Journal of Monetary Economics 90 (2017) 99–112

Contents lists available at ScienceDirect

Journal of Monetary Economics

journal homepage: www.elsevier.com/locate/jmoneco

Cash burns: An inventory model with a cash-credit choice☆

Fernando Alvarez[a], Francesco Lippi[b,*]

[a] University of Chicago, NBER, 1126 East 59th Street, Chicago, IL 60637, USA
[b] EIEF and University of Sassari, Via Sallustiana 62, Rome, Italy

ARTICLE INFO

Article history:
Received 18 February 2015
Revised 4 July 2017
Accepted 6 July 2017
Available online 12 July 2017

Keywords:
Inventory models
Cash-credit choice
Means of payment
Money demand
Phasing out cash

ABSTRACT

A dynamic cash-management model is analysed where agents choose whether to pay with cash or credit at every point in time. In the model credit usage depends on the current stock of cash, a novel result that matches recent micro evidence on households' payment choices. The optimality of such a decision rule is novel and cannot be obtained by models where cash-credit decisions are made at the "beginning" of each period. We discuss how to use the model to account for cross country-evidence on the intensity of credit usage and for several statistics on the size and frequency of cash withdrawals. The model is used to assess the household's welfare cost of phasing out cash.

1. Introduction and overview

A dynamic model of cash-management and means-of-payment choice is analysed in which optimizing households use both cash and credit. Credit is modeled as a payment instrument that involves a cost but requires no inventory at hand, such as a debit or a credit card, while the use of cash is modeled with a standard inventory setup. A central feature of the model is that at each moment the agent can choose to pay with either cash or credit. This realistic feature is an essential novelty of our model, which implies that the preferred payment instrument depends on the stock of cash holdings at the time of the purchase: agents use cash for a purchase as long as they have cash with them, and use credit otherwise. Agents behave as if "cash burns" in their hands, a pattern that has been noticed by the empirical literature.[1] Yet, agents find it optimal to intermittently replenish their cash holdings, e.g. by asking for "cash-back" after paying with credit, and so end up using both cash and credit.

Our model is the first to have both simultaneous use of cash and credit by households, as well as credit use that depends on the level of cash holdings. This feature cannot be obtained by models where cash-credit decisions are made at the "beginning" of each period since, by assumption, those models do not keep track of the dynamics of cash balances and so the use of credit cannot be conditioned on the stock of cash. The model provides a simple analytical mapping between the fundamental parameters and several observable statistics on the size and frequency of cash withdrawals, as well as on

☆ An earlier version of the paper circulated under the name: "A dynamic cash management and payment choice model". We benefited from the comments of the participants to "Consumer Finances and Payment Diaries: Theory and Empirics", Ottawa, Canada, 18–19 October 2012. Philip Barrett and David Argente provided excellent research assistance. Part of the research for this paper was sponsored by the ERC advanced grant 324008.

* Corresponding author.
E-mail addresses: falvare@uchicago.edu (F. Alvarez), f.lippi@uniss.it (F. Lippi).

[1] See, for example, Stix (2004), Mooslechner et al. (2006), Arango et al. (2011), Arango et al. (2015) and Huynh et al. (2014). See Appendix A for a brief review of this evidence.

http://dx.doi.org/10.1016/j.jmoneco.2017.07.001

the means of payment choice that determines the share of purchases paid with credit. To illustrate the applicability of the model we use it to assess the household's welfare cost of phasing out cash, a policy endorsed by Rogoff (2016) as a measure against several illegal practices.

The analytic setup consists of the Baumol–Tobin cash inventory model augmented with two features. First, we assume the agent randomly receives some *free* withdrawal opportunities, as in Alvarez and Lippi (2009). This assumption gives rise to a precautionary motive in the demand for cash, such as an above zero cash-balance at the time of cash withdrawals, a pattern systematically observed in household diary and survey evidence. Second, we allow agents to pay for their (exogenous) constant flow of expenditure using either cash or credit. Our choice of modeling expenditures as a constant flow has the interpretation of small-sized purchases. Paying with cash requires to have positive cash inventory at the time of the purchase, which is costly to accumulate and which has an opportunity cost, both standard features of an inventory model. Paying with credit entails a cost per transaction.[2] In principle it is possible for the agent to pay all of its consumption using credit but, as long as there are some free cash withdrawal opportunities, it will be optimal for the agent to also use cash.

The paper has two main predictions for the optimal choice of the means of payment. First, as long as the cost of cash withdrawals is low enough relative to the cost of credit, the agents never use credit. They take advantage of their free trips to the bank, possibly some costly trips, and use cash only. Second, if the cost of cash withdrawals is sufficiently high, then cash is withdrawn every time a *free* cash withdrawal opportunity occurs. The agent then uses cash until she runs out of it, and subsequently uses credit until the next free cash withdrawal opportunity occurs. Intuitively, for an agent with cash at hand, the cost of obtaining it is sunk. As a result, using cash is optimal since the agent pays only the opportunity cost.

We analytically derive several predictions and compare them with cross-country evidence gathered from household surveys and diary data as summarised in e.g. Bagnall et al. (2016). As mentioned, the model predicts that credit is more likely to be used when the agent is short of cash, a pattern that is robustly documented in micro studies (see footnote 1). Another related phenomenon that our model is uniquely poised to explain is obtaining "cash-back" at the register at the time of a credit purchase. This is an instance where the coexistence is clear and where, as in our model, cash is obtained because withdrawing it is particularly cheap –see our discussion of the evidence in Klee (2008) in Section 2. Moreover, the model has predictions concerning the average share of expenditures paid with cash, the frequency of cash withdrawals and their average size, the average cash holdings (both unconditional as well as at the time of cash withdrawals). We show that at low inflation rates the five moments listed above are functions of 2 parameters only: the normalized cost of credit, namely the ratio between the cost of credit and the cost of cash withdrawals, and the frequency of the free withdrawal opportunities. We discuss how these two parameters can be identified from observations on e.g. the number of cash withdrawals and the average cash holdings. We then compare with the data the model predictions about other moments not used for the parameter identification. We show that, in spite of its simplicity, the model gets the appropriate magnitudes for several observed moments.

Finally, we use our structural model to quantify the cost of a policy that limits household's cash usage, a policy endorsed by Rogoff (2016) to fight several illegal activities which are known to be cash intensive. Our objective is to quantify the welfare cost for households who are forced to move from their optimal cash-credit share to one where the cash share is zero. The analysis shows that such cost is small.

1.1. Related literature

Many papers in the literature incorporate alternative means of payments as in the seminal work of Lucas and Stokey (1987) and Prescott (1987). However such models do not have an explicit inventory theoretical model of money, so they cannot simultaneously speak to observations such as the fraction of purchases made in cash as well as cash-management statistics, such the frequency and size of cash withdrawals. Technically, in this type of models the (exogenously imposed) cash-in-advance constraint binds in every period. As a result, one withdrawal occurs every period and all cash is spent. Hence, statistics such as frequency of withdrawals, size of withdrawals, cash at the time of withdrawals, are all exogenously determined by the choice of the model's time-period length. Other models incorporate both cash management and the choice of means of payments, which ends up being dictated by the size of the purchases. Examples of such models are Whitesell (1989) or Freeman and Kydland (2000). Yet while these models introduce cash-management, those choices are all "within" the period, so that agents cannot choose *at every moment* whether to use cash or credit. Hence in these models the optimal use of credit cannot depend on cash holdings, contrary to what the data strongly suggests.[3]

The closest related models in the literature are Sastry (1970) and Bar-Ilan (1990). Sastry (1970) is one of the earliest inventory models featuring a sequential cash versus credit choice. In his deterministic Baumol–Tobin model with no

[2] The choice of cash vs credit based on the size (i.e. dollar value) of purchases has been addressed in both theoretical and empirical literature, see Arango et al. (2012) and Bouhdaoui and Bounie (2012). Empirically, smaller transactions are more likely to be paid with cash than with credit, which motivates the assumption of a fixed cost per transaction in the literature. Yet, there are many small transactions paid with either cash or credit. Our assumption of a constant flow of expenditure thus focuses on transactions that are all of the same size and small. It is thus complementary to the explanation in the literature based on size and it is able to address the choice of means of payments for small size transactions. See Alvarez and Lippi (2013) for a complementary analysis of an inventory (cash-only) model that distinguishes between large vs small sized purchases.

[3] A close analogy between our sequential formulation and these papers' simultaneous cash-credit choice is found in the difference between sequential search, as in McCall's model, versus simultaneous search, as in Stigler's search model. See Sargent's (1987) chapter 2 for a description of the two types of search models.

discounting the agent is allowed to use credit, namely an overdraft ("negative cash"), so that when cash holdings reach zero the agent may continue to consume and postpone the payment of the fixed withdrawal cost. Bar-Ilan (1990) extends this setup to a dynamic stochastic inventory model. The main difference with our setup is that the cost of credit in both of these models is assumed to be proportional to the average *stock* of credit over the holding period, completely analogue to the opportunity cost of cash. This implies that the agent using credit will periodically decide to pay the rebalancing cost in order to keep the average stock of credit bounded. Notice that under this assumption it is infinitely costly not to pay the fixed transanction cost since this implies a diverging stock of credit. In our setup, in contrast, the cost of credit is proportional to the expenditure *flow*, e.g. it is a fixed fraction of the purchase value. Using credit does not require any fixed cost to "rebalance" the cash credit stock. Credit purchases are immediately debited on to the agent's checking account. In addition to being more realistic, our assumption explicitly distinguishes the credit technology from the cash technology, and thus makes a credit-only strategy of purchases feasible for the agent.

1.2. Organization of the paper

The structure of the paper is as follows. We illustrate the model's key idea in Section 2 with a simple deterministic steady-state model. Section 3 introduces uncertainty and a proper dynamic treatment of the inventory problem with payment-choice and characterizes the conditions under which both cash and credit are used. Section 4 derives the model's implications for the frequency and size of cash withdrawals and the intensity of credit usage. We discuss some cross-country evidence to illustrate how the model can be calibrated to actual economies. In Section 5, we use our structural model to quantify the cost of a policy that imposes a zero-cash usage restriction. Section 6 extends our model by allowing for a random cost of cash withdrawals, a feature which appears desirable for empirical applications. Section 7 concludes.

2. A deterministic model with means of payment choice

This section presents a steady-state deterministic model that highlights the main mechanism of the dynamic stochastic model of Section 3. Indeed some key formulas from this simple model coincide with, or are close to, the more complex decision rules of the stochastic model. The main counterfactual prediction of the deterministic model is the lack of a precautionary motive, so that cash balances are zero at the time of a withdrawal.[4]

Consider an agent who consumes e per unit of time and can pay for this using either cash or credit. If she pays with credit she incurs a direct cost γ per unit bought. The quantity γe can be interpreted as the "handling" and "verification and authorization" costs estimated in Klee (2008) using grocery receipts data. The idea is that, as the empirical evidence in Klee suggests, using credit is more time consuming than using cash so that cash is preferred when both instruments are available. In particular, Klee uses a dataset of time stamped checkout transactions, from 99 stores of a retail chain, containing more than 6 millions checkout transactions. For each transaction at a cash register the data record the time to completion, the means of payment used for it (cash, credit, debit, check, or other), the amount of the purchase, the amount paid, whether the customer obtained cash back (i.e. whether they use it a source of cash), the number of items bought, and whether coupons were used. Table 2 on page 533 of Klee (2008) estimates the difference in time that it takes to complete it ("to ring it") transactions that are paid with different instruments: credit, debit, and checks relative to cash, controlling for other features such the number of item of transaction. These magnitudes are what Klee refers as "handling cost" and "authorization and verification costs" —see Table 3 on page 535 of Klee (2008). In Section 4 we discuss how to quantify the credit cost γ in the light of empirical evidence.

The technology to withdraw cash (from an interest bearing checking account) is as follows: at any time the agent can pay a fixed cost b and replenish her cash balances which, as in the canonical Baumol–Tobin model, are subject to an opportunity cost R which includes forgone interest on deposits as well as the probability of cash theft. Moreover, the agent has $p \geq 0$ withdrawals per period that come for free. The latter assumption is a simple parametrization of the technology for the cash withdrawals, proxying e.g. for the number of ATMs (cheap withdrawals) available to the agent.[5] To understand the nature of the optimal policies considered below notice that in a deterministic setup an agent with positive cash balances will not pay the fixed cost b to withdraw cash unless cash balances are zero. Consider now the decision of whether to purchase goods using cash or credit. For an agent with positive cash balances $m > 0$ it is not optimal to pay the cost γe to use credit, since the cost of acquiring the cash is sunk at this time.[6] For an agent with zero cash balances $m = 0$, there are two possible choices: the first one is to pay the cost γe and finance consumption using credit, waiting until the next free withdrawal opportunity to replenish cash balances. The second choice is to pay the fixed cost b and withdraw cash. We first separately describe the solution of these two cases and then analyse the best choice among the two.

[4] Readers familiar with continuous time impulse control problems may move directly to Section 3.

[5] See Appendix D for an explicit foundation of how cash theft probability and the nominal interest rate affect R. See Alvarez and Lippi (2009) for estimates of p using Italian household data and for their Appendix C for estimates of cash theft probabilities in Italy and the US.

[6] Another possibility is to use credit and deposit the cash to earn a higher interest. With a fixed cost for depositing this is not optimal unless the cash balances are very large, a situation that will not occur along an optimal path.

2.1. The deterministic cash-credit model

Consider an agent who finances her expenditures using cash, and who pays with credit once cash balances are depleted. Assume further that no costly withdrawal ever takes place so that b is never paid and the number of withdrawals, n, equals the number of withdrawals that come for free, p. After a cash withdrawal of size $W = m^*$, she spends $\tau_a = m^*/e$ units of time paying for consumption with cash, incurring an opportunity cost $Rm^*/2$, where $m^*/2$ is the average cash balance conditional on $m > 0$ and R is the opportunity cost of cash– which includes the nominal interest rate as well as the probability of cash theft. After cash balances hit zero, the remaining time until a free withdrawal opportunity, denoted by τ_r, is given by $\tau_r = 1/p - m^*/e$. Notice that $\tau_r + \tau_a = 1/p$. The steady state cost in every cycle of duration $1/p$ can be written as: $\tau_r\ \gamma\ e + \tau_a\ R\ m^*/2$. The cost per unit of time is thus $p\,\tau_r\ \gamma\ e + p\,\tau_a\ R\ m^*/2$. Thus the minimized cost of the strategy that uses both cash and credit is:

$$v_r(R, \gamma, p, e) = \min_{0 \le m^* \le e/p} p\left[(1/p - m^*/e)\ \gamma\ e\ + (m^*/e)\ R\ e\ \frac{(m^*/e)}{2}\right], \tag{1}$$

subject to the constraint that the time spent using credit is non-negative, i.e. $m^*/e \le 1/p$. We denote by s the "cash share", namely the ratio of the expenditure paid with cash to total expenditure per unit of time, given by

$$s = \frac{\tau_a}{\tau_a + \tau_r} = \min\left\{p\frac{m^*}{e}, 1\right\}. \tag{2}$$

Denoting the average real balances by M and using the cash share s we can write $M = s\ m^*/2$. The cost minimizing policy yields $\frac{m^*}{e} = \frac{W}{e} = \min\left\{\frac{1}{p}, \frac{\gamma}{R}\right\}$, $s = \min\left\{1, \frac{\gamma\ p}{R}\right\}$, $\frac{M}{e} = \min\left\{\frac{1}{2\ p}, \frac{p}{2}\left(\frac{\gamma}{R}\right)^2\right\}$ which imply

$$v_r(R, \gamma, p, e) = \begin{cases} \left(1 - \frac{\gamma\ p}{2\ R}\right)\gamma\ e & \text{if } R \ge \gamma p \\ \left(\frac{R}{2\ p}\right)e & \text{if } R < \gamma p \end{cases} \tag{3}$$

When $R \ge \gamma p$ credit is "cheap", so that both cash and credit are used. When $R < \gamma p$ credit is expensive and it is not used.

2.2. Deterministic Baumol–Tobin model with p free withdrawals

Let us consider a modified Baumol–Tobin model in which the agent pays only for the withdrawals in excess of the p free adjustments per period. The agent chooses a withdrawal of size m^* when cash balances are exhausted ($m = 0$). The policy implies an average cash balance $M = m^*/2$ and a number $n = e/m^*$ of cash withdrawals. The agent's choice of m^* gives the minimized cost function

$$v_a(R, b, p, e) \equiv \min_{m^*} \left[R\ \frac{m^*}{2} + b\max\left(\frac{e}{m^*} - p, 0\right)\right]. \tag{4}$$

where the cost is given by the sum of the opportunity cost of cash holdings and the cost associated with cash withdrawals in excess of p. The optimal policy for a technology with $p \ge 0$ is $\frac{m^*}{e} = \frac{1}{p}\sqrt{\min\left(2\frac{b\ p^2}{e\ R},\ 1\right)}$. For $p > 0$ there is no reason to have less than p withdrawals, since these are free. Hence, for $R < 2p^2 b/e$ the agent will choose a constant level money holdings: $m^* = e/p$. Note that the interest elasticity of money is zero over this range, while it is equal to 1/2 if $R > 2p^2 b/e$. The average withdrawal size W and the average cash balances satisfy: $W = m^*$, $\ M = 2\ W = 2\ m^*$. Replacing the optimal m^* choice in the cost function yields

$$v_a(R, b, p, e) = \begin{cases} \left(\sqrt{2\ R\ \frac{b}{e}} - p\frac{b}{e}\right)e & \text{if } R \ge 2\ p^2\ \frac{b}{e} \quad \text{and } n > p \\ \left(\frac{R}{2\ p}\right)e & \text{if } R < 2\ p^2\ \frac{b}{e} \quad \text{and } n = p \end{cases} \tag{5}$$

where the top branch gives the cost for the case in which the number of withdrawals exceeds p. Note that in this deterministic setup an agent with positive cash balances will not pay the fixed cost b to withdraw cash unless cash balances are zero.

2.3. The full deterministic problem

We now analyse the conditions under which it is optimal to use credit instead of withdrawing fresh cash when $m = 0$. To do so we compare the steady-state cost of the two policies computed above. The value function for the problem is then $v(R, b, \gamma, p, e) = \min\{v_a(R, b, p, e)\ ,\ v_r(R, \gamma, p, e)\}$. We define the threshold function $\underline{b}$, as the value of b that equates the two minimized costs: $v_a(R, \underline{b}, p, e) = v_r(R, \gamma, p, e)$. We have that

$$\underline{b}(R, \gamma, p, e) = \frac{\gamma^2}{2\ R}\ e. \tag{6}$$

which implies that credit is used when $b \geq \underline{b}$ provided that $\gamma p \leq R$.[7] If $b \geq \underline{b}$ and $\gamma p > R$ then credit is not used and $n = p$. Finally, for $b < \underline{b}$ credit is not used and $n > p$ since some costly withdrawals in excess of the p free withdrawals are now optimal .

The next proposition summarises the behavior of the deterministic model. It considers two cases depending on whether $\gamma \gtrless 2p\, b/e$, and for each case it analyses optimal policy as a function of R.

Proposition 1. *Let $p > 0$ and $\gamma > 0$. Then $W/M = 2/s$ and*
If $\gamma > 2\,p\,\frac{b}{e}$, then

$$\text{if } R \in \left(0\ ,\ 2\,p^2\,\frac{b}{e}\right] \qquad -\frac{\partial \log M/e}{\partial \log R} = 0 \quad \text{only cash used} \quad n = p \quad s = 1$$

$$\text{if } R \in \left(2\,p^2\,\frac{b}{e}\ ,\ \frac{\gamma^2}{2\,b/e}\right] \qquad -\frac{\partial \log M/e}{\partial \log R} = 1/2 \quad \text{only cash used} \quad n > p \quad s = 1$$

$$\text{if } R \in \left(\frac{\gamma^2}{2\,b/e}\ ,\ \infty\right) \qquad -\frac{\partial \log M/e}{\partial \log R} = 2 \quad \text{cash \& credit used} \quad n = p \quad s = \gamma\,p/R$$

Otherwise, i.e. if $\gamma \leq 2\,p\,\frac{b}{e}$, then

$$\text{if } R \in (0, \gamma\,p] \qquad -\frac{\partial \log M/e}{\partial \log R} = 0 \quad \text{only cash used} \quad n = p \quad s = 1$$

$$\text{if } R \in (\gamma\,p\ ,\ \infty) \qquad -\frac{\partial \log M/e}{\partial \log R} = 2 \quad \text{cash \& credit used} \quad n = p \quad s = \gamma\,p/R$$

The proposition illustrates three robust properties of the model. First, the model has only two parameters, γp and $p^2\,b/e$. In the modified Baumol–Tobin model the shape of the money demand depends only on $\hat{b} \equiv p^2\,b/e$. For a given value of $\hat{b}$, the cash-credit aspect of the model depends only on γp. We will see that this property continues to hold in the stochastic model analysed below. Second, for credit to be used it is necessary that the cost of a withdrawal is above the threshold defined in Eq. (6). If $b > \underline{b}$ then the agent uses both cash and credit to finance her consumption, and costly withdrawals are never used. The condition for the optimality of credit depends on a combination of the fundamental parameters R, b and γp which, intuitively, imply that the cost of cash (which is increasing in R and b) must be high relative to the cost of credit (which is increasing in γ and p). Third, the interest rate elasticity of money demand is increasing in the interest rate. There are two cases: the first corresponds to a large cost of credit ($\gamma p > 2\,p^2\,b/e$), in which case there are three qualitatively different behavior depending on the level of interest rates. If interest rates are very low, credit is not used and $n = p$, resulting in an elasticity of zero. For intermediate level of interest rates, credit is not used, but $n > p$, so the local behavior is identical to Baumol–Tobin, producing an interest rate elasticity of 1/2. For higher interest rates, both cash and credit are used. The interest rate elasticity is higher here because both the cash share as well as the size of the withdrawals react to interest rates. If instead the cost of credit is low ($\gamma \leq 2\,p\,b/e$) there is no intermediate case, since credit always dominates the Baumol–Tobin type of behavior.

3. A dynamic stochastic model with means of payment choice

In this section we solve a discounted, stochastic dynamic problem which joins the optimal cash management problem with the optimal choice of means of payment. As in the deterministic problem the agent faces a total consumption per unit of time $e > 0$ which must be paid with either cash or credit: at each instant the agent can choose to pay in cash $c \in [0, e]$ and to pay the remaining $e - c$ using credit. If the payment is made by credit, the agent pays a flow cost γ per dollar.[8] See Section 2 for a description of Klee's (2008) estimates of these costs using the time that it takes to conclude different transactions. The quantity γe can be understood as the time cost of using credit for small value transactions, see the previous section for a more detailed discussion of its nature and Section 4 for a quantification of its magnitude. The state of the agent's problem is summarised by her real cash balances $m \geq 0$. If $m = 0$ either cash must be withdrawn or credit has to be used. If $m > 0$ the agent faces a cash/credit choice. The law of motion of real balances is then $dm = -(c + m\pi)\,dt$ provided that no adjustment takes place, where π is the constant inflation rate. The agent can adjust her cash balances after paying the fixed cost $b \geq 0$. Additionally, there is a Poisson process with constant arrival rate $p \geq 0$, which describes the arrival of a free adjustment opportunity. When such an opportunity arises the agent can adjust her cash balances at no cost. We assume that holding cash m entails an opportunity cost $R\,m$ per unit of time, where R can be interpreted as the sum of the nominal interest rate plus a probability that cash is lost or stolen (see footnote 5). We assume that the agent minimises the expected discounted cost, using the constant discount rate $r \geq 0$. There are three substantive differences between the model analysed in this section and the steady-state deterministic model of Section 2. First, we take into account explicitly

[7] The expression for $\underline{b}$ comes from equating: $v_r = \gamma e[1 - \gamma p/(2R)]$ with $v_a = e\sqrt{2\ R\ \frac{b}{e}} - pb$.

[8] Considering that all expenditures are of the same size, and that the optimal policy is of the bang-bang type, our formulation can be equivalently interpreted as a fixed cost applied to each purchase.

the role of inflation, as can be seen in the law of motion. Second, the free adjustment opportunities arrive stochastically. Third, real costs are discounted by an appropriate rate r.

Formally we denote by $V(m)$ the minimum expected discounted cost of supporting a constant flow of expenditure e when the current real cash at hand is $m \geq 0$. The function V, defined in $V : \mathbb{R}_+ \to \mathbb{R}_+$ must solve the following functional equation:

$$0 = \min\Big\{ \min_{0 \leq c \leq e} Rm + \gamma[e-c] + p \min_{z \geq 0}[V(z) - V(m)] - V'(m)(c + \pi m) - rV(m), \\ b + \min_{z \geq 0} V(z) - V(m) \Big\} \quad \text{for all} \quad m \geq 0. \tag{7}$$

The outer $\min\{\cdot,\cdot\}$ in the functional Eq. (7) chooses between two strategies. The term in the first line represents the case where no costly withdrawals (that involve paying the fixed cost b) occur, although random free withdrawal opportunities may arise, and the agent chooses what fraction of her consumption to pay in cash versus credit. This is a standard continuous time Bellman equation, with flow cost $Rm + \gamma(e-c)$ and with expected changes due to either the arrival of the free adjustment opportunity or to the depletion of cash. The minimisation with respect to c describes the agent's choice of the optimal means of payment. The term in the second line corresponds to the strategy of exercising control, i.e. paying the fixed cost b and adjusting cash holdings. For each m the value function is equal to the value of the strategy that yields the minimum cost. Whenever an adjustment is made, either paying the cost b or when a free adjustment opportunity arrives, the post-adjustment quantity of cash is chosen optimally. The optimal policy for the problem in Eq. (7) consists of deciding for each $m \geq 0$ whether a costly withdrawal is made or not and, if no adjustment is made, which payment instrument (cash or credit) to use. Notice that this formulation does not impose any restriction concerning the times of cash withdrawals or concerning when different payment instruments can be used. At each point in time the agent is free to choose whether to withdraw cash and what payment instrument to use. We maintain the following assumption throughout this section.

Assumption 1. We let $b \geq 0$, $\gamma \geq 0$, $\pi \geq 0$, $p \geq 0$, and $r + p + \pi > 0$, $e > 0$, $R > 0$.

If $e = 0$ the problem becomes uninteresting since there is no expenditure to finance. The parameters b and γ are costs, and p a probability rate, so they must be non-negative. The requirement that $r + p + \pi > 0$ and $R > 0$ are important. For instance, if $r + p + \pi = 0$ there is no intertemporal incentives to use cash.[9]

3.1. Two candidate policies

It will turn out that the optimal policy of the problem depicted in Eq. (7) is one of two types, depending on parameters. The first type is a cash-burning policy, defined as follows:

Definition 1. We define a *cash burning policy* as a threshold $m^* \geq 0$ for which:

1. Credit is only used when $m = 0$, and cash is used for every $m \in (0, m^*)$.
2. Cash is only adjusted when a free adjustment opportunity arrives.
3. Immediately after a cash adjustment, cash holdings m take the value m^*.

Note that the value of following a cash-burning policy with threshold m^* is the function $V : [0, m^*] \to \mathbb{R}_+$ that satisfies:

$$(r+p)V(m) = mR + pV(m^*) - V'(m)[\pi m + e] \quad \text{for all } m \in (0, m^*] \quad \text{and} \tag{8}$$

$$(r+p)V(0) = \gamma e + pV(m^*) \tag{9}$$

The first one is the standard Bellman equation, where we assume everything is paid in cash. The second equation says that at $m = 0$ agents use credit and wait for a free withdrawal opportunity.

The second policy type is defined as follows:

Definition 2. We define a *Baumol–Tobin policy* as a threshold $m^* \geq 0$ for which:

1. Credit is never used.
2. Cash is adjusted when either a free adjustment opportunity arrives or when $m = 0$.
3. Immediately after a cash adjustment, cash holdings m take the value m^*.

We briefly comment on the differences between the two polices. There is a sense in which cash burns in the agent's hands under both policies, since in both cases, as long as it is available ($m > 0$), cash is the preferred means of payment. Note that when a Baumol–Tobin policy is followed the o.d.e. in Eq. (8) holds in the range of inaction $(0, m^*]$. However, under this policy the boundary condition at $m = 0$ is given by:

$$V(0) = b + V(m^*). \tag{10}$$

[9] While here we treat R and r, p, π as independent parameters the value of R and $r + p + \pi$ can indeed be related –for instance $r + \pi$ should be the shadow nominal interest rates. We return to this relationship in the next section.

For a cash burning policy to be optimal, i.e. to solve the problem in Eq. (7), one needs to establish that it is optimal to pay with cash at $m \in (0, m^*]$ and with credit at $m = 0$, where the optimal withdrawal m^* must be determined. Finally, it has to be shown that at $m = 0$ it is optimal to wait for a free adjustment opportunity (instead of paying b to withdraw). Likewise, for a BT policy to be optimal, i.e. to solve the problem in Eq. (7), one needs to establish that it is never optimal to pay with credit and that at $m = 0$ it is optimal to pay b and choose the optimal withdrawal level m^* (see Appendix B for a formal proof that all these properties are verified under the optimal policy).

Note that the feasible policies consistent with Eq. (7) are much broader than the two candidate policies defined above. For instance, one could consider a policy in which credit is used for some time at $m = 0$ and a costly withdrawal occurs after T periods unless a free withdrawal arrives. Such a policy is the optimal one in the models of Sastry (1970) and Bar-Ilan (1990). Interestingly, as we explain below, the optimal policy for the problem in Eq. (7) will either take the form of a modified Baumol–Tobin one (where credit is not used) or of a cash-burning policy.

3.2. Characterizing the optimal cash-credit choice

The next proposition characterizes the optimal cash vs credit choice. Appendix B provides the proof as well as a detailed analytic characterization of the decision rules, including approximate solutions for m^* in the case of zero inflation.

Proposition 2. *A cash burning policy with m^* given by Eq. (12) is optimal provided that $b \geq \underline{b}$ where the lower bound for the fixed cost of adjustment is given by*

$$\underline{b} = \frac{e}{r+p}\left[\gamma - R\,\frac{m^*}{e}\right], \tag{11}$$

and m^*/e solves:

$$0 \leq \frac{m^*}{e} = \frac{\left(1 + (r+p+\pi)\frac{\gamma}{R}\right)^{\frac{\pi}{\pi+r+p}} - 1}{\pi} \leq \frac{\gamma}{R}. \tag{12}$$

Instead if $b \leq \underline{b}$ a Baumol–Tobin policy is optimal and m^* solves:

$$\left(1 + \frac{m^*}{e}\pi\right)^{1+(r+p)/\pi} = \frac{m^*}{e}(r+p+\pi) + 1 + (r+p)(r+p+\pi)\frac{b}{eR}. \tag{13}$$

The proposition shows that there is a threshold $\underline{b}$ for the fixed cost of adjustment b above which the cash-burning policy is optimal and below which the Baumol–Tobin policy is optimal. When $b > \underline{b}$ the optimal policy consists of using both cash credit. A withdrawal of size m^*, as determined by Eq. (12), occurs every time a free withdrawal opportunity arises. When cash eventually hits $m = 0$, then it is optimal to finance consumption using credit until a free opportunity for a cash withdrawal arises. Thus, under cash-burning $n = p$ and the fixed cost b is never incurred by the agent. Conversely, when $b < \underline{b}$ credit is not used and the optimal policy at $m = 0$ consists in paying the fixed cost b to make a withdrawal of size m^*, as determined by Eq. (13). Notice that the optimal threshold $\underline{b}$ defined in Eq. (11) summarizes the effect of all fundamental parameters (γ, R, π, r, p, e) into one single function that determines the nature of the optimal policy and, in particular, the optimality of using credit. Intuitively, Eq. (11) implies that the use of credit is optimal whenever the cost of using cash (which is increasing in R and b, and decreasing in p) is high relative to the cost of credit usage (which is increasing in γ).

Finally we notice that the optimal policy is of the bang-bang type, in the sense that using credit strictly dominates, or is dominated by, the use of costly withdrawals. This result differs from the ones of Sastry (1970) and Bar-Ilan (1990) where the use of some costly credit as well as some costly withdrawals is optimal. As mentioned in the introduction, this difference depends on the way the cost of credit is modeled: we assume that the cost of credit is proportional to the expenditure *flow*: γe. They assume the cost of credit is proportional to the *stock* of accumulated credit, the exact analogue of “negative cash”, which periodically requires the agent to pay the fixed cost b to rebalance the credit cost which would otherwise diverge. We see our assumption as a reasonable description of the case of revolving credit (credit that is automatically debited on the agent’s checking account at the end of the holding period).

The next proposition analyses how the threshold $\underline{b}$ changes as a function of the parameters:

Proposition 3. *The function $\underline{b} \geq 0$ is bounded above by $e\gamma/(r+p)$ and it is homogenous of degree one in (γ, R). Moreover*

$$\frac{\partial \underline{b}}{\partial \gamma} > 0 \quad \text{with} \quad \lim_{\gamma \to 0} \underline{b} = 0 \quad \text{and} \quad \lim_{\gamma \to \infty} \underline{b} = \infty, \tag{14}$$

$$\frac{\partial \underline{b}}{\partial R} < 0 \quad \text{with} \quad \lim_{R \to 0} \underline{b} = \frac{e\gamma}{(r+p)} \quad \text{and} \quad \lim_{R \to \infty} \underline{b} = 0, \tag{15}$$

$$\lim_{r+p \to 0} \underline{b} = \frac{eR}{\pi^2}\left[\left(1 + \frac{\gamma\pi}{R}\right)\log\left(1 + \frac{\gamma\pi}{R}\right) - \frac{\gamma\pi}{R}\right] = \frac{\gamma^2}{2R}e + \pi\, o\!\left(\frac{\gamma^2}{R}\right) \geq 0, \tag{16}$$

$$\lim_{\pi \to 0} \underline{b} = \frac{e\gamma}{r+p}\left[1 - \frac{\log\left(1+(r+p)\frac{\gamma}{R}\right)}{(r+p)\frac{\gamma}{R}}\right] = \frac{\gamma^2}{2R}e \ + \ (r+p)\, o\left(\frac{\gamma^2}{R}\right) \geq 0, \tag{17}$$

$$\frac{\partial \underline{b}}{\partial \pi} < 0, \ \lim_{\pi \to \infty} \underline{b} = 0, \ \ \frac{\partial \underline{b}}{\partial (r+p)} < 0, \ \text{ and } \ \lim_{r+p \to \infty} \underline{b} = 0. \tag{18}$$

Table 1
Selected moments on cash holding patterns: data and theory.

	Country:				Model fit to US:		
	Fra	Ger	Ita	US	Cash-only	Cash-credit	Mixed
Cash balances (*median*), M/c	8.7	4.9	12.5	3.1	3.1	3.1	3.1
Number of cash withdrawals, n	96	57	55	64	64	64	61
Withdrawal size (*median*), W/M	1.7	2.1	1.3	2.3	1.8	3.7	1.9
Cash at withdrawals (*median*), $\underline{M}/M$	-	0.3	0.4	0.7	0.1	1.0	0.6
Cash share of expenditures, $s = c/e$	0.15	0.53	0.52	0.23	1.0	0.4	0.7

The data source is Bagnall et al. (2016) Tables 1 to 4. Entries are sample means (unless otherwise indicated). The Italian data are from Alvarez and Lippi (2013) for households who posses a ATM card. Cash balances M/c are measured relative to cash expenditures per day computed as $c = s\frac{e}{365}$. The number of cash withdrawals is per year. The parameters for the mixed model are chosen to the match cash holdings and number of withdrawals. The rate at which b changes is $\lambda = 250$ times per year, with a 2/3 probability the cost of withdrawal is 0.2% if daily cash expenditures, otherwise it is 1%.

The proposition shows that the critical threshold $\underline{b}$ is increasing in the credit cost γ and decreasing in the opportunity cost of using cash R. In addition, by varying γ/R the threshold $\underline{b}$ ranges from zero to infinity. The approximations in the second to last line shows that if γ^2/R is small then $\underline{b}$ coincides with the one of the deterministic model. Moreover, the threshold $\underline{b}$ is decreasing in inflation: higher inflation increases the range of parameters for which the cash burning policy is optimal. Notice however that for finite values of γ and R, $\underline{b} > 0$ which implies that there exists a sufficiently small value of $b > 0$ that makes the use of credit dominated by a cash-only policy. Also notice that in our model a credit-only policy, i.e. one where cash is not used, does not occur as long as $p > 0$. The last two results suggest that the use of cash is very resilient: technical innovations that reduce the cost of credit are also likely to reduce the cost of cash withdrawals (increase p and or lower b) so that cash usage remains convenient for agents.

4. Model predictions about observable moments

This section studies several statistics of interest generated by a household who follows the optimal policy described above. Let e be the average expenditure per unit of time, and c the average expenditure per unit of time paid with cash, so $s \equiv c/e$ denotes the cash share, namely the long run average fraction of purchases (value weighted) paid with cash. Let M be the average cash holdings of the household, namely the mean value of real balances under the invariant distribution implied by the optimal decision rule. Let n the expected number of withdrawals per unit of time and W the expected size of withdrawals under the invariant distribution. Finally, let $\underline{M}$ be the expected value of cash at the time of a withdrawal. The left panel of Table 1 reports sample means for each of these moments for selected OECD economies, with data taken from Bagnall et al. (2016) and Alvarez and Lippi (2013) computed from household surveys and diaries.[10]

One main difference across country concerns the fraction of expenditures paid with cash which is small in the US and France (around 20% of total expenditure) and larger in Germany and Italy (where it is around 50% of the total expenditure), as shown in the last row of the table. Italy also records the highest values of cash holdings, equivalent to about 12 days of cash expenditures, while the US records the lowest value, equivalent to 3 days of cash expenditures (the table entries for M are expressed relative to the daily cash expenditure, computed as $s\,e/365$).

Next, we illustrate how those statistics map into the fundamental parameters of the dynamic model considering a household who follows the cash burning policy, which is optimal when $b > \underline{b}$, as well as the case when $b < \underline{b}$ where credit is not used. We stress that the main empirical appeal of the cash-burning policy is twofold. First, it is consistent with the data as it rationalizes the use of both cash and credit. Second, the policy aligns with the empirical observation that households are more likely to use credit when their cash balances are running low, a fact that is amply documented in Arango et al. (2011), Kosse and Jansen (2012), Huynh et al. (2014) and Arango et al. (2012). Our model also provides a simple interpretation of the evidence in Wang and Wolman (2016) which shows that the share of transactions paid with cash declines steadily, from

[10] Bagnall et al. (2016) analyse data from large-scale payment diary surveys conducted between 2009 and 2012 in Australia, Austria, Canada, France, Germany, the Netherlands and the United States that enable international comparisons. Alvarez and Lippi (2013) focus on Austria and Italy. In Italy we measure e as total household expenditure per day (including durables and services) using SHIW see Alvarez and Lippi (2013), and in France, Germany and US we measure e as the total expenditure of individuals per day using diary data see Bagnall et al. (2016).

the 1st day until the 15th day of the month, with the share of debit purchases being an almost exact mirror image of this pattern (see their Section 5). Wang and Wolman interpret this pattern as consistent with households having more cash at the beginning of the month, due to a pay-day effect. Our model implies such a behavior, since our central result is that the choice of means of payment depends on the level of cash balances.[11] The next proposition summarises the main results and closed form expressions for the observables focusing on the simple case of zero inflation $\pi = 0$.

Proposition 4. *Let $s \equiv c/e$ denote the cash share, i.e. the share of purchases paid with cash. For the cash-burning and the Baumol–Tobin policy we have:*

(i) Cash-burning policy. Let $r + p > 0$, $\pi = 0$, $R > 0$ and $b \geq \underline{b}$. Then

$$n = p, \ \ s = 1 - \left(1 + \frac{\gamma(r+p)}{R}\right)^{-\frac{p}{r+p}}, \ \ \frac{M}{e} = \frac{m^*}{e} - \frac{s}{p}, \ \ W = \frac{e}{p}\, s, \tag{19}$$

thus, using Eq. (12) with $\pi \downarrow 0$ gives $\frac{m^*}{e} = \frac{1}{r+p} \log\left(1 + \frac{\gamma(r+p)}{R}\right) \geq 0$. For $r \downarrow 0$ these expressions are simple functions of $\gamma p/R$:

$$s = \frac{\frac{\gamma p}{R}}{1 + \frac{\gamma p}{R}} \ \ , \ \ \frac{M}{e} = \frac{1}{p}\left[\log\left(1 + \frac{\gamma\ p}{R}\right) - 1 + \left(1 + \frac{\gamma\ p}{R}\right)^{-1}\right] \ \text{and} \tag{20}$$

$$\frac{W}{M} = \frac{1}{\left(1 + \frac{R}{\gamma\ p}\right)\log\left(1 + \frac{\gamma\ p}{R}\right) - 1} \ \ \text{and} \ \ \underline{M}/M = 1. \tag{21}$$

(ii) Baumol–Tobin policy. Let $R > 0$, $p > 0$, $\pi = r = 0$ and $b < \underline{b}$. Then $s = 1$ and

$$n = \frac{p}{1 - e^{-m^* \frac{p}{e}}} > p, \ \ \frac{W}{M} = \frac{m^*}{M} - \frac{p}{n} \in (0, 2), \tag{22}$$

$$\frac{M}{e} = \left[\frac{1}{1 - e^{-\frac{p}{e} m^*}} \frac{m^*}{e} - 1/p\right], \ \ \text{and} \ \ \underline{M}/M = \frac{p}{n}. \tag{23}$$

Proposition 4 gives closed form expression for n, s, M, W and $\underline{M}$ under either of the two optimal policies for low-inflation ($\pi \to 0$) and low-discounting ($r \to 0$), both reasonable assumptions for developed economies. The model is over identified since it has essentially two parameters, p and $\gamma p/R$, that determine 5 observables, as we discuss below. Next we briefly comment on the economics of the proposition.

Define the scalar $\theta \equiv \gamma p/R$ as the normalized cost of credit, namely the cost of credit (γ) relative to the cost of cash (R/p). Under cash-burning, the cash share is $s = \frac{\theta}{1+\theta}$ which is monotone increasing in the normalized cost of credit, with $s \to 1$ as $\theta \to \infty$ and with $s \to 0$ as $\theta \to 0$. The equation shows that for the credit share to increase, due to e.g. the new availability of cheaper credit cards, it is necessary that the cost of credit falls faster than the cost of cash. This simple observation may account for the remarkable resilience of cash usage in several developed countries (as documented by e.g. Bagnall et al. (2016)). If technical progress in payment instruments reduces the cost of credit as well as the cost of cash, such that θ is constant, then cash usage is unaffected.

The money demand M/e under the cash-burning policy is a function of two parameters: $\theta \equiv \gamma p/R$ and p.[12] The left panel of Fig. 1 plots the average money holdings M relative to daily total expenditures $e/365$ for $p = 50$ and $p = 100$. Recall that p also equals the number of cash withdrawals n in the cash-credit model. The right panel of the figure plots the average withdrawal size relative to the average money holding, W/M, under a cash-burning policy. The ratio W/M is decreasing in θ, ranging from $W/M \to 0$ as $\theta \to \infty$, to $W/M \to \infty$ as $\theta \to 0$, as shown in Fig. 1. For comparisons, when $b < \underline{b}$ and the modified Baumol-Tobin policy is optimal, we have that $W/M \leq 2$.

The proposition allows us to interpret the variation in outcomes, such as the cross-country data reported in the left panel of Table 1. Under a cash-credit policy different frequency of cash withdrawals will then be triggered by different efficiency of the cash withdrawal technology ($n = p$). For instance the large number of cash withdrawals in France and the US requires high values of p, which one can interpret as a proxy for the density of e.g. ATM machines.[13] Likewise, different shares of cash expenditures (s) will be ascribed to different values of the normalized cost of credit, $\theta \equiv \gamma p/R$, which combines the cost of credit (a technological parameter) with the nominal interest rate. For instance, the data suggests a relatively more efficient credit cost (smaller θ) in France than in Italy.

The right panel of Table 1 presents a calibration of the model for the United States, both for the Baumol-Tobin policy as well as for the cash-burning policy. The two parameters of the model, θ and p, are chosen to match the statistics in the

[11] The deterministic model of Section 2, where deterministic cash inflows are triggered by p cash withdrawals at equally spaced intervals, is the simplest framework to see how this effect works.

[12] For the cash credit model we normalize the cash holdings by the total expenditure e. Notice the difference from the normalization used in traditional cash inventory model where the scale variable is the total *cash* expenditure $c = s\,e$.

[13] In Alvarez and Lippi (2009) we found a strong correlation between estimates of p and measures of the geographic density of bank branches and ATMs in Italy (see Table 4).

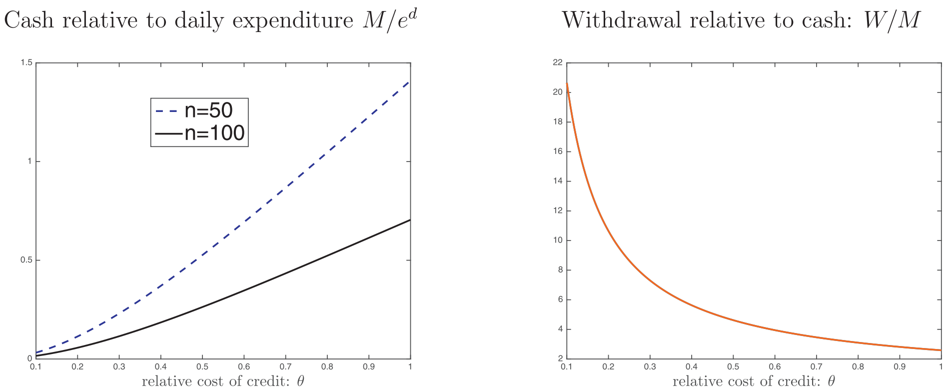

Fig. 1. Moments under cash-credit policy Note: Cash balances are measured relative to total expenditures per day, $e^d = e/365$.

first 2 rows of the data table, namely the cash holdings (normalized by cash expenditures, see footnote 12), M/c, and the number of cash withdrawals, n. The statistics in the remaining rows of the Table can thus be seen as a test of the model. The different models have mixed success depending on what statistics one focuses one. The cash-only model does a fairly good job at capturing the withdrawal size, which is 2.3 in the data and it is 1.8 in the cash-only model. However the cash-only model predicts a very low value of cash at the time of withdrawals (small precautionary motive) equal to 0.1 (versus 0.7 in the data), and by definition it is mute about the share of cash expenditures (i.e. $s = 1$ in Table 1 for the cash-only model). The cash-credit model predicts a share of cash expenditures of 0.4 (versus 0.2 in the data), and a high precautionary motive (1 versus 0.7 in the data). However the model prediction concerning the size of withdrawals is on the high side. Considering its parsimonious setup we find this results of interest. In Section 6 we discuss an extension that improves the model ability to fit the data.

4.1. Elasticities of observables with respect to θ

We conclude with a discussion of the model predictions concerning the interest rate elasticity of some key statistics which have been considered in the empirical literature, such as the interest elasticity of money demand which is essential to measure the welfare costs of inflation, see e.g. Lucas (2000). The elasticity of s with respect to $\theta \equiv \gamma p/R$ is $1/(1+\theta)$, which implies that the interest elasticity of s is given by $\frac{\partial \log s}{\partial \log \theta} = 1 - s$, or one minus the cash share, so that the elasticity is decreasing in θ, and smaller than one. We use the expression for the cash share to illustrate the tension between the use of cash vs. credit and the role of p. If $b < \underline{b}$ *only* cash is used, i.e. $s = 1$. If $b > \underline{b}$ both cash and credit are used. Yet, if p tends to zero, then $\underline{b}$ is finite, so there are values for which $b > \underline{b}$ and *only* credit is used, i.e. $s = 0$. Thus, we find that the interesting case is the one where $b > \underline{b}$ and $p > 0$, so that $0 < s < 1$.

The elasticity of Mp/e is: $0 \le \frac{\partial \log Mp/e}{\partial \log \theta} = \frac{\left(\frac{\theta}{1+\theta}\right)^2}{\log(1+\theta) - \theta/(1+\theta)} \le 2$ and is decreasing in θ. Thus money demand is decreasing in the opportunity cost R, with an interest rate elasticity *increasing* in R satisfying:

$$0 \le -\frac{\partial \log M/e}{\partial \log R} = \frac{\partial \log Mp/e}{\partial \log \theta} \le 2 \tag{24}$$

For comparison, the model with cash purchases only ($c = e$) summarized in part (ii) of Proposition 4, has an (absolute value) of the interest rate elasticity of money to *cash* consumption, M/c, that is increasing in the level R, but bounded above by 1/2. This difference reflects the elasticity of the cash share s, which is between 0 and 1, and the elasticity of the money demand relative to *cash* consumption M/c, also between 0 and 1:

$$0 \le -\frac{\partial \log M/c}{\partial \log R} = \frac{\gamma p/R}{1 + \gamma p/R} \, \frac{1}{\log(1 + \gamma p/R)} \le 1. \tag{25}$$

The higher interest elasticity of money demand produced by the model with a cash credit margin illustrates the importance of jointly modeling the cash-inventory problem and the cash-credit choice. By doing so the agent chooses both the extensive margin (how much cash vs credit to use) as well as the intensive margin (how many cash withdrawals to make). Failure to account for the interaction between these two margins, as done for instance in Alvarez and Lippi (2009) where the cash expenditure is taken as exogenous, leads to underestimating the interest rate elasticity of money demand.

Our simple model has the ability to qualitatively capture several empirical facts, such as the use of both cash and credit and the fact that credit is used when cash is low. Moreover the basic parametrization of the model, discussed above for the US, delivers magnitudes for the cash management statistics that are in the ballpark of actual data. Nonetheless our parsimonious 2 parameter model also has some clear shortcomings to match the data. For instance, notice that the average cash at the time of withdrawal $\underline{M}$ equals the mean cash holdings M under a cash-burning policy. This somewhat surprising and stark result is an immediate consequence of the fact that withdrawal times under the cash-burning policy are uncorrelated with money holdings (their arrival rate p is exogenous and independent of m). Since a withdrawal is equally likely to happen with any money balance $m \in (0, m^*)$, then the two statistics are the same. This prediction of the model is in contrast with the data where $\underline{M} < M$, yet another issue that motivates the extension of Section 6.

5. An application: how costly is it to ban cash usage?

In this section we use our model to quantify the household's cost of a policy that limits cash usage. The motivation for such a policy is that, because of its anonymous nature, cash is heavily used for illegal activities. Rogoff (2016) argues that phasing out paper money would help fighting some big problems as corruption, tax evasion, drug trade and others. Because of this argument some countries, like Sweden, have been gradually pursuing the objective of a cashless economy.

In this section we use our model to study the cost forcing all agents to be cashless, i.e. to make all payments using credit only. We use the agent's value function for a cash credit policy to quantify the welfare cost of moving from the optimal policy, where the agent chooses a certain cash-credit mix $s^* \in (0, 1)$, to a policy where cash is phased out and the agent is forced to finance all consumption using credit, i.e. using $s = 0$. Our objective is to quantify the welfare cost for agents who are forced to move from their optimal policy s^* to the mandated one where the cash share is zero ($s = 0$). For the case of zero (i.e. small) inflation and low discount rates this welfare cost has an accurate, and extremely simple, analytic approximation which depends *only* on the optimal cash share s^* and on the cost of credit γ.

Let us use the flow value function $rV(m)$ to measure the agent's minimized flow cost to finance a total annual expenditure (e), using both cash and credit. Simple analysis of the closed form expression for $V(m)$ shows that for $\pi = 0$ and $r \to 0$ gives[14]

$$v^* \equiv \lim_{\pi=0, r\to 0} rV(m) = e\frac{R}{p} \log\left(1 + \frac{\gamma p}{R}\right). \tag{26}$$

Next, let v_0 be the flow cost of financing the expenditure stream e using credit only, so that $s = 0$. It is straightforward to see that $v_0 = \gamma e$. Thus, we define the cost of implementing Rogoff's ban on cash as $\ell = (v_0 - v^*)/e$, where the normalization by e allows us to read ℓ as a fraction of yearly consumption. Using Proposition 4 to write the normalized cost of credit $\theta \equiv \gamma p/R$ as a function of the cash share $s = \theta/(1+\theta)$, we obtain

$$\ell(s, \gamma) \equiv \frac{v^* - v_0}{e} = \gamma\left(1 + \frac{1-s}{s} \log(1-s)\right). \tag{27}$$

Recall that this is a flow cost expressed as a fraction of the per-period (e.g. per year) consumption. The cost of implementing a zero-cash policy depends *only* on two parameters: it depends linearly on the cost of credit γ, and it is a convex increasing function of the share of cash purchases, s. Simple analysis of $\ell(s, \gamma)$ shows that the cost is zero at $s = 0$ and it is monotone increasing for $s > 0$. As credit becomes more convenient than cash (lower θ), the cash share s falls and the cost of moving to a no-cash system ($s = 0$) decreases. Notice also that the cost converges to γ as $s \to 1$, which is easy to understand since in an economy in which all purchases are made with cash the cost of switching to credit is given by γe.

The left panel of Fig. 2 plots the cost of implementing a zero cash policy relative to the cost of credit, namely ℓ/γ. The dashed line is the approximation given in Eq. (27), which is only a function of the cash share s, the solid line is the exact cost computed assuming inflation equal to 2% and a time discount $r = 0.02$. The convexity of the ℓ function implies that at intermediate values of s the cost (per year) of the zero-cash restriction amounts to a small fraction of the cost of credit use. For a country like Germany, where $s \approx 0.5$, the cost is approximately 30% of the yearly credit cost, γe. For the US, where $s \approx 0.25$, the value is around 15% of the yearly credit cost.

Next we quantify the cost of credit, γ, using the relation between s and $\gamma p/R$ from Proposition 4 and that $p = n$. Assuming a nominal interest rate of 2% (the opportunity cost of cash R), and $n \approx 50$ per year as measured in the US and Germany, the model gives that $\gamma = \frac{sR}{(1-s)n}$. Using this equation shows the cost of credit γ to be a tiny number, about 4 basis points for $s = 0.5$ and 1.3 basis points for $s = 0.25$. The right panel of Fig. 2 plots the implied cost per year of imposing the zero cash restriction assuming $n = 50, R = 0.02$ and computing the cost of credit γ implied by each level of s. As suggested by the estimates discussed before, the cost of implementing Rogoff's restriction appears tiny over the range of cash share values observed in the data: for a household with an annual consumption of 40K the cost is approximately 10 dollars per year in Germany, where the cash share is around 50%, it is about 2 dollars in the US where the cash share is around 25%.

[14] See Appendix B for a closed form solution for this value function.

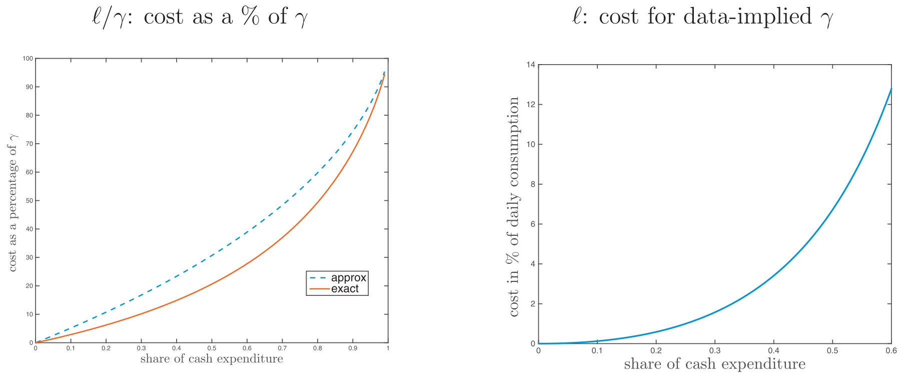

Fig. 2. Flow cost of imposing the zero-cash restriction. The "exact" cost function in the left panel assumes inflation equal to 2% and a time discount $r = 0.02$. The right panel uses, for each value of s on the x-axis, the corresponding value of γ implied by Proposition 4 under the assumption that $n = 50$ and $R = 0.02$.

Two remarks are useful to put the above figures in perspective. First, the magnitude of the cost scales proportionally with the value of γ as Eq. (27) shows. It is thus useful to note that the small estimates for γ discussed above, in the order of a few basis points, are likely affected by our specific modeling assumptions, in particular the assumption that all cash withdrawals are free under the cash-burning policy. Modifying this assumption, as done in the extension of Section 6, will likely increase the cost of credit. For arbitrary values of the cost of credit the left panel of Fig. 2 can be used to gauge the cost of the zero-cash policy at any given level of the cash share. Second, our estimated costs of phasing out cash is borne by households who already possess the credit technology. A more encompassing measure of the social cost would also include the costs borne by the currently unbanked (about 7% of households in the US, see FDIC, 2015).

6. An extension: random variation in fixed cost

The model of Section 3 is very stark in that credit is used at $m = 0$ provided that the fixed transaction costs is sufficiently high ($b > \underline{b}$) in which case we have that $\underline{M} = M$ and $n = p$. Instead, if the transaction cost is sufficiently low ($b < \underline{b}$) credit is never used, i.e. $s = 1$ and the model becomes the Baumol–Tobin with random free withdrawal opportunities discussed in Alvarez and Lippi (2009). The prediction that the cash share is either zero or 1 seem too stark against the data. In particular, there is substantive evidence, based on micro data, that the amount of cash at the time of withdrawals is smaller than the average cash balance, i.e. $\underline{M} < M$. We show in this section that allowing the fixed cost b to be random and persistent allows to account for this fact while retaining the other features of the model. The variation in b implies that agent follows a cash-burning policy when she faces a high cash withdrawal cost, while she follows a Baumol–Tobin policy when the cost of cash withdrawals is below a critical threshold. The analysis can equivalently be conducted assuming the cost of credit to be random.

In particular we assume that there is a Poisson process with constant intensity λ whose occurrence indicates that a new value of b has been drawn. This Poisson process is independent of the one for the arrival of the free adjustment opportunities. The new values for the cost $\tilde{b}$ are drawn from the cumulative distribution function $F : \mathbb{R} \to [0, 1]$. Conditional on a change in the value of the fixed cost, the new value $\tilde{b}$ is assumed to be independent of the current value b. In this case the value function has two arguments, (m, b), so we write it $V : \mathbb{R}^2_+ \to \mathbb{R}_+$. We denote by primes the derivative of V with respect to m. The value function solves the following functional equation:

$$0 = \min\left\{ \min_{0 \le c \le e} Rm + \gamma[e - c] + p \min_{z \ge 0} [V(z, b) - V(m, b)] + \lambda\left[\int V\left(m, \tilde{b}\right) dF\left(\tilde{b}\right) - V(m, b)\right] \right.$$
$$\left. - V'(m, b)(c + \pi m) - rV(m, b), \quad b + \min_{z \ge 0} V(z, b) - V(m, b) \right\} \ \forall \ (m, b) \in \mathbb{R}^2_+ \tag{28}$$

The interpretation of the terms in this functional equation is analogue to the one in Eq. (7): the outer min operator compares the value of using credit with the value of paying the fixed cost and replenishing cash balances. There are two differences. First, as mentioned above, b is also part of the state. Second, in the first term there is an extra expression given by the contribution to the expected change of the value function due to the change in the cost from b to a value drawn from the distribution with c.d.f. F.

In what follows we proceed, based on the analysis of the special cases analysed in the previous section, by constructing a solution for a type of cash-burning policy which combines the two cases analysed above. We stress that the nature of the optimal solution remains bang-bang, so that the optimal policy is a mixture of the cash-burning type and of the Baumol-Tobin type. In other words, this is the best policy there is (no restrictions are imposed on the form of the policy rule).

Definition 3. A threshold cash-burning policy is defined by a cost threshold $\underline{b}$ and a cash-target function m^*. For all $m > 0$ and all $b \ge 0$ the agent uses only cash. If $m = 0$ the agent withdraws cash when $0 \le b \le \underline{b}$, and uses credit when $b > \underline{b}$. For all

(m, b) cash balances are set equal to $m^*(b)$ every time that a free adjustment opportunity arrives. Additionally, cash balances are set to $m^*(b)$ if $m = 0$ and $b \le \underline{b}$.

Hence $\underline{b}$ is the critical threshold so that at $m = 0$ the agent uses credit if $b < \underline{b}$ and uses cash otherwise. Appendix C gives a detailed characterization of the value function for this problem under a threshold cash-burning policy. We use this extended model in a calibration that illustrates how it can produce cash management behavior featuring both cash and credit usage and where the amount of cash at the time of a withdrawal is smaller than the average cash holdings, i.e. $M < \underline{M}$. As mentioned, the latter feature is seen in the data but is not produced by the model with a constant cost of withdrawal studied in Section 3.

6.1. A quantitative assessment

As a benchmark, we use the observable statistics for US households which were summarized in Table 1. The data shows that share of cash expenditure is about 20% of total (non-durable) expenditures and that the average currency holdings is about 3.1 days of expenditures. Moreover, it shows that the amount of cash at the time of withdrawal relative to the average money balances $\underline{M}/M$ is around 0.7, that the ratio of the average withdrawal to the average money balances is about W/M is 2.3 and that the households with ATM withdraw cash about 60 times per year.

The last column of Table 1 presents a calibration of the model with a random fixed cost b. This model combines features of the two polar models described in the previous 2 columns, namely the cash-only model described in part (ii) of Proposition 4 and the cash-credit model described in part (i) of Proposition 4. We parametrize the model to the match cash holdings and number of withdrawals. The rate at which b changes is $\lambda = 250$ times per year, with a 2/3 probability the cost of withdrawal is 0.2% if daily cash expenditures, otherwise it is 1%. When b is low the agent finds it optimal not to use credit when cash is exhausted (since $b < \underline{b}$). When b is high the agent uses credit at $m = 0$ (since $b > \underline{b}$) waiting for a free withdrawal opportunity or a change in b. We assume that the rate at which the withdrawal cost changes is high so that we can interpret is as on average changing every working day.[15] Intuitively, the behavior produced by this model is close to a weighted average of the behavior of the cash-only and the cash-credit model characterized in the previous two columns.[16] The model's predictions improve upon the polar cases since the model is able to account for a cash share below 100% and for a smaller level of cash at the time of withdrawals, $\underline{M}/M < 1$.

7. Conclusions

A model was presented that combines the ingredients of the dynamic cash inventory problem with the ingredients of the cash-credit choice. The key novelty compared to the previous literature is that we allow agents to use either cash or credit at each moment. This realistic assumption implies an optimal rule for credit usage by the agent which turns out to depend on the amount of cash at hand. We find this feature interesting because it makes contact with recent evidence showing that the likelihood of using cash increases with the level of cash holdings, as documented in e.g. Arango et al. (2012, 2011) and Huynh et al. (2014) using diary data for Canada and Austria. We showed that, in spite of its simplicity, the model predictions' on credit usage, the size of cash withdrawals and the average cash holdings are aligned with the magnitudes observed in the data. We used our model to quantify the cost of phasing out cash, a policy endorsed by Rogoff (2016) to fight several cash-intensive illegal activities. We estimate that the households' cost of moving from their optimal cash-credit share to one where the cash share is zero is a small fraction of the daily consumption, for a household with a 40K the cost is approximately 10 dollars per year in Germany, where the cash share is around 50%, it is about 2 dollars in the US where the cash share is around 25%. This estimated costs assume agents already have access to both cash and credit. A more encompassing estimate of the total costs of phasing out cash must include an estimate of the cost of banking the unbanked households.

Our model abstracts from aspects of the cash credit choice that have been emphasized before: the size of purchases (e.g. Whitesell, 1989) and the acceptability of credit at the points of sale (as in e.g. Huynh et al., 2014). Future models might benefit by unifying those aspects into a single model and quantify the relative importance of each of these frictions by using the relevant micro data.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jmoneco.2017.07.001

[15] The other structural parameters are taken from, or are close to, the structural estimates in Alvarez and Lippi (2009): the number of free withdrawal per year is $p = 35$, the opportunity cost of cash is $R = 4\%$ (this includes the nominal interest rate and the probability of cash theft, as discussed in Appendix D).

[16] The option value motives which might cause the outcomes to differ from a weighted average of the two polar models are small when the value of λ is high.

References

Alvarez, F.E., Lippi, F., 2009. Financial innovation and the transactions demand for cash. Econometrica 77 (2), 363–402.
Alvarez, F.E., Lippi, F., 2013. The demand of liquid assets with uncertain lumpy expenditures. J. Monet. Econ. 60, 753–770.
Arango, C., Bouhdaoui, Y., Bounie, D., 2012. Modeling the Share of Cash Payments in the Economy. Manuscript. Bank of Canada.
Arango, C., Huynh, K., Sabetti, L., 2011. How Do You Pay? The Role of Incentives at the Point-of-Sale. Working Papers. Bank of Canada.
Arango, C.A., Hogg, D., Lee, A., 2015. Why is cash (still) so entrenched? Insights from canadian shopping diaries. Contemp. Econ. Policy 33 (1), 141–158.
Bagnall, J., Bounie, D., Huynh, K.P., Kosse, A., Schmidt, T., Schuh, S., et al., 2016. Consumer cash usage: a cross-country comparison with payment diary survey data. Int. J. Cent. Bank. 12 (4), 1–61.
Bar-Ilan, A., 1990. Overdrafts and the demand for money. Am. Econ. Rev. 1201–1216.
Bouhdaoui, Y., Bounie, D., 2012. Modeling the share of cash payments in the economy: an application to France. Int. J. Cent. Bank. 8 (4), 175–195.
FDIC, 2015. FDIC National Survey of Unbanked and Underbanked Households. Executive summary. Federal Deposit Insurance Corporation.
Freeman, S., Kydland, F.E., 2000. Monetary aggregates and output. Am. Econ. Rev. 90 (5), pp.1125–1135.
Huynh, K., Schmidt-Dengler, P., Stix, H., 2014. Whenever and Wherever: The Role of Card Acceptance in the Transaction Demand for Money. Technical Report. Free University of Berlin, Humboldt University of Berlin, University of Bonn, University of Mannheim, University of Munich.
Klee, E., 2008. How people pay: evidence from grocery store data. J. Monet. Econ. 55 (3), 526–541.
Kosse, A., Jansen, D.-J., 2012. Choosing How to Pay: The Influence of Home Country habits. DNB Working Papers. Netherlands Central Bank, Research Department.
Lucas, R.E.J., Stokey, N.L., 1987. Money and interest in a cash-in-advance economy. Econometrica 55 (3), 491–513.
Lucas, R.E.J., 2000. Inflation and welfare. Econometrica 68 (2), 247–274.
Mooslechner, P., Stix, H., Wagner, K., 2006. How are payments made in Austria? Results of a survey on the structure of Austrian households use of payment means in the context of monetary policy analysis. Monet. Pol. Econ. (Q2) 111–134.
Prescott, E.C., 1987. A multiple means-of-payment model. In: New Approaches to Monetary Economics, pp. 42–51.
Rogoff, K.S., 2016. The Curse of Cash. Economics Books. Princeton University Press.
Sargent, T.J., 1987. Dynamic Macroeconomic Theory. Harvard University Press.
Sastry, A.S.R., 1970. The effect of credit on transactions demand for cash. J. Finance 25 (4), 777–781.
Stix, H., 2004. How do debit cards affect cash demand? Survey data evidence. Empirica 31, 93–115.
Wang, Z., Wolman, A.L., 2016. Payment choice and currency use: insights from two billion retail transactions. J. Monet. Econ. 84, 94–115.
Whitesell, W.C., 1989. The demand for currency versus debitable accounts: note. J. Money Credit Bank. 21 (2), 246–251.