Urgency, Liquidity, and the Price of Convenience*

Xinhao Dong
Simon Fraser University
xinhao_dong@sfu.ca

Lucas Herrenbrueck
Simon Fraser University
herrenbrueck@sfu.ca

Zijian Wang
Wilfrid Laurier University
zijianwang@wlu.ca

October 2026

Abstract

The conventional approach to measuring a convenience yield or liquidity premium is to take two assets that differ in convenience or liquidity but are otherwise identical, and to proxy the price of convenience (PoC) with their spread. We show that this approach is insufficient. First, short-term assets are inherently liquid for non-urgent needs, therefore a spread between them reflects *urgency* of needs rather than *liquidity* of assets. Second, any spread reflects the underlying assets' loadings on convenience rather than convenience itself. To solve these problems, we develop a monetary asset pricing model and estimate it with three instruments that differ in maturity and in marketability: Treasury bills, commercial paper, and AAA corporate bonds. At our headline estimate, the PoC averages 5.1% a year and varies between 3.2% and 11.4% – an order of magnitude above conventional measures – and a private claim delivers on average about half the liquidity services of a Treasury bill. On average, only 24% of the PoC is attributable to urgency.

JEL Classification: E43, E44, G12

Keywords: convenience yield, liquidity premium, term premium, marketability, Fisher rate, Bayesian inversion, quasi-Bayesian estimation

*We acknowledge support from the Social Sciences and Humanities Research Council of Canada (Insight Grant 435-2023-1089).

1 Introduction

How much is convenience worth? A large literature in macroeconomics and finance seeks to answer this question by measuring "convenience yields" or "liquidity premia". In principle, this would be done by measuring the spread between a perfectly convenient/liquid claim and a perfectly inconvenient/illiquid claim. In practice, however, the literature proxies this with the observed yield spread between two *specific* assets that defensibly differ in convenience or liquidity and nothing else.[1]

We argue that this approach is insufficient, for two reasons. The first is an identification problem: a spread does not price convenience, it prices a *difference in the amount* of convenience multiplied by the common *price* of convenience. In the language of our model, if a fraction μ^A of asset A is **marketable** (saleable or acceptable as collateral, accounting for trading frictions and costs) when money is needed, the yield on that asset falls below its illiquid benchmark by $\mu^A \ell$, where ℓ is the aggregate marginal value of liquidity. The spread between two otherwise identical assets A and B is then:

$$r_t^A - r_t^B \;=\; \left(\mu_t^B - \mu_t^A\right) \ell_t, \tag{1}$$

and its time variation may be dominated by variation in the loadings μ_t – how much convenience each asset supplies – rather than by the price ℓ_t that the spread is meant to capture.

The second reason is fundamental: *a short-term asset is liquid by its very nature*, irrespective of market conditions. If money is needed in three months, a bond maturing in two months is as good as money. A spread between two short-term assets, precisely the comparison underlying the most widely used convenience yield measures, can therefore only price the attribute on which the two assets differ: the ability to meet needs that cannot wait.

We refer to this attribute as **urgency**. A monetary need is urgent if it cannot be deferred even for the few months it would take for a short-term asset to mature. Urgency is a discrete

[1] Prominent examples include the spread between AAA corporate bonds and long-term Treasuries (Krishnamurthy and Vissing-Jorgensen, 2012), between GC repo and T-bills (Nagel, 2016), between on-the-run and off-the-run Treasuries (Vayanos and Weill, 2008), between T-notes and TIPS (Andreasen, Christensen, and Riddell, 2021), and between Treasuries of different maturities (Juelsrud, Nenov, Schneider, and Syrstad, 2025).

penalty for lacking the means of payment at a specific moment; this distinguishes it from time preference, which is a smooth discount on deferred consumption whose cost scales with the length of the delay. What matters for liquidity needs is whether the payment is available when the need arrives, not the length of time by which it is missed.[2]

To capture these ideas, we develop a monetary asset pricing model in which safe assets differ in time to maturity and in market liquidity. It builds on the New Monetarist framework of Lagos and Wright (2005) in the tradition of indirect asset liquidity (Geromichalos and Herrenbrueck, 2016; Herrenbrueck and Geromichalos, 2017; Lagos and Zhang, 2019). A long-term asset can meet a monetary need only indirectly, through sale or use as collateral, and only to the extent that *markets are liquid*. A short asset can do that too, but may also simply mature in time, provided the *need is not urgent*. Two aggregate shocks price these attributes: the marketability of illiquid assets, μ_t, and the fraction of liquidity needs that are urgent, ω_t.[3]

We estimate the model on U.S. rates from 1971 to 2025 using three instruments that differ along exactly the two dimensions just mentioned. Treasury bills mature within the period and are fully marketable; commercial paper matures within the period but is only partly marketable; AAA corporate bonds are only partly marketable and do not mature until much later. The assumption that the second and third instrument share a marketability parameter μ_t is the identifying restriction at the centre of the paper. We defend it in Section 6, noting here only that we find it far more plausible a priori than the implicit assumption in the literature: that private assets cannot be sold or used as collateral at all.

Three quantitative findings stand out. First, the aggregate price of convenience $\bar{\ell}$ has a mean of 5.1% a year (median: 4.9%) and varies between 3.2% and 11.4%, an order of magnitude above the commercial paper spread (mean 59bps, median 44bps). Second, a private

[2] We borrow this concept from a literature far afield from ours: studying real-time congestion pricing, Bento, Roth, and Waxman (2024) find that willingness to pay is dominated by a value of urgency, a discrete penalty for late arrival that does not scale with the amount of time saved, rather than by the marginal value of travel time.

[3] This structure encapsulates the four features that make an asset "money-like": serving as a medium of exchange (obviously); serving as collateral for a loan when money is needed (Ferraris and Watanabe, 2008; Venkateswaran and Wright, 2014); the ability to sell it on a secondary market when money is needed (Geromichalos and Herrenbrueck, 2016); and the expectation that it turns into money soon, *before* that money is needed (Geromichalos, Herrenbrueck, and Salyer, 2016). The second and third depend on counterparties and can be impaired by market conditions (Benigno and Robatto, 2019; Cheng and Robatto, 2024) – the fourth does not.

claim provides about half the liquidity services of a Treasury bill: μ_t has a mean of 0.51 (median: 0.49). Third, only a minority (mean 24%, median 18%) of $\bar{\ell}$ is attributable to its urgent component. That share moves: its decade medians fall from 37% in the 1980s to 8% in the 2010s and turn back up in the 2020s.

Our empirical model has eight latent variables and five moments. To resolve the resulting indeterminacy, our baseline estimator places priors on the three least-identified dimensions, estimates one in a first stage, and solves the remaining four in closed form, quarter by quarter; this is an exactly identified inversion of the type used in the term-structure literature (Chen and Scott, 1993; Joslin, Singleton, and Zhu, 2011). This estimator produces no residuals and therefore no likelihood with which to compare alternative specifications. We therefore use a second, quasi-Bayesian estimator (Chernozhukov and Hong, 2003) for tests and model comparisons: in this estimator all latent variables have priors and the pricing equations are allowed to hold with error.

We use this second estimator for three purposes. First, we challenge our identifying assumption of equal marketability of private assets and reject the alternative assumption that private assets have no convenience at all. Second, we justify the baseline's approach to accounting for expected holding returns on long bonds when we only observe yields. Third, we examine the dynamics of the latent variables. We find that the latent variables are persistent, and that persistence matters most for the expected capital gain on long bonds. When each latent variable is allowed to jump as well as to drift, the estimates retain a jump only for urgency: urgency remains stable for several years and then changes abruptly, whereas marketability changes gradually.

Our paper relates to several literatures. The macrofinance literature on convenience yields treats the premium on Treasuries as a demand-side object to be measured from spreads (Krishnamurthy and Vissing-Jorgensen, 2012; Nagel, 2016; Del Negro, Giannone, Giannoni, and Tambalotti, 2017; Krishnamurthy and Li, 2023; Caramp and Singh, 2023; Cieslak, Li, and Pflueger, 2024); we share the object of interest but argue that spreads between short-term assets identify only the urgent component of convenience. Two results in that literature bear

directly on our approach. Liquidity premia depend on maturity: short-term Treasury bills carry a money-like premium over longer Treasuries (Greenwood, Hanson, and Stein, 2015), and the Treasury liquidity premium has a term structure whose slope turns with the business cycle (Joslin, Li, and Song, 2019). And Longstaff (2004) isolates a flight-to-liquidity premium from the Refcorp–Treasury spread, two claims alike in default risk and maturity, the cleanest existing demonstration that a long-maturity spread can price liquidity rather than credit.

The New Monetarist literature provides the microfoundations for why assets carry liquidity premia at all (Lagos and Wright, 2005; Lagos, 2010; Lagos, Rocheteau, and Wright, 2017; Lagos and Zhang, 2020). In that tradition, our model's closest predecessors (Geromichalos and Herrenbrueck, 2022; Herrenbrueck and Wang, 2026) develop the distinction between observed interest rates on tradable assets and the "Fisher rate" on a hypothetical illiquid bond, which we exploit here as the fundamental benchmark against which all convenience is priced. Whereas those earlier papers focused mainly on the implementation and effects of monetary policy (interest rates and inflation), the goal of our present paper is estimation and recovery of latent shocks from observed asset prices.

In our model, a long bond's spread over bills is $(1-\mu_t)\,\bar{\ell}_t$, the fraction of the bond that cannot be placed when a need arises times the price of convenience. On this reading, what the term-structure literature measures as a term premium is literally a liquidity premium. In this, we follow Geromichalos et al. (2016) and Caramp, Geromichalos, Lee, and Wang (2025), who separate direct from indirect liquidity as we do but discipline their models with a calibration rather than period-by-period estimation. This is in obvious tension with the affine term-structure literature (Duffee, 2002; Joslin et al., 2011; Adrian, Crump, and Moench, 2013), whose approach is to extract a term premium as the empirical residual between observed long yields and an average of model-implied expected short rates, in Gaussian no-arbitrage models estimated on the yield curve alone. Two results from that literature support our reinterpretation of the empirical residual as reflecting the inherent illiquidity of long-term assets. First, yields contain substantially less information about future excess returns than the fitted premia imply (Duffee, 2011), and the return predictability that motivates the object

is fragile in real time (Cochrane and Piazzesi, 2005; Bauer and Hamilton, 2018). A residual that is hard to forecast and hard to attribute to a risk mechanism is exactly what one would expect if it were compensation for something other than risk.

The paper is organized as follows. Section 2 presents the model and derives the asset pricing equations. Section 3 maps the model into an estimable system and defines the price of convenience. Section 4 reports our baseline Bayesian inversion estimates. Section 5 compares four estimators and conducts additional tests. Section 6 discusses identification and interpretation. Section 7 concludes. Model, data, and estimation details are collected in the appendix.

2 A model of urgency, liquidity, and asset prices

Our model is adapted from the Liquidity-Augmented Model of Geromichalos and Herrenbrueck (2022), a framework that integrates New Monetarism with neoclassical asset pricing; our innovations here are the introduction of an aggregate *urgency* shock ω_t, the distinction between three bonds differing in urgency and marketability, and the fact that we carry the stochastic Euler equations to the data rather than analyzing the theoretical effects of monetary policy.

2.1 Environment

The economy consists of a unit measure of households and a consolidated government. Each household has two members: a worker and a shopper, who make decisions jointly to maximize the household's utility. All households are anonymous, therefore they cannot make long-term promises and trade must be quid-pro-quo.

Time $t = 0, 1, \ldots$ is discrete and runs forever. Each period is divided into four subperiods: an asset market (AM), a production market (PM) further subdivided into an "early" and a "late" stage, and a centralized market (CM). There are four assets, all of them default-free. The first is fiat money, in supply M_t. The second and third are one-period nominal discount bonds, in supplies B_t^T and B_t^S, both of which mature after the early PM closes

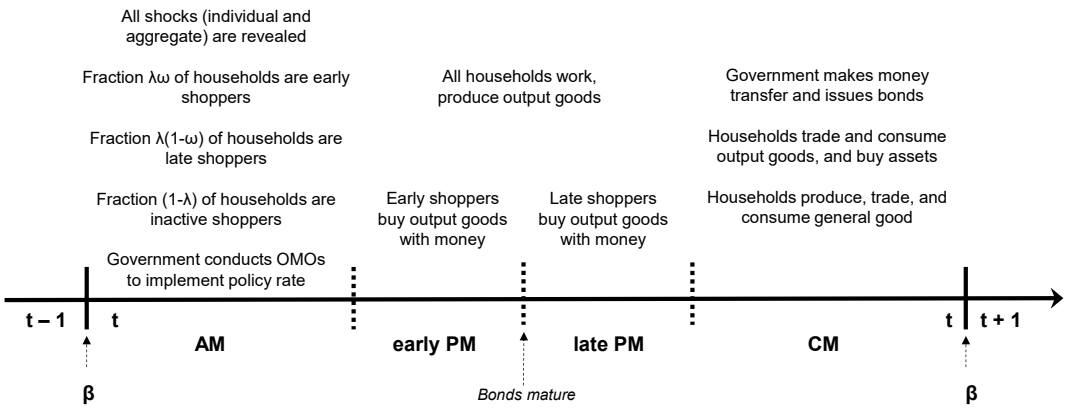

Figure 1: Timing of events.

but before the late PM opens; we call the first of these the *T-bill* and label it T, and label the second S. The fourth is a consol, in supply B_t^L and labelled L, which never matures and pays a nominal dividend d_t per unit in each CM. Money and T-bills are issued by the government, and changes in their stocks can be implemented in one of two ways: either via open-market operations in the AM, or via lump-sum transfers in the CM. The short bond S and the consol L are issued outside the household sector in exogenous supply. All redemptions, dividends and net issuance are settled in money in the CM and financed by the lump-sum transfer $\mathcal{T}_t$. This environment is illustrated in Figure 1.

At the very beginning of a period, all shocks are revealed. A randomly selected fraction $\lambda_t\omega_t$ of shoppers learn that they will enter the "early" PM where they can buy goods from workers; a fraction $\lambda_t(1-\omega_t)$ learn that they will enter the "late" PM; the remaining shoppers $(1-\lambda_t)$ are "inactive" and will not participate in the PM in that period. Thus, λ_t indexes the market-wide incidence of monetary needs, and ω_t indexes their *urgency*: the two short bonds mature after the early PM closes but before the late PM opens, so a fraction ω_t of monetary needs arrive too soon to be served by maturing bonds. This information induces differing liquidity needs among the shoppers, motivating trade in the AM. In the equilibrium we are interested in, early (urgent) shoppers will seek to hold money and sell bonds and consols, which are of no direct use to them; late shoppers will prefer to buy the short bonds, since

these mature and become money in time to be spent on their consumption opportunities; and inactive shoppers will seek to earn a return on their idle money by spending it on all three interest-bearing assets. The government can also trade money and T-bills in the AM, and we interpret the market price of T-bills in the AM as the main policy instrument.

Next, the economy proceeds to the PM stage. Each worker produces C_t units of consumption goods (which is exogenous) at zero cost, and workers can choose freely whether to sell these in the "early" or "late" PM. Due to anonymity, credit is not feasible, so shoppers must pay for the goods with a suitable medium of exchange.[4] However, since the short bonds mature partway through the PM, shoppers whose consumption opportunities arrive late will treat them as money within that period. Seen through the lens of a single asset, the parameter $1-\omega$ measures how money-like that asset is; seen through the lens of the market as a whole, ω_t is a state variable measuring how urgent the economy's liquidity needs are. The latter interpretation is the one we pursue: maturities of money market instruments are stable over time, while the urgency of the needs they serve is not.

The final sub-period, the CM, is where consumption takes place. The only consumption good is the produced good, so market clearing gives $c_t = C_t$, which we treat as exogenous for the purposes of this paper; the consol's dividend is nominal, and so transfers money rather than goods and does not enter the resource constraint. In the CM, all households can trade consumption goods among themselves, as well as all assets – hence, the CM is the 'primary' asset market, while the AM is the 'secondary' one.

In summary, the four assets differ in their liquidity as follows. Money is the most liquid: it can be traded for any other commodity, in any market. The T-bill is next: it matures partway through the PM, so the fraction $1-\omega_t$ of active shoppers whose needs are not urgent will treat it as money, and it is fully marketable in the AM, $\mu_t^T = 1$. The short bond S matures on the same timeline, and so is equally good for a non-urgent need, but only a fraction $\mu_t \leq 1$ of it can be traded in the AM (or used as collateral for a monetary loan, in which case $1-\mu_t$ is the haircut). The consol is the least liquid: only the same fraction μ_t of it can be traded,

[4] We assume that money is the only asset that can fulfill this role – an assumption here, but see Geromichalos and Herrenbrueck (2022) for a discussion of the microfoundations behind it.

and it never matures. Two of the model's three marketability parameters are thus common by construction: $\mu_t^S = \mu_t^L \equiv \mu_t$, against $\mu_t^T = 1$ for the T-bill.

Finally, there exists a "general" consumption good, $g \in \mathbb{R}$, which households can produce, consume, and trade during the CM (negative values mean production and positive ones mean consumption). This good is in zero aggregate supply; its only function in the model is to induce linear preferences for an individual household, collapse the portfolio problem into something tractable, and create a representative household. Households discount the future at rate $\beta < 1$ and have the per-period utility function

$$U_t(c_t, g_t) = u(c_t) + g_t, \tag{2}$$

where u is twice continuously differentiable with $u' > 0$ and $u'' < 0$.

2.2 The household's problem

We start with the household's problem in the CM. Consider a household with portfolio $(m_t, b_t^T, b_t^S, b_t^L, y_t)$ of money, T-bills, short bonds, consols, and output goods. The household chooses its consumption (c_t and g_t), as well as the asset portfolio $(m_{t+1}, b_{t+1}^T, b_{t+1}^S, b_{t+1}^L)$ to be carried into the next period. The prices of general goods (p_t^G), output goods (p_t), and the three interest-bearing assets (p_t^T, p_t^S, p_t^L; all in terms of money) are taken as given, and so is the transfer of money from the government ($\mathcal{T}_t$). Let V^{CM}, V^{PM}, and V^{AM} denote the value functions in the respective sub-periods. The household's CM problem is:

$$V^{CM}(m_t, b_t^T, b_t^S, b_t^L, y_t) = \max_{\substack{c_t, g_t, m_{t+1}, \\ b_{t+1}^T, b_{t+1}^S, b_{t+1}^L}} \left\{ u(c_t) + g_t + \beta\, \mathbb{E}\{V^{AM}(m_{t+1}, b_{t+1}^T, b_{t+1}^S, b_{t+1}^L)\} \right\} \tag{3}$$

$$\text{s.t.} \quad p_t c_t + p_t^G g_t + m_{t+1} + p_t^T b_{t+1}^T + p_t^S b_{t+1}^S + p_t^L b_{t+1}^L = p_t y_t + m_t + b_t^T + b_t^S + \left(d_t + p_t^L\right) b_t^L + \mathcal{T}_t.$$

Next, in the PM, active shoppers decide how much output good to purchase, and workers decide in which market to sell. Competition among workers, who can freely choose the early or late market, equalizes prices across the two markets. It does not equalize the price of goods between the PM and the CM; we denote their ratio by $q_t \leq 1$.[5] Let $\Lambda_t \in \{0, 1\}$

[5] If q_t exceeded 1, nobody would buy goods in the PM; see Geromichalos and Herrenbrueck (2022).

indicate whether an individual shopper is active, and let $\Omega_t \in \{0,1\}$ indicate whether an active shopper is early ($\Omega_t = 1$) or late ($\Omega_t = 0$); the aggregate fractions are $\Pr[\Lambda_t = 1] = \lambda_t$ and $\Pr[\Omega_t = 1 \,|\, \Lambda_t = 1] = \omega_t$. Since both short bonds mature before the late PM, a late shopper's holdings of T and S are equivalent to money at that stage. The PM problem is:

$$V^{PM}(m_t, b_t^T, b_t^S, b_t^L, \Lambda_t, \Omega_t) = \max_{y_t} \left\{ V^{CM}\big(m_t + (1-\Omega_t)(b_t^T + b_t^S) - p_t q_t y_t,\ \Omega_t b_t^T,\ \Omega_t b_t^S,\ b_t^L,\ y_t\big) \right\} \tag{4}$$

$$\text{s.t.} \quad y_t \leq \Lambda_t \, \frac{m_t + (1-\Omega_t)(b_t^T + b_t^S)}{p_t q_t}.$$

Finally, consider the AM, where shoppers rebalance after learning their types. Early shoppers sell assets other than money; late shoppers sell consols but buy short bonds (which will be money to them); inactive shoppers buy all three interest-bearing assets with their idle money. Let χ_t^j denote units of asset $j \in \{T, S, L\}$ sold and ζ_t^j the money spent buying it. AM prices, in terms of money, are s_t^T, s_t^S and s_t^L:

$$V^{AM}_{\Lambda_t=0}(m_t, b_t^T, b_t^S, b_t^L) = \max_{\zeta_t^j} \left\{ V^{CM}\big(m_t - \textstyle\sum_j \zeta_t^j,\ b_t^T + \zeta_t^T/s_t^T,\ b_t^S + \zeta_t^S/s_t^S,\ b_t^L + \zeta_t^L/s_t^L\big) \right\} \tag{5}$$

$$\text{s.t.} \quad m_t - \textstyle\sum_j \zeta_t^j \geq 0;$$

$$V^{AM}_{\Lambda_t=1,\Omega_t=1}(m_t, b_t^T, b_t^S, b_t^L) = \max_{\chi_t^j} \left\{ V^{PM}\big(m_t + \textstyle\sum_j s_t^j \chi_t^j,\ b_t^T - \chi_t^T,\ b_t^S - \chi_t^S,\ b_t^L - \chi_t^L,\ 1, 1\big) \right\} \tag{6}$$

$$\text{s.t.} \quad \chi_t^T \leq b_t^T, \ \ \chi_t^S \leq \mu_t b_t^S \ \text{ and } \ \chi_t^L \leq \mu_t b_t^L;$$

$$V^{AM}_{\Lambda_t=1,\Omega_t=0}(m_t, b_t^T, b_t^S, b_t^L) = \max_{\zeta_t^T, \zeta_t^S, \chi_t^L} \Big\{ V^{PM}\big(m_t - \zeta_t^T - \zeta_t^S + s_t^L \chi_t^L,\ b_t^T + \zeta_t^T/s_t^T,\ \ldots \\ b_t^S + \zeta_t^S/s_t^S,\ b_t^L - \chi_t^L,\ 1, 0\big) \Big\} \tag{7}$$

$$\text{s.t.} \quad m_t + s_t^L \chi_t^L \geq \zeta_t^T + \zeta_t^S \ \text{ and } \ \chi_t^L \leq \mu_t b_t^L.$$

We leave the detailed solutions to these problems to Appendix A.

2.3 Asset pricing in terms of observables

Define the nominal stochastic discount factor:

$$\Delta_{t,t+1} \equiv \beta \, \frac{u'(c_{t+1})}{u'(c_t)} \, \frac{p_t}{p_{t+1}}, \tag{8}$$

and define two liquidity premia, corresponding to the two ways an asset can serve a monetary need:

$$\ell_t^D = \lambda_t \left(\frac{1}{q_t} - 1 \right) \qquad \text{and} \qquad \ell_t^N = \lambda_t \left(\frac{s_t^T}{q_t} - 1 \right). \tag{9}$$

Here ℓ_t^D is the value of direct liquidity: the premium commanded by an asset that will be money (or will have matured into money) when a monetary need arrives. And ℓ_t^N is the value of indirect liquidity: the premium commanded by an asset that must first be sold in the secondary market, at the discount $s_t^T \leq 1$, before the proceeds can be spent.

We also define an *aggregate liquidity premium*:

$$\bar{\ell}_t \;\equiv\; (1 - \omega_t)\, \ell_t^D \;+\; \omega_t\, \ell_t^N \tag{10}$$

as the urgency-weighted average of direct and indirect liquidity. In our estimation, we identify this object with the economy-wide "price of convenience", taken as a whole rather than the convenience of a particular asset versus another.

With these definitions, the model's Euler equations take a compact form:

$$1 = \mathbb{E}\left\{ \Delta_{t,t+1} \cdot \frac{1 + \bar{\ell}_{t+1}}{s_{t+1}^T} \right\}, \tag{11}$$

$$p_t^T = \mathbb{E}\Big\{ \Delta_{t,t+1} \cdot \big(1 + \bar{\ell}_{t+1}\big) \Big\}, \tag{12}$$

$$p_t^S = \mathbb{E}\Big\{ \Delta_{t,t+1} \cdot \big[1 + (1 - \omega_{t+1})\, \ell_{t+1}^D + \omega_{t+1}\, \mu_{t+1}\, \ell_{t+1}^N\big] \Big\}, \tag{13}$$

$$p_t^L = \mathbb{E}\Big\{ \Delta_{t,t+1} \cdot \big(d_{t+1} + p_{t+1}^L\big)\, \big(1 + \mu_{t+1} \bar{\ell}_{t+1}\big) \Big\}, \tag{14}$$

The Euler equation (11) expresses that holding one dollar to next period is equivalent to selling one dollar's worth of T-bills at next period's secondary market price.[6] The other three equations are transparent: every bond is priced by its loading on the two liquidity premia, and those differences are what identify ω and μ in Section 3.

[6] If the time-$(t+1)$ secondary market price of T-bills is known at time t, then Equations (11) and (12) resolve to $p_t^T = s_{t+1}^T$ – the issue price equals the secondary market price shifted by whatever time span we interpret the CM–AM step to represent. Since T-bill rates are in fact highly persistent, Equations (11) and (12) do not provide independent moments for our estimation and we proceed with the three asset pricing equations (12)-(14).

3 Measurement

The model's Euler equations provide a mapping from observed asset prices to the latent processes of interest. This section states that mapping, inverts it, and defines the price of convenience. The assumptions required to read the mapping as a measurement are stated here and justified in Section 6; details not provided here can be found in Appendix B (data) and Appendix C (estimation).

3.1 Observables and identifying assumptions

Following Geromichalos and Herrenbrueck (2022), we define the (ex-ante) **Fisher rate** as the expected return on a **hypothetical perfectly illiquid nominal bond**, priced by the stochastic discount factor. This must be conceptually distinguished from observed returns on actual assets: by the very nature of their prices/returns being observed, real-world assets are always at least somewhat tradable. Following Canzoneri, Cumby, and Diba (2007), we construct this Fisher rate from nondurables-and-services consumption growth per capita; following Herrenbrueck and Wang (2026), we assume $\log$ utility ($u'(c) = 1/c$) as our baseline:[7]

$$1 + r_t^F = \frac{1}{\beta_t} \mathbb{E}_t \left[\frac{p_{t+1} c_{t+1}}{p_t c_t} \right]$$

Departing from previous work, we do not fix the quarterly discount factor β; we treat it as a latent object, β_t, with an economically plausible prior. Throughout, β_t is the *one-quarter* discount factor of Section 2, while we report the *annualized* discount rate $\rho_t \equiv -4 \log \beta_t$, whose implied annual discount factor is $e^{-\rho_t} = \beta_t^4$.

Crucially, r_t^F cannot be identified with realized nominal consumption growth. Equations (11)–(14) are conditional expectations, and quarterly nominal consumption growth is far too volatile to stand in for one. We therefore construct a forecast via a local-level model

[7] We consider generalized CEIS utility (as in Canzoneri et al., 2007) in Section 6.2; we do not find it identifiably superior to $\log$ utility, which wins on simplicity and the principle of growth optimality (Robson, Samuelson, and Steiner, 2023). Note that most estimates of the elasticity of intertemporal substitution in the literature (e.g., Havranek, Horvath, Irsova, and Rusnak, 2015) come from studies on Euler equations that did not account for liquidity; therefore, using such a study to justify EIS $\neq 1$ here is circular.

estimated with a one-sided Kalman filter (Appendix C.1).

For asset prices we use three instruments with a long, consistent history:

1. **Treasury bills** (3-month secondary-market rate), representing the T-bill of Section 2: a default-free claim that matures within the model period and whose secondary market is deep enough that a holder forced to liquidate before maturity can place all of it, $\mu_t^T = 1$. Bills are the only instrument in our sample for which this is a defensible description: they are the collateral of first resort, are financed and cleared on terms available to no private claim, and Greenwood et al. (2015) document the money-like premium they carry over longer Treasuries precisely because they are short.

2. **Commercial paper** (3-month, spliced; see Appendix B). This is the empirical counterpart of the model's short bond S. Like a T-bill, it matures within the period and therefore serves a non-urgent need at par; but its secondary market and acceptability as collateral pale in comparison to Treasuries, so a holder whose need arrives before maturity can place only a fraction μ_t^{CP} of it. It is therefore priced by (13).[8]

3. **Seasoned AAA corporate bonds** (Moody's index). This is the empirical counterpart of the model's consol L: it does not mature within the model period, so it is priced by (14).[9] The pricing equation requires an expected one-period holding return, whereas we observe a yield. The two differ by an expected capital gain, and we carry that gap explicitly as a *repricing wedge* ψ_t,

$$1 + r_t^{AAA} \;=\; \big(1 + y_t^{AAA}\big)\big(1 + \psi_t\big), \tag{15}$$

   with $\psi_t \equiv 0$, the random-walk restriction on the long yield, as a special case.

All four series are quarterly. The longest span over which all are available is 1971Q2–2025Q4;

[8] Nagel (2016) used GC repo as the illiquid benchmark asset; the convenience yield literature since then mainly uses commercial paper, which is available consistently since 1971.

[9] Moody's index is not a fixed portfolio of aging bonds: its constituents are replaced so that the index maintains a roughly constant time to maturity, so what it quotes a yield on is a synthetic claim that never matures. That claim is the consol of Section 2. The same reading governs the holding-period returns constructed in Appendix B.3, which follow convention in setting the coupon at the prevailing yield, so that a quarter's return is the yield plus the capital gain implied by the change in the yield.

losing the last of these to the growth forecast, the estimation uses the $T = 218$ quarters 1971Q2–2025Q3. Both money-market series are quoted as *bank discount* rates on a 360-day year rather than as effective annual rates: the bill over the 91 days of the 13-week tranche, commercial paper over 90 days. We therefore convert each quote to an effective annual rate using its own day count and then to a common quarterly rate, so that the four rates price the same period. Rates are re-annualized only for reporting. Appendix B gives the sources, the conventions in full, and the three-part splice behind the commercial-paper series.

Every asset in the model of Section 2 is default-free, and we treat both private instruments as such. Prime commercial paper and AAA-rated bonds are the two longest-running private instruments for which that is defensible, with realized credit losses of the order of a basis point a year against spreads measured in hundreds. Default aside, AAA bonds do carry repricing risk, unlike the 3-month bonds that mature within one model period. We deal with this by placing a prior on the repricing wedge ψ_t in (15), described in Section 3.4 below.

Our model assumes a common marketability $\mu_t \in [0, 1]$ for the S and L bonds, and perfect marketability for the T-bill ($\mu_t^T = 1$). Mapping the three instruments onto these assets therefore embodies our central identifying assumptions: the 3-month T-bill is perfectly marketable, while the two private instruments share a marketability parameter:

$$\mu_t^{CP} = \mu_t^{AAA} \equiv \mu_t.$$

We do not just adopt these assumptions for convenience; they are inherent to our exercise. A single asset yields a single pricing equation but (in principle) two parameters: return and marketability of the respective asset. The relevant question is therefore not whether to impose an assumption, but which assumption is least arbitrary, and on that criterion we prefer a common μ for private claims and a maximal one for T-bills. The economic content of the common μ restriction, what the data say about it, and what the literature has implicitly assumed in its place are taken up at length in Section 6.

3.2 The estimable system

Applying these assumptions to (12)–(14), adding the definition of the Fisher rate (via $\exp(\gamma_t)$, the estimated conditional forecast of gross nominal consumption growth), and combining the two definitions in (9) to eliminate q_t:

$$1 + r_t^F = \frac{e^{\gamma_t}}{\beta_t}, \tag{16}$$

$$\lambda_t + \ell_t^D = \big(1 + r_t^{TB}\big)\big(\lambda_t + \ell_t^N\big), \tag{17}$$

$$1 = \big(1 + r_t^{TB}\big)\, \mathbb{E}_t \left\{ \frac{1 + (1 - \omega_t)\, \ell_t^D + \omega_t\, \ell_t^N}{1 + r_t^F} \right\}, \tag{18}$$

$$1 = \big(1 + r_t^{CP}\big)\, \mathbb{E}_t \left\{ \frac{1 + (1 - \omega_t)\, \ell_t^D + \omega_t\, \mu_t\, \ell_t^N}{1 + r_t^F} \right\}, \tag{19}$$

$$1 = \big(1 + r_t^{AAA}\big)\, \mathbb{E}_t \left\{ \frac{1 + \mu_t\big[(1 - \omega_t)\, \ell_t^D + \omega_t\, \ell_t^N\big]}{1 + r_t^F} \right\}. \tag{20}$$

Following the convention of the term-premium and convenience-yield literatures, every object in the system is dated by the information set that identifies it, which is also the date at which the three rates are quoted. A subscript t on a *latent* therefore denotes the state prevailing over the three months that begin at t, random as of t, while a subscript t on a *rate* denotes a price quoted at t, known at that date and therefore standing outside the expectation.

Equations (16) and (17) are definitions rather than expectational restrictions: the first defines the Fisher rate given the discount factor and the conditional expectation of nominal consumption growth, the second follows from (9). The system contains eight latent variables: time preference β_t, expected nominal consumption growth γ_t, the incidence of monetary needs λ_t, urgency ω_t, marketability μ_t, the two premia (ℓ_t^D, ℓ_t^N), and, through (15), the repricing wedge ψ_t. One of the eight is not free: γ_t is supplied by the first-stage filter of Appendix C.1, and what the estimator draws for it is its *measurement* distribution, not a prior over an unknown structural object. We therefore treat γ_t as a first-stage input measured from (16), and every count of free dimensions below excludes that variable and that equation. That leaves seven free latents against four equations, so the system is underdetermined

period by period in exactly three dimensions. Two of them are inherent in the question; the third follows from the fact that the yield of a long bond is not its expected return.

3.3 From yields to expected holding-period returns

Equation (20) requires an expected holding return on AAA bonds, whereas we observe a yield. The difference between those is the repricing wedge ψ_t defined in Equation (15).

Our baseline approach is to place a prior on this wedge, disciplined by data. There are two obvious alternatives – ignoring the wedge altogether, and estimating it from *realized* holding-period returns – which we discuss and reject in Appendix C.5. While realized returns cannot serve as an estimate, they do deliver a bound on the prior: a forecasting regression produces an in-sample standard deviation of 4.9 percentage points a year, whereas pure noise generates a standard deviation of 2.7 percentage points per year (bootstrapped, see Appendix B.3). Consequently, a predictable component can have a standard deviation of at most $\sqrt{4.9^2 - 2.7^2} \approx 4.1$ percentage points a year. Two structural considerations – the horizon at which a long bond discounts, and the persistence of the convenience loading it prices – put the figure closer to one percentage point. Survey forecasts of the AAA yield itself sit in between (Appendix C.5).

We therefore draw ψ_t from a prior centred at zero, scaled in units of the *yield* rather than the return:

$$\psi_t \;=\; -\,D\,\mathbb{E}_t\big[\Delta y^{AAA}_{t+1}\big], \qquad \mathbb{E}_t\big[\Delta y^{AAA}_{t+1}\big] \sim N\big(0, \sigma_y^2\big), \qquad \sigma_y = 5\text{bp}, \tag{21}$$

with $D = 11.1$ years the effective duration of the index in log yields, measured from the same realized returns (Appendix B.3). A standard deviation of five basis points on the expected quarterly yield change corresponds to 2.2 percentage points a year on the expected holding return. That is inside the bound, above the structural channels, and slightly below the survey evidence.

3.4 Inverting the system

Our baseline approach is to place priors on the free dimensions and let the equations do the rest. We draw the trend growth rate γ_t from its (one-sided, ex-ante) measurement distribution, β_t from a prior whose median is a 2% annual discount rate (Appendix C.4), λ_t from a uniform prior on $[0, 1]$, and the repricing wedge ψ_t from (21); (16)–(20) then determine r_t^F, μ_t, ω_t, ℓ_t^D and ℓ_t^N exactly and in closed form.

For any gross rate $1 + r$, define the *model premium*:

$$\Delta_t(r) \;\equiv\; \frac{1 + r_t^F}{1 + r} - 1, \tag{22}$$

the shortfall of that yield below the Fisher rate, which the model must explain with some form of convenience. Writing $A_t \equiv (1 - \omega_t)\ell_t^D$ for the direct/non-urgent and $B_t \equiv \omega_t \ell_t^N$ for the indirect/urgent component, the three pricing equations (18)–(20) become:

$$\Delta_t(r_t^{TB}) = A_t + B_t, \qquad \Delta_t(r_t^{CP}) = A_t + \mu_t B_t, \qquad \Delta_t(r_t^{AAA}) = \mu_t\,(A_t + B_t). \tag{23}$$

The T-bill is fully liquid and so must be compensated for nothing; its shortfall is the whole convenience premium. Commercial paper matures within the period, so it still collects the direct premium in full, but collects the urgent premium only to the extent μ_t that it can be sold. The AAA bond does not mature, so it collects *both* only to the extent μ_t. The solution is now straightforward:

$$\mu_t = \frac{\Delta_t(r_t^{AAA})}{\Delta_t(r_t^{TB})}, \qquad B_t = \frac{\Delta_t(r_t^{TB}) - \Delta_t(r_t^{CP})}{1 - \mu_t}, \qquad A_t = \Delta_t(r_t^{TB}) - B_t. \tag{24}$$

Given A_t and B_t, Equation (17) is a quadratic in ω_t with a unique positive root, and whenever $A_t > 0$ and $B_t > 0$ that root lies in $(0, 1)$ and delivers $\ell_t^D \geq \ell_t^N \geq 0$ automatically; the theoretical restrictions are satisfied by construction rather than imposed.

Because Δ_t is decreasing in the yield, the admissibility conditions $0 \leq \mu_t \leq 1$, $A_t \geq 0$ and $B_t \geq 0$ collapse into a single ordering restriction:

$$r_t^{TB} \;\leq\; r_t^{CP} \;\leq\; r_t^{AAA} \;\leq\; r_t^F. \tag{25}$$

The model requires the four rates to line up by liquidity: bills are fully liquid, commercial paper matures but is only partly marketable, AAA bonds are only partly marketable and do not mature, and an asset with no liquidity value at all would earn the Fisher rate. Note that the third rate in the chain is the expected holding return, not the yield, so (25) is a restriction on ψ_t as much as on the latents.

Proposition 3.1 (What the three observed rates identify)**.** *Fix $(\gamma_t, \beta_t, \lambda_t, \psi_t)$, so that the Fisher rate r_t^F and hence $\Delta_t(\cdot)$ are determined, and let the two private instruments carry marketabilities μ_t^{CP} and μ_t^{AAA} that need not be equal. Then, from (23):*

(i) The price of convenience and the marketability of the long bond are point-identified, $\bar{\ell}_t = \Delta_t(r_t^{TB})$ and $\mu_t^{AAA} = \Delta_t(r_t^{AAA})/\Delta_t(r_t^{TB})$, and neither involves the commercial-paper yield.

(ii) The commercial-paper equation identifies the product $(1-\mu_t^{CP})B_t$ and nothing more, so the urgent component and the marketability of the short private claim are identified only jointly.

(iii) The requirement $A_t \geq 0$ restricts μ_t^{CP} to $\big[0,\ \Delta_t(r_t^{CP})/\Delta_t(r_t^{TB})\big]$, a set that is non-empty if and only if $r_t^{TB} \leq r_t^{CP} \leq r_t^F$.

(iv) Writing $\kappa_t \equiv (1-\mu_t^{CP})/(1-\mu_t^{AAA})$, so that $\kappa_t = 1$ is the common-μ restriction, the urgent share is $B_t/\bar{\ell}_t = s_t/\kappa_t$, where s_t is its value at $\kappa_t = 1$. At the upper end of the set $A_t = 0$ and the share equals 1.[10]

Three things follow that our analysis leans on later. The level of the price of convenience and the marketability of a long private bond are free of the $\mu^{CP} \equiv \mu^{AAA}$ assumption entirely, by (i). The urgent share is not, since by (iv) it is exactly proportional to $1/\kappa_t$. And the middle link of the ordering chain (25) is, by (iii), a restriction on the identifying assumption rather than on the model: a quarter in which commercial paper yields more than a long AAA bond is admissible at a small enough μ_t^{CP}, and only a quarter in which it yields more than the Fisher rate is not. Section 6 provides more details.

[10] *Proof.* The third equation of (23) gives (i) immediately, since $A_t + B_t = \Delta_t(r_t^{TB})$ by the first. The second gives $\Delta_t(r_t^{TB}) - \Delta_t(r_t^{CP}) = (1-\mu_t^{CP})B_t$, which is (ii); solving it for B_t and substituting into $A_t = \Delta_t(r_t^{TB}) - B_t$ gives $A_t \geq 0 \iff \mu_t^{CP}\Delta_t(r_t^{TB}) \leq \Delta_t(r_t^{CP})$, and Δ_t is decreasing in the yield, which is (iii). For (iv), B_t is proportional to $1/(1-\mu_t^{CP})$ with a factor that does not involve μ_t^{CP}, and $\bar{\ell}_t$ does not involve it at all. □

3.5 The price of convenience

We therefore define *the* price of convenience as the premium on an asset that collects convenience in full, relative to the perfectly-inconvenient benchmark Fisher rate. We identify T-bills as being that asset, or certainly close enough for our purposes and for the level of precision our estimation attempts. Its liquidity premium is $\bar{\ell}_t$ of (10):

$$\bar{\ell}_t \;=\; \Delta_t(r_t^{TB}) \;=\; \frac{1+r_t^F}{1+r_t^{TB}} - 1 \;=\; \underbrace{(1-\omega_t)\,\ell_t^D}_{A_t,\ \text{direct}} \;+\; \underbrace{\omega_t\,\ell_t^N}_{B_t,\ \text{urgent}}. \tag{26}$$

Our baseline estimate depends on (γ_t, β_t) and the T-bill yield alone: no prior on ω_t, λ_t, μ_t or the premia enters it, and neither do the commercial-paper or AAA yields. What the two private instruments do is split it. Rearranging (23):

$$\big(1-\mu_t\big)\,\omega_t\,\ell_t^N \;=\; \frac{r_t^{CP} - r_t^{TB}}{1+r_t^{TB}} \cdot \frac{1+r_t^F}{1+r_t^{CP}}, \tag{27}$$

$$\big(1-\mu_t\big)\,\bar{\ell}_t \;=\; \frac{r_t^{AAA} - r_t^{TB}}{1+r_t^{TB}} \cdot \frac{1+r_t^F}{1+r_t^{AAA}}. \tag{28}$$

A short-maturity private spread – commercial paper minus bills, or repo minus bills (Nagel, 2016) – prices illiquidity times the *urgent* component of convenience alone; a long-maturity private spread prices the same illiquidity times the *whole* of it. The traditional measure of the convenience yield is not merely silent about the direct component A_t, which both short instruments share and which nets out of any short-maturity spread; it also understates the urgent component it does purport to measure, by the factor $1/(1-\mu_t)$, because it treats an asset that can be partly sold as one that cannot be sold at all.

Dividing (27) by (28) to obtain the *share of the price of convenience attributable to urgency*, the Fisher rate cancels and with it nearly every prior in the system:

$$\frac{\omega_t \ell_t^N}{\bar{\ell}_t} \;=\; \frac{\dfrac{1}{1+r_t^{TB}} - \dfrac{1}{1+r_t^{CP}}}{\dfrac{1}{1+r_t^{TB}} - \dfrac{1}{1+r_t^{AAA}}}. \tag{29}$$

This share is a function of the three rates only and notably does not contain trend growth, the discount factor, incidence, or marketability. It is not entirely prior-free: r_t^{AAA} is the expected holding return, not the yield, so once the wedge is drawn the identity is free of every prior in the system *except* the one on ψ_t. Unlike the other priors, however, that one is not externally chosen: its scale is set by holding-period returns constructed from these same yields (Appendix B.3).

3.6 Reading the results

What our baseline estimation reports, quarter by quarter, is a *certainty equivalent*: the value of each latent that reproduces the observed price when substituted into the pricing equation, rather than the forecast mean of that latent.[11] All series are reported at annual rates; shaded regions are 50% and 90% posterior bands. The three rates enter our estimation without error and each quarter is exactly identified. The bands are therefore entirely *propagated prior uncertainty*, quantifying uncertainty about assumptions rather than about sampling.

In our estimates, nothing imposes persistence (the latents are independent across quarters) until Section 5.3, where we investigate just that. Where a number below is a statistic across quarters (e.g., a correlation), the sampling interval that goes with it is in Appendix C.8, from a stationary block bootstrap that allows for the serial dependence of the series.

4 Empirical results

Figure 2 shows the four rates of our system. The T-bill and commercial paper yields are read off market prices, the other two are estimated; the Fisher rate carries the priors on trend growth and the discount factor, and the AAA expected holding return carries the prior on the wedge.

Plotted at their posterior medians, the four line up by liquidity in all 218 quarters. That

[11] The distinction is immaterial for the premia, which enter (18)–(20) linearly, so that $\bar{\ell}_t$ is recovered as a conditional expectation exactly; it bites, weakly, on μ_t, which is a ratio of two such expectations, and on ω_t, which solves a quadratic. Appendix C.1 bounds all three of the resulting corrections; the largest is worth half a basis point a year.

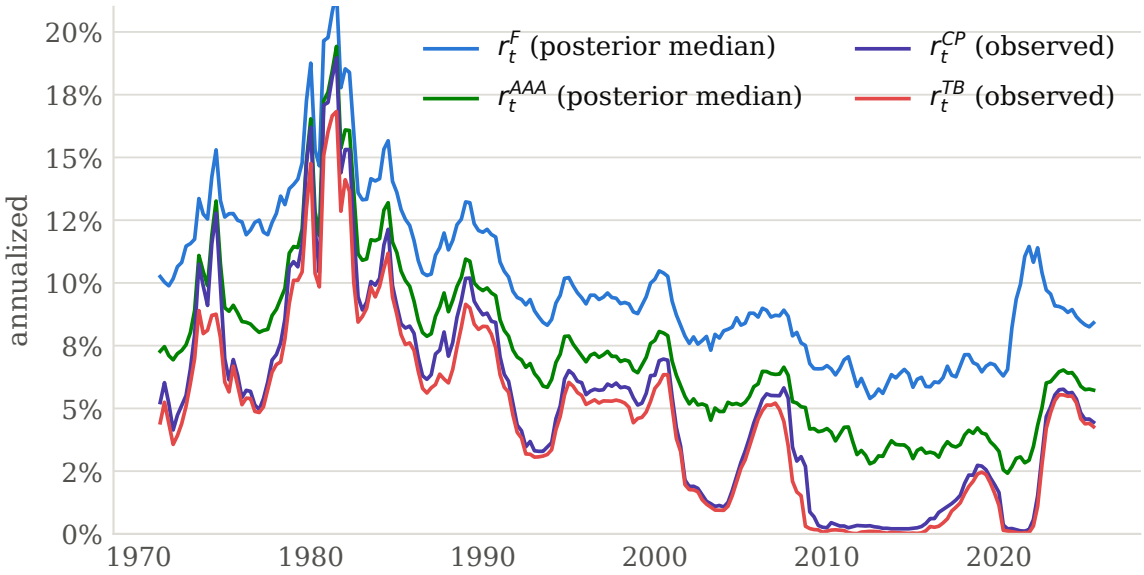

Figure 2: The four rates of the system. The T-bill rate r_t^{TB} and the commercial-paper rate r_t^{CP} are observed. The other two are plotted at their posterior medians: r_t^{AAA} is the estimation-inferred expected holding return via (15), and r_t^F is built from trend nominal consumption growth and the discount factor via (16).

is not imposed on the medians – the admissibility penalty is soft and the wedge prior is centred at zero, so a posterior median that broke the chain was available to the estimator. This ordering matches our model's restriction on these four rates. The bill is cheapest to the issuer because it is the most convenient thing to hold; commercial paper costs a little more because a holder who needs money before it matures must find a counterparty; the AAA bond costs more again because its holder must find a counterparty whatever the timing of the need; and the Fisher rate – what a claim with no monetary usefulness at all would have to pay – sits above all three.

The commercial-paper spread over T-bills has a median of 44 basis points and the AAA spread over T-bills a median of 2.54 percentage points (in means, 59 basis points against 2.41 points). Under the traditional reading, the first of these measures convenience and the second measures term. Under our reading, the first of these measures only the urgent component of convenience, and "term" is another word for the essential inconvenience of long-term bonds.

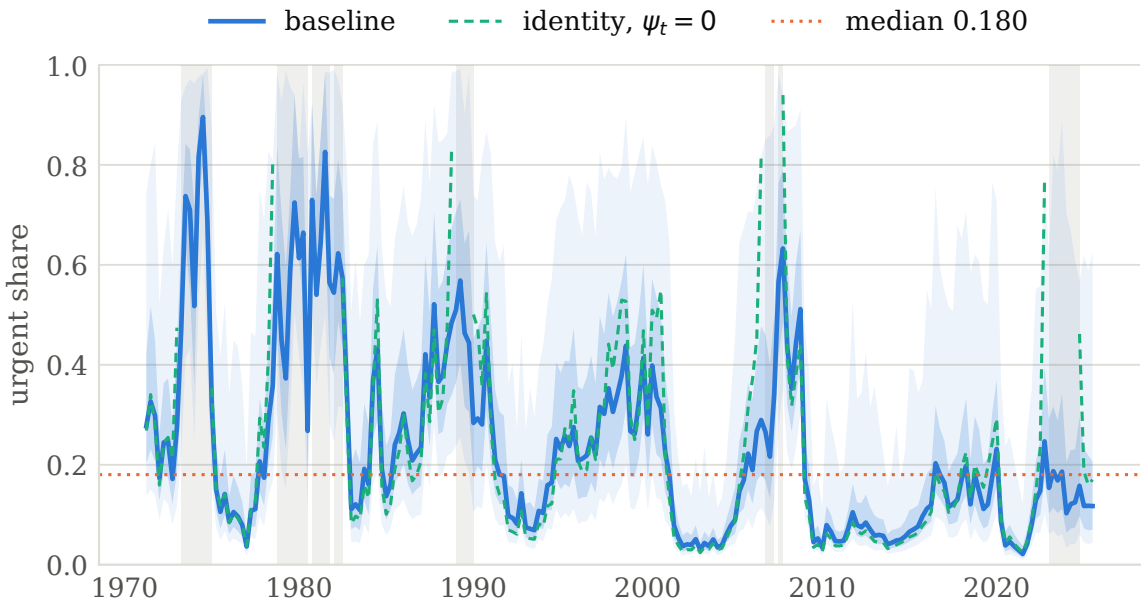

Figure 3: The composition of convenience: the share of the price of convenience attributable to urgency, $\omega_t \ell_t^N / \bar{\ell}_t$. Baseline posterior median (solid) with 50% and 90% credible bands (shaded); the horizontal line marks the baseline's sample median of 0.180, taken over all 218 quarters. The dashed line is the prior-free identity (29) evaluated at $\psi_t = 0$, whose own median is 0.149 over the same 218 quarters. It is broken, and the quarters are shaded, where r^{CP} exceeds the raw yield y^{AAA}; elsewhere, the lines functionally coincide.

4.1 The composition of convenience

We continue with the result that is least sensitive to the priors. The *share of convenience attributable to urgency* is determined by the three asset returns alone, via the identity (29), and can therefore be read before any of the estimates below. We show this share in Figure 3. In the baseline, over all 218 quarters, it has a median of 0.180 (mean: 0.236), and decade medians of 0.28, 0.37, 0.24, 0.17, 0.08 and 0.12, falling over the four decades to 2019 and then rising again. That profile is not independent of the level of nominal interest rates – it follows the rate cycle, and conditionally on the bill rate it has no time trend. What the estimates imply is therefore a statement about the composition of convenience at a given opportunity cost of holding money, rather than a secular change in the nature of liquidity needs.

The only prior that does affect the urgent share is the one on the AAA repricing wedge, and it matters when the raw AAA yield falls below the commercial-paper rate, breaking

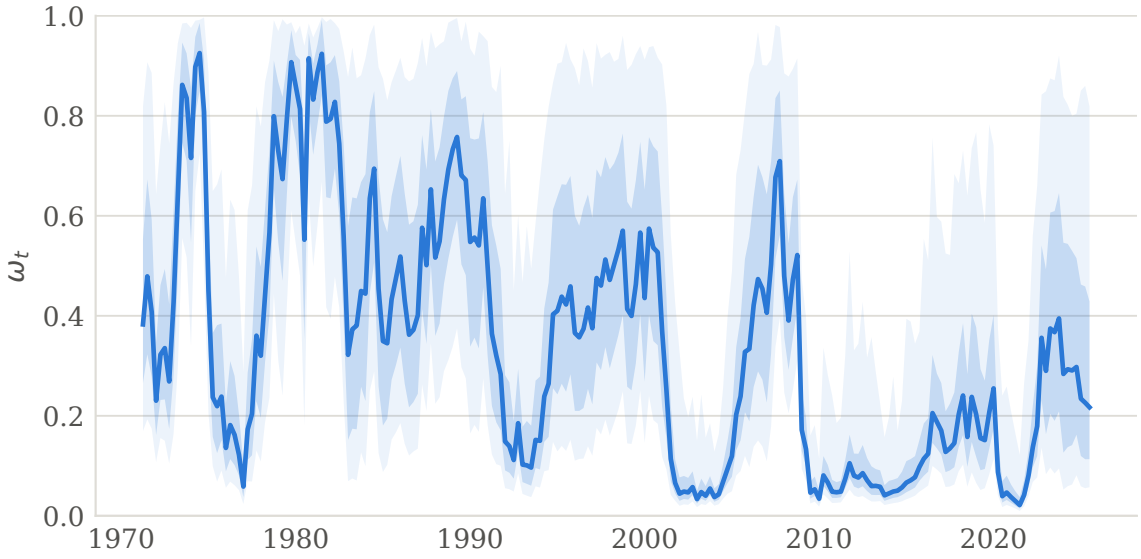

Figure 4: Urgency ω_t: the share of monetary needs that arrive before short-term paper matures. Posterior median with 50% and 90% credible bands. Unlike λ_t and μ_t, urgency carries no prior of its own in the baseline – it is solved from (17) – so there is no prior band to compare against.

the ordering restriction (25). In those quarters, the identity (29) evaluated at $\psi_t = 0$ does not merely become noisy, it becomes undefined. In the 21 quarters in which the AAA yield falls below the bill rate, the denominator of (29) changes sign and the identity returns a negative number; in 13 further quarters where it sits between the bill rate and the commercial paper yield, it is well defined but returns a value *above one*. In Figure 3, those quarters are shaded and the $\psi_t = 0$ line is broken in them; the baseline estimate survives by inferring an appropriate value of ψ_t. In the other 184 quarters, the two lines functionally coincide.

4.2 Urgency and marketability

There is a difference in kind between our two keystone concepts. Urgency is a property of the economy's monetary needs: what fraction of them cannot wait three months. Marketability is a property of the market for private claims: what fraction of a claim can be placed on demand, or equivalently, the probability that all of it can be placed. Indeed, our estimates show that they move in different ways and follow distinct patterns.

Urgency ω_t (Figure 4) moves a great deal: from a median of 0.93 in 1974Q3 to 0.02 in 2021Q3. In the baseline estimation, ω_t carries no prior: it is not drawn but solved from (17) given $(\gamma_t, \beta_t, \lambda_t, \psi_t)$ and the observed rates, so its posterior spread (averaging 0.24 in interquartile range) is entirely propagated uncertainty from the drawn latents and not a prior that the data have narrowed. Like the urgent share in the previous section, it is highly correlated with the T-bill rate ($+0.86$) and does not trend independently. When we investigate persistence in Section 5.3 below, we find that among all our latents, urgency alone requires occasional jumps rather than smooth variation over time.

Marketability μ_t (Figure 5) has a median of 0.49 (mean: 0.51) and an average posterior interquartile range of 0.39, so it is estimated less sharply than urgency – unsurprisingly, since it inherits both the AAA repricing wedge and the uncertainty about the Fisher rate that the urgent share does not. It also moves, but less dramatically than urgency: marketability falls to 0.33 in 2008Q4, the second lowest reading in the sample, drifts around 0.5 through the 1990s and 2000s, and rises in the 2020s on, never falling below 0.58 thereafter.[12] Unlike urgency, it is uncorrelated with the T-bill rate ($+0.01$) and with the price of convenience ($+0.05$); a private claim is not systematically easier to place when money is cheap. Marketability and urgency are also uncorrelated with one another (-0.10, with a 90% confidence interval of $[-0.41, +0.22]$).

The natural follow-up is to test the restriction of constant versus time-varying μ_t. However, a Bayes factor is not available here: in the baseline μ_t is solved rather than drawn, so it carries no prior, and the two specifications do not assign commensurable probabilities to the AAA yield (see Appendix C.7). What *is* well-posed is to ask how large of a wedge a constant-μ restriction forces. The single best fit is obtained at $\mu = 0.47$ (against a baseline median of 0.49); it implies expected quarterly yield movements with a median of -1.6 basis points and a standard deviation of 3.5, against a prior standard deviation of 5. Thus, constant marketability is not something our estimation can reject.[13] Setting $\mu = 0$, however, forces a median

[12] Only one reading is lower: 0.32 in 1974Q3, at the peak of monetary and funding-market stress during the first oil-shock recession. That quarter is also where urgency and the urgent share attain their sample maxima.

[13] We keep μ_t time-varying through the rest of the paper on the a priori reasoning that dealer capacity is not a constant of nature, but report the constant alongside the baseline throughout Appendix C.9.

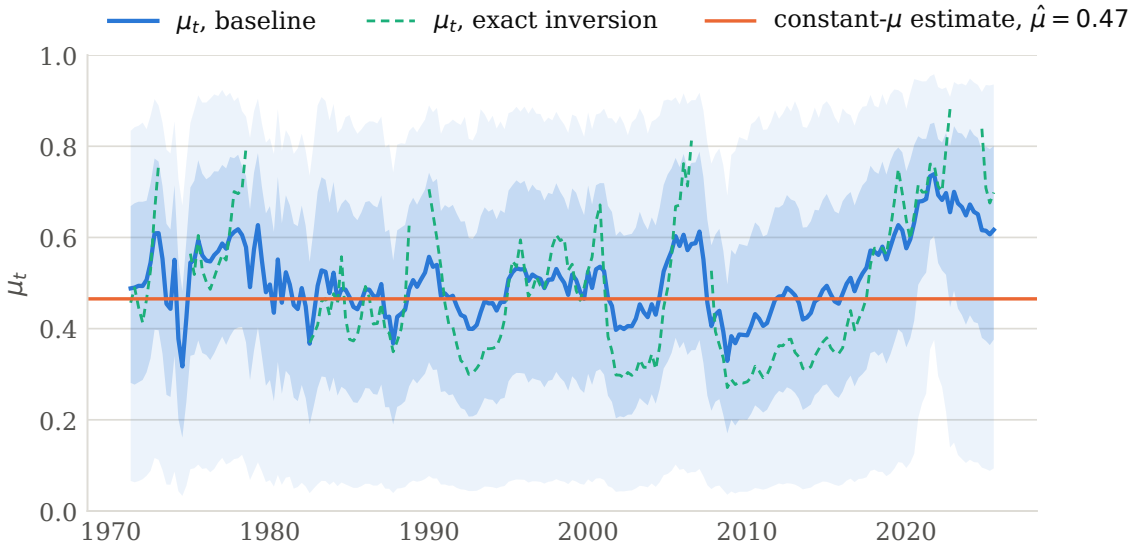

Figure 5: Marketability μ_t: the fraction/probability of placing a private claim in the secondary market. Baseline posterior median with 50% and 90% credible bands, against the exact inversion median and the best single constant. The exact inversion is the $\sigma_y \to 0$ limit of the baseline, in which the wedge is zero; this inversion is undefined in 34 quarters (see Appendix C.3), so its line is broken there, whereas the baseline is defined throughout.

of $+1.9$ basis points with a standard deviation of 5.0: a significant displacement, costing 56 log points against the constant on a like-for-like comparison of one-parameter models, with a block-bootstrap 90% confidence interval of $[23,\ 103]$ log points. The literature's default of assuming no convenience to private assets is not close to tenable.

4.3 The Fisher rate and the price of convenience

The Fisher rate estimate derives from trend consumption growth and the discount factor alone; no asset return enters it directly, though our baseline estimation does infer a high discount rate, and with it a high Fisher rate, in the quarters where the ordering restriction (25) would otherwise fail (Figure 6; see Section 4.4). Via $\bar{\ell}_t = (r_t^F - r_t^{TB})/(1 + r_t^{TB})$, the price of convenience inherits that independence: it depends on trend growth, the discount factor and the T-bill rate, and neither the commercial-paper rate nor the AAA yield enter it directly.

The price of convenience $\bar{\ell}_t$ is large (Figure 7). Its posterior median has a sample median of 4.9% a year (mean: 5.1%) and ranges from 3.2% in 2024Q1 to 11.4% in 2021Q4. Since the

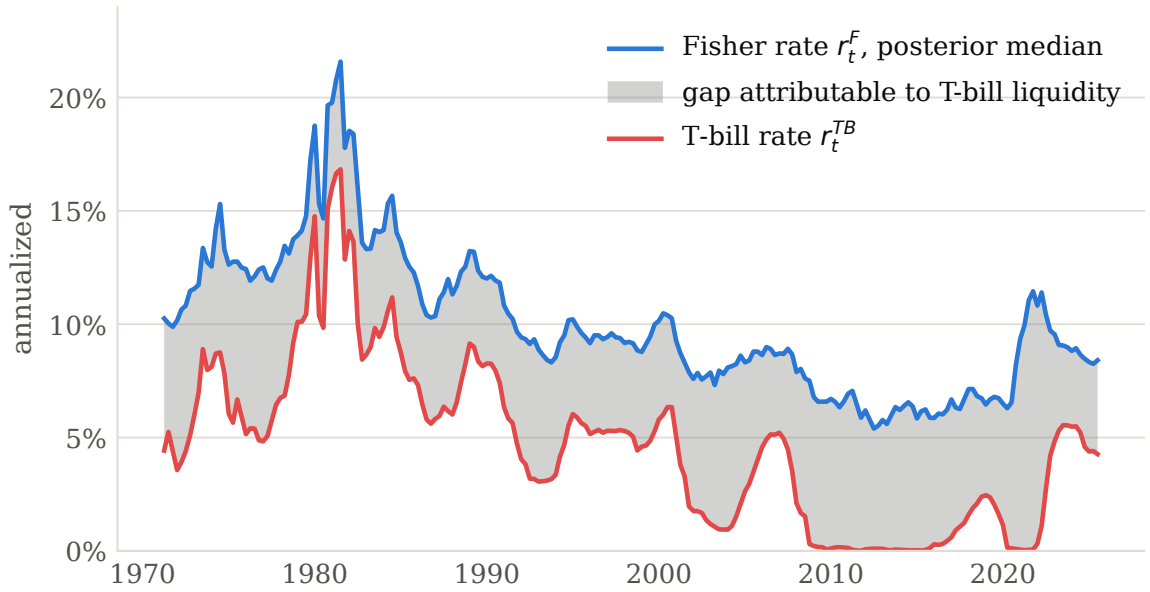

Figure 6: The Fisher rate and its decomposition. Posterior medians, annualized. The shaded region is the arithmetic difference $r_t^F - r_t^{TB} = (1 + r_t^{TB})\,\bar{\ell}_t$, hence "gap attributable"; the $\bar{\ell}_t$ quoted in the text is the exact object.

Fisher rate is much smoother in the medium and long run than interest rates, $\bar{\ell}_t$ and r_t^{TB} move in opposite directions, with a correlation of -0.61 ($[-0.76, -0.52]$).

Against this, the commercial-paper spread has a median of 44 basis points (mean: 59), correlates positively with the T-bill rate ($+0.57$, $[+0.37, +0.71]$), but negatively with our price of convenience measure (-0.23, $[-0.39, -0.07]$). The intervals are wide but do not intersect across the two measures, so the disagreement in sign is not a matter of sampling error.

Our interpretation of this disagreement is straightforward: the commercial-paper spread prices only the urgent component of convenience, and only a fraction $1 - \mu_t$ of it at that. The consequences go far beyond measurement: convenience yields are commonly used as summaries of financial conditions, so the fact that our measure and the traditional one comove with the T-bill rate *in opposite directions* is a red flag for this kind of use. The 2010s are the sharpest case: our decade median is 5.8%, the highest of any decade, while the commercial-paper spread over the same years was the narrowest in the sample. A spread-based measure would have called convenience cheap through a decade our measure calls expensive.

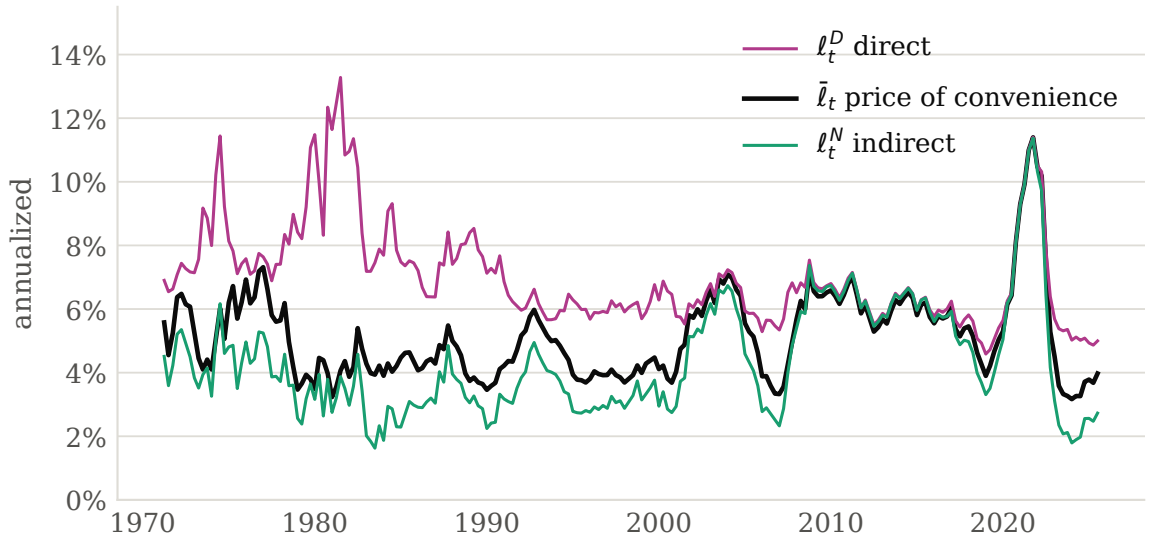

Figure 7: The price of convenience and the two premia. ℓ_t^D is the premium on an asset that will be money when a monetary need arrives; ℓ_t^N is the premium on an asset that must first be sold in the secondary market; and $\bar{\ell}_t = (1-\omega_t)\ell_t^D + \omega_t \ell_t^N$. Posterior medians, annualized; credible bands are omitted because what they would report about the level of $\bar{\ell}_t$ is the prior on ρ_t, on which see Section 4.4 and Appendix C.4.

Figure 7 also shows the two premia out of which the price of convenience is built, at medians of 6.6% for direct liquidity and 3.9% for indirect. The gap between them follows nearly mechanically from the T-bill rate: via Equation (17), $\ell_t^D - \ell_t^N = r_t^{TB}(\lambda_t + \ell_t^N)$. It is therefore widest in the early 1980s and closes almost exactly at the zero lower bound. This is a property of the model: with no interest forgone, being forced to sell early costs nothing, and the two ways of holding liquidity become equivalent.

4.4 Liquidity needs and time preference

The incidence of monetary needs λ_t is not identified: its posterior interquartile range is indistinguishable from the prior's on the same grid. The reason is structural, visible in (63): λ_t enters only through the product $\lambda_t r_t^{TB}$, so it is invisible whenever the T-bill rate is low and weakly informative even when it is high.

The discount rate ρ_t (Figure 8) is only slightly better off, and it matters more, because it sets the levels of the Fisher rate and thereby the price of convenience. Its posterior tracks

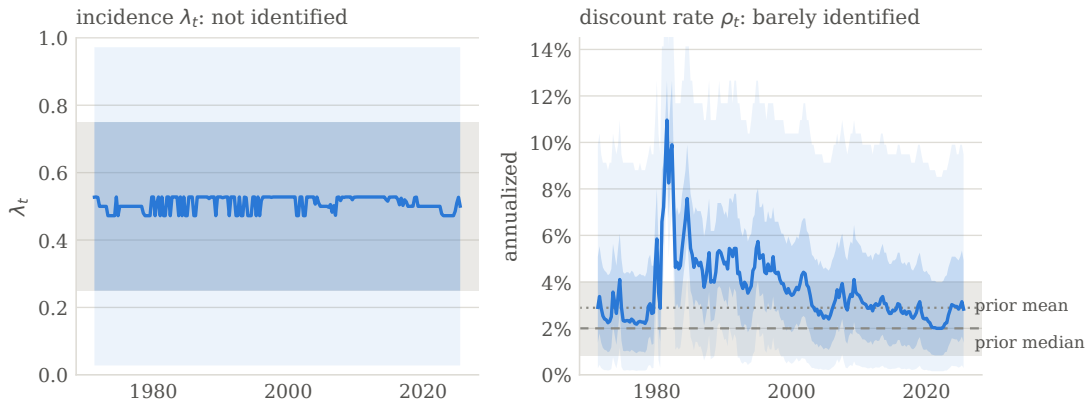

Figure 8: The two objects the data do the least to discipline. Left: the incidence of monetary needs λ_t, posterior median with 50% and 90% credible bands; the grey band is the interquartile range of the $U[0,1]$ prior, which the posterior reproduces exactly. Right: the annualized discount rate $\rho_t = -4 \log \beta_t$, posterior median with 50% and 90% credible bands; the grey band is the prior's interquartile range, and the dashed and dotted lines its median and mean. The 90% band is clipped at the top in 1981–82.

the prior except in the 1980s and 1990s, where the ordering $r_t^F \geq r_t^{AAA}$ pushes it up (see Appendix C.3). Conversely, nothing in the system pushes back against a *high* discount rate except our preconceptions; put in Bayesian terms, our priors.

We report these null results because they have their own scientific relevance, especially the one regarding time preference. We could tighten all the credible bands in our paper, superficially increasing 'precision', simply by narrowing our prior on ρ; taken to the extreme, this is just the common calibration of $\rho = 3\%$ or 4%. Instead, we regard our posterior uncertainty about ρ, r^F, and $\bar{\ell}$ as a finding about what asset pricing equations in a monetary economy can and cannot deliver. See Section 6.2 for more details.

5 Tests and model comparison

Section 4 reported our baseline results; this section reports the specification tests and model comparisons. They are more than robustness checks: they are the only place in this paper where anything can be rejected.

The reason for this is that the baseline is an *exactly identified inversion*: Priors go on the

dimensions the system leaves free (β_t, λ_t and ψ_t), then every equation is imposed exactly, and the remaining latents are solved in closed form.[14] Inverting pricing equations exactly for latent states is the standard construction of the affine term-structure literature (Chen and Scott, 1993; Joslin et al., 2011); pushing a prior over the dimensions the restrictions do not pin down, and imposing inequalities by a one-sided penalty on draws rather than by rejecting them, is the practice of the Bayesian set-identified literature (Uhlig, 2005). The combination of these two methods is motivated by our model: T-bills and commercial paper mature inside the period and are priced without error, while long-term bonds carry an unknown amount of repricing risk and the Fisher rate reflects an unknown degree of time preference.

An exactly identified system, however, has no residuals, and therefore nothing to conduct tests and model selection on. Every test below is accordingly conducted in an alternative, *quasi-Bayesian* specification (Chernozhukov and Hong, 2003): the pricing equations are allowed to fail by residuals, the residuals are given a Gaussian pseudo-likelihood at a stated tolerance, and the product with the prior is a quasi-posterior. Put differently, our instrument is calibrated and justified in an environment (the quasi-Bayesian one) where it is allowed to be wrong, and then used in one (the baseline) where it is not.

5.1 Four estimators, compared

Table 1 compares four estimators. What separates them is how many of the system's free dimensions each one draws, which is the first block of the table. Without the repricing wedge ψ_t, the system has four equations and six free latents, leaving two free dimensions per quarter; including the wedge makes it seven and three. An estimator that draws exactly as many dimensions as the system leaves free solves for everything else and has no residuals; one that draws fewer, or imposes a restriction across quarters, must let equations fail.

The *baseline* draws $(\beta_t, \lambda_t, \psi_t)$ and solves the rest (Appendix C.5). The *exact inversion* is its $\sigma_y \to 0$ limit, so $\psi_t \equiv 0$ (Appendix C.2). *Constant* μ restricts μ to its profile maximum

[14] Trend nominal consumption growth γ_t is not one of those dimensions and carries no prior of its own: throughout our paper, we filter it from nominal consumption growth as a first stage, and what enters each draw is a value from the measurement distribution the filter itself produces.

and forces whatever wedge ψ_t that requires, quarter by quarter (Appendix C.7). Finally, the *equation-error* specification carries no wedge and draws four latents from priors against two degrees of freedom, so Equations (19) and (20) must absorb the excess as residuals penalized at 25 basis points a year (Appendix C.9); that is what allows the data to reject it.

Table 1: Four estimators. All statistics are sample medians (across the per-quarter posterior medians), annualized where applicable.

	baseline (wedge)	exact ($\psi \equiv 0$)	constant μ	equation-error
Latents drawn	β, λ, ψ	β, λ	β, λ (μ imposed)	$\beta, \lambda, \omega, \mu$
Drawn / free	3 / 3	2 / 2	2 / 2	4 / 2
Price of convenience $\bar{\ell}$	**4.9%**	4.8%	4.0%	4.2%
direct component $(1-\omega)\ell^D$	**3.7%**	4.3%	2.7%	3.3%
indirect/urgent component $\omega\ell^N$	**0.87%**	0.78%	0.78%	0.67%
urgent share $\omega\ell^N/\bar{\ell}$	**0.180**	0.159	0.208	0.199
Direct liquidity premium ℓ^D	**6.6%**	6.5%	5.8%	5.5%
Indirect liquidity premium ℓ^N	**3.9%**	4.1%	2.9%	2.9%
Urgency ω	**0.33**	0.23	0.41	0.34
Marketability μ	**0.49**	0.46	0.47	0.41
Annual discount factor $e^{-\rho}$	**0.968**	0.971	0.974	0.975
Quarters with an estimate	218	184	218	218

The point of the table is the agreement. Four estimators that differ in which equations are imposed exactly, which latents are drawn, and whether marketability moves at all put the median price of convenience between 4.0% and 4.9% and median marketability between 0.41 and 0.49. The urgent share is tighter still, 0.159 to 0.208: the composition is governed by observed yields whatever the estimator does with the overall level.[15]

Figure 9 shows that this agreement holds over time as well, not just on average. For urgency and the urgent-share (left panels), the four series share every turn and every spike, including the four large ones; what separates them is the height of those peaks and a modest level difference through the 1980s and 1990s. The price of convenience (bottom right) varies slightly in levels and hardly at all in shape.

Marketability is the object on which the specifications differ most, which has a mechani-

[15] Table C.3 varies the priors, tolerances and ξ within the baseline itself. The urgent share barely moves; the level moves with ξ and with the prior on ρ.

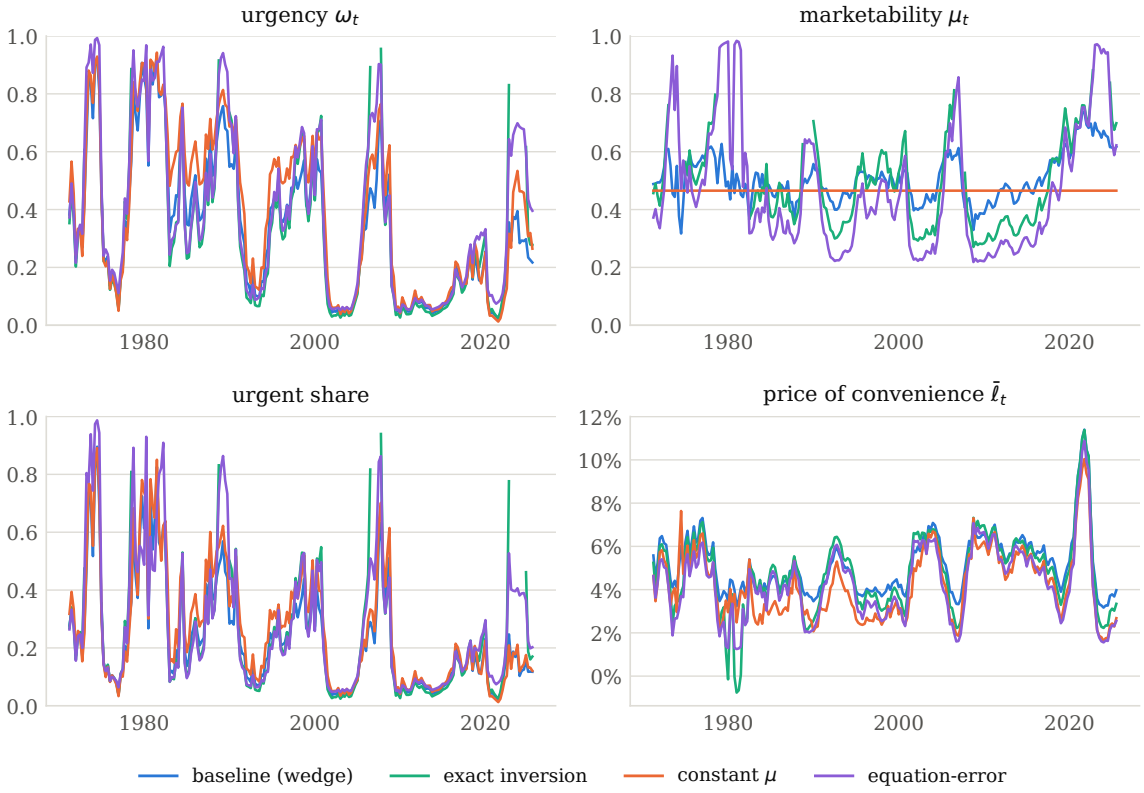

Figure 9: The four estimators of Table 1, posterior medians, each over the quarters in which it is defined: the exact inversion has urgency and marketability undefined in the 34 quarters in which $r^{CP} > y^{AAA}$. The horizontal line in the top right panel is the constant-μ estimator.

cal source: μ_t is solved from (20) where the system is exactly identified and drawn against a residual where it is not, so it is the object an equation error is most able to absorb. Even so, every measurement in this paper (Tables 1, C.2, and C.3) places median μ between 0.41 and 0.57. Among them, varying the prior on the repricing wedge ψ_t by over an order of magnitude only moves median μ_t by 0.06 (Table C.2). A constant μ can also be profiled because the latents are independent across quarters and the quasi-likelihood factorizes: the profile is peaked at $\widehat{\mu} = 0.465$, below the baseline median of 0.494 but well inside the latter's posterior interquartile range, with a block-bootstrap 90% interval of $[0.36,\ 0.54]$ (Appendix C.7).

What the data say about marketability is therefore not that it can be anything. It is that a private claim delivers a significant fraction of the liquidity services of a Treasury bill, whichever way the system is estimated, that the best single value for the whole sample is just

under one half, and that the corner of zero where much of the literature resides is excluded by a wide margin.[16]

5.2 An encompassing test of the AAA repricing wedge

The equation-error estimator can be run with the wedge ψ_t drawn *in addition to* the residual on the AAA equation (20). If the wedge, expressing uncertainty about next period's AAA yield, is the right model of any error in the equation, then the residual on that equation should collapse once the wedge is available.

We find that this is exactly what happens. Without the wedge, the absolute residual at the posterior median has a median across quarters of 4.2bps per year on (19) and 4.8bps on (20) (means: 14.5bps and 16.9bps): the two private-asset equations fit about equally well, which is what one expects if their errors have a common source. With the wedge drawn from (21) they are 7.0bps and 1.3bps (means: 7.1bps and 1.6bps). The AAA equation's residual falls by a large amount, while the commercial-paper equation does not share in the improvement: its typical residual rises, and only its mean, driven by a few large misses, falls. The asymmetry makes the point. The wedge does not relax the system uniformly; it draws the AAA equation's error onto itself.

We report this as a test and not as a robustness column, because in this specification the wedge and the residual are two devices for the same discrepancy in (20); thus, marketability and the urgent share are identified only jointly. The purpose of the specification to assess the change in residuals is not affected by this problem.

5.3 Persistence

Two further comparisons are available in the equation-error specification (Appendix C.10). First, latents that move smoothly beat independently drawn ones by a large margin. Among all latents, persistence matters most for the wedge ψ_t, with an implied lag-1 autocorrelation

[16] How wide of a margin depends on the prior on the wedge: at $\sigma_y = 2$bps the corner $\mu_t \equiv 0$ loses by 107 log points, at our baseline 5bps by 56 log points, and at 20bps still by 2 log points. As explained above, the latter represents an *enormously* agnostic view of what investors might have assumed about yield movements.

of 0.96, and a constant wedge is firmly rejected. Second, when we allow each latent to jump, this is selected for urgency ω_t only: the data want it near-constant between episodes and completely redrawn about once every ten quarters, which is the shape of a liquidity scramble rather than of a drift.

As above, these are diagnostics and not rival estimates, since the additional flexibility of the estimation procedure costs identification. Nor do they serve as a test *against* our baseline estimates, which are already smooth with no kernel imposed: the baseline posterior-median ω_t has a realized lag-1 autocorrelation of 0.93 and μ_t of 0.89, against 0.83 and 0.99 under the fitted kernel.

6 Identification and interpretation

Section 3 stated the assumptions under which the three observed rates measure the objects of Section 2, and Sections 4 and 5 reported what follows from them. Here we challenge and defend the most critical ones: that the two private instruments are equally marketable, and that a consumption Euler equation can be asked to price a claim nobody trades. Additionally, we discuss whether urgency is just a relabelling of the nominal interest rate (it is not), and whether illiquidity is a term premium under another name (it is, and much more).

6.1 Commercial paper, AAA bonds, and their marketability

Our key identifying assumption is that commercial paper and AAA bonds share a common marketability parameter μ. We make this assumption foremost because *something* has to be assumed, and *among all simple restrictions* we find this one a priori most plausible. However, that does not insulate it from challenges, and here we review (some of) them in depth.

The economic content of a common μ is that what constrains the secondary-market sale of a private claim is the capacity of the dealers who intermediate it (Lagos, Rocheteau, and Weill, 2011; Cohen, Kargar, Lester, and Weill, 2024; Fleskes, 2025; Wittwer and Allen, 2026), not the identity of the security. Against that, AAA bonds trade in an over-the-counter, dealer-inventoried market, while commercial paper rarely does; on the other hand, both are fi-

nanced on the same balance sheets, and (crucially) both stand well outside the arrangements that make Treasuries far more marketable. Those arrangements have changed over the years. In the 1970s, the Treasury market alone had a designated dealer network quoting continuously and a central bank transacting in it as a matter of routine; today, the same asymmetry is written into rules, a T-bill being eligible at the Standing Repo Facility and as a Level 1 liquid asset for regulatory purposes whereas a high-grade corporate bond is Level 2A and commercial paper is not considered a liquid asset at all. The modern designations postdate most of our sample and cannot themselves tell us anything about 1975; the common throughline is that the constraint on selling a private claim runs through dealer capacity rather than through the identity of the security.

What a common μ determines, versus distinct values of μ^{CP} and μ^{AAA}, is the split of a given price of convenience into direct and urgent parts (Proposition 3.1). The alternative implied in traditional convenience yields, $\mu_t^{CP} = 0$, is the floor of the identified set, asserting that commercial paper is literally unmarketable (on top of not having a monetary value for non-urgent needs). Allowing μ^{CP} and μ^{AAA} to differ by a constant rescales the urgent share without reshaping it, and we can ask how far that could plausibly go (Appendix C.6): for the urgent share of convenience to rise to 0.5 in the median quarter, μ^{CP} would have to exceed μ^{AAA} by about 0.27. At the floor of $\mu_t^{CP} = 0$, conversely, the urgent share falls to 0.087 instead of 0.180 in the baseline. Hence we find that our assumption of common μ, made on a priori grounds, also delivers implications that do not fail any obvious plausibility test.

In our model, $1 - \mu_t$ can be interpreted as the probability of not being able to trade in the secondary market, or equivalently, as the fraction of an asset that cannot be used as collateral – in other words, the haircut. Taken literally, the second interpretation does not work: repo haircuts on money-market instruments and corporate bonds are typically about five percent (Copeland, Duffie, Martin, and McLaughlin, 2012; Choi, Jang, Kang, and Kim, 2026), not the 51% our estimate implies. However, the two are fractions of different things. A haircut is the price of pledging a security *conditional* on already having a counterparty, a line (Lebeau, 2020) and a security the schedule admits; transaction-cost measures put the median

round-trip cost of a corporate bond at some five times its quoted bid–ask spread (Bao, Pan, and Wang, 2011; Juelsrud et al., 2025), and commercial paper trades in a secondary market accounting for under a tenth of its activity (Gross, Li, and Wang, 2022). Thus, our μ_t is to be interpreted as an *effective* pledgeability, to be read against market depth and dealer capacity – a friction that search theory prices (Duffie, Gârleanu, and Pedersen, 2005; Weill, 2008; Lagos and Rocheteau, 2009; Uslu, 2019; Uslu and Velioglu, 2025) and that measurement finds larger for corporate than for government bonds, and larger still under stress (Kargar, Lester, Lindsay, Liu, Weill, and Zuniga, 2021; Pinter, Uslu, and Wijnandts, 2025).

Spread-based measures of convenience in the literature almost all assume a corner: Krishnamurthy and Vissing-Jorgensen (2012) read the AAA–Treasury spread as the convenience yield on Treasuries, which is $\mu^{AAA} = 0$; Nagel (2016) takes the general-collateral repo rate as the illiquid benchmark, $\mu^{repo} = 0$; van Binsbergen, Diamond, and Grotteria (2022) recover a risk-free rate from option box spreads on the assumption that risky assets earn no convenience yield.[17] The same literature assumes the opposite when the question changes: Sunderam (2015), Greenwood et al. (2015) and Krishnamurthy and Li (2023) all require $\mu^{CP} > 0$, and materially so, as do Cipriani and La Spada (2021), who price the moneyness that prime money-market funds lost in the 2014 reform at 20–30 basis points. The magnitudes are compatible with ours: van Binsbergen et al. (2022) put the Treasury convenience yield at about 65 basis points inside three months, against our commercial-paper spread's mean of 59, which is what (27) implies for a short-dated claim.

6.2 The consumption Euler equation

The level of the price of convenience is a statement about the Fisher rate, the return on a hypothetically perfectly illiquid benchmark asset, and the data have very little to say *directly* on the subject. The consumption Euler equation, after all, is the restriction in macrofinance with the longest record of failure. The yield on AAA bonds supplies a lower bound in our model, since the Fisher rate must exceed their expected returns, but no upper bound.

[17] Studies that avoid the corner include Geromichalos, Herrenbrueck, and Lee (2023a,b) and Geromichalos, Herrenbrueck, Lee, and Lee (2024); these provide calibrations and not estimates.

The first term in the Fisher rate is the time discount rate. Here, we report a pointwise mapping: over the sample, the median price of convenience is 1.0% at a zero discount rate, 4.1% at $\rho = 3\%$ and 7.3% at $\rho = 6\%$ (Appendix C.4). Replacing our prior (exponential with median 2%) by an exponential with median 3% moves the posterior median discount rate from 3.3% to 4.5%. A reader who believes $\rho = 1\%$ and a reader who believes $\rho = 6\%$ will therefore take away prices of convenience of 2.0% and 7.3%. Outside evidence does not narrow the range from above: microeconomic studies consistently find discount rates of 10% and higher (Andersen, Harrison, Lau, and Rutström, 2008; Andreoni and Sprenger, 2012), and not constant (Harrison, Lau, and Yoo, 2025). Macroeconomists conventionally honour such findings by disregarding them. If they did not, a much higher prior on the time discount rate and a much worse violation of the traditional Equler equation would follow – and so would a much larger implied price of convenience.

Even granting our cautious approach to the time discount rate, our paper is still open to a challenge that our identification of marketability, and our finding $\mu \approx 0.5$, is an artifact of a mismeasured Fisher rate and the literature's $\mu = 0$ therefore survives. Half of this challenge is fair: μ_t is not identified without r_t^F. Of the model's objects, only functions of the ratio of the two spreads are free of the Fisher rate, and $\mu_t = \Delta_t(r_t^{AAA})/\Delta_t(r_t^{TB})$ depends on the Fisher rate through both of its arguments. The 56 log points by which a constant μ beats $\mu = 0$ in Section 5 are obtained under the prior on β_t, and are not free of it either.

What does not follow is that the objection is unanswerable, because $\mu_t = 0$ is not a vague skepticism about the Fisher rate but an exact statement about it. Setting $\Delta_t(r_t^{AAA}) = 0$ means $r_t^F = r_t^{AAA}$: the return on a claim of no monetary usefulness whatsoever *is* the yield on a long corporate bond. Since $1 + r_t^F = e^{\gamma_t}/\beta_t$ by (16), that pins the discount rate at $\rho_t = 4\big[\log(1 + r_t^{AAA}) - \gamma_t\big]$, and the objection reduces to a claim about the time discount rate. The table below (reproducing Table C.1 in part) reads that mapping over the whole sample.

ρ	median μ_t, all 218 quarters	median μ_t, ordering holds at that ρ	quarters out of order
0%	$\mathbf{0.25}$	0.32	174
1%	0.19	0.40	140
2%	0.25	0.40	108
3%	0.41	0.43	78
4%	0.53	0.48	47
6%	0.66	0.61	34

No defensible discount rate delivers $\mu = 0$: over the whole sample, median marketability never falls below 0.17, which it reaches at $\rho \approx 1.1\%$. (It is not monotone in ρ, because below the T-bill rate both premia change sign.) Quarter by quarter, holding $\mu_t = 0$ throughout forces ρ_t below zero in 63 quarters, with 8 changes of sign and a swing of 4.8 percentage points across decade medians against 2.8 for our baseline ρ_t estimates.

A reader unwilling to grant the Euler equation will be unwilling to grant $u'(c) = 1/c$ with it, but relaxing it cuts both ways. Under CEIS preferences with elasticity of intertemporal substitution $1/\xi$ (or, equivalently here, CRRA utility with coefficient ξ), the Fisher rate is $\rho + \xi g_t + \pi_t$, where g_t is trend *real* consumption growth, so the discount rate that $\mu_t = 0$ requires in each quarter shifts by $(1 - \xi)g_t$; g_t has a median of 1.8% a year on the same filter. At $\xi = 2$, an elasticity of one half, $\mu = 0$ requires a median ρ_t of -0.4%, negative in 122 of 218 quarters, and the median ρ_t corresponding to $\mu = 0$ reaches zero at $\xi = 1.73$. In this sense, $\xi = 1$ is a scientifically cautious assumption that makes it harder for us to conclude that $\mu > 0$ and that the price of convenience is large.

But a level is not all that ξ moves. Because g_t and π_t have to be filtered separately once $\xi \neq 1$, and because trend real growth fell from about 2.5% to about 1.2% over this sample while trend inflation went from 11% to 1% and part of the way back, a different ξ changes the *path* of the Fisher rate and with it the whole time profile of μ_t. We therefore re-estimate at $\xi \in \{1/2,\ 2/3,\ 1,\ 3/2,\ 2\}$, filtering real growth with the same local-level Student-t filter, the same one-sided extraction and the same lead as the nominal trend:

ξ	$\bar{\ell}$	μ	urgent share	μ_{2008Q4} (percentile)	μ_{2022} (percentile)
1/2	4.6%	0.49	0.190	0.33 (0)	0.64 (95)
2/3	4.7%	0.49	0.187	0.33 (0)	0.66 (96)
1	**4.9%**	**0.49**	**0.180**	**0.33 (0)**	**0.69 (98)**
3/2	5.3%	0.51	0.173	0.33 (0)	0.73 (97)
2	5.9%	0.53	0.168	0.32 (0)	0.76 (96)

Three things follow. The levels move as much as the discount rate moves them – a factor of ξ from a half to two carries the price of convenience from 4.6 to 5.9 percent – so ξ belongs with ρ on the list of things a prior must supply. The composition is far steadier, the urgent share running 0.190 to 0.168 across the same range. The exercise also says about ξ the same one-sided thing the model says about ρ: with the wedge switched off and $\rho = 3\%$ imposed, the ordering fails in 107 quarters at $\xi = 1/2$ and 96 at $2/3$, against 78 at log utility, 52 at $\xi = 3/2$ and 43 at $\xi = 2$. Admissibility pushes the elasticity *down* exactly the way it pushes the discount rate up, because both raise the Fisher rate and the only binding restriction is that the Fisher rate clear the AAA yield.

Outside evidence points in the same direction. Di Tella, Hébert, Kurlat, and Wang (2026) recover a close counterpart to our Fisher rate, the zero-beta rate, from equity returns, and find that it tracks aggregate consumption where observed safe rates do not; Di Tella, Hébert, and Kurlat (2024) trace the difference to a large and volatile liquidity premium on safe assets. Their estimates favour an elasticity of one half or less, that is $\xi \geq 2$, which on our own reading raises the Fisher rate and the price of convenience further still. However, their inference runs on the identifying assumption that equities carry no convenience at all, which we find implausible: equities are certainly less acceptable as collateral than high-grade bonds, but they trade in highly liquid markets, and Geromichalos and Herrenbrueck (2022) found that long-run equity returns track the T-bill rate more closely than the Fisher rate (which is, via our model, the signature of a liquid asset).

Ultimately, we maintain log utility because the elasticity is not a parameter this literature has settled – published estimates span an order of magnitude, and the disagreement is about

identification rather than about samples – and we record that it is an identifying assumption rather than a normalization. Mechanically, it is what makes nominal consumption growth a sufficient statistic for the Fisher rate, with an attendant robustness benefit: the estimator never has to separate real growth from inflation.

6.3 Urgency and the level of nominal rates

Urgency is strongly correlated with the T-bill rate, at $+0.86$, and so is the urgent share, at $+0.73$ with a 90% interval of $[+0.60, +0.82]$ (Appendix C.8). The share's correlation is the cleaner evidence: the share is the identity (29) in observed spreads, whereas ω_t is solved from an equation in which the bill rate enters directly. Both load on the *nominal* bill rate rather than the real one. One plausible interpretation of this is that urgency reflects money demand, where the natural opportunity cost is the nominal rate and not the real one. Treating urgency as endogenous is a natural extension, but beyond the scope of the present paper.

The correlation is not so perfect as to amount to mere relabelling, however. The bill rate accounts for about half of the variance of the urgent share, and the mapping is not monotone across decades: the 1990s carry the same urgency as the 1970s at a lower nominal rate, and the 2000s nearly the same as the 2020s at half the rate. Conditional on the bill rate there is no residual time trend in either series. The decade profile of Section 4.1 therefore tracks the rate cycle rather than a secular change in the nature of liquidity needs.[18] Appendix C.8 gives the decade medians, the first differences and the subsample checks.

6.4 The term premium as a liquidity premium

In our system a long AAA bond earns more than a Treasury bill for one reason: a holder who needs money cannot use it as money, and can sell only a fraction μ_t of it into a dealer market. Equation (28) converts that into an accounting identity – the AAA–bill spread equals $(1-\mu_t)$ times the entire price of convenience. This is an accounting identity, not a restriction or a

[18] We have also checked the slope of the yield curve, the other obvious candidate for a common driver: regressing urgency on said slope and the T-bill rate, the coefficient on the slope is significant but the coefficient on the T-bill rate is unchanged, so the slope does not displace the level.

finding. There is no residual left over for a term premium in the usual sense, and the model has no mechanism for generating one; our estimates put $1 - \mu_t$ at about 0.51 and the price of convenience at about 4.9%, so a long–short spread of $\approx 2.5\%$ is half the price of convenience rather than a risk premium of unexplained origin.

That raises the question of what the usual sense is. Why do we believe long bonds carry term premia? Almost entirely because long bonds outearn short ones on average and the expectations hypothesis says they should not: a measurement, not a mechanism. The literature names the residual and supplies mechanisms afterwards – covariance with consumption growth, with a stochastic discount factor, with inflation – but the central object remains a residual with a name. Thus, "term" is a labelling choice, and arguably a less defensible one than "liquidity" given that the convenience of Treasuries in particular is perfectly well established. Tacitly, this choice is made by every study that fits the whole yield curve with a pricing kernel in which every maturity carries the same μ, namely zero. The liquidity mechanism has a normative counterpart as well, in which optimal debt management makes long-term government debt deliberately less tradable than short-term debt, and so prices it at a discount (Andolfatto, 2020; Andolfatto and Martin, 2023).

Two qualifications apply. First, our claim is about the average level of the long–short spread and its low-frequency movement, not about the day-to-day dynamics the affine term-structure literature is built to fit; we have four series at quarterly frequency and no ambition to compete on that ground. Second, the identity holds by construction, but reading it as a liquidity premium requires μ_t to be a fraction, and so rests on the assumptions discussed in Sections 6.1 and 6.2.

7 Conclusion

Convenience has a price and it has a quantity, and any spread between two assets is the product of the two. In order to separate the product, we construct a monetary asset pricing model with claims that differ by liquidity and maturity. Maturity matters because a short-term asset is liquid by nature and not by the grace of its market, in that it matures before

urgent payment needs arrive. Market liquidity matters for the rest: turning an asset into money whenever money *is* needed urgently.

What the separation reveals is that conventional measures of the price of convenience are small because they are measuring a corner of the object. A spread between two short assets can only price urgent liquidity needs, and only the unmarketable fraction of those. Urgent needs are in the minority, representing only a median of 18% of the aggregate price of convenience (mean: 24%), and the marketability of a private claim is about half of that of a Treasury bill on top of that. Accordingly, our estimated price of convenience for the U.S. since 1971 has a median of 4.9% a year (mean: 5.1%), against a median of 44 basis points (mean: 59) of the commercial-paper–T-bill spread, the conventional measure of the convenience yield. The conventional measure and ours are negatively correlated, and thereby disagree about when convenience has been most valued.

APPENDIX

Appendix A collects the derivations omitted from Section 2. Appendix B describes the data, including the commercial-paper splice and the construction of the synthetic holding-period returns used to bound the wedge. Appendix C elaborates upon the estimation approaches our paper did (or did not) adopt.

A Omitted derivations

Solutions to the agents' problems. From the CM problem (3), linearity in g_t implies the envelope conditions:

$$\partial_m V^{CM} = \partial_{b^T} V^{CM} = \partial_{b^S} V^{CM} = \frac{\partial_{b^L} V^{CM}}{d_t + p_t^L} = \frac{\partial_y V^{CM}}{p_t} = \frac{u'(c_t)}{p_t}. \tag{30}$$

The PM problem (4) implies that, as long as $q_t \leq 1$, the payment constraint binds: $y_t = \Lambda_t\big(m_t + (1-\Omega_t)(b_t^T + b_t^S)\big)/(p_t q_t)$. The first-order conditions of the AM problems (5)–(7), with θ_t^M denoting the multiplier on the money constraint and θ_t^j, $j \in \{T, S, L\}$, the multipliers on the marketability constraints, are:

$$\text{inactive:} \qquad \partial_m V^{CM} + \theta_t^M = \frac{\partial_{b^j} V^{CM}}{s_t^j}, \qquad j \in \{T, S, L\}, \tag{31}$$

$$\text{early:} \qquad \partial_{b^j} V_{11}^{PM} + \theta_t^j = s_t^j \, \partial_m V_{11}^{PM}, \qquad j \in \{T, S, L\}, \tag{32}$$

$$\text{late:} \qquad \partial_m V_{10}^{PM} + \theta_t^M = \frac{\partial_{b^j} V_{10}^{PM}}{s_t^j}, \qquad j \in \{T, S\}, \tag{33}$$

$$\partial_{b^L} V_{10}^{PM} + \theta_t^L = s_t^L \, \frac{\partial_{b^T} V_{10}^{PM}}{s_t^T}, \tag{34}$$

where, for ease of exposition, we write $V_0^{AM} = V_{\Lambda_t=0}^{AM}$, $V_{11}^{AM} = V_{\Lambda_t=1,\Omega_t=1}^{AM}$ (early/urgent), and $V_{10}^{AM} = V_{\Lambda_t=1,\Omega_t=0}^{AM}$ (late). Absence of arbitrage among the three interest-bearing assets for inactive shoppers pins down the relative AM prices:

$$\frac{\partial_{b^T} V^{CM}}{s_t^T} = \frac{\partial_{b^S} V^{CM}}{s_t^S} = \frac{\partial_{b^L} V^{CM}}{s_t^L} \quad \Longleftrightarrow \quad s_t^S = s_t^T \text{ and } s_t^L = \big[d_t + p_t^L\big]\, s_t^T. \tag{35}$$

The two short bonds therefore trade at the same AM price, as they must: they pay the same amount of money at the same date, and they differ in a quantity constraint rather than in price. The marginal values of the four assets, by shopper type, are then:

$$\partial_m V_0^{AM} = \frac{\partial_{b^T} V^{CM}}{s_t^T} = \frac{u'(c_t)}{s_t^T p_t}, \tag{36}$$

$$\partial_m V_{11}^{AM} = \partial_m V_{11}^{PM} = \frac{\partial_y V^{CM}}{p_t q_t} = \frac{u'(c_t)}{p_t q_t}, \tag{37}$$

$$\partial_m V_{10}^{AM} = \frac{\partial_{b^T} V_{10}^{PM}}{s_t^T} = \frac{u'(c_t)}{s_t^T p_t q_t}, \tag{38}$$

$$\partial_{b^T} V_0^{AM} = \partial_{b^S} V_0^{AM} = \partial_{b^T} V^{CM} = \frac{u'(c_t)}{p_t}, \tag{39}$$

$$\partial_{b^T} V_{11}^{AM} = s_t^T \, \partial_m V_{11}^{PM} = \frac{s_t^T}{q_t} \, \frac{u'(c_t)}{p_t}, \tag{40}$$

$$\partial_{b^S} V_{11}^{AM} = \mu_t s_t^T \, \partial_m V_{11}^{PM} + (1-\mu_t) \, \partial_{b^S} V_{11}^{PM} = \left(\mu_t \frac{s_t^T}{q_t} + 1 - \mu_t \right) \frac{u'(c_t)}{p_t}, \tag{41}$$

$$\partial_{b^T} V_{10}^{AM} = \partial_{b^S} V_{10}^{AM} = \partial_{b^T} V_{10}^{PM} = \frac{u'(c_t)}{p_t q_t}, \tag{42}$$

$$\partial_{b^L} V_0^{AM} = \partial_{b^L} V^{CM} = \left(d_t + p_t^L \right) \frac{u'(c_t)}{p_t}, \tag{43}$$

$$\partial_{b^L} V_{11}^{AM} = \mu_t s_t^L \, \partial_m V_{11}^{PM} + (1-\mu_t) \, \partial_{b^L} V_{11}^{PM} = \left[1 - \mu_t + \frac{\mu_t s_t^T}{q_t} \right] \left(d_t + p_t^L \right) \frac{u'(c_t)}{p_t}, \tag{44}$$

$$\partial_{b^L} V_{10}^{AM} = \mu_t \frac{s_t^L}{s_t^T} \, \partial_{b^T} V_{10}^{PM} + (1-\mu_t) \, \partial_{b^L} V_{10}^{PM} = \left[1 - \mu_t + \frac{\mu_t}{q_t} \right] \left(d_t + p_t^L \right) \frac{u'(c_t)}{p_t}. \tag{45}$$

The first-order conditions of the CM problem, taking expectations over next period's shopper types (where a fraction $\lambda_{t+1}\omega_{t+1}$ are early/urgent, $\lambda_{t+1}(1-\omega_{t+1})$ are late, and $1-\lambda_{t+1}$ are inactive), are:

$$\frac{u'(c_t)}{p_t} = \beta \, \mathbb{E}\left\{ \lambda_{t+1}(1-\omega_{t+1}) \, \partial_m V_{10}^{AM} + \lambda_{t+1}\omega_{t+1} \, \partial_m V_{11}^{AM} + (1-\lambda_{t+1}) \, \partial_m V_0^{AM} \right\}, \tag{46}$$

$$\frac{u'(c_t) \, p_t^T}{p_t} = \beta \, \mathbb{E}\left\{ \lambda_{t+1}(1-\omega_{t+1}) \, \partial_{b^T} V_{10}^{AM} + \lambda_{t+1}\omega_{t+1} \, \partial_{b^T} V_{11}^{AM} + (1-\lambda_{t+1}) \, \partial_{b^T} V_0^{AM} \right\}, \tag{47}$$

$$\frac{u'(c_t) \, p_t^S}{p_t} = \beta \, \mathbb{E}\left\{ \lambda_{t+1}(1-\omega_{t+1}) \, \partial_{b^S} V_{10}^{AM} + \lambda_{t+1}\omega_{t+1} \, \partial_{b^S} V_{11}^{AM} + (1-\lambda_{t+1}) \, \partial_{b^S} V_0^{AM} \right\}, \tag{48}$$

$$\frac{u'(c_t) \, p_t^L}{p_t} = \beta \, \mathbb{E}\left\{ \lambda_{t+1}(1-\omega_{t+1}) \, \partial_{b^L} V_{10}^{AM} + \lambda_{t+1}\omega_{t+1} \, \partial_{b^L} V_{11}^{AM} + (1-\lambda_{t+1}) \, \partial_{b^L} V_0^{AM} \right\}. \tag{49}$$

Substituting (35) and (36)–(45) into (46)–(49) yields the Euler equations for the four assets:

$$\frac{u'(c_t)}{p_t} = \beta \, \mathbb{E}\left\{ \frac{u'(c_{t+1})}{p_{t+1}} \left[\frac{\lambda_{t+1}(1-\omega_{t+1})}{s_{t+1}^T q_{t+1}} + \frac{\lambda_{t+1}\omega_{t+1}}{q_{t+1}} + \frac{1-\lambda_{t+1}}{s_{t+1}^T} \right] \right\}, \tag{50}$$

$$\frac{u'(c_t)\,p_t^T}{p_t} = \beta\,\mathbb{E}\left\{\frac{u'(c_{t+1})}{p_{t+1}}\left[\frac{\lambda_{t+1}(1-\omega_{t+1})}{q_{t+1}} + \frac{\lambda_{t+1}\omega_{t+1}s_{t+1}^T}{q_{t+1}} + 1 - \lambda_{t+1}\right]\right\}, \tag{51}$$

$$\frac{u'(c_t)\,p_t^S}{p_t} = \beta\,\mathbb{E}\left\{\frac{u'(c_{t+1})}{p_{t+1}}\left[\frac{\lambda_{t+1}(1-\omega_{t+1})}{q_{t+1}} + \lambda_{t+1}\omega_{t+1}\left(1 - \mu_{t+1} + \frac{\mu_{t+1}s_{t+1}^T}{q_{t+1}}\right) + 1 - \lambda_{t+1}\right]\right\}, \tag{52}$$

$$\frac{u'(c_t)\,p_t^L}{p_t} = \beta\,\mathbb{E}\left\{\left(d_{t+1} + p_{t+1}^L\right)\frac{u'(c_{t+1})}{p_{t+1}}\left[1 - \lambda_{t+1}\mu_{t+1} + \frac{\lambda_{t+1}\mu_{t+1}\left[\omega_{t+1}s_{t+1}^T + 1 - \omega_{t+1}\right]}{q_{t+1}}\right]\right\}, \tag{53}$$

The Euler equations share a common structure: the cost of obtaining the asset is matched on the right-hand side by a fundamental return multiplied by a liquidity premium term that accounts for the asset's usefulness in AM or PM trade. Money serves urgent needs at full value ($1/q$); for non-urgent needs it is dominated by first buying maturing bonds ($1/(s^T q) \geq 1/q$). Both short bonds serve non-urgent needs at full value; the T-bill also serves urgent needs in full, because it can always be sold, while S serves them only to the extent μ. The consol serves any monetary need only to the extent μ that it can be sold.

AM market clearing. Equating the demand for money in the AM to the supply of saleable assets gives:

$$\frac{1}{s_t^T} = \frac{\lambda_t\omega_t}{1-\lambda_t\omega_t} \cdot \frac{B_t^T + \mu_t B_t^S + \mu_t\left(d_t + p_t^L\right)B_t^L}{M_t}. \tag{54}$$

Thus, the AM price of T-bills depends on the demand for immediate liquidity, $\lambda_t\omega_t/(1-\lambda_t\omega_t)$, which arises from *urgent* needs only, and on the supply of saleable assets relative to money. Equation (54) can be inverted to obtain the money supply consistent with any target for the policy rate. This is how we think of the implementation of monetary policy: the monetary authority sets a *policy rate*, which for our purposes is close to the T-bill rate, and never the Fisher rate (Andolfatto, 2015; Herrenbrueck, 2019; Rocheteau, Wright, and Xiao, 2018).

Money spent in the PM and the price level. In the AM, late shoppers obtain $1/s_t^T$ short bonds per unit of money. The total money (including matured bonds) spent in the PM is:

$$M_t^{PM} = \underbrace{\lambda_t\omega_t M_t}_{\text{money carried by early shoppers from the CM}}$$

$$+ \underbrace{s_t^T \lambda_t \omega_t \left[B_t^T + \mu_t B_t^S + \mu_t (d_t + p_t^L)\, B_t^L \right]}_{\text{money obtained by early shoppers in the AM}}$$

$$+ \underbrace{\frac{\lambda_t (1-\omega_t) M_t + s_t^T \lambda_t (1-\omega_t)\, \mu_t (d_t + p_t^L)\, B_t^L}{s_t^T}}_{\text{short bonds obtained by late shoppers in the AM}}$$

$$+ \underbrace{\lambda_t (1-\omega_t) \left[B_t^T + B_t^S \right]}_{\text{short bonds carried by late shoppers from the CM}} \tag{55}$$

Substituting the AM market clearing condition (54) and simplifying:

$$M_t^{PM} = \left[1 + \frac{1-\omega_t}{\omega_t} \cdot \frac{1}{s_t^T} \right] M_t + \lambda_t (1-\omega_t)(1-\mu_t)\, B_t^S. \tag{56}$$

The T-bill leaves no residual term because it is fully marketable; only the partly marketable short bond does. Since all workers together sell C_t at the PM price $q_t p_t$, the price level satisfies $p_t = M_t^{PM}/(q_t C_t)$.

Steady state. Although our empirical strategy works with the stochastic Euler equations, the steady state is useful for intuition and for comparison with the predecessor papers. Assume the money stock, and thereby nominal consumption, grows at rate $\bar{\gamma}$ per period. Dropping time subscripts and treating $1 + r^F$, s^T as constants, the Euler equation for money (50) becomes:

$$\frac{1+\bar{\gamma}}{\beta} = \frac{(1-\omega)\lambda}{q\, s^T} + \frac{\omega\lambda}{q} + \frac{1-\lambda}{s^T}, \tag{57}$$

which solves for the PM relative price:

$$q = \frac{\omega\lambda\, s^T + (1-\omega)\lambda}{(1+\bar{\gamma}) s^T/\beta - 1 + \lambda}. \tag{58}$$

Finally, combining (58), (56), and $p = M^{PM}/(qC)$ gives the steady-state price level, and the AM clearing condition (54) then determines the money supply required to implement a given policy rate.

B Data

All series are quarterly, 1971Q2–2025Q4, and all are drawn from FRED (Federal Reserve Bank of St. Louis). Monthly and daily series are averaged to quarterly frequency before use. The start date is set by commercial paper, the series with the shortest history; 1971Q1 is retained in the raw file so that a growth rate can be formed for 1971Q2.

B.1 The four rates

Treasury bills, r_t^{TB}. `TB3MS`, the 3-month Treasury bill secondary-market rate, monthly, averaged to quarterly. The rates are closing quotations on the *most recently auctioned* bill in each tranche, and the 3-month tranche is the 13-week bill, auctioned weekly; the instrument quoted is therefore at most a few days old. The quote is a *bank discount* rate on a 360-day year and not an effective annual rate: the price per unit of face is $1 - d\,n/360$ with $n = 91$ days, and it is the gross return implied by that price, converted to a common quarterly basis, that enters the system.

AAA corporate bonds, r_t^{AAA}. `AAA`, Moody's Seasoned Aaa Corporate Bond Yield, monthly, averaged to quarterly. Per Moody's, this is an index "based on bonds with maturities 20 years and above"; our constructed holding-period returns assume par coupons and a tenor of 25 years, implying an effective duration of about ten years. We could find no compounding convention documented by Moody's, and we read the yield as an effective annual rate.

Commercial paper, r_t^{CP}. This series is spliced from three FRED series. All three legs are *primary-market offer rates on new issuance*, since the Federal Reserve's series are built from the offer side rather than from secondary bids, so each is directly the issue price of (13). All three are quoted on a discount basis over a 360-day year, at 90 days. No single commercial-paper series spans the sample:

FRED ID	description	quarters	coverage
CP3M	3-Month Commercial Paper Rate (discontinued)	105	1971Q2–1997Q2
CPF3M	3-Month AA Financial CP Interest Rate	117	1997Q1–2025Q4
RIFSPPAAAD90NB	90-Day AA Asset-Backed CP Interest Rate	101	2001Q1–2025Q4

All three are Federal Reserve commercial-paper rates quoted on the same basis, so the splice does not mix quoting conventions; it mixes issuers. The two outer series never overlap, so they cannot be compared directly; the middle one, financial paper, overlaps both. It also lies systematically below both: over the 101 quarters of common coverage the asset-backed rate exceeds the financial rate by 8.62 basis points on average, and over the two quarters where the discontinued series overlaps the financial one the gap is 11.0 basis points. The pooled estimate is 8.67 basis points, and it is always positive in the overlap, with a standard deviation of 8.3 basis points.

We therefore shift the financial series up by that constant, $\widetilde{\texttt{CPF3M}} = \texttt{CPF3M} + 0.0867$, and splice by preference: RIFSPPAAAD90NB where available, else CP3M, else $\widetilde{\texttt{CPF3M}}$. The asset-backed series is preferred for the modern period on grounds of market depth: AA asset-backed paper is issued in several times the volume of AA nonfinancial paper, so its rate is fitted from far more trades and is the less noisy of the two. The adjusted financial series is therefore used in only 14 quarters, 1997Q3–2000Q4, bridging the gap between the discontinued series and the asset-backed one; the other 204 quarters come from the two unadjusted outer series.

Three remarks are in order. First, the offset is small but not negligible relative to the object it enters. The commercial-paper spread over bills averages 59 basis points over the estimation sample, with a median of 44, so 8.67 basis points is about 15% of the mean; in the fourteen quarters to which the adjustment is actually applied the spread averages 69 basis points, so it is nearer 13% of those. Second, its standard deviation across the overlap is as large as its mean, so it is a noisy estimate of a spread that is itself time-varying. Third, it is estimated predominantly from the *post*-2001 overlap and applied to 1997–2000, which assumes that the financial–asset-backed spread was stable across the break. That spread is

not constant elsewhere: it averages 3.5 basis points before 2007 and 20.2 through 2008–09, when asset-backed paper was the instrument under run. We regard this as the least bad of the available splices; the alternative of leaving the 14 quarters unadjusted would put a visible 9-basis-point step in the middle of the series.

The Fisher rate, r_t^F. Constructed rather than observed, from nominal personal consumption expenditure on nondurables and services per capita (`PCND` + `PCES`, divided by population, `B230RC0Q173SBEA`; quarterly) via (16) and the trend model of Appendix C.1. It is the only rate in the system that is not read from a market price.

B.2 Frequency and units

The model period is a quarter, and the system is not invariant to annualization: the product form $(1 + r_t^{TB})(1 + \bar{\ell}_t) = 1 + r_t^F$ survives raising gross returns to the fourth power, but the linear loading $(1 - \omega_t)\ell_t^D + \omega_t \ell_t^N$ in (18) does not. Every quote is therefore put on a common quarter before the system is solved, and results are re-annualized only for reporting, premia as $(1 + \ell)^4 - 1$. For the bond yield that is $1 + i_q = (1 + i_a)^{1/4}$. For the two discount quotes it is two steps: the price per unit of face is $P = 1 - d\,n/360$ over the instrument's own n days, which gives an effective annual rate $(1/P)^{365/n} - 1$ and then $1 + i_q$. Reading d itself as an effective annual rate understates the bill by 27 basis points on average and by 178 in 1981Q3, one-signed and growing with the level of rates.

B.3 Holding-period returns on AAA bonds and the forecastability of the wedge

The system (16)–(17) requires an expected one-period holding return on AAA bonds, whereas the data provide a quoted yield. This subsection constructs a realized holding-period return from the yield series, uses it to measure the duration that converts the wedge prior (21) into return units, and then shows that the realized returns cannot be used to estimate the wedge itself. The returns enter the estimation only through the bound they place on the scale of the wedge prior, which Appendix C.5 sets out.

Construction. We price a par coupon bond of 25 years at quarterly compounding: the coupon d_t is set at the purchase yield, the bond is repriced one quarter later at the then-prevailing yield with one period less to run, and:

$$\mathrm{HPR}_{t\to t+1} \;=\; \frac{P_{t+1}+d_t}{P_t}-1.$$

Regressed on the yield change it reflects, $y_{t+1}^{AAA}-y_t^{AAA}$, the series has $R^2=0.89$ and a slope implying an effective duration of 10.2 years. Its quarterly standard deviation is 4.2%, in line with a long corporate index.

Duration. Effective duration here is the empirical price sensitivity, the percentage capital loss from a one-point rise in yields. It is far shorter than the 25-year tenor because a par coupon bond returns most of its present value through coupons paid long before maturity. Three numbers appear in the paper, and they are the same object under three conventions: 10.2 is the regression slope above, on quoted yields; the $D=11.1$ used in (21) is that same regression on *log* yields, which is the convention Δy_t^{AAA} carries throughout, the two differing by a factor of $1+\bar{y}$; and 11.7 is the textbook modified duration of a 25-year quarterly-pay par bond at the sample-mean yield of 7.1%, which is what makes the measured value the right order. We hold D constant, although the duration of a long bond is not, because 10.2 is an average price sensitivity estimated over the whole sample and substituting a per-quarter par-bond duration would mix an empirical object with a textbook one. The approximation is not neutral – par duration is shortest at high yields, and the quarters in which the wedge matters are mostly high-yield quarters – but we have run the state-dependent alternative, and it moves no estimate in this paper by more than one unit in its second significant figure.

Forecasting the wedge from realized returns. An alternative to assuming the wedge away or carrying it as a latent is to estimate it: regress the realized return on variables in the date-t information set and use the fitted value as $\mathbb{E}_t[1+r_t^{AAA}]$. This approach keeps the expected return an observable, so the system remains exactly identified. It does not succeed, and

the reasons apply to any such attempt rather than to our specification in particular. The computations are in `code/scripts/hpr_forecast.py`.

Write $w_t \equiv r_t - y_{t-1}^{AAA}$ for the realized wedge, the return over quarter t less the yield at which the position was opened, and regress it on the AAA level, the ten-year–bill spread, the AAA–ten-year spread and the thirty–ten-year slope, all dated $t-1$. The fit appears respectable: $R^2 = 0.086$, fitted values with a standard deviation of 4.9 percentage points a year, ranging from -10.8 to $+14.7$, with a first-order autocorrelation of 0.91. On inspection, however, the fitted values are a set of decade dummies. Their decade means are -1.3, $+2.9$, $+1.3$, $+2.7$, $+1.0$ and -5.6 against realized decade means of -3.3, $+1.9$, $+1.6$, $+2.9$, $+3.3$ and -5.8, a correlation of $+0.93$ across six numbers (Figure B.1, left). Fitting six decade averages is what four highly persistent regressors and 218 quarters permit: the effective sample size $T(1-\hat{\rho}_1)/(1+\hat{\rho}_1)$ is 1.1 observations for the AAA level, and 11.5, 7.9 and 7.0 for the three spreads.

Two checks make this precise. First, we bootstrap under the null of *no* predictability, in which the regressors follow their fitted VAR(1) and the return innovations are resampled jointly with the regressor innovations so that the Stambaugh (1999) correlation is preserved. The realized R^2 of 0.086 lies at the 99th percentile of that null, so some predictability is present; but the null's own fitted values have a median standard deviation of 2.7 percentage points a year, 56% of what we estimated. More than half of the apparent variation in expected excess returns is what pure noise produces on regressors this persistent. A predictable component consistent with the in-sample fit therefore has a standard deviation of at most about four percentage points a year, the square root of the difference between the two variances; this is the bound used in Appendix C.5. Second, we forecast out of sample. Expanding-window forecasts, one-sided in the same way as the trend model of Appendix C.1, give an R^2 against the random-walk restriction of $+0.014$, -0.003, -0.021 and -0.004 for the term-spread specification at burn-ins of 24, 40, 60 and 80 quarters, and negative at every burn-in for the four-regressor specification (Figure B.1, right). The largest entry is under one and a half percent of the variance, it changes sign with the burn-in, and the

forecasts that produce it have a standard deviation of five points a year. This reproduces, on our instrument, the conclusion that the bond-predictability literature has reached about its own forecasts (Thornton and Valente, 2012; Bauer and Hamilton, 2018).

Underlying both checks is the standard error of the mean. The realized wedge has a quarterly standard deviation of 16.6 points a year and a mean of $+0.62$ with a standard error of 1.12. These returns do not identify even the average wedge, so it is not surprising that they fail to identify a time-varying one.

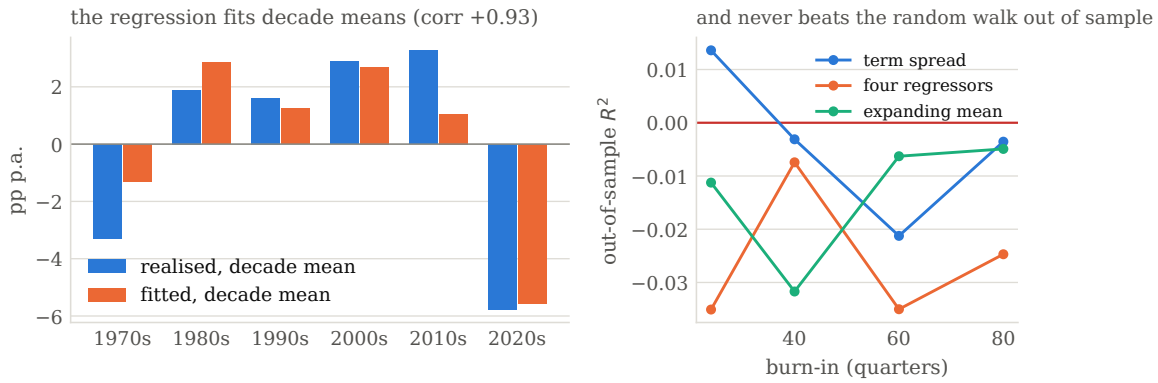

Figure B.1: Why the wedge cannot be forecast from realized returns. Left: decade means of the realized wedge and of the full-sample regression's fitted values; the regression reproduces them almost exactly, which is what four regressors with effective sample sizes between one and twelve observations can do. Right: out-of-sample R^2 against the random-walk restriction, expanding window, by burn-in length. No entry is reliably above zero.

C Estimation details

This appendix sets out the estimators in three groups. The first, C.1 to C.4, builds the baseline: the trend model, the closed-form inversion, the ordering restriction and the discount rate. The second, C.5 to C.7, defends the priors and the identifying assumption. The third, C.8 to C.11, reports sampling uncertainty, the equation-error estimator, the persistence of the latents, and the numerical implementation.

C.1 The Fisher-rate trend

Let $g_t \equiv \log\left[(p_t c_t)/(p_{t-1}c_{t-1})\right]$ be quarterly log nominal consumption growth. We model it as a multiplicative local level,

$$g_t = \gamma_t + e_t^1, \qquad e_t^1 \sim \mathcal{T}_\nu(0, \sigma_1^2), \tag{59}$$

$$\gamma_t = \gamma_{t-1} + u_t, \qquad u_t \sim N(0, \sigma_u^2), \tag{60}$$

where $\exp(\gamma_t)$ is the *expected* gross growth factor. This interpretation is what permits the conditional expectations (18)–(20) to be imposed on the trend rather than on realized growth: because γ_t is the expectation, e_t^1 *is* the one-step-ahead expectation error.

The observation density is Student-t with $\nu = 4$, fitted by iteratively reweighted Kalman recursions. A Gaussian specification is not usable: 2020Q2 and 2020Q3 are -10.6 and $+6.4$ standard deviations from trend, and a Gaussian filter either follows them or is forced to inflate σ_1 for the whole sample. Under the Student-t specification, estimated by concentrated maximum likelihood with exact diffuse initialization, $\widehat{\sigma}_1 = 0.0034$ per quarter (1.35% annualized) and $\widehat{\sigma}_u = 0.0013$, so the trend is very stiff; ten quarters are flagged at $|z| > 3$: 1980Q2, 1986Q2, 2008Q4–2009Q1 and six in 2020–22.

The Kalman recursions deliver $m_t = \mathbb{E}[\gamma_t \mid \mathcal{I}]$ and $V_t = \mathrm{Var}[\gamma_t \mid \mathcal{I}]$ for two information sets: *one-sided*, $\mathcal{I} = \{g_s\}_{s<t}$, an ex-ante expectation with no look-ahead, and *two-sided*, $\mathcal{I} = \{g_s\}_{s\le T}$, a better estimate of the latent trend that uses future data. We take the one-sided version as our baseline, because the object the model requires is a conditional expectation,

and an estimate that uses future data is not one. The two-sided version changes no conclusion and appears in Table C.3. The cost of the one-sided choice is width rather than location: the prediction variance V_t exceeds the smoothing variance, so the credible bands are wider. Figure C.1 plots the two estimates against realized growth.

One indexing convention requires care. The recursions are indexed by the quarter whose growth is being predicted, whereas the estimator is indexed by the date of the information set, as Section 3 sets out. The estimate carried at date t therefore uses the prediction of the trend for $t+1$ formed from data through t, and we write that object as m_t from here on, in this appendix and in the main text. Leading the filter's output by one quarter in this way gains 1971Q2 at the start of the sample and loses 2025Q4 at the end, which is why the estimation covers 1971Q2–2025Q3 with the same $T = 218$ quarters.

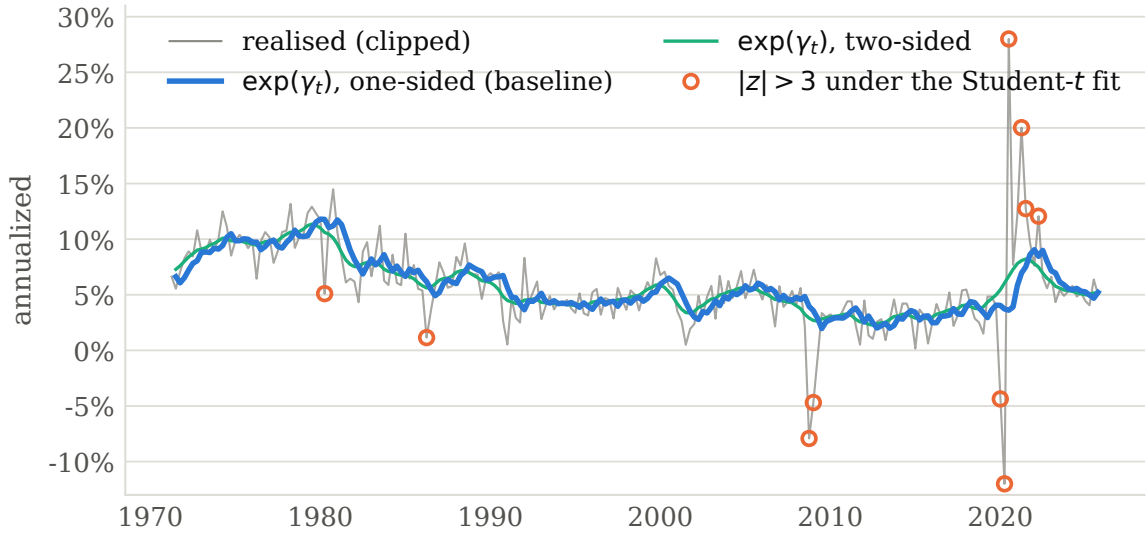

Figure C.1: The trend model. Realized annualized nominal consumption growth (clipped at -12% and 28% so that 2020 does not compress the axis), against the one-sided estimate of $\exp(\gamma_t)$ used throughout and the two-sided estimate reported as a robustness check, each plotted at the quarter whose growth it predicts. Circles mark the ten quarters at $|z| > 3$ under the Student-t fit; the two 2020 observations lie at -10.6 and $+6.4$ standard deviations and are the reason a Gaussian observation density is not usable.

Throughout, γ_t enters the estimator as a draw from $N(m_t, V_t + \sigma^2_{\text{tol}})$. The extra term is an explicit allowance for what the system does not model: measurement error in the quoted yields, and slack for the maintained assumptions of Section 3. It is *not* an allowance for the Jensen terms that arise when a point value replaces an expectation; those are bounded directly and are worth under a basis point a year each, against the 25 basis points of σ_{tol}.[19] The tolerance is set at 25 basis points at an annual rate, the same convention as the tolerances σ_A and σ_{AAA} below. The filter's own uncertainty about the trend averages 0.92 percentage points a year, so adding σ_{tol} in quadrature widens the γ_t prior to 0.96: the allowance is a modest addition to what the filter already implies, not a replacement for it.

The consequences of this choice run counter to intuition, so we state them. Quadrupling σ_{tol} to 100 basis points a year leaves the *width* of the posterior essentially unchanged, since the average interquartile range of the price of convenience moves from 0.0377 to 0.0380 and the 90% band from 0.103 to 0.105, by one and two percent respectively, while it moves the *level* by 0.15 percentage points, from 4.9% to 5.1%.

Neither part of this result is accidental. The width does not move because the posterior for r^F_t is dominated by the prior on the discount rate, whose standard deviation is three percentage points a year; against that, a change in the prior standard deviation of γ_t from 0.96 to 1.36 enters in quadrature and nearly vanishes. The level moves because the admissibility restrictions bind r^F_t from below only: a wider γ_t prior supplies more high-growth draws that satisfy $r^F_t \geq r^{AAA}_t$ and none that are excluded for being too high, so the posterior mass shifts upward. A symmetric prior widened symmetrically moves a posterior median only when the constraints are one-sided, as they are here. The level moves only once σ_{tol} exceeds 50 basis points.

[19] The inversion returns certainty equivalents – the value of each latent that reproduces the observed price – rather than conditional means. The premia enter (18)–(20) linearly, so $\bar{\ell}_t$ is recovered as a conditional expectation exactly; three places carry a correction. Writing the Euler object in (16) as an expectation is exact only up to $s^2/2$, worth 0.23 basis points a year on r^F_t and 0.0004 percentage points on the price of convenience. Fixing r^F_t at its certainty equivalent displaces $\bar{\ell}_t$ by a covariance worth at most 0.3 basis points a year. The ratio bias on μ_t is under 0.005 and negative in sample, so it understates marketability and with it the urgent component. The right scale for all three is the *household's* conditional uncertainty about the state, which is the filter's own observation error, $\widehat{\sigma}_1$, not the posterior spread of the latents, which is our uncertainty rather than theirs and is an order of magnitude larger.

C.2 The closed-form inversion

Fix a draw of $(\gamma_t, \beta_t, \lambda_t)$ and write $\beta_{q,t} = \beta_t^{1/4}$. Equation (16) gives $1 + r_t^F = \exp(\gamma_t)/\beta_{q,t}$, and the three pricing equations then read as in (23). Because the middle equation is linear in μ_t and the third is proportional to it, the system solves in one pass:

$$\mu_t = \frac{\Delta_t(r_t^{AAA})}{\Delta_t(r_t^{TB})}, \qquad \omega_t \ell_t^N = B_t = \frac{\Delta_t(r_t^{TB}) - \Delta_t(r_t^{CP})}{1 - \mu_t}, \qquad (1-\omega_t)\ell_t^D = A_t = \Delta_t(r_t^{TB}) - B_t. \tag{61}$$

Two rearrangements are used in the main text. Eliminating μ_t from the last two,

$$A_t \;=\; \Delta_t(r_t^{TB})\,\frac{\Delta_t(r_t^{CP}) - \Delta_t(r_t^{AAA})}{\Delta_t(r_t^{TB}) - \Delta_t(r_t^{AAA})}, \qquad B_t \;=\; \Delta_t(r_t^{TB})\,\frac{\Delta_t(r_t^{TB}) - \Delta_t(r_t^{CP})}{\Delta_t(r_t^{TB}) - \Delta_t(r_t^{AAA})}. \tag{62}$$

Substituting (61) into the identity (17) and clearing denominators gives a quadratic:

$$\lambda_t r_t^{TB}\,\omega_t^2 \;+\; \left[A_t - \lambda_t r_t^{TB} + (1 + r_t^{TB})B_t\right]\omega_t \;-\; (1 + r_t^{TB})B_t \;=\; 0, \tag{63}$$

after which $\ell_t^D = A_t/(1-\omega_t)$ and $\ell_t^N = B_t/\omega_t$.

Proposition C.1. *Let $r_t^{TB} < r_t^{CP} < r_t^{AAA} \le r_t^F$. Then the solution (61) satisfies $\mu_t \in [0,1)$, $A_t > 0$ and $B_t > 0$; the quadratic (63) has exactly one root in $(0,1)$; and at that root $\ell_t^D \ge \ell_t^N \ge 0$. Conversely, each of the three inequalities is necessary: $B_t > 0 \iff r_t^{CP} > r_t^{TB}$, $A_t > 0 \iff r_t^{AAA} > r_t^{CP}$, and $\mu_t \ge 0 \iff r_t^F \ge r_t^{AAA}$.*

Proof. Δ_t is strictly decreasing in its argument, so the hypothesis gives $\Delta_t(r_t^{TB}) > \Delta_t(r_t^{CP}) > \Delta_t(r_t^{AAA}) \ge 0$, and in particular $\Delta_t(r_t^{TB}) > 0$. Hence $\mu_t = \Delta_t(r_t^{AAA})/\Delta_t(r_t^{TB}) \in [0,1)$, and $B_t > 0$, $A_t > 0$ follow from (62) because both numerators and the common denominator are positive. For the root, write the quadratic as $f(\omega) = a\omega^2 + b\omega + c$ with $a = \lambda_t r_t^{TB} \ge 0$ and $c = -(1+r_t^{TB})B_t < 0$. If $a > 0$ the discriminant is $b^2 - 4ac > 0$, so the roots are real and, since $c/a < 0$, of opposite sign; there is exactly one positive root. If $a = 0$ the equation is linear with the single root $-c/b$, positive because $b = A_t + (1 + r_t^{TB})B_t > 0$. In either case $f(0) = c < 0$ and $f(1) = a + b + c = A_t > 0$, so the positive root lies in $(0,1)$. Non-negativity of the premia is then immediate, and $\ell_t^D \ge \ell_t^N$ follows from (17) because $\ell_t^D - \ell_t^N = r_t^{TB}(\lambda_t + \ell_t^N) \ge 0$. The

converse statements read off (62) and the definition of μ_t, using that Δ_t is decreasing. □

Corollary C.1 (the composition of convenience is free of the Fisher rate)**.** *Under the hypotheses of Proposition C.1,*

$$\frac{B_t}{A_t + B_t} = \frac{\Delta_t(r_t^{TB}) - \Delta_t(r_t^{CP})}{\Delta_t(r_t^{TB}) - \Delta_t(r_t^{AAA})} = \frac{(1 + r_t^{TB})^{-1} - (1 + r_t^{CP})^{-1}}{(1 + r_t^{TB})^{-1} - (1 + r_t^{AAA})^{-1}},$$

which depends on neither γ_t, β_t, λ_t nor μ_t.

Proof. The first equality is (62). For the second, $\Delta_t(x) - \Delta_t(y) = (1 + r_t^F)\big[(1+x)^{-1} - (1+y)^{-1}\big]$, so the factor $(1 + r_t^F)$ appears in numerator and denominator and cancels. □

Corollary C.1 is the analytical content of Equation (29) and the methodological gain from adding a long asset. Its one qualification is that r_t^{AAA} is an expected holding return, not observed, and through (15) inherits the prior on ψ_t, the only prior in the system that affects the identity. The dependence is small: over the 184 quarters in which commercial paper yields less than the AAA bond, the typical (median absolute) per-quarter gap between the posterior share and the identity evaluated at $\psi_t = 0$ is 0.020, and the gap between the two sample medians is 0.003. The corollary has no counterpart in the system where $\mu = 0$, where the ratio becomes $\big[\Delta_t(r_t^{TB}) - \Delta_t(r_t^{CP})\big]/\Delta_t(r_t^{TB})$ and the Fisher rate does not cancel. Figure C.2 plots the decomposition against the two raw spreads.

C.3 The ordering restriction and the admissibility penalties

By Proposition C.1 the model's admissibility conditions $0 \leq \mu_t \leq 1$, $A_t \geq 0$ and $B_t \geq 0$ collapse into the single chain (25). This subsection shows what the chain looks like before the estimator is applied; catalogues the quarters in which it fails and identifies which link is responsible, because the failures of different links carry different information; and sets out the penalties through which the estimator handles them. What the baseline makes of the failures is reported with the wedge, in Appendix C.5.

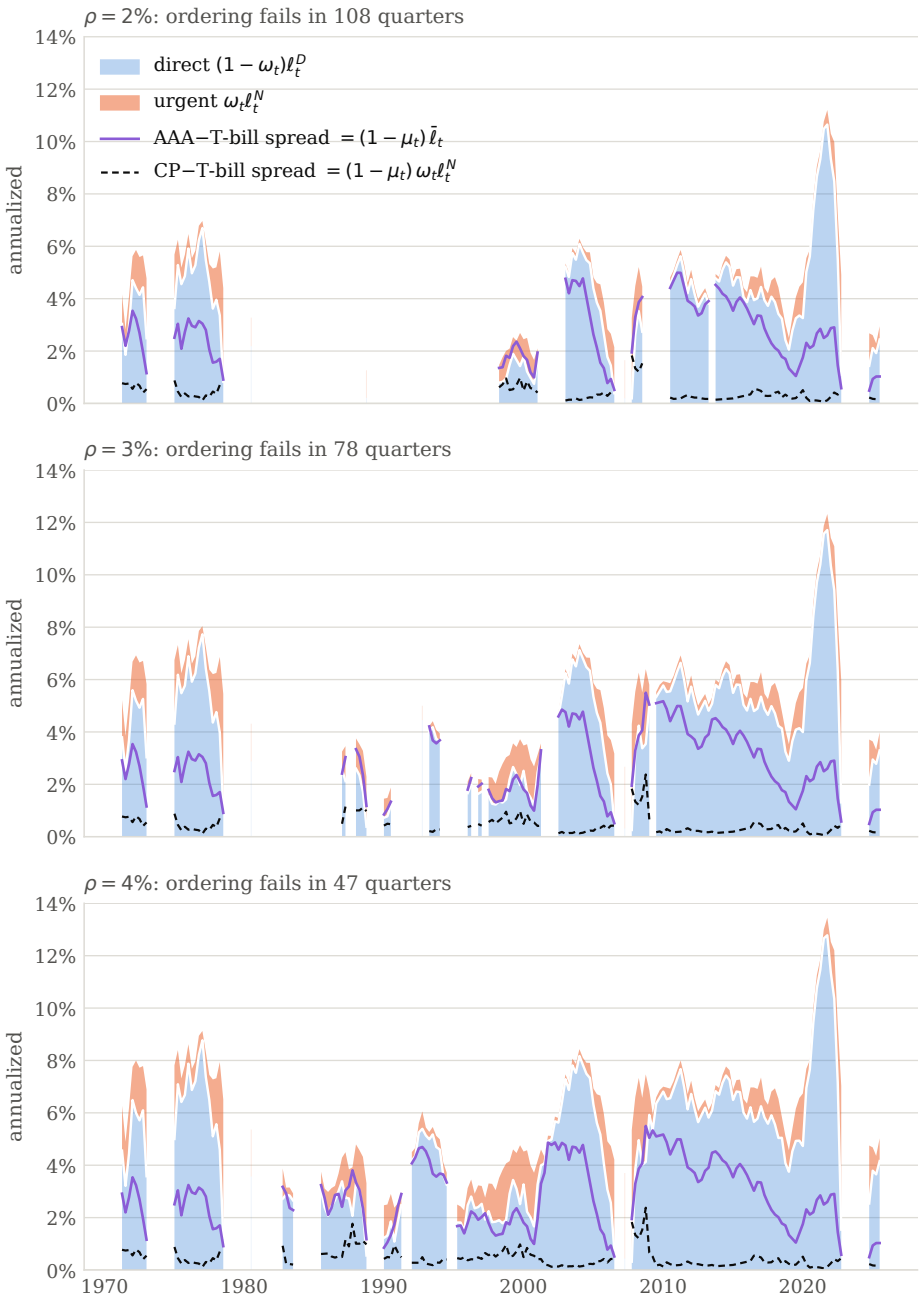

Figure C.2: The exact decomposition of the price of convenience and the two spread identities, from the closed form (62) with the wedge switched off and the Fisher rate $r_t^{F\rho}$ of (64) at $\rho = 2\%, 3\%$ and 4% (top to bottom). The stack is $(1-\omega_t)\ell_t^D$ and $\omega_t \ell_t^N$, summing to $\bar{\ell}_t$; quarters in which the ordering fails at given ρ are left blank. The two lines are the raw AAA–T-bill and commercial-paper–T-bill spreads, which by (27)–(28) equal $(1 - \mu_t)$ times the whole of the stack and $(1 - \mu_t)$ times its upper block respectively.

The chain before the estimator. Figure C.3 is the main text's Figure 2 with both estimated series replaced by benchmarks that do not depend on the posterior. In place of the expected holding return r_t^{AAA} it plots the observed yield y_t^{AAA}, that is, the wedge switched off, and in place of the posterior median Fisher rate it plots:

$$r_t^{F\rho} \;\equiv\; \exp\big(\gamma_t + \rho/4\big) - 1 \qquad \text{at} \qquad \rho = 2\%,\ 3\%,\ 4\%, \tag{64}$$

the Fisher rate of (16) with a fixed discount rate imposed and the filtered trend in place of a draw.

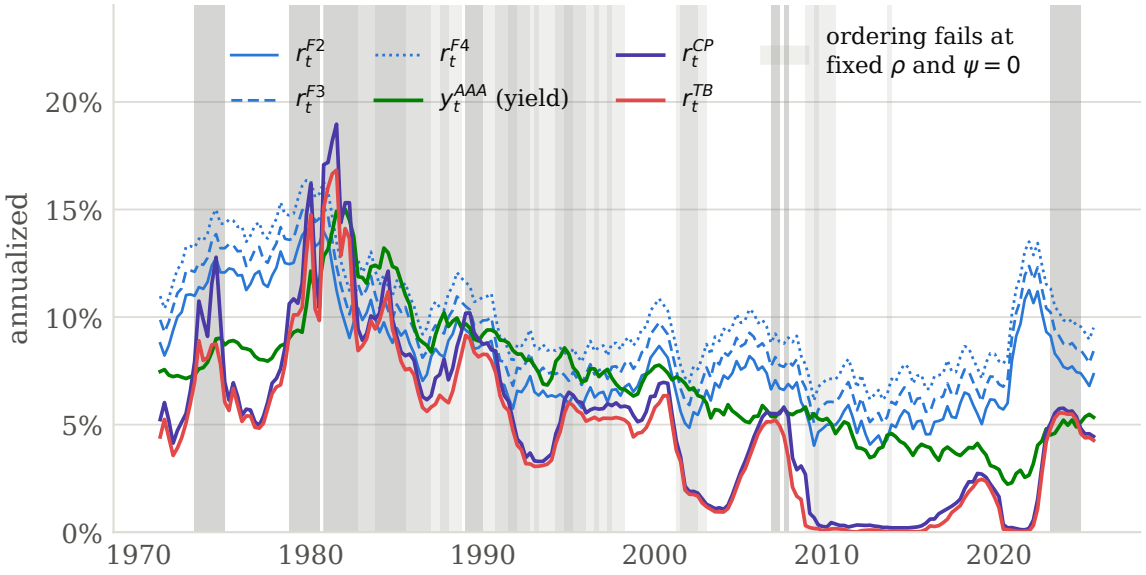

Figure C.3: The four rates before the estimator touches them. Showing the observed AAA yield y_t^{AAA} in place of the expected holding return, and r_t^{F2}, r_t^{F3} and r_t^{F4} of (64) (solid, dashed, dotted), the Fisher rate at an imposed ρ of 2%, 3% and 4%, in place of the posterior median. Shaded columns mark the quarters in which the chain (25) fails on *these* series, darker the more of the three discount rates it fails at: 108 quarters at 2%, 78 at 3% and 47 at 4%. Compare Figure 2, where both estimated rates are at their posterior medians and the chain holds in every quarter.

Which link fails, and when. At $(\psi_t, \rho) = (0, 3\%)$ the chain holds in 140 of 218 quarters (110 at 2%, 171 at 4%). We separate the failures by link.

The first link, $r_t^{TB} \le r_t^{CP}$, **never fails**. Commercial paper is dearer to the issuer than a bill in every quarter of the sample.

The second link, $r_t^{CP} \leq y_t^{AAA}$, fails in **34 quarters**: seven consecutive quarters through 1973Q2–1974Q4, thirteen scattered quarters through 1978Q4–1982Q2, four through 1989, three through 2006Q4–2007Q3, and seven consecutive quarters through 2023Q1–2024Q3. The median shortfall is 92 basis points and the largest 426. All 21 quarters of yield-curve inversion in the sample – quarters in which $y_t^{AAA} < r_t^{TB}$ – are among these failures, and all of them fall inside the three long episodes; the 1989 and 2006–07 episodes contain no inversion at all. Equivalently, these are the quarters in which the estimated direct premium A_t would have to be negative. The Fisher rate does not enter this link, so no value of β_t or γ_t can repair it.

The two kinds of failure differ in what the identity (29) does. In the 21 inversion quarters its denominator changes sign and the identity returns a negative number; in the remaining 13 it is well defined but returns a value above one, a median of 1.90, which is a rejection rather than an absence, since a share above one is $A_t < 0$.

The third link, $y_t^{AAA} \leq r_t^{F3}$, fails in **48 quarters** – most of 1981Q2–1987Q4 and 1990Q4–1997Q2, and 2001Q3–2002Q2 and 2009Q2 – by a median of 71 basis points; at 2% and 4% it fails in 82 and 16. These are the quarters in which marketability would have to be negative at the respective level of ρ. Unlike the second link, this one involves the discount rate, so it represents information about ρ_t rather than a rejection: with the trend at its filtered mean, $\mu_t \geq 0$ holds exactly when $\rho_t \geq \underline{\rho}_t \equiv 4\big[\log(1 + y_t^{AAA}) - \gamma_t\big]$, a lower bound computed from observables alone. This lower bound has a median of 1.3% across quarters and peaks at 6.9% in 1982Q2.

The second and third links fail together in four quarters, 1981Q2–Q3 and 1982Q1–Q2, which is why the $34 + 48$ violations come to 78 distinct quarters rather than 82. The second-link failures, and their concentration in the inversion episodes, are the case for the repricing wedge of Appendix C.5.

The weight a quarter carries. Every specification in this paper attaches to each draw a scalar weight W_t, the quasi-likelihood's contribution from quarter t, and the same object that Appendix C.10 calls the emission b_t once the quarters are linked by a transition kernel. What

distinguishes the specifications is which restrictions are imposed exactly and which are left to W_t, and two forms of term appear. A restriction $x_t \geq 0$ that is imposed *softly* contributes the one-sided penalty:

$$\mathcal{P}\big(x_t;\sigma\big) \;\equiv\; -\tfrac{1}{2}\left[\frac{4\min\{x_t,\,0\}}{\sigma}\right]^2, \tag{65}$$

which is zero wherever the restriction holds and grows quadratically in the violation; the factor 4 converts x_t, a quarterly rate, into the annualized units in which the tolerances σ are quoted. An *equation* allowed to fail by a signed residual contributes the two-sided term $-\tfrac{1}{2}\big(4\varepsilon_t/\sigma\big)^2$ instead; the only such residual in this subsection is $\varepsilon_t \equiv \mu_t\,\Delta_t(r_t^{TB}) - \Delta_t(r_t^{AAA})$, the amount by which (20) fails. Writing the three specifications of this subsection out in full, with $\sigma_A = \sigma_{AAA} = 25$ basis points annualized:

$$\text{exact inversion:}\quad \log W_t \;=\; \mathcal{P}\big(\Delta_t(r_t^{AAA});\,\sigma_A\big), \tag{66}$$

$$\text{baseline (wedge):}\quad \log W_t \;=\; \mathcal{P}\big(\Delta_t(r_t^{AAA});\,\sigma_A\big) \;+\; \mathcal{P}\big(\Delta_t(r_t^{CP}) - \Delta_t(r_t^{AAA});\,\sigma_A\big), \tag{67}$$

$$\text{penalized } \mu\text{:}\quad \log W_t \;=\; -\tfrac{1}{2}\left[\frac{4\varepsilon_t}{\sigma_{AAA}}\right]^2 \;+\; \mathcal{P}\big(A_t;\,\sigma_A\big). \tag{68}$$

The argument of each $\mathcal{P}$ is the quantity whose sign the restriction controls. $\Delta_t(r_t^{AAA}) \geq 0$ is $\mu_t \geq 0$, equivalently $r_t^F \geq r_t^{AAA}$; and since $\Delta_t(r_t^{CP}) - \Delta_t(r_t^{AAA}) = (1-\mu_t)A_t$, the second term of (67) is the $A_t \geq 0$ leg, equivalently $r_t^{AAA} \geq r_t^{CP}$.

The exact inversion carries one term only: under $\psi_t \equiv 0$, A_t has the same sign at every point of the grid within a quarter, so the term could not be relieved. The baseline carries both terms, because a drawn ψ_t moves r_t^{AAA} within a quarter and so makes the A_t term vary across the grid; this is what it means for both links to discipline the wedge. In the penalized specification the A_t term is active for a different reason: μ_t is drawn there, and lowering μ_t raises A_t.

Only two of these three are estimators the paper reports. The penalized-μ weight (68) is not a column of Table 1 and was never a candidate for one: it is the alternative to drawing the wedge that Appendix C.5 sets out and rejects, and it is written here because the three weights side by side are what make the rejection legible. Table 1's other two estimators place

their weight elsewhere and are set out where they are used. In Appendix C.7, the constant-μ estimator forces the wedge and weights it by the prior (21); in Appendix C.9, the equation-error estimator gives both private-asset equations two-sided residuals.

What the penalties do. Under (66) the quasi-posterior tilts toward a lower β_t and a higher γ_t in the offending quarters, at a cost set by their own priors.

Draws that violate $A_t > 0$ or $B_t > 0$ retain their weight, but Proposition C.1 does not apply to them: the root of (63) leaves $(0, 1)$ and $\ell_t^D = A_t/(1 - \omega_t)$ loses its meaning. We record ω_t, ℓ_t^D and ℓ_t^N as missing on those draws, so the reported marginals are conditional on the model holding, and we report the violating mass separately.

C.4 The discount rate

Because β_t is the prior that moves the level of the price of convenience most, we report the mapping from the discount rate to the level rather than a point estimate. Table C.1 holds everything else fixed and varies the annualized discount rate $\rho = -4 \log \beta$.

Table C.1: The price of convenience as a function of the discount rate, from the exact inversion. Sample medians, annualized. The components, μ and the urgent share are medians over the quarters in which the ordering (25) holds at that ρ; the last column counts the others, so the subsample changes down the table and the components are not comparable across rows in the way $\bar{\ell}$ is.

ρ	β (ann.)	$\bar{\ell}$	direct	urgent	μ	quarters out of order
0.00	1.000	1.0%	2.8%	0.43%	0.32	174
0.01	0.990	2.0%	3.4%	0.55%	0.40	140
0.02	0.980	3.1%	4.0%	0.61%	0.40	108
0.03	0.970	4.1%	4.7%	0.67%	0.43	78
0.04	0.961	5.1%	5.0%	0.72%	0.48	47
0.06	0.942	7.3%	6.3%	1.10%	0.61	34
0.08	0.923	9.4%	8.1%	1.43%	0.70	34

The urgent *share* barely moves with ρ across Table C.1, at 0.153, 0.153, 0.149, 0.153, 0.159, 0.159 and 0.159: by Corollary C.1 it is invariant quarter by quarter, and what movement remains comes from the admissible subsample changing down the table. Figure C.4 plots

the same exercise: the level and its two components against ρ on the left, and the posterior mass on an inadmissible direct component, quarter by quarter, on the right.

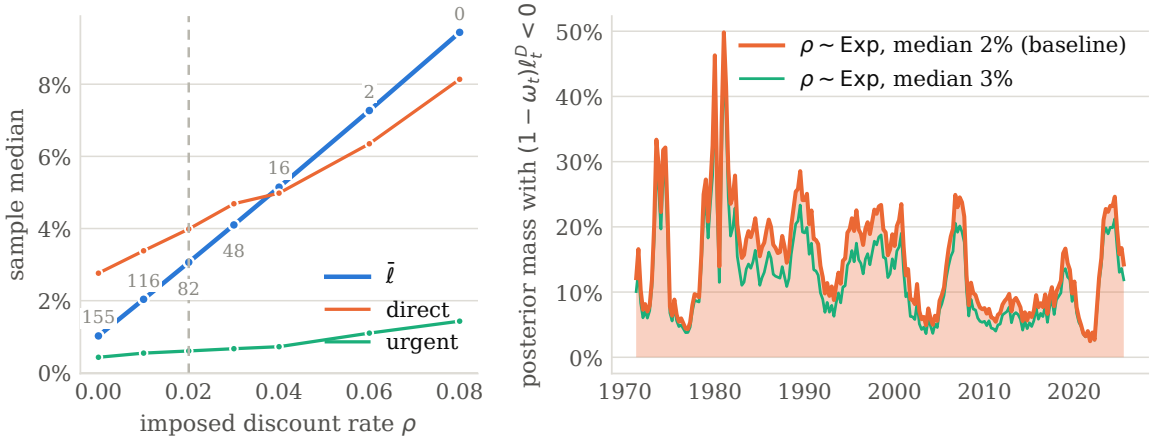

Figure C.4: Left: the sample-median price of convenience and its two components as a function of an imposed discount rate ρ; grey numbers count the quarters with $y_t^{AAA} > r_t^{F\rho}$, and the dashed line marks the prior median. Right: the posterior mass on an inadmissible direct component, quarter by quarter, under the baseline prior on ρ and the alternative with median 3%.

How much the data say about ρ. The data say little about ρ, and we state here how little. Throughout, "the posterior median discount rate" means the median across the 218 quarters of the per-quarter quasi-posterior medians: the latents are independent across quarters, so there is no single pooled posterior to report. The mean across quarters is a little higher in each case below, 3.68% and 4.91% against the medians quoted.

In the baseline our prior on ρ is exponential with median 2% (mean 2.89%), and the posterior median is 3.28%. The natural alternative is the same family with the conventional median of 3% (mean 4.33%); there, the posterior median is 4.53%. Two conclusions follow. First, the data are not silent: from either starting point the admissibility restrictions push the discount rate *up*, by 1.28 points under the baseline and 1.53 under the alternative. But they cannot choose between the two priors: the posterior medians differ by 1.25 points, when the priors differ by 1.00 at the median and 1.44 at the mean. A lower bound shifts the distribution without narrowing the gap, and the level continues to be set largely by the prior. Table C.3 reports the alternative in full.

That the movement is upward under both priors is not accidental, and it is the same one-sidedness that Section 6.2 notes. The third link of the chain (25) binds from below on r_t^F and therefore from below on ρ: a discount rate that is too low leaves the Fisher rate too low to price the AAA bond. Nothing in the system pushes back from above except the prior. We regard this as a scientific obstacle rather than a statistical one, as we have said before (Herrenbrueck and Wang, 2026): a stochastic Euler equation for a marketable asset does not discipline the rate at which households discount. We think an exponential prior fairly represents our uncertainty: unbounded support, but concentrated on percentage points in the low single digits.

C.5 The AAA repricing wedge

Section 3.3 treats the gap between the AAA yield and the expected one-period holding return on the index as a latent variable, the repricing wedge ψ_t, drawn with the prior (21). We also investigated two a priori reasonable alternatives.

Assuming the gap away, by setting $\psi_t \equiv 0$, is the exact inversion of Appendix C.2. $\psi_t \equiv 0$ is right on average: the constructed returns of Appendix B.3 give a realized excess return over bills of 2.92 points a year (standard error 1.15), against a yield spread of 2.30 on the same convention. But the system must then hold quarter by quarter, and in the 34 quarters in which commercial paper yields more than the AAA bond it violates the ordering restriction (25). The Fisher rate does not enter that link, so no prior on β_t or γ_t can repair it, and the exact inversion accordingly has no estimate of marketability or of the composition in those quarters (Table 1). Most of them fall in the three episodes that contain every quarter of yield-curve inversion in the sample (Appendix C.3), which is when the expected capital gain on a long bond should be largest. A lower μ_t^{CP} could absorb most of these failures (Section 3.4), but not the way they cluster in episodes of yield-curve inversion.

Estimating the gap from realized returns fails for the reasons Appendix B.3 explained: the fitted expected returns vary with a standard deviation of 4.9 points a year, but they reproduce six decade means and little else, and they have no out-of-sample power.

What remains is to carry the gap as a latent variable. The rest of this subsection describes how we do this and what it makes of the ordering failures; derives the two channels through which the model itself generates a wedge, with a one-parameter relaxation of each, the second of which, an anchoring of the Fisher rate, is the only alternative in the paper that moves a conclusion of the main text; sets the scale of the prior and reports the sensitivity to it; and closes with a third alternative, which leaves the yield alone and lets (20) fail instead.

How our baseline treats the wedge. In the baseline, $1 + r_t^{AAA} = (1 + y_t^{AAA})(1 + \psi_t)$; ψ_t is drawn on the grid with the prior (21), and μ_t is solved from (24) exactly as the closed form solves it. Relative to the exact inversion, one latent has been added and no equation carries a residual, so the system remains exactly identified conditional on the draws; no prior is placed on marketability, urgency or the premia; and as $\sigma_y \to 0$ the specification returns the exact inversion. Because a drawn ψ_t moves r_t^{AAA}, the admissibility penalty on $r_t^{CP} \leq r_t^{AAA}$ in (67) can now be relieved within a quarter, so both links of the ordering that involve r_t^{AAA} discipline the wedge, and the posterior on the expected quarterly yield change has a standard deviation of 2.4 basis points against a prior standard deviation of 5, with a worst-quarter magnitude of 8.7 (Figure C.5, centre panel). The wedge does its work where the ordering needs it: its posterior median is an expected quarterly yield decline, and so a higher expected holding return, of 3.4 basis points in the 34 quarters in which commercial paper yields more than the AAA bond, against a rise of 0.6 elsewhere.

What the baseline does with the failures. The share of each quarter's posterior mass lying inside (25) has a median of 0.86 across the 218 quarters, a mean of 0.85, a tenth percentile of 0.75 and a minimum of 0.50 (Table C.2 reports the same distribution at other values of σ_y). We report this distribution and nothing else: admissibility is continuous here, and a count of quarters past a threshold would put a cliff where the data have none.

At the posterior medians plotted in Figure 2 the chain holds in all 218 quarters. This is a property of the estimates rather than a construction, since those medians are taken over all draws and are not conditioned on admissibility. The 48 quarters that Appendix C.3 attributes

to the third link should therefore be read as quarters in which the discount rate had to move to keep the model admissible, not as quarters in which the model failed; and the 34 quarters attributed to the second link as quarters in which the wedge had to move.

How much the wedge moves the composition. Evaluated at $\psi_t = 0$ the identity (29) gives a median urgent share of 0.149 against the baseline's 0.180 over all 218 quarters, and 0.159 against 0.162 over the 184 in which commercial paper yields less than the AAA bond. By decade, the identity runs 0.14, 0.29, 0.23, 0.15, 0.07 and 0.05, against the baseline's 0.28, 0.37, 0.24, 0.17, 0.08 and 0.12.

Ranked by the gap between them the split is clean: the three largest gaps (0.136 in the 1970s, 0.083 in the 1980s and 0.070 in the 2020s) are the three decades that contain 31 of the 34 second-link failures, the ones only the wedge can repair, and the other three differ by 0.014, 0.022 and 0.013. The gap is therefore not a discrepancy but a reading of how much work the wedge is doing, and it appears in the decades where one would predict it and nowhere else. The decade medians are taken over all 218 quarters and are insensitive to the five failing quarters in which the grid is not converged (Appendix C.11).

Where the wedge comes from. A claim that does not mature within the period is priced off the whole expected path of the one-period required return rather than off the current quarter's. To first order, for a fixed cash-flow stream,

$$\log\left(1 + y_t^{AAA}\right) \;=\; \sum_{j \geq 0} \varphi_j \, \mathbb{E}_t\left[\log\left(1 + r_{t+j}^{AAA}\right)\right], \qquad \varphi_j \geq 0, \;\; \textstyle\sum_j \varphi_j = 1, \tag{69}$$

with weights whose mean lag is the index's duration. Equation (20) gives $\log(1 + r_t^{AAA}) = \log(1 + r_t^F) - L_t$ with $L_t \equiv \log\left(1 + \mu_t \bar{\ell}_t\right)$, so the wedge has exactly two sources. First, the convenience loading it prices may differ from the current quarter's. Second, the Fisher rate at which the bond discounts may differ from the current quarter's. The baseline sets both differences to zero. Each is an assumption rather than an oversight, and each has a one-parameter relaxation that nests the baseline.

The loading channel. Suppose L_t is stationary with quarterly persistence ϕ. Then Equation (69) gives $\log(1+r_t^F) - \log(1+y_t^{AAA}) = \bar{L} + \kappa_L\,(L_t - \bar{L})$ with $\kappa_L = (1-\varrho)/(1-\varrho\phi)$ and $\varrho = (1+r^F)^{-1}$. The baseline sets $\kappa_L = 1$, that is, $\phi = 1$, a permanent convenience loading on long private claims, so that this channel contributes nothing to ψ_t and the whole of the wedge is left to the prior. Appendix C.10 estimates kernel autocorrelations of 0.89 for ω_t and 0.99 for μ_t at the joint optimum, so the assumption deserves a check.

The quasi-likelihood cannot provide one. Because κ_L rescales the μ axis, the profile is dominated by a Jacobian term of order $T\log(1/\kappa_L)$, and depending on which side of (20) the rescaling is applied to, the raw profile either rewards or punishes $\kappa_L < 1$ by hundreds of log points; neither number is informative. One scale-free criterion remains: the bound of Appendix B.3 requires $\kappa_L \geq 0.38$ for the implied wedge to remain within what realized returns can accommodate. Taking the lag-1 autocorrelation of the AAA model premium, 0.967, at face value as ϕ would imply $\kappa_L = 0.18$, which that bound rejects. We retain the nested corner, and moving away from it matters little: at $\kappa_L = 0.8$ the implied wedge has a standard deviation of 0.55 percentage points a year, an order of magnitude below what the forecasting regression of Appendix B.3 claims.

The Fisher rate channel. The second source is larger than the first, and it has two halves, because $\log(1+r_t^F) = \gamma_t + \rho_t$ by (16) and Equation (69) asks for the expected path of both. Under the local-level trend model, γ_t is a random walk, so $\mathbb{E}_t[\gamma_{t+j}] = \gamma_t$ at every horizon and that half contributes nothing. A local-level filter, however, is a statement about how we extract the trend from quarterly consumption data, not a claim that a bondholder in 1980 expected 12% nominal growth to continue for twenty-five years.

The other half is a convention. The baseline has the household expect the discount rate it faces today to persist, $\mathbb{E}_t[\rho_{t+j}] = \rho_t$, so that half contributes nothing either. This is the same permanence convention as $\kappa_L = 1$ above, imposed for the same reason. It should be distinguished from the per-quarter prior on ρ_t, which expresses our ignorance of its level and is not a claim that the household sees a fresh draw each quarter.

Anchoring the Fisher rate. A one-step relaxation of the random-walk expectation for γ_t lets $\bar{\gamma}_t$ be a one-sided exponentially weighted anchor for long-run nominal growth with gain Γ, and lets the AAA equation discount at:

$$\log\left(1+\bar{r}_t^F\right) \;=\; \log\left(1+r_t^F\right)+\alpha_t, \qquad \alpha_t \;=\; \Theta\,(\bar{\gamma}_t - m_t), \tag{70}$$

while (18) and (19), which price instruments that mature inside the quarter, keep r_t^F. Setting $\Theta = 0$ in (70) returns the baseline. Because α_t is built from the filtered trend m_t rather than the drawn γ_t, it is a per-quarter observable, and the Fisher rate still cancels from the ratio behind Corollary C.1. The anchoring therefore adds no prior dependence to the urgent share beyond the wedge; $1/(1+r_t^{AAA})$ is replaced by $e^{-\alpha_t}/(1+y_t^{AAA})$ in the denominator of (29).

Two considerations recommend $\Theta = 1$ with Γ of the order of $1/(4D)$, a memory equal to the index's duration. First, the quasi-likelihood rises. The comparison is clean in this case, since α_t shifts the equation rather than rescaling an axis. The gain is modest whichever way it is measured: against the exact inversion, which has no wedge, it is 24 log points, and against the baseline, which has one, it is 18. Of the two, 18 is the relevant number. Second, the correction has some outside support. The series α_t is built from no yield data, and it has a correlation of -0.46 with the spread between the ten-year Treasury yield and the bill rate. The implied expected holding return exceeds the yield by 2.2 points a year in the 1970s, falls 0.4 to 1.4 points below it through the 1990s and 2000s, and turns positive again in the 2020s, which is the pattern of an investor who expected the inflation of the 1970s to subside, expected the disinflation to reverse, and was wrong both times. Survey data are consistent with the first of these episodes: subjective premia on long Treasuries were high in the late 1970s and early 1980s (Piazzesi, Salomao, and Schneider, 2015), although the same surveys treat the level of rates as more persistent than statistical models do, which is closer to our baseline. The pattern is the reverse of what the regression of Appendix B.3 asserts.

We do not adopt this specification, for two reasons. First, a gain of 18 log points against the baseline does not justify a substantive change of specification, and the choice of (Θ, Γ) rests on criteria that are weak individually and do not agree exactly. Second, the anchoring

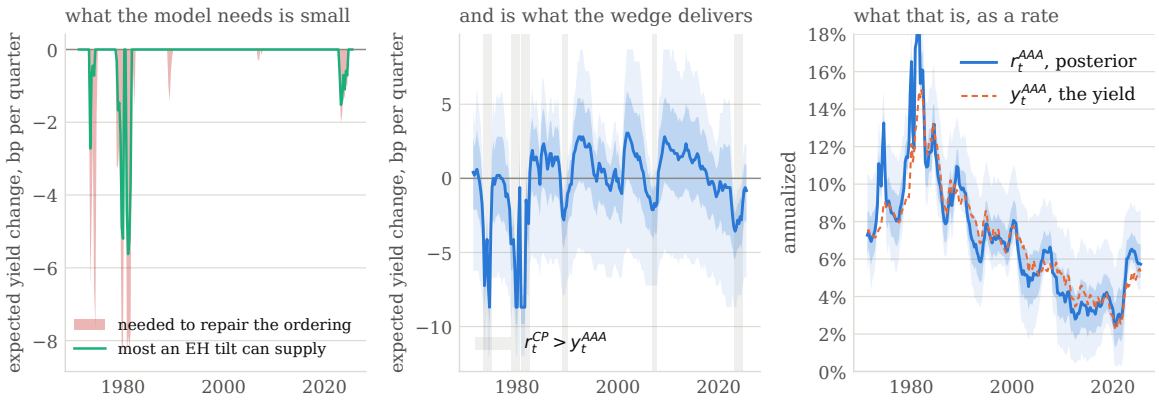

Figure C.5: The wedge, in yield space and in rate space. Left: the expected yield change over the quarter that would restore the ordering (25), and the most that an interpolation between the random walk and the expectations hypothesis can supply. Centre: the posterior on the expected quarterly yield change in the baseline of Table C.2, median with interquartile and 90% bands; shading marks the quarters in which the commercial paper rate exceeds the raw AAA yield. Right: the same posterior expressed as a rate, the expected holding return r_t^{AAA}, median with the same two bands, against the observed yield y_t^{AAA} from which it is built. This is the credible band omitted from Figure 2; the wedge moves the long rate by less than the width of the line over most of the sample, and visibly only in the inversion episodes.

changes little. Applied on top of the baseline it lowers the price of convenience from 4.9% to 4.7% and raises marketability from 0.49 to 0.51, both within the range that Table C.3 already spans, and it leaves the urgent share at 0.18.

One conclusion of the main text does move. The 2008Q4 trough in marketability remains in the bottom two percent of the sample distribution under either specification, but the 2022 elevation reverses, from the 98th percentile of that distribution in the baseline to the 12th under the anchoring. Section 6.4 draws out the implication.

What the ordering requires of the wedge. The wedge that the model requires is small, which is apparent only when it is expressed in yield units. Over the 34 quarters in which $y_t^{AAA} < r_t^{CP}$, the wedge required to restore the ordering has a median of 0.86 log points a year and a maximum of 3.75 (the 92 and 426 basis points of Appendix C.3 are the same shortfalls as differences of effective annual yields). At a duration of 11.1 years these are expected yield movements of 1.9 and 8.4 basis points over the quarter (Figure C.5, left panel).

Two further facts about the required wedge bear on how we treat it. First, no interpolation between the random-walk restriction and the expectations hypothesis can restore (25). The family $\mathbb{E}_t[1 + r_t^{AAA}] = (1 + y_t^{AAA})^{1-\theta}(1 + r_t^{TB})^{\theta}$ for $\theta \in [0, 1]$ nests both. At $\theta = 1$ the expected holding return is the T-bill rate, so the largest wedge the family can deliver in a violating quarter is $r_t^{TB} - y_t^{AAA}$, a median of 0.30 points against the 0.86 needed. The shortfall is exactly the spread between commercial paper and bills, which is non-negative in every quarter of our sample. The model requires the expected return on the long private claim to exceed the commercial-paper yield, and the expectations hypothesis raises it at most to the bill rate. Second, realized returns point in the wrong direction. In the same quarters the realized wedge over the following year averaged -5.0 percentage points a year, because yields rose sharply in 1974, in 1980–81 and in 2023–24. A wedge fitted to realized returns would therefore be negative exactly where the model needs it positive, and would deepen the violations rather than repair them. Whether this reflects a failure of the model or episodes in which investors were surprised is not a question that realized returns can settle, and it is the reason the wedge enters as a prior rather than as an estimate.

The scale of the prior. Four figures bear on σ_y. We state each both as a standard deviation of the expected quarterly yield change, in basis points, and as one of the expected holding return, in percentage points a year; the two differ by the factor $4D$. Realized returns imply that the standard deviation of any predictable component of the wedge is at most about four percentage points a year, or 9 basis points (Appendix B.3). The two structural channels above imply between 0.6 (the loading channel at $\kappa_L = 0.8$) and 1.8 (the standard deviation of the anchoring correction α_t at $\Gamma = 1/(4D)$), or 1 to 4 basis points. Survey forecasts of the AAA yield itself give the most direct figure. In the Survey of Professional Forecasters, the median forecaster's expected change in the quarterly-average Moody's Aaa yield, taken from the next quarter to the one after so that neither average is partly observed at the survey date, has a standard deviation of 5.9 basis points over 1990Q4–2025Q3 (before 1990Q4 the survey asks about a different index), net of the sampling noise in the median. Our sample includes the more volatile 1970s and 1980s, and scaled by realized yield volatility to the whole of it the

figure is 7.7. Surveys of Treasury yields give a figure of the same order: subjective expected excess returns on ten-year Treasuries over a one-year horizon have a standard deviation of 3.2 points a year (Piazzesi et al., 2015), although these are premia over the short rate rather than over the bond's own yield. Our $\sigma_y = 5$ basis points is 2.2 points a year: above the structural channels, slightly below the survey figures, and well inside the realized-return bound.

The ordering restriction is *not* among these criteria. The chain (25) is what carries the data's information about the latents, and a prior chosen to make it hold would let the prior do that work instead. We therefore fix σ_y on the external figures above and report admissibility as an outcome, in the last three rows of Table C.2. We take $\sigma_y = 5$ because it lies well inside the statistical bound, above the structural channels, which are only two of many plausible sources of a wedge, and close to the low end of the survey figures. The choice does not depend on what it does to the last three rows.

Table C.2: Sensitivity to the wedge prior. Sample medians over all 218 quarters; σ_y is the prior standard deviation of the expected quarterly yield change. The first column is the exact inversion (the $\sigma_y \to 0$ limit); the fourth is the baseline. The last three rows are the distribution across quarters of the share of each quarter's posterior mass lying inside (25); the exact inversion has no wedge to integrate over, so the distribution is degenerate there.

σ_y =	0	1bp	2bp	**5bp**	10bp	20bp
Price of convenience $\bar{\ell}$	4.8%	4.9%	4.8%	**4.9%**	5.1%	5.5%
urgent share	0.159	0.187	0.198	**0.180**	0.171	0.158
μ	0.46	0.49	0.52	**0.49**	0.48	0.46
ω	0.23	0.29	0.32	**0.33**	0.27	0.24
β (ann.)	0.971	0.971	0.970	**0.968**	0.965	0.963
sd of the wedge (bp/qtr)	0	0.7	1.3	**2.4**	3.1	3.6
admissible mass, median	–	0.90	0.90	**0.86**	0.85	0.86
mean	–	0.80	0.82	**0.85**	0.85	0.86
worst quarter	–	0.00	0.00	**0.50**	0.70	0.72

Table C.2 reports the sensitivity of the baseline to σ_y. The posterior standard deviation of the wedge rises with the prior but remains below it at every setting. Read as an outcome rather than a criterion, the last three rows show that a wider prior does not make the sample uniformly easier to price; it *trades* median admissibility for worst-case admissibility. At $\sigma_y =$

1 basis point the median quarter carries 0.90 of its posterior mass inside the chain, the highest figure in the table (with 2bp), while at least one quarter carries none at all; by $\sigma_y = 20$ the median has fallen to 0.86 and the worst quarter has risen to 0.72. The mean rises from 0.80 to 0.86 and is nearly flat from the baseline on.

The penalized alternative. A third alternative leaves the AAA yield alone, promotes μ_t to a drawn latent with a uniform prior on $[0, 1]$, as λ_t is drawn, and lets (20) fail by a residual ε_t penalized at a stated tolerance; its weight is (68), and as $\sigma_{AAA} \to 0$ it returns the closed form (61). The residual lands where the wedge does. Its posterior median is 3 basis points annualized over the sample; it exceeds 50 basis points in absolute value in 22 quarters, all of them in the episodes of curve inversion, where it reaches -1 to -2 percentage points. Read as the model intends, these are estimates of the expected capital gain on a long AAA bond in a quarter in which the curve is inverted; read skeptically, they are the size of the misspecification that the yield-as-forecast assumption imposes. The two readings cannot be separated with the data in hand.

As a specification, however, the penalized alternative is a larger change than it looks. When μ_t is drawn on a grid and pinned by a penalty of scale σ_{AAA}, the band of μ_t compatible with the data has a width proportional to $\sigma_{AAA}/\Delta_t(r_t^{TB})$, so the quasi-marginal likelihood of every quarter carries a factor $1/\Delta_t(r_t^{TB}) = 1/\bar{\ell}_t$. Given that model this is a correct Occam factor: a higher price of convenience pins μ_t more tightly, so more of a flat prior's mass is wasted. But the price of convenience rises with the Fisher rate, so the factor penalizes a high discount rate, and it does so in every one of the 218 quarters. It drags the posterior median discount rate from 3.3% to 2.4%, and with it the price of convenience from 4.9% to 4.2% and marketability from 0.49 to 0.35.

The mechanism can be checked directly. Adding $\log \Delta_t(r_t^{TB})$ back to the penalized specification's weights – undoing the bandwidth factor draw by draw, and nothing else – moves the posterior median discount rate from 2.41% to 2.89%, the price of convenience from 4.2% to 4.6%, and marketability from 0.354 to 0.390, against a baseline of 3.28%, 4.9% and 0.494. The factor therefore accounts for 61% of the gap in the price of convenience and 26% in mar-

ketability. In the discount rate it accounts for 56%, but that figure is the least reliable of the three, because the penalized specification draws the discount rate on a fifteen-point axis and its median moves between grid points. The factor is the largest single component of the gap that we can identify, and it is a pure artifact of where the prior was put; the remainder is whatever else drawing μ_t from a flat prior does. For marketability, where a prior is placed matters more than how wide it is: widening the prior on ψ_t tenfold, from $\sigma_y = 2$ to 20 basis points, moves the price of convenience from 4.8% to 5.5% and marketability from 0.52 to 0.46 (Table C.2), while moving the prior from ψ_t to μ_t moves them from 4.9% to 4.2% and from 0.49 to 0.35.

None of this is a statement about liquidity. It is the shadow that a flat prior over marketability, the parameter we set out to estimate, casts on the discount rate, which the data barely identify (Section 6.2). The hazard is general: Baumeister and Hamilton (2015) show that where the data identify only a set, a prior adopted for convenience continues to shape the quantities one goes on to report, however much data there is. The baseline therefore puts its prior on the wedge, a nuisance, and solves for marketability. Nor does the baseline put a residual on any of the other three equations, because there is nothing for one to represent: (17) is an identity given the definitions of ℓ_t^D and ℓ_t^N, (18) is the definition of the bill rate once $\mu_t^T = 1$, and (19) prices an instrument that matures within the period, so that its quoted yield is its holding return.

C.6 The identifying assumption of common marketability

Section 6.1 defends the assumption of $\mu^{CP} = \mu^{AAA}$ and gives the distance a departure would have to cover. This subsection carries the arithmetic.

By Proposition 3.1(iv) the urgent share is s_t/κ_t with $\kappa_t \equiv (1 - \mu_t^{CP})/(1 - \mu_t^{AAA})$, so a departure from $\kappa_t = 1$ rescales the series and cannot reshape it. Two reference points fix the scale. At the literature's corner, $\mu_t^{CP} = 0$ and $\kappa_t = 1/(1-\mu_t^{AAA}) \approx 2$, the sample median share is 0.087 against 0.180 under the assumption. Going the other way, urgency accounts for half of convenience in the median quarter when commercial paper is about 0.27 more marketable

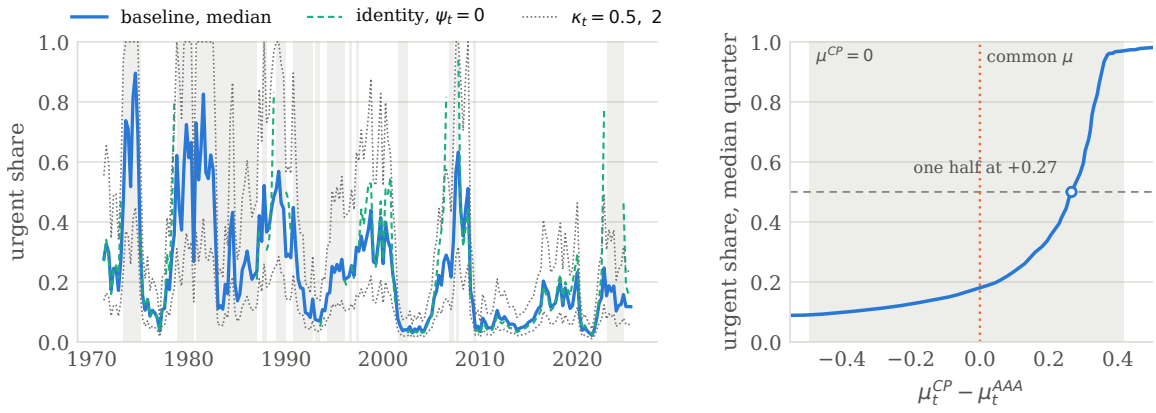

Figure C.6: The urgent share across the identified set for μ_t^{CP}. Left: the baseline posterior median of Figure 3 (solid), and the same series at $\kappa_t = 0.5$ and $\kappa_t = 2$ (dotted); the three are one curve rescaled, and the upper line clips where the rescaling would exceed one. Right: the share in the median quarter when $\mu_t^{CP} = \mu_t^{AAA} + d$, clipped to each quarter's identified set, which is the shaded region. The share passes one half at $d \approx 0.27$.

than the AAA bond, μ_t^{CP} near 0.76 against 0.49. Figure C.6 plots the whole curve.

Two cautions on reading it, both of which follow from Proposition 3.1(iii). The identified set is a *per-quarter* bound, $\mu_t^{CP} \leq \Delta_t(r_t^{CP})/\Delta_t(r_t^{TB})$, with a sample median of 0.92, so a displacement applied uniformly leaves the set in some quarters and the curve is clipped there: a constant $d = 0.27$ is outside the set in 36 of the 218 quarters. And the set is not a useful indication of anything, because its upper end holds by construction – at $A_t = 0$, the share equals 1.

C.7 A constant marketability parameter

Because the latents are independent across quarters, the quasi-likelihood factorizes and the profile over a single constant μ is available exactly:

$$\log L(\mu) \;=\; \sum_t \log \int W_t(\gamma, \beta, \lambda; \mu)\, \mathrm{d}\Pi(\gamma, \beta, \lambda),$$

each inner integral being the same 21^3 quadrature used in the exact inversion. Under the wedge convention, fixing μ forces the wedge, one value per quarter, through:

$$\big(1 + y_t^{AAA}\big)\big(1 + \psi_t\big) \;=\; \frac{1 + r_t^F}{1 + \mu\,\Delta_t(r_t^{TB})}, \tag{71}$$

and W_t is the prior (21) evaluated at that forced value. Because the baseline is exactly identified quarter by quarter and leaves nothing to select on, the $T-1$ over-identifying restrictions that a constant μ imposes are what make a likelihood comparison possible.

What the profile can and cannot test. It cannot deliver a Bayes factor of constant against time-varying μ, for the same reason that determined the baseline. In the baseline μ_t is solved rather than drawn, so it carries no prior; a specification in which μ_t moves freely therefore assigns no probability to the AAA yield that can be set against the probability assigned by one in which it does not, and a log Bayes factor between the two would measure that mismatch rather than the evidence. Two comparisons *are* available on the same measure, however, and both are interpretable.

The first compares one restrictions with another. Constant μ and $\mu = 0$ are both evaluated by (71) on the same prior, so their likelihoods are directly comparable. The profile (Figure C.7, left) is smooth and sharply peaked at $\widehat{\mu} = 0.47$ with a 90% posterior interval of $[0.40, 0.51]$ from its curvature, which a block bootstrap widens to about $[0.36, 0.54]$ (Appendix C.8), and the best single constant lies somewhat below the median of the time-varying series, 0.49. Against $\mu = 0$, it is preferred by 56 log points.

The second comparison reads the forced wedges in the units in which the prior is stated, signed so that a positive entry is an expected yield *decline* over the quarter, that is $-\mathbb{E}_t[\Delta y_{t+1}^{AAA}]$:

restriction	median	sd	max $\lvert\cdot\rvert$
$\mu = 0.47$ (highest likelihood)	-1.6bp	3.5bp	7.6bp
$\mu = 0$ (commensurate with literature)	$+1.9$bp	5.0bp	18.7bp
prior (21)	0	5.0bp	–

A constant marketability is well inside what the prior already admits about the wedge. Setting μ to zero instead forces a wedge to be as dispersed as the prior itself, with extremes near four prior standard deviations.

For the highlighted results in our paper, we retain time-varying μ_t because there is no a priori reason to make it constant. We report the constant case as an alternative in Table 1, where it lowers the price of convenience by 0.9 percentage points and raises the urgent share by 3 percentage points.

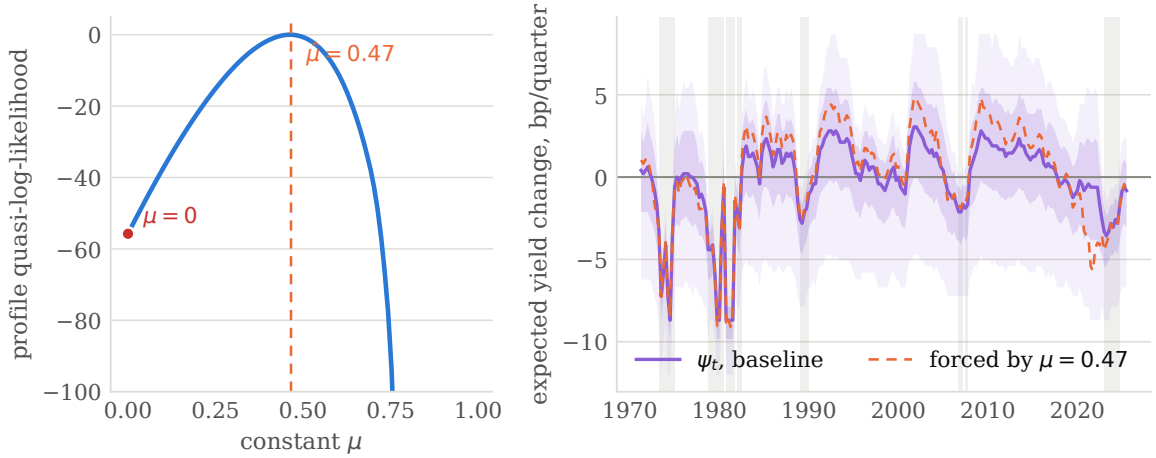

Figure C.7: Left: the profile quasi-log-likelihood over a single constant μ, normalized to zero at its maximum; the red point is $\mu = 0$, commensurate with the absence of secondary markets in the convenience yield literature. Right: the repricing wedge, in basis points of expected quarterly yield change, posterior median with 50% and 90% bands, and, dashed, the wedge that a constant μ would force. The baseline wedge is negligible outside the shaded episodes of curve inversion, and both series stay inside the prior.

C.8 Sampling uncertainty

Every number this paper reports across quarters – a sample median, a correlation, the profile quasi-log-likelihood over a constant μ – is a function of 218 serially dependent observations. The lag-1 autocorrelations are 0.93 for the price of convenience, 0.93 for urgency, 0.89 for marketability, 0.88 for the wedge and 0.85 for the urgent share. An interval computed as if those quarters were independent is therefore too narrow, and in this subsection we report by how much.

We use the stationary bootstrap of Politis and Romano (1994): quarters are resampled in blocks of geometric length, which preserves the local dependence without imposing a model for it, and intervals are percentiles of 10,000 replications. The expected block length is

chosen by the Politis and White (2004) rule, which returns 9 to 22 quarters for the series here; because that selector is itself estimated, every interval below was also computed at fixed expected lengths of 4, 8 and 16 quarters, and none of the conclusions turns on the choice. One caveat belongs with the numbers: they are sampling intervals for statistics computed at the posterior medians, not posterior objects, so they do not incorporate the estimation uncertainty that the credible bands in the figures show.

Levels. The price of convenience has a median of 4.9% a year with a 90% interval of $[4.4,\ 5.6]$ and a mean of 5.1% with $[4.8,\ 5.5]$; the urgent share a median of 0.180 with $[0.122,\ 0.250]$; marketability a median of 0.49 with $[0.47,\ 0.54]$; urgency a median of 0.33 with $[0.17,\ 0.43]$.

Comovements. The correlations quoted in Sections 4 and 6, with 90% intervals. The spread of commercial paper over bills enters net, as $(1 + r_t^{CP})^4 - (1 + r_t^{TB})^4$, and gross, as $\left[(1 + r_t^{CP})/(1 + r_t^{TB})\right]^4 - 1$:

	correlation	90% interval
urgency, bill rate	$+0.86$	$[+0.79,\ +0.91]$
urgent share, bill rate	$+0.73$	$[+0.60,\ +0.82]$
price of convenience, bill rate	-0.61	$[-0.76,\ -0.52]$
commercial-paper spread (net), bill rate	$+0.57$	$[+0.37,\ +0.71]$
commercial-paper spread (gross), bill rate	$+0.54$	$[+0.34,\ +0.69]$
price of convenience, commercial-paper spread (net)	-0.23	$[-0.39,\ -0.07]$
price of convenience, commercial-paper spread (gross)	-0.21	$[-0.38,\ -0.06]$
marketability, bill rate	$+0.01$	$[-0.31,\ +0.37]$
marketability, price of convenience	$+0.05$	$[-0.55,\ +0.35]$
marketability, urgency	-0.10	$[-0.41,\ +0.22]$

Urgency and the level of nominal rates. Section 6.3 reads the first two rows of the table above as a comovement with the level of nominal rates rather than as a trend or an artifact, and the checks behind that reading belong here. Both correlations survive controls for the commercial-paper spread and for the price of convenience, which are the two components

of the ratio that defines ω_t, so neither is doing the work. Neither is common trending: conditional on the bill rate there is no residual time trend in either series, and in first differences, taken over consecutive quarters in which commercial paper yields less than the AAA bond, the correlations with the bill rate are $+0.54$ for ω_t, $+0.39$ for the urgent share and $+0.56$ for the identity (29). (Quarters in which commercial paper yields more than the AAA bond are left out of the differencing, because in five of them the per-quarter median of the share moves with the resolution of the ψ grid; see Appendix C.11.) Nor is it an artifact of the zero lower bound, where ω_t might have been expected to become unidentified: the correlation survives dropping every quarter with a T-bill rate below one percent ($+0.81$) and holds within the high-rate half of the sample alone ($+0.73$). Finally, the loading is on the *nominal* rate and not the real one – in a regression with both regressors standardized, over the 215 quarters for which a trailing four-quarter inflation rate is available, ω_t takes $+0.22$ on the nominal rate against -0.015 on the real one, and the urgent share $+0.16$ against -0.040. The decade medians behind the non-monotonicity noted there are 0.41, 0.57, 0.41, 0.22, 0.08 and 0.23 for ω_t, against decade medians of the bill rate of 6.0, 8.9, 5.2, 2.1, 0.13 and 4.2 percent.

The constant-μ profile. The profile of Appendix C.7 is a sum over quarters of the per-quarter terms in (21) evaluated at a forced wedge, so a block bootstrap over quarters applies to it directly: resample the quarters, re-sum the rows, and re-maximize; the selected block length applies the same rule to the per-quarter terms at $\mu = 0$. The peak is stable – 0.465 in the sample, and a bootstrap mean of 0.455 – but the interval is not the one the profile's own curvature implies:

expected block length	90% interval for $\widehat{\mu}$	90% interval for the gap against $\mu = 0$
1 quarter (independent)	$[0.436,\ 0.495]$	$[44,\ 69]$
4 quarters	$[0.396,\ 0.525]$	$[32,\ 90]$
8 quarters	$[0.366,\ 0.535]$	$[27,\ 98]$
14 quarters (selected)	$[0.356,\ 0.544]$	$[23,\ 103]$
16 quarters	$[0.346,\ 0.544]$	$[22,\ 104]$

The interval on $\widehat{\mu}$ widens from $[0.40,\ 0.51]$, which is what the profile's curvature gives, to about $[0.36,\ 0.54]$; the main text quotes the wider one. But the comparison against $\mu = 0$ is not overturned by serial dependence at all. The gap is a sum of 218 per-quarter contributions, and resampling in blocks widens its sampling distribution without moving its centre: 56 log points in the sample, with a 90% interval of $[23,\ 103]$.

The wedge prior. Re-running the profile at other values of σ_y:

σ_y	$\widehat{\mu}$	90% profile interval	gap against $\mu = 0$
2bp	0.465	$[0.417,\ 0.501]$	107
5bp (baseline)	0.465	$[0.398,\ 0.509]$	56
10bp	0.366	$[0.279,\ 0.440]$	20
20bp	0.178	$[0.041,\ 0.291]$	2

The strength of the evidence against $\mu = 0$ is therefore a statement about how tight a prior we are willing to place on the repricing wedge, and not only about the data: at a prior four times the baseline's width the gap is two log points rather than fifty-six, and the best constant falls to 0.18. That is a larger qualification than the sampling one, and it is not visible in any interval. So σ_y matters far more for the *evidence* against $\mu = 0$ than for the *estimates*, which move by at most 0.7 points in the price of convenience across Table C.2; the grounds on which σ_y is fixed are set out in Appendix C.5.

The credible bands are not confidence intervals. The shaded regions in every figure in this paper are quantiles of a quasi-posterior: they describe how much the priors on $(\beta_t, \lambda_t, \psi_t)$ and the measurement distribution of γ_t leave undetermined within a quarter, propagated through an exactly identified inversion. They are not repeated-sampling statements, unlike the intervals in this subsection.

C.9 The equation-error estimator and sensitivity within the baseline

Section 5 compares four estimators in Table 1. The exact inversion is constructed in Appendix C.2, and the baseline in Appendix C.5, which adds the wedge to it; the constant-μ

estimator is in Appendix C.7. This subsection constructs the fourth, the equation-error estimator, and reports in Table C.3 the sensitivity of the baseline to its priors and tolerances.

The Sobol equation-error estimator draws β_t, λ_t, ω_t and μ_t from their priors, together with the trend γ_t, which every specification draws from its measurement distribution, and lets (19) and (20) fail by residuals that carry the two-sided Gaussian penalty of Appendix C.3 at a tolerance of 25 basis points a year. Its Sobol sequence therefore has five dimensions, of which four are structural: on the bookkeeping fixed in Section 3, γ_t is a measured first-stage input and is excluded from every count of free and drawn structural dimensions, so this estimator draws four where the no-wedge system leaves two. The Fisher rate is determined by (16) once γ_t and β_t are, and the two liquidity premia are determined by (18) and (17) once ω_t and λ_t are. Substituting (17) into (18) and solving,

$$\ell_t^N \;=\; \frac{\Delta_t(r_t^{TB}) - (1-\omega_t)\,\lambda_t\, r_t^{TB}}{1 + (1-\omega_t)\, r_t^{TB}}, \qquad \ell_t^D \;=\; (1 + r_t^{TB})\big(\lambda_t + \ell_t^N\big) - \lambda_t. \tag{72}$$

The premia therefore carry no priors of their own in this estimator, since insisting that (18) and (17) hold exactly leaves no freedom in which to place one. The only restriction imposed on them directly is the theoretical ordering $0 \le \ell_t^N \le \ell_t^D \le r_t^F$, which enters as an indicator that zeroes the weight rather than as a density, and which Proposition C.1 shows to be implied rather than additional wherever the model holds.

This also makes precise the sense in which the estimator is over-identified. The system leaves two dimensions free per quarter, so a coherent inversion may choose at most two latents and must solve for the rest, which is what the baseline does. The equation-error estimator chooses four, β_t, λ_t, ω_t and μ_t, two more than the system permits. Two equations must therefore fail, and (19) and (20) are the two we allow to fail. Consequently the data can reject this estimator in a way they cannot reject the baseline. If the best-fitting admissible point in a quarter still leaves a large residual, the model has failed in that quarter rather than merely strained.[20]

[20] In this estimator the weight concentrates on a codimension-two surface in five dimensions, which a product lattice intersects poorly; a lattice comparable to the baseline grid returned effective sample sizes in the tens. We therefore draw 2^{18} points from a scrambled Sobol sequence, which gives several thousand effective draws per

Table C.3: Sensitivity within the baseline. Sample medians over all 218 quarters. Every row draws the same latents and imposes the same equations; only a prior, a tolerance, or the curvature of utility changes.

	$\bar{\ell}$	urgent share	μ	ω	β (ann.)
Baseline ($\sigma_y = 5$bp)	4.9%	0.180	0.49	0.33	0.968
$\sigma_y = 2$bp	4.8%	0.198	0.52	0.32	0.970
$\sigma_y = 10$bp	5.1%	0.171	0.48	0.27	0.965
ρ exponential, median 3%	6.2%	0.167	0.57	0.28	0.956
$\xi = 2$ (Section 6.2)	5.9%	0.168	0.53	0.27	0.975
Two-sided trend	4.8%	0.180	0.50	0.33	0.968
$\sigma_{\text{tol}} = 100$bp p.a.	5.1%	0.180	0.50	0.32	0.968

Table C.3 varies the priors, tolerances, and ξ within the baseline, one at a time. Two features stand out. First, the urgent share is nearly invariant, at 0.167 to 0.198 across every row of Table C.3 and 0.159 to 0.208 across every column of Table 1, which is what Corollary C.1 implies; the residual movement comes from the wedge, the one prior that enters the identity. Second, the level moves with the prior on ρ by about 1.3 percentage points and with the wedge prior by about 0.3 of a point over a factor of five in σ_y; the total range across Tables 1, C.3 and C.2 is 4.0% to 6.2%. Only the two alternatives to the baseline's economic assumptions, the prior median of ρ and ξ, move the level by more than half a percentage point, and both raise it.

C.10 The persistence of the latents

Every specification reported so far treats the latents as independent across quarters. This subsection asks whether they move slowly, by comparing path priors. It does not produce an alternative estimate of the price of convenience, and the levels at the end of the subsection should not be read as one.

Why the equation-error specification. We conduct the exercise in the equation-error estimator of Appendix C.9 rather than in the baseline. In the baseline every equation is imposed exactly, so the emission is a constraint rather than a fit, since the only thing a quarter's weight

quarter and agreement across seeds to four decimal places. In the exactly identified specifications a lattice is efficient, because every draw is admissible, and Sobol sampling offers no advantage.

measures is whether the draw is admissible. Profiling a transition kernel there does not ask how smoothly the latents move but how easily a smooth path can satisfy admissibility, and the answer is pathological. Run in the exact inversion, which shares this property, the profile prefers near-constancy in ρ_t and pins the discount rate at a single high value in every one of the 218 quarters, which drives the price of convenience well above its baseline level. The mechanism is the one-sidedness of the penalties of Appendix C.3. Admissibility binds ρ_t from below only, so with independent draws only the eight strained quarters of 1981–84 need a high value, whereas with a smooth path one value must serve all 218 quarters and is pushed to the sample-wide maximum. Table C.1 shows the same fact from the other direction, since an imposed constant ρ cannot satisfy admissibility at a low value.

The equation-error estimator does not have this problem, because (19) and (20) carry residuals against observed yields, so that a smooth path which misprices an asset is penalized for mispricing it. The version used here adds the wedge back as a sixth latent, so that $(\gamma_t, \beta_t, \lambda_t, \omega_t, \mu_t, \psi_t)$ all carry priors on their own supports and each has a kernel estimable on the same footing; the Fisher rate and the two premia are solved from (16) and (72) and have no kernels of their own. A second reason points the same way: in the baseline ω_t is solved from (17) rather than drawn, so the persistence of urgency is not a well-posed statement about a prior there.

The kernel. Each latent is represented by the n equal-probability quantile points of its prior, indexed $0, \ldots, n-1$. The transition kernel is a Gaussian random walk on the index, with step standard deviation τn index units, reflected at both boundaries. Thus τ is a step size on the prior's own grid, expressed as a fraction of the grid's range, and not an autocorrelation. Setting $\tau = 0$ pins the latent at its initial draw for the whole sample, and $\tau \to \infty$ returns independence. Reflection makes the kernel symmetric and doubly stochastic. Because the grid points are equal-probability quantiles, the uniform distribution over ranks is the prior, and a doubly stochastic kernel leaves it stationary. The marginal prior on every latent in every quarter is therefore identical for every τ, and only the joint distribution over paths changes. Differences in the marginal likelihood across τ are consequently Bayes factors between path

priors, uncontaminated by any shift in the per-period prior.

For a vector $\tau = (\tau_\gamma, \tau_\rho, \tau_\lambda, \tau_\omega, \tau_\mu, \tau_\psi)$,

$$\log L(\tau) \;=\; \log \int \prod_t b_t(\theta_t)\, \mathrm{d}\Pi_\tau(\theta_1, \ldots, \theta_T), \tag{73}$$

where b_t is the emission, the two residual penalties of Appendix C.9 together with the admissibility indicator, and Π_τ is the path prior. Since the grid is discrete, this is a hidden Markov model whose transition matrix is a Kronecker product of the six kernels, so (73) is the forward-pass likelihood and costs $O\big(TG\sum_j n_j\big)$ with $G = \prod_j n_j$ states. The emissions are computed once and reused across the whole profile, so each additional value of τ costs one forward pass. We use $9 \times 9 \times 7 \times 17 \times 17 \times 7 = 1{,}147{,}041$ states; the mean effective sample size of the lattice is about $45{,}000$ per quarter and its minimum $1{,}700$, so a product grid represents this posterior adequately. The lattice of Appendix C.9 did not, because there both private-asset equations bind at 25 basis points; here, the drawn wedge lets the AAA equation be met across a band as wide as the wedge prior, so only the commercial-paper equation binds tightly. Because τ maps monotonically but non-linearly to the autocorrelation of the latent's value, we report throughout the implied lag-1 autocorrelation under the stationary law rather than τ itself.

Results. Table C.4 reports two profiles. The first scans each τ_j with the other five held at independence; the second scans each with the other five at the joint optimum, located by coordinate ascent. The second is the profile that matters, because the emission constrains the latents jointly: a smooth μ_t is useful only if ω_t is also smooth, since the two together must reconcile the same two spreads. Consistent with this, the gains are super-additive; the six separate gains sum to 420 log points and the joint optimum gives 480.

Four features of Table C.4 and Figure C.8 carry the message. First, persistence matters most for the wedge. Its separate gain over independence is 151 log points, the largest of the six latents, at an implied lag-1 autocorrelation of 0.96, and a constant ψ_t falls a further 300 points below independence, so both ends are rejected.

Table C.4: Persistence of the latents in the equation-error specification, six axes. "Separate" scans each axis with the other five independent; "joint" scans each with the other five at the joint optimum. $\Delta \log L$ is against independence in that same configuration. Under "diffusion" the kernel is the pure random walk $\mathrm{RW}(\tau_j)$; under "mixture" it is (74) with (τ_j, p_j) both fitted. Autocorrelations are lag-1 at quarterly frequency, implied by the fitted kernel.

	separate, diffusion		separate, mixture		joint (diffusion)	
latent	$\widehat{\tau}$	$\Delta \log L$	$(\widehat{\tau}, \widehat{p})$	$\Delta \log L$	$\widehat{\tau}$	autocorr.
γ_t	0.04	26.9	$(0.00,\ 0.05)$	30.0	0.04	1.00
ρ_t	0.08	64.0	$(0.04,\ 0.05)$	66.1	0.04	0.99
λ_t	0.00	35.8	$(0.00,\ 0)$	35.8	0.00	1.00
ω_t	0.15	82.2	$(0.00,\ 0.10)$	96.3	0.15	0.89
μ_t	0.04	60.7	$(0.00,\ 0.05)$	61.3	0.04	0.99
ψ_t	0.08	150.7	$(0.08,\ 0.05)$	152.5	0.08	0.96
all six		(sum 420)		(sum 442)		**480.1**
mixture, joint						**492.7**

Second, every latent is persistent. The implied autocorrelations at the joint optimum run from 0.89 to 1.00 at quarterly frequency, and independence is rejected by between 27 and 151 log points depending on the latent. The size of the joint gain should not be over-read. The emission constrains roughly two of the six dimensions to a band of a few basis points, and under independence each quarter pays the prior-volume cost of landing in that band afresh; the gain is therefore about the number of constrained dimensions times the log width ratio per quarter, which at 2.2 log points a quarter is what we observe. A large Bayes factor against independence mainly establishes that independence was an implausible null; the informative comparisons are those within the persistent family.

Third, urgency is rejected at both ends. Scanned separately, ω_t has an interior optimum at an autocorrelation of 0.89, as at the joint optimum; independence is rejected by 82 log points and a constant ω by a further 114 below independence. Urgency moves, at a speed the data can determine, and it is neither noise nor a constant. The wedge is also rejected at both ends, more sharply still, but for the wedge this is a statement about a measurement object rather than about the economy. Marketability has an interior optimum at 0.99, and a constant μ falls 31 log points short of it.

Fourth, the γ_t row is a consistency check rather than a result. The kernel acts on the

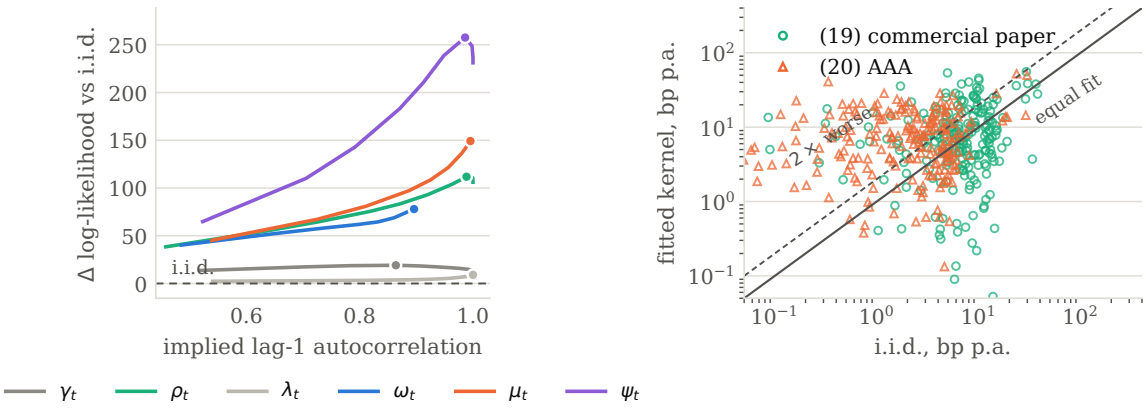

Figure C.8: Left: the profile marginal likelihood of each latent against its implied lag-1 autocorrelation, with the other five at the joint optimum; dots mark the optima, the dashed line is independence, and the right-hand edge is a latent held constant for the whole sample. Right: posterior-median absolute residuals under independence and under the fitted mixture kernel, quarter by quarter, on log axes. The AAA equation's residuals move decisively above the 45-degree line and the commercial-paper equation's straddle it, which is the asymmetry discussed in the text.

latent's rank within its own prior, and the prior on γ_t is the time-varying measurement distribution $N(m_t, V_t + \sigma^2_{\text{tol}})$ of Appendix C.1. Persistence on that axis is therefore persistence of the trend's measurement error, not of the trend, which is a random walk by construction. An autocorrelation near one is the expected answer, since the one-sided filter's prediction error $\gamma_t - m_t$ equals last period's filtered error plus this period's innovation and is serially correlated by construction.

Allowing the latents to jump. A Gaussian random walk is a strong prior for a financial state variable. Because its log-density is quadratic in the size of the move, it does not merely make large jumps unlikely but effectively forbids them: at the fitted step sizes a one-quarter fall in marketability of a third of its range, the kind of move one might expect in September 2008, is assigned a prior probability of order 10^{-15}. Dealer capacity and the urgency of liquidity needs are states one expects to be stable for years and then to move a long way in a short time, and a kernel that rules this out is not a neutral way of asking whether they are persistent.

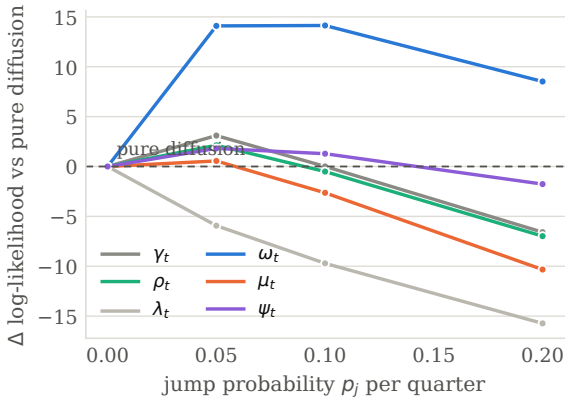

Figure C.9: What each latent does with a jump. For every axis and every jump probability p_j, the profile takes the best step size τ_j; the curve is therefore the profile in p_j alone, measured against that axis's own pure diffusion. Five of the six axes are nearly flat or falling: γ_t, ρ_t, μ_t and ψ_t gain at most three log points at $p = 0.05$, and λ_t loses immediately. Urgency is the exception, gaining more than ten log points and preferring a complete redraw roughly once every ten quarters. The mixture nests the diffusion at $p = 0$, so this is a choice the data made and not a restriction we imposed.

We therefore re-estimate with a mixture kernel,

$$T_j \;=\; (1 - p_j)\,\mathrm{RW}(\tau_j) \;+\; p_j\,\tfrac{1}{n_j}\mathbf{1}\mathbf{1}', \tag{74}$$

under which the latent takes a small random-walk step with probability $1 - p_j$ and is redrawn from its own prior, a jump of arbitrary size, with probability p_j. Both components are symmetric and doubly stochastic, so the mixture is as well, and the marginal prior is again unchanged, so the Bayes-factor logic above survives. The pure random walk is $p_j = 0$ and independence is $p_j = 1$, so (74) nests both endpoints of Figure C.8 and separates two things the Gaussian kernel confounds, how far a state drifts in a normal quarter and how often it moves abruptly.

We fit (τ_j, p_j) separately on each of the six axes, so that each latent is free to indicate whether it drifts or switches. On λ_t the fitted p is exactly zero; on γ_t, ρ_t, μ_t and ψ_t a small jump probability is marginally preferred but gains at most three log points apiece. At the joint optimum p is zero on five of the six axes, and the whole gain of the mixture over the pure diffusion, 492.7 against 480.1, or 12.5 log points, comes from urgency. For urgency the

description changes qualitatively. Its diffusion profile is negative at the two smallest step sizes, -114 and -104 log points, so near-constancy is far worse than independence for ω_t; yet $\tau_\omega = 0$ combined with $p_\omega = 0.10$ is the best point on the axis. Read literally, this is a latent that is constant between shocks and completely redrawn about once every ten quarters. At the joint optimum the fitted mixture is milder, $(\tau_\omega, p_\omega) = (0.04, 0.10)$, but the shape of the preference is the same, and it accords with the economics of a liquidity scramble, which begins abruptly rather than gradually.

The cost and the benefit of persistence. The right panel of Figure C.8 plots the posterior-median absolute residual in each private-asset equation, quarter by quarter, under independence and under the fitted mixture kernel. Over all 218 quarters, the median absolute residual on (19) barely moves, 7.0 basis points a year under independence against 7.0 under the fitted kernel, while the residual on (20) rises from 2.4 to 7.1. Smoothing costs the AAA equation nearly threefold and the commercial-paper equation next to nothing, which is to be expected, since (20) is the equation in which the wedge enters, and a wedge constrained to move smoothly cannot track a single quarter's yield. The cost is already accounted for in the likelihood, because the residual penalties are part of the emission b_t, so the 493 log points of Table C.4 are net of the deteriorated fit.

Persistence does move the discount factor. Its posterior median (reported, as everywhere in this paper, as the annual discount factor $e^{-\rho_t}$) is 0.980 under independence and 0.950 under the fitted kernel, past the baseline's 0.968.

The evidence, however, is not concentrated in the episodes the model cannot price. Decomposing (73) into its one-step-ahead predictive increments, 306 of the 493 log points accrue over the 139 quarters in which the ordering (25) holds at $\rho = 2.9\%$ (the prior mean) and 187 over the 79 in which it fails, or 2.20 against 2.36 log points per quarter. Dropping the problem episodes would leave the conclusion of this subsection intact.

Figure C.10 shows the two series under both specifications, and Table C.5 below quantifies the narrowing of the bands.

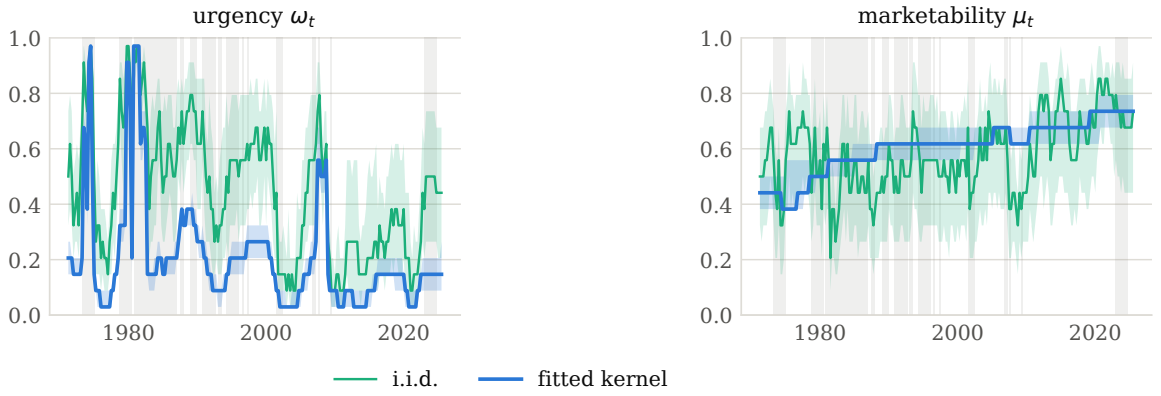

Figure C.10: Urgency and marketability under independent latents and under the fitted mixture kernel, equation-error specification. Posterior medians with interquartile bands. Shaded columns mark quarters in which the ordering fails at $\rho = 2.9\%$ (the prior mean). The bands are the point of the figure; pooling across quarters increases precision, and it also moves the levels.

How much of this the baseline already had. The baseline series of Section 4 are already smooth without any transition kernel. Table C.5 measures the realized lag-1 autocorrelation of each posterior-median series, which describes the estimate where the kernel autocorrelation of Table C.4 describes the prior.

Table C.5: Properties of the estimated series. Every column uses every quarter with an estimate, which is all 218 except for the exact inversion, undefined in the 34 quarters in which $r^{CP} > y^{AAA}$. The last two rows are the equation-error specification of this subsection, which adds ψ_t as a sixth latent. The mixture kernel is (74) at its fitted (τ_j, p_j). "Autocorr." is the realized lag-1 autocorrelation of the posterior-median series – an output, not the kernel input of Table C.4. $|\Delta|$ is the mean absolute quarter-on-quarter change in the median; IQR is the average posterior interquartile range.

	urgency ω_t				marketability μ_t			
	autocorr.	$\lvert\Delta\rvert$	IQR	median	autocorr.	$\lvert\Delta\rvert$	IQR	median
Wedge baseline (Sec. 4)	0.93	0.064	0.240	0.33	0.89	0.027	0.394	0.49
Exact inversion	0.84	0.085	0.093	0.23	0.92	0.042	0.304	0.46
Equation-error, i.i.d.	0.91	0.065	0.346	0.44	0.79	0.058	0.399	0.56
Equation-error, mixture kernel	0.83	0.039	0.084	0.15	0.99	0.002	0.096	0.62

Three readings of the table together make the most important qualification in this subsection. First, the persistence is in the data, not in the prior. The wedge baseline, which imposes no kernel, already delivers ω_t with a realized autocorrelation of 0.93 and μ_t with 0.89; the fit-

ted kernel moves these to 0.83 and 0.99. Both series inherit their smoothness from the yields, which are themselves close to random walks, through equations that are nearly deterministic functions of them. What Table C.4 establishes is that a path prior which matches the smoothness the data already exhibit is far more probable than one which contradicts it.

Second, for urgency the kernel mostly repairs the equation-error estimator rather than improving on the baseline. The baseline's posterior interquartile range for ω_t is 0.240; the equation-error estimator at independence widens it to 0.346, which is the price of drawing ω_t from a prior instead of solving it from (17); the kernel then narrows it to 0.084. Much of the 493 log points is therefore the equation-error model recovering, through a smoothness prior, information that the exactly identified system obtains directly from an identity, which is a fact about the estimator and not about the economy. The kernel also goes past the baseline, cutting the mean quarter-on-quarter movement in ω_t from 0.064 to 0.039 and its median from 0.33 to 0.15, and the movement removed is movement the identity says is there. Since the baseline already produces an ω_t as autocorrelated as any kernel would make it, we see no case for imposing a kernel on urgency.

Third, for marketability the exercise is informative. μ_t is less precisely identified than ω_t, with a baseline average interquartile range of 0.394, because it is identified only through the AAA equation. Pooling across quarters cuts the range to 0.096; this gain, however, comes with a different series: 0.62 under the kernel against 0.49 in the baseline, the two correlating only 0.22. The one place where persistence adds information is therefore the one place where it changes the answer, and it is the parameter identified by the AAA yield, the observable whose interpretation Appendix C.5 treats with the most caution.

Why these are diagnostics and not estimates. Over all 218 quarters, the fitted kernel reports a price of convenience of 8.0% against 3.6% under independence, marketability of 0.62 against 0.56, urgency of 0.15 against 0.44, and an urgent share of 0.142 against 0.263. Except for marketability, these are large movements, and we do not treat them as an alternative measurement, for two reasons in addition to the deterioration of the AAA fit already shown.

First, they are movements within a different estimator. At independence, the equation-

error specification used here already reports 3.6% and 0.56 against the baseline's 4.9% and 0.49. These figures also differ from the equation-error column of Table 1, 4.2% and 0.41, because that column carries no wedge.

Second, the kernel's hyperparameters are profiled rather than integrated. We choose (τ_j, p_j) to maximize the same quasi-likelihood from which the posterior is then read. This is ordinary empirical Bayes, but it means that the bands above condition on a point estimate of the prior, which is most of why they collapse by a factor of about four. An estimate offered as a rival to Section 4 would have to carry the uncertainty in (τ_j, p_j) as well, and would be much less precise. Borrowing strength across quarters is an appropriate way to learn about dynamics; it is a poor trade when the object wanted is the quarter-by-quarter level, because we give up quarter-by-quarter fit.

C.11 Numerical implementation

The priors are $\gamma_t \sim N(m_t, V_t + \sigma^2_{\text{tol}})$, $\beta_t \sim \text{Beta}(200 \ln 2,\ 1)$, $\lambda_t \sim U[0,1]$ and, in the baseline, ψ_t from (21); marketability is solved from (24) and carries no prior of its own. We report rates in annual units, but the estimator works in quarters: β_t is the one-quarter discount factor of (16), and the object we place a prior on is the *annualized* discount rate $\rho_t \equiv -4 \log \beta_t$, drawn as $\rho_t \sim \text{Exp}$ with median 0.02 a year (mean $\theta = 0.02/\ln 2 \approx 0.0289$). Since $\rho \sim \text{Exp}(\text{mean } \theta)$ implies $e^{-\rho/4} \sim \text{Beta}(4/\theta, 1)$, that prior is $\beta_t \sim \text{Beta}(200 \ln 2,\ 1) \approx \text{Beta}(138.6,\ 1)$ on the quarterly factor. Its median and mean in annual-discount-factor units, which is how Table 1 reports it, are 0.980 and 0.972; in quarterly units they are 0.995 and 0.993.

Each latent variable is represented by the equal-probability points of its prior, that is, the $\big((i - \frac{1}{2})/n\big)$-quantiles, so that every grid point carries equal prior weight. The baseline uses 15 points on the γ axis, 201 on ρ, 18 on λ and 61 on ψ, or $3{,}310{,}470$ points in all; the exact inversion uses 21 points on each of its three axes, γ, ρ and λ. Sample medians and means are invariant to the grid at the precision we report. Moving γ from 15 points to 31 shifts the median price of convenience by 1bp and leaves the urgent share unchanged; moving ψ from 61 points to 121, on a coarser grid in γ and λ, moves the sample medians and means

of the price of convenience, marketability, urgency and the urgent share, and their decade medians, by at most two units in the last digit printed. Individual quarters are less robust: in five quarters in which the commercial paper rate exceeds the AAA yield (1979Q4, 1980Q1, 1980Q4, 1981Q1 and 1981Q2), the median urgent share moves by 0.05 to 0.15.

The grid is visible in one figure. We set $n_\lambda = 18$ because $n_\lambda \equiv 2 \pmod 4$ places the quartiles of a uniform grid exactly on 0.25 and 0.75; since the posterior of λ_t almost exactly reproduces its prior, this is convenient for a reader comparing the two in Figure 8.

Latents are independent across quarters in every specification the paper reports. The machinery for persistence, used in the equation-error estimator in Appendix C.10, is the following: a reflecting Gaussian random walk on the equal-probability quantile grid that is doubly stochastic and therefore leaves each marginal prior unchanged, applied by Kronecker contraction. That same appendix also explains why it is not applied to the baseline.

We summarize each quarter's posterior by its median and its 5th, 25th, 75th and 95th percentiles, not by its mean, since $\ell_t^D = A_t/(1-\omega_t)$ has a tail heavy enough that its posterior mean does not exist. Unless stated otherwise, a statistic across quarters, such as a sample median, mean or extreme, is computed from the per-quarter medians.

References

Adrian, T., R. K. Crump, and E. Moench (2013). Pricing the term structure with linear regressions. *Journal of Financial Economics 110*(1), 110–138.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2008). Eliciting risk and time preferences. *Econometrica 76*(3), 583–618.

Andolfatto, D. (2015). A model of US monetary policy before and after the great recession. *Federal Reserve Bank of St. Louis Review 97*(3), 233–56.

Andolfatto, D. (2020). Maturity structure and liquidity risk. Working Paper 2020-008, Federal Reserve Bank of St. Louis.

Andolfatto, D. and F. M. Martin (2023). Welfare-Enhancing Inflation and Liquidity Premia. *Review of Economic Dynamics 51*, 1036–1047.

Andreasen, M. M., J. H. Christensen, and S. Riddell (2021). The TIPS liquidity premium. *Review of Finance 25*(6), 1639–1675.

Andreoni, J. and C. Sprenger (2012). Estimating time preferences from convex budgets. *American Economic Review 102*(7), 3333–56.

Bao, J., J. Pan, and J. Wang (2011). The illiquidity of corporate bonds. *The Journal of Finance 66*(3), 911–946.

Bauer, M. D. and J. D. Hamilton (2018). Robust bond risk premia. *Review of Financial Studies 31*(2), 399–448.

Baumeister, C. and J. D. Hamilton (2015). Sign restrictions, structural vector autoregressions, and useful prior information. *Econometrica 83*(5), 1963–1999.

Benigno, P. and R. Robatto (2019). Private Money Creation, Liquidity Crises, and Government Interventions. *Journal of Monetary Economics 106*, 42–58.

Bento, A., K. Roth, and A. Waxman (2024). The value of urgency: Evidence from real-time congestion pricing. *Journal of Political Economy Microeconomics 2*(4), 786–851.

Canzoneri, M. B., R. E. Cumby, and B. T. Diba (2007). Euler equations and money market interest rates: A challenge for monetary policy models. *Journal of Monetary Economics 54*(7), 1863–1881.

Caramp, N., A. Geromichalos, S. Lee, and Y. Wang (2025). Why do (so many) corporate bonds pay coupons? Working paper.

Caramp, N. and S. R. Singh (2023). Bond Premium Cyclicality and Liquidity Traps. *Review of Economic Studies 90*(6), 2822–2879.

Chen, R.-R. and L. Scott (1993). Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates. *The Journal of Fixed Income 3*(3), 14–31.

Cheng, Y. and R. Robatto (2024). Fire Sales and Liquidity Requirements. Bank of Canada Staff

Working Paper 24-18; formerly "Financial Intermediation and Fire Sales with Liquidity Risk Pricing".

Chernozhukov, V. and H. Hong (2003). An MCMC approach to classical estimation. *Journal of Econometrics 115*(2), 293–346.

Choi, S., I. Jang, K.-Y. Kang, and H. Kim (2026). Haircut, Interest Rate, and Collateral Quality in the Tri-Party Repo Market. *Review of Economic Dynamics 61*, 101351.

Cieslak, A., W. Li, and C. Pflueger (2024). Inflation and Treasury Convenience. NBER Working Paper 32881.

Cipriani, M. and G. La Spada (2021). Investors' Appetite for Money-Like Assets. *Journal of Financial Economics 140*(1), 250–269.

Cochrane, J. H. and M. Piazzesi (2005). Bond risk premia. *American Economic Review 95*(1), 138–160.

Cohen, A., M. Kargar, B. Lester, and P.-O. Weill (2024). Inventory, market making, and liquidity in OTC markets. *Journal of Economic Theory 222*, 105917.

Copeland, A., D. Duffie, A. Martin, and S. McLaughlin (2012). Key mechanics of the U.S. tri-party repo market. *Federal Reserve Bank of New York Economic Policy Review 18*(3), 17–27.

Del Negro, M., D. Giannone, M. P. Giannoni, and A. Tambalotti (2017). Safety, liquidity, and the natural rate of interest. *Brookings Papers on Economic Activity 2017*(1), 235–316.

Di Tella, S., B. Hébert, and P. Kurlat (2024). Aggregation, Liquidity, and Asset Prices with Incomplete Markets. NBER Working Paper 32268.

Di Tella, S., B. M. Hébert, P. Kurlat, and Q. Wang (2026). The zero-beta interest rate. *Journal of Political Economy 134*(7), 2074–2118.

Duffee, G. R. (2002). Term premia and interest rate forecasts in affine models. *The Journal of Finance 57*(1), 405–443.

Duffee, G. R. (2011). Information in (and not in) the term structure. *Review of Financial Studies 24*(9), 2895–2934.

Duffie, D., N. Gârleanu, and L. H. Pedersen (2005, November). Over-the-counter markets. *Econometrica 73*(6), 1815–1847.

Ferraris, L. and M. Watanabe (2008). Collateral secured loans in a monetary economy. *Journal of Economic Theory 143*(1), 405–424.

Fleskes, N. (2025). Liquidity Crises and Endogenous Dealer Capacity in Over-the-Counter Markets. Working paper.

Geromichalos, A. and L. Herrenbrueck (2016). Monetary policy, asset prices, and liquidity in over-the-counter markets. *Journal of Money, Credit, and Banking 48*(1), 35–79.

Geromichalos, A. and L. Herrenbrueck (2022). The liquidity-augmented model of macroeconomic aggregates: A New Monetarist DSGE approach. *Review of Economic Dynamics 45*,

134–167.

Geromichalos, A., L. Herrenbrueck, C. Lee, and S. Lee (2024). What's So Inconvenient About TIPS? Working paper.

Geromichalos, A., L. Herrenbrueck, and S. Lee (2023a). Asset safety versus asset liquidity. *Journal of Political Economy 131*(5), 1172–1212.

Geromichalos, A., L. Herrenbrueck, and S. Lee (2023b). The strategic determination of the supply of liquid assets. *Review of Economic Dynamics 49*, 1–36.

Geromichalos, A., L. Herrenbrueck, and K. Salyer (2016). A search-theoretic model of the term premium. *Theoretical Economics 11*(3), 897–935.

Greenwood, R., S. G. Hanson, and J. C. Stein (2015). A comparative-advantage approach to government debt maturity. *The Journal of Finance 70*(4), 1683–1722.

Gross, M., Y. Li, and A. Wang (2022). Dealer intermediation in the primary market of commercial paper. FEDS Notes, Board of Governors of the Federal Reserve System, 7 October.

Harrison, G. W., M. I. Lau, and H. I. Yoo (2025). Constant discounting, temporal instability, and dynamic inconsistency in Denmark: A longitudinal field experiment. *International Economic Review 66*(1), 363–392.

Havranek, T., R. Horvath, Z. Irsova, and M. Rusnak (2015). Cross-country heterogeneity in intertemporal substitution. *Journal of International Economics 96*(1), 100–118.

Herrenbrueck, L. (2019). Frictional asset markets and the liquidity channel of monetary policy. *Journal of Economic Theory 181*, 82–120.

Herrenbrueck, L. and A. Geromichalos (2017). A tractable model of indirect asset liquidity. *Journal of Economic Theory 168*, 252–260.

Herrenbrueck, L. and Z. Wang (2026). Interest rates, moneyness, and the Fisher equation. *Canadian Journal of Economics, in press*.

Joslin, S., W. Li, and Y. Song (2019). The Term Structure of Liquidity Premium. Working paper.

Joslin, S., K. J. Singleton, and H. Zhu (2011). A new perspective on Gaussian dynamic term structure models. *Review of Financial Studies 24*(3), 926–970.

Juelsrud, R., P. Nenov, F. Schneider, and O. Syrstad (2025). Transaction costs, the value of convenience, and the cross-section of safe asset returns. Working paper, Bank of Canada.

Kargar, M., B. Lester, D. Lindsay, S. Liu, P.-O. Weill, and D. Zuniga (2021). Corporate Bond Liquidity During the COVID-19 Crisis. *Review of Financial Studies 34*(11), 5352–5401.

Krishnamurthy, A. and W. Li (2023). The demand for money, near-money, and Treasury bonds. *The Review of Financial Studies 36*(5), 2091–2130.

Krishnamurthy, A. and A. Vissing-Jorgensen (2012). The aggregate demand for Treasury debt. *Journal of Political Economy 120*(2), 233–267.

Lagos, R. (2010, November). Asset prices and liquidity in an exchange economy. *Journal of Monetary Economics 57*(8), 913–930.

Lagos, R. and G. Rocheteau (2009). Liquidity in asset markets with search frictions. *Econometrica 77*(2), 403–426.

Lagos, R., G. Rocheteau, and P.-O. Weill (2011). Crises and liquidity in over-the-counter markets. *Journal of Economic Theory 146*(6), 2169–2205.

Lagos, R., G. Rocheteau, and R. Wright (2017). Liquidity: A New Monetarist perspective. *Journal of Economic Literature 55*(2), 371–440.

Lagos, R. and R. Wright (2005). A unified framework for monetary theory and policy analysis. *Journal of Political Economy 113*(3), 463–484.

Lagos, R. and S. Zhang (2019). A monetary model of bilateral over-the-counter markets. *Review of Economic Dynamics 33*, 205–227.

Lagos, R. and S. Zhang (2020). Turnover Liquidity and the Transmission of Monetary Policy. *American Economic Review 110*(6), 1635–1672.

Lebeau, L. (2020). Credit Frictions and Participation in Over-the-Counter Markets. *Journal of Economic Theory 189*, 105100.

Longstaff, F. A. (2004). The flight-to-liquidity premium in U.S. Treasury bond prices. *The Journal of Business 77*(3), 511–526.

Nagel, S. (2016). The liquidity premium of near-money assets. *The Quarterly Journal of Economics 131*(4), 1927–1971.

Piazzesi, M., J. Salomao, and M. Schneider (2015). Trend and cycle in bond premia. Unpublished manuscript, Stanford University.

Pinter, G., S. Uslu, and J.-C. Wijnandts (2025). Comparing Search and Intermediation Frictions Across Markets. BIS Working Paper 1283.

Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical Association 89*(428), 1303–1313.

Politis, D. N. and H. White (2004). Automatic block-length selection for the dependent bootstrap. *Econometric Reviews 23*(1), 53–70.

Robson, A., L. Samuelson, and J. Steiner (2023). Decision theory and stochastic growth. *American Economic Review: Insights 5*(3), 357–376.

Rocheteau, G., R. Wright, and S. X. Xiao (2018). Open market operations. *Journal of Monetary Economics 98*, 114–128.

Stambaugh, R. F. (1999). Predictive regressions. *Journal of Financial Economics 54*(3), 375–421.

Sunderam, A. (2015). Money creation and the shadow banking system. *The Review of Financial Studies 28*(4), 939–977.

Thornton, D. L. and G. Valente (2012). Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *Review of Financial Studies 25*(10), 3141–3168.

Uhlig, H. (2005). What are the effects of monetary policy on output? Results from an agnostic identification procedure. *Journal of Monetary Economics 52*(2), 381–419.

Uslu, S. (2019). Pricing and Liquidity in Decentralized Asset Markets. *Econometrica 87*(6), 2079–2140.

Uslu, S. and G. Velioglu (2025). Risk and Illiquidity Premia in the Cross Section of OTC Assets. Working paper.

van Binsbergen, J. H., W. F. Diamond, and M. Grotteria (2022). Risk-free interest rates. *Journal of Financial Economics 143*(1), 1–29.

Vayanos, D. and P.-O. Weill (2008). A search-based theory of the on-the-run phenomenon. *The Journal of Finance 63*(3), 1361–1398.

Venkateswaran, V. and R. Wright (2014). Pledgability and liquidity: A New Monetarist model of financial and macroeconomic activity. *NBER Macroeconomics Annual 28*(1), 227–270.

Weill, P.-O. (2008). Liquidity premia in dynamic bargaining markets. *Journal of Economic Theory 140*(1), 66–96.

Wittwer, M. and J. Allen (2026). Market Power and Capital Constraints. *American Economic Review 116*(4), 1309–1339.