THE NATURE OF FIRM GROWTH*

Benjamin W. Pugsley
Federal Reserve Bank of New York

Petr Sedláček
University of Bonn

Vincent Sterk
University College London

July 2016

Abstract

There are vast differences in the growth patterns of firms: high-growth, young businesses, or "gazelles", account for the vast majority of employment growth at incumbent firms. Using a large administrative panel data set for the United States, we provide evidence that ex-ante differences in the growth potential of firms account for most of the size heterogeneity across firms of a given age. First, we estimate a reduced-form employment process, allowing for heterogeneity in steady-state levels and deriving parameter identification from the autocovariance function of employment. Next, we estimate a general equilibrium firm dynamics model, in which firms gradually learn about their growth potential. We explore the implications for firm selection and the macro effects of tax policies.

*VERY PRELIMINARY.** We thank Richard Blundell, Christian Bayer, Urban Jermann, Greg Kaplan, Fabien Postel-Vinay and Kjetil Storesletten for helpful comments.

1 Introduction

There are vast differences in the growth patterns of firms: high-growth, young businesses, sometimes called "gazelles", account for the vast majority of employment growth at incumbent firms (see e.g. Haltiwanger *et al.* (2014)). One potential explanation for these growth differences is that, following entry, firms are exposed to idiosyncratic shocks, to for example productivity or demand. According to this view, a firm outgrows its peers when it is hit by relatively benign shocks during its lifetime. An alternative view is that there are ex-ante differences between firm startups, with some types poised for growth and others destined to stay small. Under this view, heterogeneity in firms' growth paths are predictable, given their initial characteristics.

While it seems plausible that both views on firm growth are –to some extent– grounded in reality, little is known about their relative empirical relevance. However, the precise nature of growth differences may have large consequences for the selection of firms, as well as for the impact of firm selection on aggregate outcomes. For example, if a firm's growth pattern is largely deterministic (conditional on being of a certain type), then a key factor in a firm's entry and exit decisions may be its beliefs about its type rather than its current state of productivity or demand. Moreover, distortions to the firm selection process, induced by for example policies or market frictions, may have large aggregate effects if they cause firms with high growth potential to exit at an early stage, or if they prevent such firms from entering in the first place.

This paper presents direct empirical evidence on the importance of a deterministic component in firms' growth patterns, vis-à-vis post-entry shocks. Towards this end, we propose a model framework for employment dynamics in which firms are subject to both post-entry shocks and ex-ante differences in their long-run steady state employment levels. While the latter form of heterogeneity is determined before startup, it may take time for firms to actually converge towards their steady states. We estimate the model using large administrative data sets on firm-level employment in the United States, and then quantify the importance of both sources of firm heterogeneity. We find that ex-ante heterogeneity drives the majority of size heterogeneity among firms of a given age.

Our analysis can be broken down into two parts. In the first part (Section 2), we estimate a reduced-form stochastic process for employment dynamics. At this stage, we confine the confine the analysis to firms which survive up to a certain age, which enables us to derive analytical formulas for the model parameters as functions of the autocovariance function of firm-level employment. Specifically, we derive a simple and intuitive formulate for the key parameter, which controls the degree of steady-state heterogeneity across firms.

The reduced-form estimates imply that ex-ante heterogeneity in steady-state levels account for most of the firm size heterogeneity across firms of a given age, and that this importance increases as firms age. We obtain this finding in both the reduced-form model and the structural model. Three patterns in the data are crucial in driving this result. First, size differences between firms are very persistent, as indicated by high autocorrelations, even at long lags. For example, the autocorrelation of employment at a lag of ten years is 0.42 at age ten, growing to 0.88 at age fifteen. Second, the cross-sectional mean variance of employment are increasing functions of firm age, up to roughly ten years after entry. Thus, not only are size differences persistent, they also grow larger in the years following entry. Finally, we find evidence that firms revert to their individual steady-state levels relatively quickly. This is strongly suggested by the fact that, after ten years following entry, the cross-sectional moments mostly flattened out as functions of age. The relatively modest degree of *within*-firm employment persistence naturally leads to the conclusion that the strong degree persistence of differences *between* firms, as measured by the autocorrelations, is driven by ex-ante differences in their steady-state levels of firm employment. We corroborate this conclusion using a battery of panel data estimators. We further reach the same conclusion when we repeat the analysis within sectors.

In the second part (Section 3), we propose and estimate a general equilibrium firm dynamics model with endogenous entry and exit, broadening the analysis to all firms. In the model, firms are monopolistically competitive, à la Melitz (2003), but face heterogeneous demand curves. Specifically, the demand base of a firm evolves according to a process which is affected by (i) endogenous marketing investments made by firms, (ii) exogenous post-entry shocks, and (iii) a firm-specific parameter which controls the steady-state level of demand. This parameter is drawn upon entry from an exogenous distribution, but can only be observed imperfectly via a noisy signal, similar to Jovanovic (1982). As a firm ages, its demand base gradually grows towards the steady-state level, while at the same time the firm slowly learns about the precise magnitude of the steady-state level.

The model captures the reality that demand is intrinsically different across goods: some goods are be relatively well-liked, or appeal to a particularly large subset of consumers. At the same time, it takes time and resources to make consumers aware of a newly offered goods. During the early stages of a firm's life-cycle, it is difficult to predict how much the demand for their goods will be scaled up, once consumers have become fully aware of its existence. At the same time, running and growing a business is costly, which confronts young firms with the very difficult decision whether or not to continue, in the face of uncertain long-run demand growth. The purpose of the structural model is therefore not only to estimate the importance of heterogeneity in the long-run growth potential of firms, but also the extent to which selection of entrants and young firms is determined by their true growth potential, as opposed to temporary shocks and misinformation. Further, we use the structural model to shed light of the effect of size- and age-dependent tax rates on the selection of firms and aggregate outcomes.

The estimated structural model closely aligns with the reduced-form results, as heterogeneity in the steady-state parameters is found to be the most important contributor to size heterogeneity across firms of a given age. Additionally, we find that [to be completed].

Relation to the literature. The reduced-form part of our analysis (Section 2) is inspired by a large empirical literature on earnings dynamics of workers, which traditionally derives identification from the autocovariance structure of earnings, see e.g. MaCurdy (1982), Abowd and Card (1989). A common assumption in this literature that earnings are the sum of an individual fixed effect, an age fixed effect, an AR(1) process with zero mean, and an i.i.d. shock. Some authors, however, have argued that allowing in addition for individual-specific trends helps to capture the autocovariance structure of earnings, see Guvenen (2009) for a discussion of this branch of the literature. The possible presence of such "Heterogeneous Income Profiles" (Guvenen (2007)), has received much attention since they may have large implications for the extent to which income changes should be expected to transmit to consumption, from the perspective of standard life-cycle models (see e.g. Guvenen and Smith (2014)). Somewhat surprisingly, the literature on firm-level employment dynamics does not have a similar tradition of estimating reduced-form processes.[1] To the best of our knowledge, even the basic autocovariance structure of employment dynamics has not been systematically documented.[2] Our analytical results demonstrate how the autocovariance function can be exploited as a key source of parameter identification, which is also relevant to the structural model as it turns out that the employment dynamics in the model closely align with the reduced-form process.

The structural model (Section 3) builds on a large literature on firm dynamics models, following the seminal contributions of Jovanovic (1982), Hopenhayn (1992) and Hopenhayn and Rogerson (1993). Like in their models, firm dynamics are ultimately driven by the evolution of a firm-level fundamental.[3] In Hopenhayn and Rogerson (1993), the fundamental follows an AR(1) process which is common across firms. As in our model, the growth of young firms is partly driven by a convergence of the funda-

[1]Employment equations are a common application in the panel data literature, see e.g. Arellano and Bond (1991). However, we are not aware of any studies trying to estimate the importance of profile heterogeneity in accounting for employment heterogeneity across firms. One reason for this may be that panel data estimators are typically difference out the firm-specific fixed effect, which is typically left unestimated. A main purpose of our analysis is exactly to estimate the variance in this fixed effect, as we will explain in more detail in the next section.

[2]The precise way in which we model profile heterogeneity is somewhat different from the earnings literature. Instead of introducing a firm-specific trend, we include a firm-specific constant effect in an AR(1) process for productivity/demand. As the firm ages, the effect of the constant accumulates through the autoregressive component of the process, creating a tendency to converge towards a firm-specific steady-state level. Our choice to model heterogenous steady-state levels rather than firm-specific trends is motivated by our desire to relate closely to popular structural models of firm dynamics, as well as the empirical observation that cross-sectional moments appear to become stationary beyond a certain age.

[3]In their models, the fundamental is productivity, whereas in our model it is the demand base. However, it can be shown that under our setup demand and productivity are observationally equivalent.

mental towards a steady-state level. However, they abstract from ex-ante heterogeneity in the steady-state sizes of firms, a key feature of our model. In Jovanovic (1982), the fundamental is the sum of a fixed firm-level parameter and an i.i.d. shock. His model thus allows for steady-state heterogeneity, but since the shock is i.i.d. he abstracts from transitional dynamics in the fundamental. We show that allowing for such dynamics is important to account for the empirical autocovariance structure of employment. Further, transitional dynamics affects firm selection, as young firms tend to be small and are not yet fully impacted by their steady-state parameter, which makes their entry/continuation decision relatively sensitive to transitory shocks and misinformation.

More recently, Abbring and Campbell (2005) and Arkolakis *et al.* (2015) have both introduced learning à la Jovanovic (1982) in firm dynamics models with monopolistic competition. Abbring and Campbell (2005) present a rich partial equilibrium model with both transitory and persistent shocks, and an information asymmetry between economic agents and the econometrician. They conduct a Maximum Likelihood estimation using sales data from 305 bars in Texas. We instead design a model and estimation strategy suited for application to the entire cross-section of U.S. employers. Further, we develop a general equilibrium model, which is important for macro policy analysis. Arkolakis *et al.* (2015) also present a general equilibrium model, but do not disentangle the importance of pre- and post entry sources of heterogeneity and abstract from transitional dynamics in the demand base. They calibrate their model towards data from Colombian manufaturing plants and show how the presence of learning improves the model's ability to match growth patterns conditional on age and size. They also demonstrate that the combination of learning and monopolistic competition renders equilibrium outcomes inefficient.

? document the presence of strong cohort effects in employment data and estimate a firm dynamics model with ex-ante demand heterogeneity and aggregate shocks. They find that much of the differences *across* cohorts born can be attributed to the state of the economy in the year of startup. In the present paper, we instead estimate the degree to which heterogeneity *within* cohorts can be attributed to the startup phase and, accordingly, model idiosyncratic post-entry shocks rather than aggregate shocks. Moreover, we estimate the model with cross-sectional data rather than aggregated cohort-level data, thus exploiting a very different source of information. That said, the two sets of findings are complementary in the sense that in both cases the composition of entrants with respect to their growth potential emerges as a prime determinant of later outcomes. The importance of the composition of the firm population is also emphasized by **?**, who document a strong trend in the U.S. towards older firms, which is the result of accumulating startup deficits. They provide evidence that firm ageing can explain the emergence of jobless recoveries.

2 Reduced-form estimations

2.1 Data description

We use data on firm-level employment in the United States, take from the from Census Longitudinal Business Database. [*to be completed*]

We construct a balanced panel of employment in firms surviving up to age 15. Prior to the analysis, we take out a fixed effect for the birth year of the firm and the firm's 4 digit industry.

2.2 The autocovariance structure of firm-level employment

Figure 1 presents basic cross-sectional moments of logged employment by firm age.[4] The upper left panel shows that, at younger ages, average log employment is an increasing of age, growing from an average of 0.25 (1.30 in levels) in the year of startup to 4.1 at age fifteen (59.1 in levels). After age ten, there is relatively little average size growth. The upper right panel present shows the standard deviation by age. During the first 10 years, employment levels fan out. After this age, the cross-sectional standard deviation stabilizes as a function of age.

[4] We present moments for logged employment rather than levels, since we will use logged employment in the estimations.

Observed jointly the upper panels reveal two important patterns. First, differences between firms of the same age grow larger during the first ten years of their lifetime. Second, by age 10, the cross-sectional employment distribution appears close to a stationary end point.

The lower left panels of Figure 1 present the autocorrelation functions by age. Two important patterns stand out. First, autocorrelations are high, and decay only slowly as the lag length is increased. As a result, differences in employment across firms are very persistent, even at very long lags. Second, for a given lag length, the degree of persistence increases with firm age. For example, the correlation of logged employment at age 0 and 10 is 0.42, whereas the correlation between age 5 and age 15 is 0.88. Finally, the lower right panel presents the autocovariance function by age. Although the information in this panel is already summarized by the standard deviation an autocorrelation, we present it for the sake of completeness and because we will use the autocovariance function directly in the estimations to be presented below.

Figure 1: Log employment: basic moments in the data.

2.3 Employment process

We now estimate a reduced-form process for firm employment. Let $n_{i,a}$ be the employment level of an individual firm of age a. We consider the following process for logged employment:

$$\ln n_{i,0} = \rho \ln \widetilde{n}_i + \theta_i, \tag{1}$$

$$\ln n_{i,a} = \rho \ln n_{i,a-1} + \theta_i + \varepsilon_{i,a}, \text{ for } a = 1, 2, 3, ... \tag{2}$$

Here, $\rho \in (-1, 1)$ is a persistence parameter which is common across firms, and $\ln \widetilde{n}_i$ is an i.i.d. shock, drawn before entry from a distribution with mean $\mu_{\widetilde{n}}$ and standard deviation $\sigma_{\widetilde{n}}$. Further, θ_i is a firm-specific parameter, drawn before entry, from a distribution from an i.i.d. distribution with mean μ_θ, and standard deviation σ_θ. Finally, $\varepsilon_{i,a}$ is a post-entry shock which is drawn at age a from an i.i.d. distribution with mean zero and standard deviation σ_ε.

Given the above process, one can decompose the firm's logged employment level at age a into three components:

$$\ln n_{a,i} = \rho^{a+1} \ln \widetilde{n}_i + \sum_{k=0}^{a} \rho^k \theta_i + \sum_{k=1}^{a} \rho^{a-k} \varepsilon_k \tag{3}$$

The first term on the right-hand side of the above equation captures the effect of the pre-entry shock $\ln \widetilde{n}_i$. Given that ρ is smaller than one in absolute value, this component decays to zero as the firm ages. The second component captures the effect of the pre-entry parameter θ_i and does not die out with age. Letting age approach infinity, the second component converges to $\frac{\theta_i}{1-\rho}$. The third component captures the effect of the post-entry shocks, and has an unconditional mean of zero. Thus, $\frac{\theta_i}{1-\rho}$ is the steady-state level of employment of firm i, i.e. the mean of its old-age stationary employment distribution. It now follows that when $\sigma_\theta = 0$, there is no heterogeneity in steady state level across firms. That is, under this assumption all firms would fluctuate around the same employment levels at old age. Hence, σ_θ is a key parameter of interest.

Equation (3) directly leads to the following expression for the cross-sectional variance of logged employment, conditional on age:

$$Var(\ln n_a) = \left(\rho^{a+1}\right)^2 \sigma_{\widetilde{n}}^2 + \left(\sum_{k=0}^{a} \rho^k\right)^2 \sigma_\theta^2 + \sum_{k=1}^{a} \rho^{2(a-k)} \sigma_\varepsilon^2 \tag{4}$$

Note that the expression is a function of parameters only and that again it is the sum of the three components mentioned above. Thus, once the parameters of the process are known it is straight-forward to decompose the cross-sectional employment variance of firms at a certain age into the contributions of the pre-entry transitory draw, the steady-state parameter (also drawn pre-entry), and the post-entry shocks. Letting age approach infinity, we arrive at a particularly simple expression for the cross-sectional employment variance, which is a function of only three parameters, ρ, σ_θ^2 and σ_ε^2, as the transitory pre-entry component no longer plays a role:

$$Var(\ln n_\infty) = \frac{\sigma_\theta^2}{(1-\rho)^2} + \frac{\sigma_\varepsilon^2}{1-\rho^2}. \tag{5}$$

2.4 Parameter identification

We now show that key parameters ρ, σ_θ and σ_ε and $\sigma_{\tilde{n}}$ can be identified from the autocovariance matrix of logged employment. In the appendix, we derive the following moment conditions:

$$Cov\left(\ln n_{a,i}, \ln n_{a-j,i}\right) = \gamma_{j-1}\gamma_{a-j}\sigma_\theta^2 + \rho^j Var\left(\ln n_{a-j,i}\right), \tag{6}$$

$$Var\left(\ln n_{a,i}\right) = \rho^{2j} Var\left(\ln n_{a-j,i}\right) + \gamma_{j-1}^2\sigma_\theta^2 + \phi_{j-1}\sigma_\varepsilon^2 + 2\rho^j\gamma_{j-1}\gamma_{a-j}\sigma_\theta^2, \tag{7}$$

where $\gamma_j \equiv \sum_{k=0}^{j} \rho^k$ and $\phi_j \equiv \sum_{k=0}^{j} \rho^{2k}$. It now directly follows that the parameters ρ, σ_θ^2 and σ_ε^2 can be identified by three elements of the autocovariance matrix: $Var\left(\ln n_{a,i}\right)$, $Var\left(\ln n_{a-j,i}\right)$, and $Cov\left(\ln n_{a,i}, \ln n_{a-j,i}\right)$ for any arbitrary age $a > 0$ and lag length $j < a$. To identify $\sigma_{\tilde{n}}$ it is useful to the variance of the initial employment level as:

$$Var\left(\ln n_{0,i}\right) = \rho^2\sigma_{\tilde{n}}^2 + \sigma_\theta^2.$$

It follows that the four key parameters can be identified off the autocovariance matrix. Also, any element of the autocovariance matrix is useful for parameter identification. For his reason, we will exploit the entire matrix in the estimation.

While the autocovariance matrix of logged employment levels is sufficient to identify the parameters, other moments may be useful as well. In particular, a large literature on dynamic panel data estimation has developed moment estimators that can be used for the parameter ρ, see e.g. Anderson and Hsiao (1982), Arellano and Bond (1991). Moreover, given that in our case the initial employment levels do not have the same mean as the old stationary distribution, we can also identify ρ directly from following moment condition:

$$E(\ln n_{i,a}) - E(\ln n_{i,a-1}) = \rho E(\ln n_{i,a-1}) - \rho E(\ln n_{i,a-2}). \tag{8}$$

Intuitively, firms grow on average during the first years of their lives. The speed at which average growth decays is pinned down by the persistence parameter ρ.[5]

2.5 Estimation results

2.5.1 Baseline estimates

	ρ	σ_θ	σ_ε	$\sigma_{\tilde{n}}$	θ contrib to Var [ln$_\infty$]
Exactly identified	0.33	0.49	0.11	5.1	89.8%
	(1e-3)	(1e-3)	(1e-3)	(0.06)	
EWMD	0.51	0.21	0.21	3.9	75.5%
	(8e-4)	(8e-4)	(4e-4)	(0.01)	

- Baseline GMM estimation: match covariance matrix only, estimate ρ, σ_θ and σ_ε and $\sigma_{\tilde{n}}$
- Results using all moments in autocovariance matrix
- Variance decomposition, see Figure 2.

[5] For age groups old enough for employment to have reached its stationary distribution, the condition is not useful to identify ρ, as stationarity implies that $E(\ln n_{i,a}) - E(\ln n_{i,a-1}) = E(\ln n_{i,a-1}) - E(\ln n_{i,a-1-2}) = 0$. For younger firms, however, the data feature substantial employment growth, see Figure 1.

	ρ	σ_θ	σ_ε	$\sigma_{\tilde{n}}$	contr. θ to $Var(\ln n_\infty)$
point estimate	0.70	0.30	0.25	0.87	0.89
std. error	0.01	..	..	..	..
95% confidence band	[0.693,0.716]	..	..	..	..

Figure 2: Log employment: variance decomposition.

	ρ	σ_θ	σ_ε	$\sigma_{\tilde{n}}$	contr. θ to $Var(\ln n_\infty)$
moment condition (8) for firms $\leq$ age 3	0.70	0.30	0.25	0.87	0.89
moment condition (8) for firms $\leq$ age 5	0.70	..	..	..	..
Fixed-effects	0.65	..	..	..	..
Anderson-Hsiao (1981)	0.65	..	..	..	..
Arellano-Bond (1991)	0.65	..	..	..	..

2.5.2 The importance of ρ

- Discuss possibility that $\sigma_\theta^2 = 0$ – very high ρ needed.
- Estimate ρ in variety of alternative ways:
- Re-estimate σ_θ and σ_ε and $\sigma_{\tilde{n}}$ and variance decomposition conditional on alternative estimates for ρ.

2.5.3 Within-industry results

- Distribution of parameter estimates and variance decomposition

3 Structural model

We now investigate how the estimated ex-ante heterogeneity in long-run firm size interacts with the firm selection process and what this implies for aggregate productivity, employment and output dynamics. Towards this end, we build a general equilibrium model of monopolistically competitive firms for which the underlying driving force, interpreted either as productivity or demand shocks, exhibits not only transitory post-entry shocks, but also ex-ante heterogeneity. In addition, it is assumed that firms must learn about the true structure of the driving force over time and that they growth subject to convex adjustment costs in market size.

The model economy consists of a representative household and heterogeneous firms. The goods market is monopolistically competitive, whereas the labor market is perfectly competitive. For now, we abstract from government policies, but we will consider tax policies at a later stage. Time is discrete and there is no aggregate uncertainty.

3.1 Firms

There is an endogenous measure of profit-maximizing firms, indexed by i and owned by the representative household. The production technology is linear:

$$y_i = n_i, \tag{9}$$

where $y_{i,t}$ is the output level of firm i in and n_i is the firm's employment level which comes at a wage W per unit. Both variables are continuous. Each firm produces a unique goods variety, and faces a time-invariant flow cost of operation $c_f > 0$, denominated in units of the aggregate consumption basket. Households have CES preferences over these varieties, which we will describe in more detail in the next subsection. For now, let us take as given that preferences lead to the following demand constraint:

$$y_i = \phi_i \left(\frac{p_i}{P}\right)^{-\eta} Y, \tag{10}$$

where p_i is the price set by firm i, P is an aggregate price index, Y is aggregate output, and $\eta > 1$ is the elasticity of price substitution between individual goods. Further, ϕ_i is a firm-specific demand- shifter which stems from the utility level that the representative household derives from the variety i. We refer to ϕ_i as *marketing capital.* The demand-shifter evolves according to the following law of motion:

$$\ln \phi_i' = \theta_i + (1-\delta) \ln \phi_i + m_i' + \epsilon_i', \tag{11}$$

where primes denote next period's variables. Here, θ_i is a parameter which is time-invariant specific to the variety produced by firm i, and draw before entry from distribution $N(\mu_\theta, \sigma_\theta^2)$. Moreover, $\delta \in (0, 1)$ is the rate at which marketing capital depreciates, m_i is a marketing investment made by the firm, which comes at a quadratic cost given by $\frac{\psi}{2}m_i^2$, with $\psi > 0$, and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ is an demand shock which is i.i.d. across firms and over time. Startup firms enter the economy without any marketing capital.

The sequence of events within a period is as follows. First, the demand shock realizes. Next, a firm decides whether to exit or not. Optimal decision making dictates that a firm exits once its value becomes negative. If the firm decides to continue, it decides on the levels of m_i, n_i, p_i and y_i. The real value of a continuing firm can be expressed as

$$V_i = \max_{m_i, y_i, p_i} \left\{ \frac{p_i}{P_t} y_i - W_t y_i - \frac{\psi}{2} m_i^2 - c_f + \beta \mathbb{E}_i \max\left[0, V_i'\right] \right\}, \tag{12}$$

subject to constraints (10) and (11), where $\beta \in (0, 1)$ is the subjective discount factor of the household, and $\mathbb{E}_i$ is the firm's expectation operator given its beliefs and information. The firm's optimal pricing decision is static and can be shown to imply $\frac{p_i}{P} = \frac{\eta}{\eta - 1} W$.

3.1.1 Learning

We assume that, while a firm can observe its demand level ϕ_i, it cannot observe θ_i and ϵ_i individually. Note that the firm can observe $\epsilon_i + \theta_i$, which it uses to gradually learn about θ_i. Note also that θ_i is a key parameter in determining a firm's long-run growth potential. In the absence of demand shocks and demand investment, the firm's marketing capital would gradually grow from $\ln \phi_i = 0$ before entry towards a steady-state level $\ln \phi_i = \frac{\theta_i}{1-\delta}$ as the firm ages. We can think of this steady-state level as the firm's long-run market size. At the startup phase, a firm has only an imperfect idea about its long-run growth potential, but it learns more about this as it ages.

Before entry, the firm receives some information about its θ_i. In particular, we assume that upon entry (after paying an entry cost) the firm observes a signal $s_{\theta,i}$ given by

$$s_{\theta,i} = \theta_i + \upsilon_i, \tag{13}$$

where υ_i is an i.i.d. noise shock which is distributed as $\upsilon_i \sim N(0, \sigma_\upsilon^2)$. Note that σ_υ^2 controls the severity of the information friction. By setting $\sigma_\upsilon^2 = 0$, we obtain a full-information version of the model, whereas $\sigma_{\upsilon}^2 \to \infty$ corresponds to a version in which the firm has no information about its long-run growth potential prior to entry. An intermediate case, in which aspiring startups have some imperfect idea about their growth potential seems realistic, and we will discipline the level of σ_{υ}^2 using data.

We assume that firms use a Bayesian learning algorithm to update their beliefs about θ_i. Let us define marketing level before new investment (but after the shock) as $\ln \overline{\phi}_i \equiv \ln \phi_i - m_i$ and note that the amount of marketing capital generated by endogenous investment is irrelevant for the firm's learning about θ_i. Let τ_i denote the age of the firm i. The Appendix shows that the firm's posterior belief about the distribution of $\ln \overline{\phi}_i$ in the next period satisfies:

$$\ln \overline{\phi}_i' \sim N(\mu_{\theta,i} + (1-\delta)\ln \phi_i, \sigma_{\theta,i}^2 + \sigma_{\epsilon,i}^2) \tag{14}$$

where $\mu_{\theta,i} = \frac{\tau_i \overline{x}_i \sigma_{\upsilon}^2 \sigma_{\theta}^2 + s_{\theta,i}\sigma_{\epsilon}^2\sigma_{\theta}^2 + \mu_{\theta}\sigma_{\epsilon}^2\sigma_{\upsilon}^2}{\tau_i\sigma_{\upsilon}^2\sigma_{\theta}^2 + \sigma_{\epsilon}^2\sigma_{\theta}^2 + \sigma_{\epsilon}^2\sigma_{\upsilon}^2}$ and $\sigma_{\theta,i}^2 = \frac{\sigma_{\epsilon}^2\sigma_{\upsilon}^2\sigma_{\theta}^2}{\tau_i\sigma_{\upsilon}^2\sigma_{\theta}^2 + \sigma_{\epsilon}^2\sigma_{\theta}^2 + \sigma_{\epsilon}^2\sigma_{\upsilon}^2}$ and $\overline{x}_i \equiv \sum_{k=1}^{\tau} \frac{\ln \overline{\phi}_{i,\tau-k} - (1-\delta)\ln \phi_{i,\tau-k-1}}{\tau}$, where $\overline{\phi}_{i,\tau-k}$ ($\phi_{i,\tau-k-1}$) denotes the level of $\overline{\phi}_i$ (ϕ_i) during the period when the firm was $\tau - k$ ($\tau - k - 1$) periods old. Note that the variable $\overline{x}_i$ is the average level of $\theta_i + \epsilon_i$ that the firm has observed over the course of its lifetime.

Note that there are four state variables that the firm uses to make optimal decision given the available information. The first state variable is the observed level of demand (before new investment), $\overline{\phi}_i$, the second is $\overline{x}_i$, and the third state variable is the pre-entry signal $s_{\theta,i}$. The final state variable is the firm's age, τ_i.

3.1.2 Entry

Starting up a firm requires a fixed amount of resources given by χ and denominated in units of the aggregate consumption basket. After paying this cost, the firm receives the signal $s_{\theta,i}$ and learns about its initial level of demand $\ln \overline{\phi}_i = \theta_i + \epsilon_i$. At this point the firm can decide to shut down operations without every producing. Free entry dictates the following condition:

$$\chi = \mathbb{E} \max \left\{ V\left(\phi_i, s_{\theta,i}\right), 0 \right\}$$

where the expectation operator $\mathbb{E}$ takes into account the distributions from which θ_i, ϵ_i and $s_{\theta,i}$ are drawn.[6]

3.2 Households

There is a representative household which derives utility from a CES basked of goods varieties as well as disutility from labor. The household's utility flow is given by:

$$\mathcal{U} = \left(\left(\int_{i \in \Omega} \phi_i^{\frac{1}{\eta}} C_i^{\frac{\eta-1}{\eta}} di \right)^{\frac{\eta}{\eta-1}} \right) - \nu N$$

where C_i denotes consumption of variety i, Ω denotes the total set of available varieties, N denotes labor supply, and ν is labor disutility parameter. As noted above, η is the elasticity of substitution between goods varieties. Moreover, ϕ_i is a variety-specific preference parameter which evolves according to Equation (11). The budget constraint of the household is given by:

$$\int_{i \in \Omega} p_i c_i di = PWN + \Pi$$

where Π denotes aggregate nominal firm profits, net of wages, fixed operation costs and entry costs.

The household optimally chooses the level of consumption of each variety, C_i, and the level of labor supply, N. The first-order conditions of the household's decision problem give rise to the demand constraint, Equation (10) and the aggregate price index can be shown to satisfy $P \equiv \left(\int_{i \in \Omega_i} \phi_i p_i^{1-\eta} dj \right)^{\frac{1}{1-\eta}}$.

3.3 Stationary equilibrium

We consider a stationary equilibrium in which the distribution of firm-level state variables remains constant over time, households and firms solve their decision problems, and the labor and goods markets clear.

[6]Note that, before entry, $\overline{x}_i = \ln \overline{\phi}_i$ and hence we do not include $\overline{x}_i$ as a separate state variable in the above equation.

3.4 Calibration

To be done.

In principle, we can calibrate the model by simulating it and matching the autocovariance with that in the data. I'm trying something a bit more rough right now, the idea is below (using targets from the document, I guess for the UK data?):

- annual frequency: $\beta = 0.96$
- estimated $\rho = 0.6$?
- σ_θ: cross-sectional variance of old firms (var($n_{15} = 1.05$, Figure 1)
- μ_θ: average size of old firms ($\mathbb{E}[n_{15}] = 60$, Figure 1)
- c_f: exit rate of 1 year old firms (about 20%)
- σ_ϵ: exit rate of young firms (5 years, about 10%) - even without learning you get a gradual decline in the exit rate depending on how large the idiosyncratic shocks are
- ψ: average size of young (new) firms (about 1.5, Figure 1) - the adjustment costs will determine the growth profile This is, however, conditional on ρ. Moreover, the empirical part is based on no adjustment costs (otherwise employment is an AR(2)). I'm not entirely sure how to deal with that nicely right now...
- σ_υ: cross-sectional variance of employment in new firms (about 0.7, Figure 1). The idea is that for a given σ_θ and σ_ϵ, any remaining dispersion in new firms' employment levels must be because of dispersion in beliefs (unless there are additional shocks at birth or large measurement error).

References

Jaap Abbring and Jeffrey Campbell. A firm's first year, 2005. Tinbergen Institute Discussion Paper 05-046/3.

John Abowd and David Card. On the covariance structure of earnings and hours changes. *Econometrica*, 57(2):4111–445, 1989.

T.W. Anderson and Cheng Hsiao. Formulation and estimation of dynamic models using panel data. *Journal of Econometrics*, 18:47–82, 1982.

Manuel Arellano and Stephen Bond. Some tests of specification for panel data: Monte carlo evidence and application to employment equations. *Review of Economic Studies*, 58(2):277–297, 1991.

Costas Arkolakis, Theodore Papageorgiou, and Olga Timoshenko. Firm learning and growth, 2015. mimeo.

Fatih Guvenen and Anthony Smith. Inferring labor income risk and partial insurance from economic choices. *Econometrica*, 82(6):2085–2129, 2014.

Fatih Guvenen. Learning your earning: Are labor income shocks really very persistent? *American Economic Review*, 97(3):687–712, 2007.

Fatih Guvenen. An empirical investigation of labor income processes. *Review of Economic Dynamics*, 12(1):58–79, 2009.

John Haltiwanger, Ron Jarmin, Robert Kulick, and Javier Miranda. High growth young firms: Contribution to job, output and productivity growth, 2014. working paper.

Hugo Hopenhayn and Richard Rogerson. Job turnover and policy evaluation: A general equilibrium analysis. *Journal of Political Economy*, 101(5):915–938, 1993.

Hugo Hopenhayn. Entry, exit and firm dynamics long run equilibrium. *Econometrica*, 60(5):1127–1150, 1992.

Boyan Jovanovic. Selection and the evolution of industry. *Econometrica*, 50(3):649–670, 1982.

Thomas MaCurdy. The use of time-series processes to model the error structure of earnings in a longitudinal data analysis. *Journal of Econometrics*, 18:83–114, 1982.

Marc Melitz. The impact of trade on intra-industry reallocations and aggregate industry productivity. *Econometrica*, 71:1695–1725, 2003.

4 Appendix

4.1 Learning

This Appendix derives the Bayesian updating formulas presented in the main text. The underlying problem for the firm is the following. Individual firms face the following demand constraint

$$y_i = \phi_i \left(\frac{p_i}{P}\right)^{-\eta} Y, \tag{15}$$

where p_i is the price set by firm i, P is an aggregate price index, Y is aggregate output, and $\eta > 1$ is the elasticity of price substitution between individual goods. Further, ϕ_i is a firm-specific demand- shifter which stems from the utility level that the representative household derives from the variety i. We refer to ϕ_i as *marketing capital.* The demand-shifter evolves according to the following law of motion:

$$\ln \phi_i' = \theta_i + (1-\delta)\ln \phi_i + m_i' + \epsilon_i', \tag{16}$$

where primes denote next period's variables. Here, θ_i is a parameter which is time-invariant specific to the variety produced by firm i, and draw before entry from distribution $N(\mu_\theta, \sigma_\theta^2)$. Moreover, $\delta \in (0,1)$ is the rate at which marketing capital depreciates, m_i is a marketing investment made by the firm, which comes at a quadratic cost given by $\frac{\psi}{2}m_i^2$, with $\psi > 0$, and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ is an demand shock which is i.i.d. across firms and over time. Startup firms enter the economy without any marketing capital.

Importantly, the firm does not observe ϕ_i directly, but must learn about it over time. We assume that firms can observe $\epsilon_i + \theta_i$. Note also that In addition, before entry, the firm receives some information about its θ_i. In particular, we assume that upon entry (after paying an entry cost) the firm observes a signal $s_{\theta,i}$ given by

$$s_{\theta,i} = \theta_i + \upsilon_i, \tag{17}$$

where υ_i is an i.i.d. noise shock which is distributed as $\upsilon_i \sim N(0, \sigma_\upsilon^2)$. Finally, we assume that firms use a Bayesian learning algorithm to update their beliefs about θ_i. In particular, they combine their prior knowledge about θ_i, the known distribution $N(\mu_\theta, \sigma_\theta^2)$, the obtained signal at entry, distributed as $N(0, \sigma_\upsilon^2)$, and the observed realizations of demand over the course of their life-cycle. For the latter piece of the puzzle it is useful to define a marketing level before new investment (but after the shock) as $\ln \overline{\phi}_i \equiv \ln \phi_i - m_i$ and let τ_i denote the age of the firm i. The posterior distribution of θ_i can then be written as

$$P(\theta_i|\mathbb{I}_\tau) \propto \exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma_\epsilon^2}\sum_{t=1}^{\tau}(x_t-\theta_i)^2 + \frac{1}{\sigma_\upsilon^2}(s_\theta-\theta_i)^2 + \frac{1}{\sigma_\theta^2}(\theta_i-\mu_\theta)^2\right)\right\}, \tag{18}$$

where $\mathbb{I}_\tau$ denotes the information available to the firm, $x_t = \ln \overline{\phi}_t - (1-\delta)\phi_{t-1}$. This latter term is distributed as $N(\theta_i, \sigma_\epsilon^2)$. Thus, the first term in the brackets on the RHS is the likelihood of all the observations from $t = 1$ to $t = \tau$, the second term is the likelihood of the signal and the last term is the prior. Expanding the brackets, collecting terms and completing the square gives the following expression

$$P(\theta_i|\mathbb{I}_\tau) \propto \exp\left\{-\frac{1}{2\sigma_{\theta|\tau}^2}(\theta_i-\mu_{\theta|\tau})^2\right\}, \tag{19}$$

where $\mu_{\theta|\tau} = \frac{\tau\overline{x}\sigma_\upsilon^2\sigma_\theta^2 + s_\theta\sigma_\epsilon^2\sigma_\theta^2 + \mu_\theta\sigma_\epsilon^2\sigma_\upsilon^2}{\tau\sigma_\upsilon^2\sigma_\theta^2 + \sigma_\epsilon^2\sigma_\theta^2 + \sigma_\epsilon^2\sigma_\upsilon^2}$ and $\sigma_{\theta|\tau}^2 = \frac{\sigma_\epsilon^2\sigma_\upsilon^2\sigma_\theta^2}{\tau\sigma_\upsilon^2\sigma_\theta^2 + \sigma_\epsilon^2\sigma_\theta^2 + \sigma_\epsilon^2\sigma_\upsilon^2}$ are the mean and variance of θ conditional on having τ observations ($\overline{x} = \sum_{t=1}^{\tau} x_t/\tau$). We can rewrite the conditional mean as

$$\mu_{\theta|\tau} = \frac{\sigma^2_{\theta|\tau}}{\sigma^2_\epsilon \sigma^2_\theta}(\sigma^2_\theta \tau \overline{x} + \sigma^2_\epsilon \mu_\theta) + \sigma^2_{\theta|\tau} s_\theta / \sigma^2_\upsilon,$$

which shows that the effect of the initial signal dies out as τ grows large (since the conditional variance declines with τ).

Finally, the firm's optimal decisions will be based on the conditional expectations of $\ln \overline{\phi}_{\tau+1}$, given the learning it was able to do up until τ. Using the above,

$$\ln \overline{\phi}_{\tau+1|\tau} \sim N(\mu_{\theta|\tau} + (1-\delta)\ln \phi_\tau, \sigma^2_{\theta|\tau} + \sigma^2_\epsilon), \tag{20}$$

showing that tomorrows expected demand (prior to the endogenous marketing investment) is the sum of the conditional mean of the permanent component together with the autoregressive term and its variance is the sum of the conditional variance of the permanent component and that of the transitory shock.

The Appendix shows that the firm's posterior belief about the distribution of $\ln \overline{\phi}_i$ in the next period satisfies:

$$\ln \overline{\phi}'_i \sim N(\mu_{\theta,i} + (1-\delta)\ln \phi_i, \sigma^2_{\theta,i} + \sigma^2_{\epsilon,i}) \tag{21}$$

4.2 GMM estimation and overidentification from autocovariance

Repeating from above: (CHANGE NOTATION TO MATCH PAPER)

Generally, the autocovariance function for $a, j \geq 0$ is:

$$\begin{aligned} \text{Cov}\left[\log z_a, \log z_{a+j}\right] &= \sigma^2_\varepsilon \sum_{k=0}^{a-1} \rho^{2k+j} + \sigma^2_\theta \sum_{k=0}^{a} \rho^k \sum_{k=0}^{a+j} \rho^k + \rho^{2(a+1)+j}\sigma^2_z & (22) \\ &= \sigma^2_\varepsilon \rho^j \frac{1-\rho^{2a}}{1-\rho^2} + \sigma^2_\theta \frac{1-\rho^{a+1}}{1-\rho}\frac{1-\rho^{a+j+1}}{1-\rho} + \rho^{2(a+1)+j}\sigma^2_z & (23) \end{aligned}$$

Exact identification of $\rho, \sigma_\theta$and σ_z follows from the first three autocovariances, $\text{Cov}\left[\log z_0, \log z_a\right]$ $0 \leq a \leq 2$, which are a nonlinear function of only ρ, σ_θ and σ_z. These can be solved in closed form for the parameters:

$$\begin{aligned}\frac{\mathrm{Cov}\left[\log z_0, \log z_2\right] - \mathrm{Cov}\left[\log z_0, \log z_1\right]}{\mathrm{Cov}\left[\log z_1, \log z_0\right] - \mathrm{Cov}\left[\log z_0, \log z_0\right]} &= \rho \frac{\rho\sigma_\theta^2 + \rho\sigma_z^2}{\rho\sigma_\theta^2 + \rho^3\sigma_z^2} \\ &= \rho\end{aligned}$$

$$\mathrm{Cov}\left[\log z_1, \log z_0\right] - \rho \mathrm{Cov}\left[\log z_0, \log z_0\right] = \sigma_\theta^2$$

$$\frac{\mathrm{Cov}\left[\log z_0, \log z_0\right] - \sigma_\theta^2}{\rho^2} = \sigma_z^2$$

Adding a fourth moment identifies σ_ε

$$\mathrm{Cov}\left[\log z_1, \log z_1\right] - \rho^2 \mathrm{Cov}\left[\log z_0, \log z_0\right] = \sigma_\varepsilon^2 + (1 + 2\rho)\,\sigma_\theta^2.$$

4.2.1 Nonlinear GMM estimation

Let $\theta = \left(\rho, \sigma_\theta^2, \sigma_z^2, \sigma_\varepsilon^2\right)'$ be an arbitrary parameter vector in compact parameter space $\mathbb{P}$. Since we use ages 0 to A, we define the $\frac{A*(A+1)}{2}$ length vector valued function

$$f\left(n_i, \theta\right) = \left[\left(\log n_{ia} - E\left[\log n_{ia}\right]\right) \log n_{ij} - \mathrm{Cov}\left[\log n_{ia}, \log n_{ia+j}; \theta\right]\right]$$

where $a = 0, \ldots, A$ and $j = a,\ a{+}a,\ \ldots,\ A$. Let θ_0 be the true parameter vector, so that identification follows from $E\left[f\left(n_i; \theta\right)\right] = 0$ iff $\theta = \theta_0$. The term $\mathrm{Cov}\left[\log n_{ia}, \log n_{ia+j}; \theta\right]$ is a constant and equal to the formula from equation **??** computed for an arbitrary parameter vector θ.

Writing out some of the elements of f:

$$f(n_i;\theta_0) = \begin{bmatrix} (\log n_{i0} - E[\log n_{i0}]) \log n_{i0} - \frac{\sigma_{\theta 0}^2 + \rho^2 \sigma_z^2}{(1-\alpha)^2} \\ (\log n_{i0} - E[\log n_{i0}]) \log n_{i1} - \frac{(1+\rho_0)\sigma_\theta^2 + \rho^3 \sigma^2}{(1-\alpha)^2} \\ \vdots \\ (\log n_{i0} - E[\log n_{i0}]) \log n_{iA} - \frac{\sigma_\theta^2 \sum_{k=0}^{10} \rho^k + \rho^{12} \sigma_z^2}{(1-\alpha)^2} \\ (\log n_{i1} - E[\log n_{i1}]) \log n_{i1} - \frac{\sigma_\varepsilon^2 + (1+\rho)^2 \sigma_\theta^2 + \rho^4 \sigma_z^2}{(1-\alpha)^2} \\ \vdots \\ (\log n_{iA} - E[\log n_{iA}]) \log n_{i1} - \frac{\sigma_\varepsilon^2 \sum_{k=0}^{A-1} \rho^{2k+A} + \sigma_\theta^2 \sum_{k=0}^{A} \rho^k \sum_{k=0}^{2A} \rho^k + \rho^{2(A+1)+A} \sigma_z^2}{(1-\alpha)^2} \end{bmatrix}$$

Define the sample analog to $E[f(n_i;\theta)]$

$$g_N(\theta) \equiv \frac{1}{N} \sum_i f(n_i;\theta).$$

A law of large numbers implies $g_N(\theta) \to^p E[f(n_i;\theta)]$. Define the GMM estimator

$$\tilde{\theta}_N = \underset{\theta \in \mathbb{P}}{\operatorname{argmin}}\, g_N(\theta)' W g_N(\theta)$$

for an arbitrary symmetric positive definite weighting matrix W. The asymptotic distribution of the estimator $\tilde{\theta}_N$ is:[7]

$$\sqrt{N}\left(\tilde{\theta}_N-\theta_0\right)\rightarrow_d N\left(0,\Sigma\right)$$

where

$$\begin{aligned}\Sigma &\equiv \left(d'Wd\right)^{-1}\left(d'WVWd\right)\left(d'Wd\right)^{-1}\\ d &\equiv \frac{\partial Ef\left(n_i;\theta_0\right)}{\partial\theta'}\\ V &\equiv E\left[f\left(n_i;\theta_0\right)f\left(n_i;\theta_0\right)'\right].\end{aligned}$$

Note V is not a covariance matrix for $\log n_{ia}$ since $E\left[f\right]$ is the unique elements of the covariance matrix.

Estimation To operationalize the estimator we have to estimate both V and the means $E\left[\log n_i\right]$. Define

$$\tilde{f}\left(n_i,\theta\right)\equiv\left[\left(\log n_{ia}-\frac{1}{N}\sum_{i'}\log n_{ia}\right)\log n_{ij}-\mathrm{Cov}\left[\log n_{ia},\log n_{ia+j};\theta\right]\right]$$

[7] To see this, write $g_N\left(\tilde{\theta}_N\right)$ as

$$g_N\left(\tilde{\theta}_N\right)\approx g_N\left(\theta_0\right)+\frac{\partial g_N\left(\theta_0\right)}{\partial\theta'}\left(\tilde{\theta}_N-\theta_0\right).$$

Multiplying through by $\frac{\partial g_N\left(\tilde{\theta}_N\right)}{\partial\theta'}W$ so that the LHS is equal to the first order condition $\frac{\partial g_N\left(\tilde{\theta}_N\right)}{\partial\theta'}Wg_N\left(\tilde{\theta}_N\right)=0$, then

$$\begin{aligned}0 &\approx \frac{\partial g_N\left(\tilde{\theta}_N\right)}{\partial\theta'}Wg_N\left(\theta_0\right)+\frac{\partial g_N\left(\tilde{\theta}_N\right)}{\partial\theta'}W\frac{\partial g_N\left(\theta_0\right)}{\partial\theta'}\left(\tilde{\theta}_N-\theta_0\right)\\ \left(\tilde{\theta}_N-\theta_0\right) &\approx -\left(\frac{\partial g_N\left(\tilde{\theta}_N\right)}{\partial\theta'}W\frac{\partial g_N\left(\theta_0\right)}{\partial\theta'}\right)^{-1}\frac{\partial g_N\left(\tilde{\theta}_N\right)}{\partial\theta'}Wg_N\left(\theta_0\right).\end{aligned}$$

Letting $N\rightarrow\infty$ then

$$\sqrt{N}\left(\tilde{\theta}_N-\theta_0\right)\rightarrow^d-\left(\frac{\partial g_N\left(\theta_0\right)}{\partial\theta'}W\frac{\partial g_N\left(\theta_0\right)}{\partial\theta'}\right)^{-1}\frac{\partial g_N\left(\theta_0\right)}{\partial\theta'}W\sqrt{N}g_N\left(\theta_0\right)$$

since $\tilde{\theta}_N\rightarrow^p\theta_0$. And from the CLT $\sqrt{N}g_N\left(\theta_0\right)\rightarrow^d N\left(0,V\right)$.

$$\tilde{g}_N(\theta) = \frac{1}{N}\sum_{i=1} \tilde{f}(n_i, \theta),$$

where $a = 0, \ldots, 10$ and $j = a, a+a, \ldots, 10$. Define the $A(A+1)/2 \times A(A+1)/2$ moment covariance matrix

$$\tilde{V}_N = \frac{1}{N}\sum_{i=1}\left[\tilde{f}(n_i,\theta)\,\tilde{f}(n_i,\theta)'\right] = \frac{1}{N}\sum_{i=1}\left(h(n_i) - \bar{h}\right)h(n_i)'.$$

Note that since $\tilde{f}(n_i,\theta) = h(n_i) - q(\theta) = 0$ then $E\left[\tilde{f}_i\tilde{f}_i\right] = \text{Cov}\left[\tilde{f}_i, \tilde{f}_i\right] = \text{Cov}\left[h(n_i), h(n_i)\right]$.

To deal with missing data. We can create an indicator variable λ_{iaj} for whether or not the observation is missing and then define

$$\tilde{f}(n_i,\theta,\lambda_i) \equiv \left[\lambda_{iaj}\left(\left(\log n_{ia} - \frac{1}{N}\sum_{i'}\log n_{ia}\right)\log n_{ij} - \text{Cov}\left[\log n_{ia}, \log n_{ia+j};\theta\right]\right)\right]$$

and use weighting matrix

$$A = \Pi^{-1}\Pi^{-1}$$

where

$$\Pi = \begin{bmatrix} \frac{N_{00}}{N} & & & \\ & \frac{N_{01}}{N} & & \\ & & \ddots & \\ & & & \frac{N_{AA}}{N} \end{bmatrix}$$

and N_{aj} is the number of non missing observations for that moment. This is equivalent to equally weighting but uses the correct number of observations when computing $\tilde{V}$

4.3 Autocovariance of employment

Provide the full autocovariance matrix for reference