{t-1}\big] &= \bar{\theta} + \rho \cdot \big(\mathbb{E}\big[\theta_{t-1} \mid \mathcal{I}_{t-1}, y_{t-1}\big] - \bar{\theta}\big) \\ Var\big[\theta_t \mid \mathcal{I}_{t-1}, y_{t-1}\big] &= \rho^2\big[\Sigma_{t-1}^{-1} + \sigma_a^{-2}\big]^{-1} + \sigma_\theta^2\end{aligned}$$

Then, use these as priors and update them with $\bar{s}_t$:

$$\widehat{\mu}_t = \mathbb{E}\big[\theta_t \mid \mathcal{I}_t\big] = \frac{\left[\rho^2\left[\Sigma_{t-1}^{-1} + \sigma_a^{-2}\right]^{-1} + \sigma_\theta^2\right]^{-1} \cdot \mathbb{E}\big[\theta_t \mid \mathcal{I}_{t-1}, y_{t-1}\big] + n_t\sigma_\epsilon^{-2}\bar{s}_t}{\left[\rho^2\left[\Sigma_{t-1}^{-1} + \sigma_a^{-2}\right]^{-1} + \sigma_\theta^2\right]^{-1} + n_t\sigma_\epsilon^{-2}} \tag{3}$$

$$\Sigma_t = Var\big[\theta \mid \mathcal{I}_t\big] = \left\{\left[\rho^2\left[\Sigma_{t-1}^{-1} + \sigma_a^{-2}\right]^{-1} + \sigma_\theta^2\right]^{-1} + n_t\sigma_\epsilon^{-2}\right\}^{-1} \tag{4}$$

Multiply and divide equation (3) by Σ_t as defined in equation (4) to get

$$\widehat{\mu}_t = (1 - n_t\sigma_\epsilon^{-2}\Sigma_t)\left[\bar{\theta}(1-\rho) + \rho\left((1-M_t)\mu_{t-1} + M_t\tilde{y}_{t-1}\right)\right] + n_t\sigma_\epsilon^{-2}\Sigma_t\bar{s}_t, \tag{5}$$

where $M_t = \sigma_a^{-2}(\Sigma_{t-1} + \sigma_a^{-2})^{-1}$.

Equations (4) and (5) constitute the Kalman filter describing the firm dynamic information problem. Importantly, note that Σ_t is deterministic.

Modeling quadratic-normal signals from output. When y_{t-1} is observed, agents know their capital k_{t-1}. Therefore, they can back out A_{t-1} exactly. To keep the model simple for a short paper, we assumed that when agents see A_{t-1}, they also learn whether the quadratic term $(a_{t-1} - \theta_{t-1} - \epsilon_{a,t-1})^2$ had a positive or negative root. An interpretation is that they can figure out if their action a_t was too high or too low.

Relaxing this assumption complicates the model because, when agents do not know which root of the square was realized, the signal is no longer normal. One might solve a model with binomial distribution over two normal variables, perhaps with other simplifying assumptions. For numerical work, a good approximate solution would be to simulate the binomial-normal and then allows firms to observe a normal signal with the same mean and same variance as the true binomial-normal signal. This would capture the right amount of information flow, and keep the tractability of updating with normal variables.

Deterministic Dynamic Production Choice

Consider the firm sequential problem:

$$\max \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t \left(P_t A_t k_t^\alpha - r k_t\right)$$

We can take a first order condition with respect to a_t and get that at any date t and for any level of k_t, the optimal choice of technique is

$$a_t^* = \mathbb{E}[\theta_t | \mathcal{I}_t].$$

Given the choice of a_t's, using the law of iterated expectations, we have:

$$\mathbb{E}[(a_t - \theta_t - \epsilon_{a,t})^2 | \mathcal{I}_s] = \mathbb{E}[Var[\theta_t | \mathcal{I}_t] | \mathcal{I}_s],$$

for any date $s \leq t$. We will show that this object is not stochastic and therefore is the same for any information set that does not contain the realization of θ_t.

Lemma. The sequence problem of the firm can be solved as a non-stochastic recursive problem with one state variable.

Proof of Lemma 1. We can restate the sequence problem recursively. Let us define the

value function as:

$$V_t(\{s_{t,m}\}_{m\in[1:n_t]}, y_{t-1}, \hat{\mu}_{t-1}, \Sigma_{t-1}) = \max_{k_t, a_t} \mathbb{E}\left[A_t k_t^\alpha - r k_t + \beta V_{t+1}(\{s_{t+1,m}\}_{m\in[1:n_{t+1}]}, y_t, \hat{\mu}_t, \Sigma_t)|\mathcal{I}_{t-1}\right]$$

with $n_t = k_{t-1}^\alpha$. Taking a first order condition with respect to the technique choice conditional on $\mathcal{I}_t$ reveals that the optimal technique is $a_t^* = \mathbb{E}[\theta_t|I_t]$. We can substitute the optimal choice of a_t into A_t and rewrite the value function as

$$V_t(\{s_{t,m}\}_{m\in[1:n_t]}, y_{t-1}, \hat{\mu}_{t-1}, \Sigma_{t-1}) = \max_{k_t} \mathbb{E}\Big[\big(\bar{A} - (\mathbb{E}[\theta_t|I_t] - \theta_t - \epsilon_{a,t})^2\big)k_t^\alpha - r k_t + \beta V_{t+1}(\{s_{t+1,m}\}_{m\in[1:n_{t+1}]}, y_t, \hat{\mu}_t, \Sigma_t)|\mathcal{I}_{t-1}\Big].$$

Note that $\epsilon_{a,t}$ is orthogonal to all other signals and shocks and has a zero mean. Thus,

$$V_t(\{s_{t,m}\}_{m\in[1:n_t]}, y_{t-1}, \hat{\mu}_{t-1}, \Sigma_{t-1}) = \max_{k_t} \mathbb{E}\Big[\big(\bar{A} - ((\mathbb{E}[\theta_t|I_t] - \theta_t)^2 + \sigma_a^2)\big)k_t^\alpha - r k_t + \beta V_{t+1}(\{s_{t+1,m}\}_{m\in[1:n_{t+1}]}, y_t, \hat{\mu}_t, \Sigma_t)|\mathcal{I}_{t-1}\Big].$$

Notice that $\mathbb{E}[(\mathbb{E}[\theta_t|\mathcal{I}_t] - \theta_t)^2|\mathcal{I}_{t-1}]$ is the time-t conditional (posterior) variance of θ_t, and the posterior variance of beliefs is $\mathbb{E}[(\mathbb{E}[\theta_t|\mathcal{I}_t] - \theta_t)^2] := \Sigma_t$. Thus, expected productivity is $\mathbb{E}[A_t] = \bar{A} - \Sigma_t - \sigma_a^2$, which determines the within period expected payoff. Additionally, using the Kalman system equation (4), this posterior variance is

$$\Sigma_t = \left[\left[\rho^2(\Sigma_{t-1}^{-1} + \sigma_a^2)^{-1} + \sigma_\theta^2\right]^{-1} + n_t \sigma_\epsilon^{-2}\right]^{-1}$$

which depends only on Σ_{t-1}, n_t, and other known parameters. It does not depend on the realization of the data. Thus, $\{s_{t,m}\}_{m\in[1:n_t]}, y_{t-1}, \hat{\mu}_t$ do not appear on the right side of the value function equation; they are only relevant for determining the optimal action a_t. Therefore, we can rewrite the value function as:

$$V_t(\Sigma_t) = \max_{k_t} \left[(\bar{A} - \Sigma_t - \sigma_a^2)k_t^\alpha - r k_t + \beta V_{t+1}(\Sigma_{t+1})\right]$$
$$\text{s.t.} \qquad \Sigma_{t+1} = \left[\left[\rho^2(\Sigma_t^{-1} + \sigma_a^2)^{-1} + \sigma_\theta^2\right]^{-1} + z k_t^\alpha \sigma_\epsilon^{-2}\right]^{-1}$$

■

Equilibrium with Two Types of Firms

Here we re-introduce the dependence on firm index i. Assume there are two types of firms, $i = L, H$, were L (H) stands for low (high)-tech firm. High tech firms have high data efficiency: $z_{new} > z_{old}$. There is a mass M of firms with high data efficiency.

The first order condition of each firm in this environment is:

$$\frac{\partial V_{i,t}(\Sigma_{i,t})}{\partial k_{i,t}} = \alpha P_t \left(\bar{A} - \Sigma_{i,t} - \sigma_A^2\right) k_{i,t}^{\alpha-1} - r_t + \beta V'_{i,t+1}(\Sigma_{i,t+1}) \frac{\partial \Sigma_{i,t+1}}{\partial k_{i,t}} = 0 \tag{6}$$

where $\frac{\partial \Sigma_{i,t+1}}{\partial k_{i,t}} = -\Sigma_{i,t+1}^2 \sigma_\epsilon^{-2} z_i \alpha k_{i,t}^{\alpha-1}$. Substitute the latter expression into equation (6) to get the firm euler equation for dynamically optimal choice of capital. Note that the Euler equation is a 2^{nd} order equation in posterior variance: it involves $\Sigma_{i,t-1}$ (through substituting for $k_{i,t}$ from equation (4)) , $\Sigma_{i,t}$, and $\Sigma_{i,t+1}$. Alternatively, it can be expressed a 2^{nd} order equation in firm capital: it involves $k_{i,t-1}$, $k_{i,t}$, and $k_{i,t+1}$.

At each date t, the equilibrium is characterized by the following 8 equations in 8 unknowns $(A_{old}, A_{new}, k_{old}, k_{new}, \Sigma_{old}, \Sigma_{new}, Y, P)$:

$$\begin{aligned}
\alpha P_t\left(\bar{A} - \Sigma_{old,t} - \sigma_a^2\right) + \alpha\beta z_{old}\Sigma_{old,t+1}^2\sigma_\epsilon^{-2}P_{t+1}k_{old,t+1}^\alpha &= r_t k_{old,t}^{1-\alpha}\\
\alpha P_t\left(\bar{A} - \Sigma_{new,t} - \sigma_a^2\right) + \alpha\beta z_{new}\Sigma_{new,t+1}^2\sigma_\epsilon^{-2}P_{t+1}k_{new,t+1}^\alpha &= r_t k_{new,t}^{1-\alpha}\\
\Sigma_{old,t} &= \left[\left[\rho^2(\Sigma_{old,t-1}^{-1} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + n_{old,t}\sigma_\epsilon^{-2}\right]^{-1}\\
\Sigma_{new,t} &= \left[\left[\rho^2(\Sigma_{new,t-1}^{-1} + \sigma_a^{-2})^{-1} + \sigma_\theta^2\right]^{-1} + n_{new,t}\sigma_\epsilon^{-2}\right]^{-1}\\
A_{old,t} &= \bar{A} - \sigma_a^2 - \Sigma_{old,t}\\
A_{new,t} &= \bar{A} - \sigma_a^2 - \Sigma_{new,t}\\
Y_t &= A_{old,t}k_{old,t}^\alpha + MA_{new,t}k_{new,t}^\alpha\\
P_t &= \bar{P}Y_t^{-\gamma}
\end{aligned}$$

The first pair of equations are the Euler equations for firm optimal choice of production. The second pair determine firm posterior variance given the prior variance and production choice last period. The third pair determine the within period productivity given firm's posterior variance. The last two equations determine the aggregate output and market price given individually optimal choice of production.

Figure 2 provides the transition path from a steady state with a unit measure of old firms, with $z_{old} = 1$, to a new steady state with unit measure of old firms and measure M of new data-savvy firms, with $z_{new} = 1.5$. Parameter values are: $\bar{A} = 3$, $\hat{A} = 2.5$, $\alpha = 0.5$, $\beta = 0.98$, $\gamma = 0.5$, $\bar{P} = 10$, $r = 0.2$, $\rho = 0.99$, $\sigma_a^2 = \sigma_e^2 = \sigma_\theta^2 = 1$. The initial steady-state is given by: $k_{old} = 8.17$, $k_{new} = 0.01$, $\Sigma_{old} = 0.45$, $\Sigma_{new} = 1.49$, $A_{old} = 3.13$, $A_{new} = 0$, $\pi_{old} = 87.96$, $\pi_{new} = 0$, $P = 0.33$, $Y = 8.96$. The ending steady-state is: $k_{old} = 6.95$, $k_{new} = 8.64$, $\Sigma_{old} = 0.48$, $\Sigma_{new} = 0.33$, $A_{old} = 3.05$, $A_{new} = 3.50$, $\pi_{old} = 79.18$, $\pi_{new} = 101.2$, $P = 0.31$, $Y = 10.11$. π_i represents the profit of firm type i.

Data

The only data used in this paper were the twelve firm sizes and their market shares from the 2016 Longitudinal Business Database. Those 12 data points are:

Table 1—Firm Size Data

Firm Size Bin	Average Size	Market Share
1 to 4	2.10	0.55831
5 to 9	6.62	0.19961
10 to 19	13.68	0.12026
20 to 49	30.77	0.07704
50 to 99	69.76	0.02365
100 to 249	153.36	0.01305
250 to 499	343.96	0.00403
500 to 999	674.24	0.00192
1000 to 2499	1448.80	0.00118
2500 to 4999	3054.95	0.00045
5000 to 9999	5618.17	0.00024
10000 +	25436.08	0.00027