Simple Models and Biased Forecasts[*]

Pooya Molavi[†]

September 23, 2024

This paper proposes a framework in which agents are constrained to use simple models to forecast economic variables and characterizes the resulting biases. It considers agents who can only entertain state-space models with no more than d states, where d measures the intertemporal complexity of a model. Agents are boundedly rational in that they can only consider models that are too simple to capture the true process, yet they use the best model among those considered. Using simple models adds persistence to forward-looking decisions and increases the comovement among them. This mechanism narrows the gap between business-cycle theory and data. In a new neoclassical synthesis model, the assumption that agents use simple models fits the data much better than the rational-expectations hypothesis. Moreover, simple models simultaneously resolve the Barro-King and forward guidance puzzles while improving the propagation of TFP shocks.

[*] This paper subsumes and substantially extends the second half of my job-market paper, titled "Macroeconomics with Learning and Misspecification: A General Theory and Applications." I am grateful to Daron Acemoglu, Marios Angeletos, Larry Christiano, Ian Dew-Becker, Théo Durandard, Marty Eichenbaum, Drew Fudenberg, Laura Gáti, Lars Hansen, Ben Hébert, Ali Jadbabaie, Diego Känzig, Chen Lian, Alessandro Pavan, Giorgio Primiceri, Bruno Strulovici, Alireza Tahbaz-Salehi, Harald Uhlig, Iván Werning, and numerous seminar and conference participants for their helpful comments.

[†] Northwestern University, pmolavi@kellogg.northwestern.edu.

1 Introduction

When faced with the difficult task of forecasting in a complex world, people tend to rely on simple models and past experiences. Yet the rational-expectations hypothesis assumes that agents can forecast the future as if they knew the true model of the economy. The unrealistic nature of the rational-expectations assumption would not be of great concern if the predictions of the standard macro models were robust to alternative specifications of expectations. However, the answers to many important questions in macroeconomics, ranging from the power of forward guidance to the response of the economy to aggregate supply and demand shocks, are sensitive to how agents form their expectations.

This paper studies the forecasting biases arising from individuals' reliance on simple models and the macroeconomic implications of those biases. I introduce a framework in which the true data-generating process features complex intertemporal relationships among variables, but agents can only entertain stochastic models with a bound on their intertemporal complexity. Specifically, they only consider stochastic processes that can be represented using a d-dimensional state variable, where d is a parameter that measures the complexity of an agent's model. Agents are boundedly rational in that they can only entertain models that are too simple to capture the true process, but they find the best d-dimensional approximation to it.

The framework has sharp predictions for agents' forecasts and forward-looking actions. I show that agents misperceive *intertemporal* statistical relationships among time-series variables. In particular, while agents can accurately forecast the most persistent components of those observables, they miss the dynamics of the other components. This bias increases the *persistence* and *comovement* of agents' forward-looking choices. Agents' forecasts and forward-looking actions are anchored to the most persistent state variables in the economy. This increases the persistence of those forecasts and actions. Furthermore, different agents—with different payoffs, facing different decisions, and using models with different dimensions—all agree on the most persistent components. Since forward-looking decisions of such agents are influenced by a limited set of common components, those decisions comove more than they would under rational expectations.

This paper focuses on bounded rationality manifested as a reduction in intertemporal complexity. I make two main assumptions, which allow me to abstract from errors that agents might make when confronting other forms of complexity. First, I assume that agents are capable of entertaining any stochastic model that has a d-dimensional linear-Gaussian state-space representation. Second, the best model is defined as the stochastic process that minimizes the Kullback–Leibler divergence from the true process. While this approach deviates from the utility-based notion of model optimality prevalent in the rational-inattention literature, it aligns with the emerging literature on model misspecification in game theory (e.g., Esponda and Pouzo (2016)). These simplifying assumptions result in a useful linear-invariance property: Expectations formed using simple models respect linear *intratemporal* relationships among observables. Moreover, these assumptions enhance the framework's tractability, rendering it applicable even in large-scale macro models without adding to the computational burden or introducing additional degrees of freedom (beyond d).

I apply the framework to study the propagation of aggregate shocks in a general-equilibrium economy. Many business-cycle models have difficulty generating empirically plausible degrees of persistence and comovement in endogenous variables in response to shocks. A common solution pursued in the literature is to introduce auxiliary frictions such as habit formation and adjustment costs (e.g., Christiano, Eichenbaum, and Evans (2005)). However, the resulting DSGE models often require implausibly large frictions to generate realistic business cycles. This paper proposes a novel and parsimonious alternative that relies on agents' bounded rationality. By replacing rational expectations with simple models, this approach increases persistence and comovement in a way that is universally applicable across applications.

The paper's main applied contribution is to quantify the extent to which the additional persistence and comovement delivered by simple models can improve the empirical fit of business-cycle models. It does so in the context of a standard model economy, which combines elements of the New Keynesian and real business cycle models. The model economy features price and wage rigidities, endogenous capital formation subject to neoclassical adjustment costs, and realistic monetary and fiscal policies. However, it does not contain any of the add-ons often introduced in DSGE models to increase persistence and comovement: external habit formation in consumption, investment-adjustment costs, price and wage indexation, endogenous capital utilization, or a monetary policy that responds to the level and growth rate of the output gap.

Without these add-ons, and under rational expectations, the model economy does not generate realistic impulse-response functions (IRFs). The IRFs to productivity shocks are essentially monotonic, not hump-shaped. Small monetary expansions lead to unrealistically large increases in output, consumption, investment, and inflation, a manifestation of the forward-guidance puzzle. Investment-demand shocks lead to a negative comovement between consumption and investment, the Barro and King (1984) puzzle. These observations suggest that standard rational-expectations New Keynesian models cannot provide a satisfactory account of business cycles.

Replacing rational expectations with simple models addresses these shortcomings. The responses to productivity shocks become hump-shaped. Expansionary monetary-policy shocks lead to much smaller increases in output, consumption, investment, and inflation—well within the range of estimated responses to identified monetary-policy shocks. Investment shocks generate a positive comovement between investment and consumption. Overall, the framework of simple models emerges as a plausible and parsimonious substitute for DSGE add-ons. Nonetheless, the extent to which simple models can narrow the gap between the business-cycle theory and data is ultimately an empirical question.

To answer this question, I use Bayesian estimation techniques. The parameters of the model economy are estimated separately under rational expectations and under the assumption that agents use one-dimensional simple models. A comparison of marginal likelihoods finds overwhelming evidence (by more than 150 log points) in favor of simple models. Perhaps more strikingly, no single standard DSGE add-on can increase the marginal likelihood by as much as simple models. The key to the empirical success of the boundedly rational model is its ability to reduce the impact of aggregate demand shocks on inflation and to generate comovement in quantities in response to those shocks. Posterior variance decomposition confirms this explanation. At business-cycle frequencies, demand shocks account for the vast majority of the variance in output, consumption, and investment, and a negligible portion of the variance in inflation or interest rates.

Related Literature. This paper belongs to the literature in macroeconomics on deviations from full-information rational expectations (FIRE)—see Woodford (2013) for a survey. The literatures on dispersed information, e.g., Lucas (1972), noisy information, e.g., Orphanides (2003) and Angeletos and La'O (2009), sticky information, e.g., Mankiw and Reis (2002), or costly attention, e.g., Sims (2003), Woodford (2003a), Maćkowiak and Wiederholt (2009), and Gabaix (2014), deviate from the FIRE benchmark by imposing imperfect knowledge of the payoff-relevant variables.[1] This paper abstracts from the difficulty of observing a large cross-section of variables and instead focuses on the difficulty of comprehending complex time-series (or intertemporal) relationships. The predictions of this framework also distinguish it from the literature mentioned above: In my model, agents fully uncover cross-sectional relationships among variables, but their expectations could deviate from rational expectations even if the economy has a single exogenous shock.

The paper also contributes to the literature that quantifies the consequences of bounded rationality using DSGE models. Slobodyan and Wouters (2012) evaluate the empirical performance of a medium-scale DSGE model with agents who do adaptive learning and show that the adaptive learning model fits the data better than the rational-expectations model. Maćkowiak and Wiederholt (2015) consider a model with rationally inattentive agents, whereas Angeletos, Collard, and Dellas (2018) consider an island economy where agents experience shocks to their higher-order expectations. Chahrour, Nimark, and Pitschner (2021) consider a multisector economy in which the deviation from FIRE arises from selective reporting by public media. And Bianchi, Ilut, and Saijo (2023) estimate a New Keynesian model in which agents have diagnostic expectations. This paper differs from these earlier contributions in its focus on intertemporal complexity and dimension reduction as bounded rationality. Furthermore, the quantitative macro model consid-

[1] See also Nimark (2008), Lorenzoni (2009), Alvarez, Lippi, and Paciello (2015), Angeletos and Lian (2018), and Angeletos and Huo (2021).

ered here is a standard New Keynesian model with a single belief parameter: the dimension d of agents' models. The remaining parameters are standard preference, technology, and policy parameters, which are estimated using standard Bayesian techniques.

A large literature studies the question of whether households, firms, and professional forecasters under- or over-extrapolate from new information. Coibion and Gorodnichenko (2015) provide evidence of under-extrapolation in consensus forecasts for professional forecasters. Bordalo, Gennaioli, Ma, and Shleifer (2020) show that individual forecasts of professional forecasters over-extrapolate from recent news. Angeletos, Huo, and Sastry (2021) find evidence of under-extrapolation at short horizons and over-extrapolation at longer horizons, whereas Afrouzi, Kwon, Landier, Ma, and Thesmar (2021) find evidence of over-extrapolation in a lab setting. More recently, Broer and Kohlhas (2024) find evidence for both under- and over-extrapolation depending on the aggregate variable being studied. In parallel with this empirical literature, many papers have proposed theoretical models of under- and over-extrapolation. Natural expectations, e.g., Fuster, Laibson, and Mendel (2010) and Fuster, Hebert, and Laibson (2012), and diagnostic expectations, e.g., Bordalo, Gennaioli, and Shleifer (2018), are examples of models where agents over-extrapolate from the recent past. Cognitive discounting of Gabaix (2020) and level-k thinking, e.g., García-Schmidt and Woodford (2019) and Farhi and Werning (2019), are examples of models that feature under-extrapolation. The framework proposed in this paper is neither a model of under-extrapolation nor of over-extrapolation; agents who use simple models always under-extrapolate some observables and over-extrapolate others.

This paper also contributes to the literature that studies the properties of pseudo-true models. The term pseudo-true model originates in the pioneering work of Sawa (1978), who proposes using the Kullback–Leibler divergence as a model-selection criterion when models are misspecified. Agents in the restricted-perceptions equilibrium of Bray (1982) and Bray and Savin (1986), Rabin and Vayanos (2010)'s model of the gambler's fallacy, the natural-expectations framework of Fuster, Laibson, and Mendel (2010) and Fuster, Hebert, and Laibson (2012), the Berk–Nash equilibrium of Esponda and Pouzo (2016, 2021), and the constrained rational-expectations equilibrium of Molavi (2019) all use pseudo-true models to forecast payoff-relevant variables. Agents in Krusell and Smith (1998) also have a misspecified model of the economy since they believe that current and future prices do not depend on anything but the first few moments of the wealth distribution. However, despite this long history, surprisingly few general results on the properties of pseudo-true models have appeared in the literature. Such results are almost exclusively derived (with the notable exception of Rabin and Vayanos (2010)) in settings where the set of models is sufficiently restricted that the pseudo-true model can be estimated using OLS regression and the bias in agents' forecasts reduces to the omitted-variable bias. I contribute to this literature by characterizing the set of pseudo-true state-space models of a given dimension.

The state-space models used in this paper are relatives of dynamic-factor models, e.g., Stock and Watson (2011, 2016). However, the two offer two distinct ways of decomposing time-series data. Dynamic factor models decompose data into common factors and idiosyncratic disturbances, whereas state-space models decompose it into persistent and transitory components. The two approaches thus suggest two different simplifications of large time-series data: using a small number of common factors in the former case and a small number of persistent states in the latter.[2]

Finally, in a follow-up paper, Molavi, Tahbaz-Salehi, and Vedolin (2024) use a closely related framework to study the implications of model misspecification for asset prices and returns. They show that constraining the complexity of investors' models leads to return and forecast-error predictability and provides a parsimonious account of several puzzles in the asset-pricing literature.

Outline. The rest of the paper is organized as follows: Section 2 presents the framework of simple models and formally defines and discusses the notion of fit used in the paper. Section 3 contains the paper's characterization results for simple models. Section 4 discusses the implications of using simple models for agents' forecasts and choices. Section 5 presents a business-cycle application. Section 6 concludes. Omitted details and additional results for the business-cycle application are relegated to three appendices. Some additional theoretical results as well as the proofs can be found in the online appendices.

2 Framework

In this section, I present the general framework and the main behavioral assumption of the paper.

2.1 Environment

Time is discrete and is indexed by $t \in \mathbb{Z}$. An agent observes a sequence of variables over time and uses her past observations to forecast their future values. I let $y_t \in \mathbb{R}^n$ denote the time-t value of the vector of observables, or simply the *observable*. Vector y_t follows a mean-zero stochastic process $\mathbb{P}$ with the corresponding expectation operator $\mathbb{E}[\cdot]$. I start by taking $\mathbb{P}$ as a primitive, but the process will be an endogenous outcome of agents' actions in the business-cycle application studied in Section 5.

I make several technical assumptions on the true process. First, $\mathbb{P}$ is purely non-deterministic, stationary, and ergodic, and has a finite second moment. Second, there exists a subspace $\mathcal{W}$ of $\mathbb{R}^n$ (possibly equal to $\mathbb{R}^n$ itself) such that y_t is supported on $\mathcal{W}$ with density $\mathbb{f}$.[3] Finally, the true

[2] The sets of time series that can be represented by dynamic-factor and state-space models are not nested. Instead, any finite dynamic-factor model has a state-space representation, and any finite state-space model has a dynamic-factor representation. See Forni and Lippi (2001) for a representation result for the (generalized) dynamic factor models.

[3] This assumption is weaker than the assumption that $\mathbb{P}$ has full support over $\mathbb{R}^n$ because it allows for the possibility that the true process is degenerate. This additional level of generality will be useful in applications where the elements of y_t may be linearly dependent.

process has finite entropy rate, i.e., $\lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left[-\log \mathbb{f}(y_1, \ldots, y_t)\right] < \infty$. These assumptions are all quite weak. For instance, they are satisfied if y_t follows a stationary vector ARMA process with Gaussian innovations.

The agent has perfect information about the past realizations of the observable; her time-t information set is given by $\{y_t, y_{t-1}, \ldots\}$. However, she may use a misspecified model to map her information to her forecasts. This model misspecification leads to deviations in the agent's forecasts from those that arise in the rational-expectations benchmark.

2.2 Simple Models

As the paper's main behavioral assumption, I assume that the agent is constrained to use state-space models with a small number of state variables to forecast the vector of observables. She can only entertain models of the form

$$\begin{aligned} z_t &= A z_{t-1} + w_t, \\ y_t &= B' z_t + v_t, \end{aligned} \tag{1}$$

where z_t is the d-dimensional vector of *subjective latent states*, $A \in \mathbb{R}^{d\times d}$, $w_t \in \mathbb{R}^d$ is i.i.d. $\mathcal{N}(0, Q)$, $B \in \mathbb{R}^{d\times n}$, $v_t \in \mathbb{R}^n$ is i.i.d. $\mathcal{N}(0, R)$, and w_t and v_t are independent. While the integer d is a primitive of the model that parameterizes the dimension of the agent's models, matrices A, B, Q, and R are parameters that are determined endogenously by maximizing the fit to the true process. Formally, I define a *d-state model* as a stationary stochastic process over $\{y_t\}_{t=-\infty}^{\infty}$ that has a representation of the form (1) such that (i) the dimension of vector z_t is d, (ii) A is a convergent matrix, (iii) Q is positive definite, and (iv) R is positive semidefinite.[4] I let P^θ denote the d-state model parameterized by the collection of matrices $\theta \equiv (A, B, Q, R)$, let $E^\theta[\cdot]$ denote the corresponding expectation operator, and let Θ_d denote the set of all d-state models.[5] Whenever there is no risk of confusion, I use the term d-state model to refer both to the stochastic process P^θ for y_t and the parameters $\theta = (A, B, Q, R)$ of its state-space representation.

The integer d captures the agent's sophistication in modeling the stochastic process for the vector of observables, with larger values of d indicating agents who can entertain more complex models. When d is sufficiently large, the agent can approximate the unconditional and conditional second moments of any purely non-deterministic covariance-stationary process arbitrarily well using a model in her set of models. On the other hand, when d is small relative to the number of states required to model the true process, no model in the agent's set of models will provide a good approximation to $\mathbb{P}$. The agent then necessarily ends up with a misspecified model of the

[4] A matrix is *convergent* if all of its eigenvalues are smaller than one in magnitude. A being convergent and Q being positive definite are sufficient for a model (A, B, Q, R) to define a stationary ergodic process.

[5] One can define the set of d-state models without any reference to the latent state z_t. Stochastic process P for $\{y_t\}_{t=-\infty}^{\infty}$ with expectation operator E is a d-state model if $E[y_t y'_{t-l}] = CA^{l-1}\overline{C}'$ for all $l = 1, 2, \ldots$, some convergent $d \times d$ matrix A, and some $C, \overline{C} \in \mathbb{R}^{n\times d}$. See, for instance, Faurre (1976) or Katayama (2005, Chapter 7). I opt for the definition that uses the subjective latent state since z_t will have an intuitive interpretation as agents' view of the state of the economy in the macro application I consider in this paper.

true process and biased forecasts—regardless of which model in the set Θ_d she uses to make her forecasts. Characterizing this bias is the focus of the next section of the paper.

My preferred rationale for the constraint on the number of states is to capture the agent's bounded rationality, but the constraint can also arise from the agent's rational fear of overfitting. Models with a large number of parameters and many degrees of freedom are prone to overfitting. Such concerns may lead rational agents to prefer more parsimonious statistical models, especially if they only have a short time series to draw upon when estimating the parameters of their model. In the remainder of the paper, I abstract away from any issues arising from small samples and instead consider the long-run limit where the sampling error vanishes.

2.3 The Notion of Fit

I assume that the agent forecasts using a model in the family of d-state models that provides the best fit to the true process. I use the Kullback–Leibler divergence rate of process P^θ from the true process $\mathbb{P}$ as the measure of the fit of model θ.[6] The *Kullback–Leibler divergence rate* (KLDR) of P^θ from $\mathbb{P}$ is denoted by $\mathrm{KLDR}(\theta)$ and defined as follows. Recall that the true process is supported on a subspace $\mathcal{W}$ of $\mathbb{R}^n$. If P^θ is also supported on $\mathcal{W}$, then

$$\mathrm{KLDR}(\theta) \equiv \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left[\log\left(\frac{\mathbb{f}(y_1,\ldots,y_t)}{f^\theta(y_1,\ldots,y_t)}\right)\right],$$

where f^θ denotes the density of P^θ; if P^θ is not supported on $\mathcal{W}$, then $\mathrm{KLDR}(\theta) \equiv +\infty$.

The Kullback–Leibler divergence rate is the natural generalization of Kullback–Leibler (KL) divergence to stationary stochastic processes. In the i.i.d. case, the KL divergence of a candidate model from the true model captures the difficulty of rejecting the candidate model in favor of the true model using a likelihood-ratio test. That is why the KL divergence is commonly used as a measure of a model's fit.[7] Similarly, $\mathrm{KLDR}(\theta)$ captures the rate at which the power of a test for separating a stochastic process P^θ from the true process $\mathbb{P}$ approaches one as $t \to \infty$.[8] The KLDR is also tightly linked to asymptotics of Bayesian learning, as discussed in the following subsection.

Model $\theta \in \Theta_d$ is a *pseudo-true* d-state model if $\mathrm{KLDR}(\theta) \le \mathrm{KLDR}(\tilde{\theta})$ for all $\tilde{\theta} \in \Theta_d$. If the agent's set of models contains a model θ such that $f^\theta(y_1,\ldots,y_t) = \mathbb{f}(y_1,\ldots,y_t)$ almost everywhere and for all t, then any pseudo-true d-state model is observationally equivalent to the true process.[9] The set of models Θ_d is then correctly specified. When no such d-state model exists, $\mathrm{KLDR}(\theta) > 0$ for any model $\theta \in \Theta_d$, and the set of models is misspecified. The following proposition states that the pseudo-true models are observationally equivalent to the true process when the set of models is correctly specified:

[6] The mean-squared forecast error is another commonly used notion of fit. In Online Appendix D, I define the weighted mean-squared forecast error and show that it is equivalent to the Kullback–Leibler divergence rate under an appropriate choice of the weighting matrix.

[7] See, for instance, Hansen and Sargent (2008).

[8] See, for instance, Shalizi (2009).

[9] Processes P and $\tilde{P}$ are *observationally equivalent* if all their finite-dimensional marginal distributions are identical.

Proposition 1. *Suppose the set Θ_d of d-state models is correctly specified. Then any pseudo-true d-state model P^θ is observationally equivalent to the true process $\mathbb{P}$.*

The paper's focus is the misspecified case, where d is small relative to the number of states required to capture the true process. This statement is about d being smaller than the "true d," not necessarily smaller than n, the dimension of y_t. However, it is often natural to also think of d as much smaller than n. Approximating the true process by a pseudo-true d-state model then corresponds to using a parsimonious time-series model to capture the essential features of a large data set. Unless otherwise specified, I assume throughout the paper that $d \leq n$. However, the paper's characterization results easily generalize to the $d > n$ case.

2.4 Learning Foundation

Pseudo-true models arise naturally as the long-run outcome of learning by Bayesian agents with misspecified priors. Consider an agent who starts with prior μ_0 with full support over the points in the set $\mathbb{R}^d \times \Theta_d$, each corresponding to an initial value of the subjective states z_0 and a d-state model θ, which describes how states and the observable evolve over time. Suppose the agent observes y_t over time and updates her belief using Bayes' rule. Let μ_t denote the agent's time-t Bayesian posterior over $\mathbb{R}^d \times \Theta_d$. Berk (1966)'s theorem establishes that, in the limit $t \to \infty$, the agent's posterior will assign a probability of one to the set of pseudo-true models.[10]

This result offers an "as if" interpretation of the pseudo-true d-state models. One can assume that the agent has a subjective prior—which may be different from the true distribution—and updates her belief in light of new information using Bayes' law. By Berk's theorem, as long as the agent's prior is supported on the set of d-state models, she will forecast the observable in the long run *as if* she were using a pseudo-true d-state model. Focusing on pseudo-true models allows me to abstract away from learning dynamics and focus on the asymptotic bias caused by misspecification.[11]

The set of pseudo-true d-state models is independent of the agent's preferences. Instead, it only depends on the number of states the agent can entertain and the true stochastic process. The independence of the agent's pseudo-true models from her preferences is evident given the "as if" interpretation discussed above. Two agents who start with identical priors, observe the same sequence of observations, and update their beliefs using Bayes' rule will end up with identical posteriors at any point in time—irrespective of their preferences. Berk's theorem goes a step

[10] While Berk (1966) only covers the case of i.i.d. observations and parametric models, the result has been extended much more generally. Bunke and Milhaud (1998) and Kleijn and Van Der Vaart (2006) substantially extend Berk (1966) by providing conditions for the weak convergence of posterior distributions and considering infinite-dimensional models. Shalizi (2009)'s extension of Berk's theorem covers the case of non-i.i.d. observations and hidden Markov models.

[11] One can alternatively consider agents who estimate the parameters of their d-state models using a quasi-maximum-likelihood estimator. Such agents also will asymptotically forecast *as if* they relied on the pseudo-true d-state models. See, for instance, Theorem 2 of Douc and Moulines (2012).

further by establishing that, in the long run, the posterior only depends on the support of the prior (not its fine details) and the distribution of observations (not their realizations).

The independence of the agent's pseudo-true models from her preferences has a significant consequence: The set of pseudo-true d-state models is generically disjoint from the set of d-state models that maximize the agent's payoff. However, this disparity is a feature, not a bug, of a positive theory of bounded rationality. While finding the payoff-maximizing model requires knowledge of the true process, one arrives at the set of pseudo-true models simply by following Bayes' rule—no knowledge of the true process is necessary. Following Bayes' rule would have led the agent to the truth had her model been correctly specified, but it can lead her astray in the presence of model misspecification.

Agents' use of pseudo-true models should therefore be viewed as a positive statement—not a normative one. A pseudo-true d-state model is not what an agent should use for forecasting in order to maximize her payoff. It is what she will use to forecast in the long run if she starts with a prior over the set of d-state models and updates her belief using Bayes' rule.

3 Characterization of Pseudo-True Models

In this section, I characterize the set of pseudo-true d-state models, beginning with the case where $d = 1$. As a preliminary step, I discuss a useful property of the pseudo-true models, which is of independent interest.

3.1 The Invariance Property

I begin with a result that shows the invariance of the pseudo-true d-state models to linear transformations of the observable. Consider an agent who, instead of observing vector $y_t \in \mathbb{R}^n$, observes vector $\tilde{y}_t = Ty_t \in \mathbb{R}^m$, where T denotes an $m \times n$ matrix. As long as T is a rank-n matrix, y_t and $\tilde{y}_t$ convey the exact same information. Thus, one might expect that the agent's beliefs when she observes y_t are consistent with her beliefs when she instead observes $\tilde{y}_t$.

The following definition formalizes the notion that two probability distributions are consistent with each other given a linear transformation of the observable. Let $T \in \mathbb{R}^{m\times n}$ be a matrix and P be a probability distribution over infinite sequences in $\mathbb{R}^n$. The probability distribution over infinite sequences in $\mathbb{R}^m$ induced by T and P is denoted by $T(P)$ and defined as $T(P)(\mathcal{Y}) \equiv P\left(\{y_t\}_{t=-\infty}^{\infty} : \{Ty_t\}_{t=-\infty}^{\infty} \in \mathcal{Y}\right)$ for any measurable set $\mathcal{Y} \subseteq \mathbb{R}^{m\mathbb{Z}}$.[12] If the observable y_t follows the stochastic process $\mathbb{P}$, then its linear transformation $\tilde{y}_t = Ty_t$ follows the transformed process $T(\mathbb{P})$. The following result establishes that transforming the observable by a rank-n matrix leads the set of pseudo-true models to be transformed accordingly:

[12] The probability distribution induced by a mapping is formally known as the *pushforward measure*.

Theorem 1 (linear invariance)**.** *Suppose $T \in \mathbb{R}^{m\times n}$ is a rank-n matrix. Then P^{θ} is a pseudo-true d-state model given true model $\mathbb{P}$ if and only if $T\left(P^{\theta}\right)$ is a pseudo-true d-state model given true model $T\left(\mathbb{P}\right)$.*

The result shows that an agent using simple models can discern all linear intratemporal relationships among the observables while facing significant constraints in understanding complex intertemporal relationships. While arguably stark, this dichotomy highlights the paper's premise that forecasting is challenging because it requires forecasters to recognize stochastic patterns that unfold over time. The result makes it possible to abstract from the cognitive costs of acquiring information about a large cross-section of variables and the mistakes individuals make when dealing with cross-sectional complexity, allowing me to instead concentrate on time-series complexity.

The linear invariance property makes the predictions of the framework invariant to the exact specification of the variables included in the vector of observables. The agent's pseudo-true models and forecasts only depend on the observables' information content, not on how that information is presented. For instance, whether the agent observes the nominal interest rate and the inflation rate or the real interest rate and the inflation rate is immaterial to how she forms her expectations. Likewise, the agent's expectations remain unchanged if the vector of observables is augmented with linear combinations of variables already in her information set.

The theorem thus suggests that it is without loss to assume that the vector of observables is free of redundant variables. Define the lag-l autocovariance matrix of the true process as follows:

$$\Gamma_l \equiv \mathbb{E}[y_t y_{t-l}']. \tag{2}$$

When y_t includes redundant variables, the variance-covariance matrix Γ_0 is singular, and the true process is degenerate.[13] In such cases, a lower-dimensional vector $\tilde{y}_t$ and a full-rank matrix T exist such that $\mathbb{E}[\tilde{y}_t\tilde{y}_t']$ is non-singular and $y_t = T\tilde{y}_t$. Therefore, by Theorem 1, the pseudo-true models given y_t can be found by first finding the pseudo-true models given $\tilde{y}_t$ and then applying transformation T. This observation implies that there is no loss of generality in assuming that the variance-covariance matrix Γ_0 is non-singular and that the agent only considers subjective models with non-singular variance-covariance matrices.[14] I maintain these assumptions throughout the rest of the paper.

3.2 Pseudo-True One-State Models

I start the analysis of pseudo-true models by considering the case where the agent can only entertain one-state models. In this case, a complete characterization of the agent's pseudo-true

[13] A probability distribution on a space is said to be *degenerate* if it is supported on a manifold of lower dimension.

[14] Whenever the true variance-covariance matrix Γ_0 is non-singular, any subjective model with a singular variance-covariance matrix is dominated in terms of the fit to the true process by every subjective model with a non-singular variance-covariance matrix. Therefore, no subjective model with a singular variance-covariance matrix can be a pseudo-true model.

models is possible. The insights from the single-state case generalize to the d-state case, as discussed later in this section.

The agent's pseudo-true one-state forecasts turn out to depend on the true process only through the unconditional variance and the autocorrelation structure of the vector of observables. The autocorrelations are measured by a novel set of objects, which I refer to as autocorrelation matrices. I define the lag-l *autocorrelation matrix* of the observable under the true process as follows:[15]

$$C_l \equiv \frac{1}{2}\Gamma_0^{\frac{-1}{2}}\left(\Gamma_l + \Gamma_l'\right)\Gamma_0^{\frac{-1}{2}}. \tag{3}$$

The concept of autocorrelation matrices naturally extends the idea of autocorrelation functions. If the observable y_t is a scalar, C_l simplifies to the standard autocorrelation function at lag l. However, when the observable is an n-dimensional vector, C_l is an $n \times n$ real symmetric matrix with eigenvalues inside the unit circle.[16] Autocorrelation matrices capture the extent of serial correlation in the vector of observables. When the spectral radius of C_l is close to zero for all l, the process is close to being i.i.d., whereas when the spectral radius of C_l is close to one, then the process is close to being unit root.[17]

With the definition of autocorrelation matrices at hand, I can state the general characterization result for the $d = 1$ case:

Theorem 2. *Under any pseudo-true one-state model θ, the agent's s-period-ahead forecast is given by*

$$E_t^{\theta}[y_{t+s}] = a^s(1-\eta)qp'\sum_{\tau=0}^{\infty} a^{\tau}\eta^{\tau}y_{t-\tau}, \tag{4}$$

where a and η are scalars in the $[-1,1]$ and $[0,1]$ intervals, respectively, that maximize $\lambda_{max}(\Omega(\tilde{a},$ $\tilde{\eta}))$, the largest eigenvalue of the $n \times n$ real symmetric matrix

$$\Omega(\tilde{a},\tilde{\eta}) \equiv -\frac{\tilde{a}^2(1-\tilde{\eta})^2}{1-\tilde{a}^2\tilde{\eta}^2}I + \frac{2(1-\tilde{\eta})(1-\tilde{a}^2\tilde{\eta})}{1-\tilde{a}^2\tilde{\eta}^2}\sum_{\tau=1}^{\infty}\tilde{a}^{\tau}\tilde{\eta}^{\tau-1}C_{\tau},$$

and $p = \Gamma_0^{\frac{-1}{2}}u$ and $q = \Gamma_0^{\frac{1}{2}}u$, where u is an eigenvector of $\Omega(a,\eta)$ with eigenvalue $\lambda_{max}(\Omega(a,\eta))$, normalized so that $u'u = 1$.

The endogenous variables a, η, p, and q have intuitive meanings. The scalar a represents the persistence of the subjective latent state. If $a = 0$, the subjective state is i.i.d., whereas if $a = 1$, it follows a unit-root process.[18] The scalar η captures the perceived noise in the agent's observations of the subjective state. When η is small, the agent believes recent observations to be

[15] Here and throughout the paper, I follow the usual convention that, for a symmetric positive definite matrix X, the square-root matrix $X^{\frac{1}{2}}$ is the unique symmetric positive definite matrix that satisfies $X^{\frac{1}{2}}X^{\frac{1}{2}} = X$.

[16] See Lemma G.2 in the Online Appendix for a proof.

[17] The spectral radius $\rho(X)$ of matrix X denotes the maximum among the magnitudes of eigenvalues of X.

[18] The theorem does not rule out the possibility that $|a| = 1$, in which case the corresponding state-space model might not be stationary ergodic. However, Lemma G.3 in the Online Appendix establishes that any pseudo-true one-state model inherits the stationarity and ergodicity of the true process.

highly informative of the value of the subjective state. As a result, her expectations respond more to recent observations and discount old observations more. The vector p determines the agent's relative attention to different components of the vector of observables. When p_i is larger than p_j, the agent puts more weight on $y_{i,t-\tau}$ relative to $y_{j,t-\tau}$ for all τ when forming her estimate of the subjective state. Finally, the vector q captures the relative sensitivity of the agent's forecasts of different observables to changes in her estimate of the subjective state. When q_i is larger than q_j, then a change in the estimated value of the state at time t leads the agent to change her forecast of $y_{i,t+s}$ by more than her forecast of $y_{j,t+s}$ for all s.

It follows standard Kalman filter results that the agent's forecasts take the form of equation (4) for *some* a, η, p, and q. The substance of the result is rather characterizing the (a,η,p,q) tuple that lead to a model with minimal KLDR from the true process. The theorem suggests a tractable way of computing the pseudo-true one-state forecasts in any stationary and ergodic environment given only the knowledge of the true autocorrelation matrices.

The theorem significantly reduces the computational complexity of finding the set of pseudo-true models. It concentrates out all parameters in the agent's models except for two scalars. As a result, the optimization problem simplifies from a problem over a $2n$-dimensional non-compact manifold to a much simpler problem over a two-dimensional compact rectangle.[19] Furthermore, since the size of the problem is independent of n, it can be solved efficiently in any application, regardless of the dimension of the vector of observables.

The next result characterizes the perceived variance-covariance matrix of the observable under the pseudo-true one-state models:

Theorem 3. *Given any pseudo-true one-state model θ, the subjective variance-covariance of the vector of observables, $E^\theta[y_t y_t']$, coincides with the true variance-covariance matrix, $\Gamma_0 \equiv \mathbb{E}[y_t y_t']$.*

The theorem hinges on two main assumptions: First, there are no constraints on the agent's set of models other than the bound on the number of subjective state variables; put differently, matrices A, B, Q, and R of representation (1) are unrestricted other than the constraint on their dimension. This flexibility allows the agent to represent any cross-sectional correlation pattern by an appropriate selection of matrices A, B, Q, and R. Second, the agent uses a model that minimizes the KLDR from the true process. This leads her to a set of such matrices that perfectly capture the true cross-sectional correlations.

Theorems 2 and 3 fully characterize the pseudo-true one-state models in terms of the true variance-covariance matrix Γ_0 and the tuple (a,η,p,q), which in turn only depends on the true autocorrelation matrices $\{C_l\}_{l=1}^{\infty}$. Any unconditional or conditional moment of the pseudo-true one-state model can, in turn, be found in terms of Γ_0 and (a,η,p,q).

[19] The set of all d-state models is a non-compact manifold of dimension $2nd$ (Gevers and Wertz, 1984). Additionally, the KLDR is a non-convex function of $\theta=(A,B,Q,R)$.

3.3 Pseudo-True One-State Models Under Exponential Ergodicity

The pseudo-true one-state models can be found in closed form given a class of true stochastic processes that naturally arise in applications. The appropriate class turns out to be the following:

Definition 1. A stationary ergodic process $\mathbb{P}$ is *exponentially ergodic* if $\rho(C_l) \leq \rho(C_1)^l$ for all $l \geq 1$, where $\rho(C_l)$ denotes the spectral radius of C_l.

Exponential ergodicity is stronger than ergodicity. Ergodicity requires that the serial correlation at lag l decays to zero as $l \to \infty$. Exponential ergodicity requires the rate of decay to be faster than $\rho(C_1)$. Although exponentially-ergodic processes only constitute a subset of the class of stationary ergodic processes, many standard processes are exponentially ergodic. For instance, the vector of observables follows an exponentially-ergodic process if it is a spanning linear combination of n independent AR(1) shocks.

The following result characterizes the agent's pseudo-true one-state forecasts when the true process is exponentially ergodic. It links the agent's forecasts to the eigenvalues and eigenvectors of the true autocorrelation matrix at lag one:

Theorem 4. *Suppose the true process is exponentially ergodic. Under any pseudo-true one-state model θ, the agent's s-period-ahead forecast is given by*

$$E_t^{\theta}[y_{t+s}] = a^s q p' y_t, \tag{5}$$

where a is an eigenvalue of C_1 largest in magnitude, u denotes the corresponding eigenvector normalized so that $u'u = 1$, and $p = \Gamma_0^{\frac{-1}{2}} u$ and $q = \Gamma_0^{\frac{1}{2}} u$.

A remarkable feature of the characterization in Theorem 4 is that the agent's forecasts only depend on the last realization of the observable (and not its lags). In other words, the pseudo-true one-state model is *Markovian* if the true process is exponentially ergodic. This property might come as a surprise in light of the fact that in the correctly-specified case forecasts obtained using the stationary Kalman filter generically use the entire history of the observable. The seeming discrepancy between the two results is due to misspecification of the agent's set of models in Theorem 4, as illustrated by the following example:

Example 1. Suppose the observable is scalar and follows an AR(∞) process: $y_{t+1} = \sum_{\tau=1}^{\infty} \phi_\tau y_{t+1-\tau}$.[20] It is then immediate that the one-step-ahead forecast of the observable under the true, correctly-specified model is given by

$$\mathbb{E}_t[y_{t+1}] = \sum_{\tau=1}^{\infty} \phi_\tau y_{t+1-\tau}.$$

[20] Such a representation exists for generic processes in the class of mean-zero, purely non-deterministic, and stationary processes.

Contrast this with what an agent can do when she is constrained to use (misspecified) one-state models. Under any such model θ, the agent's one-step-ahead forecast takes a similar form:

$$E_t^{\theta}[y_{t+1}] = \sum_{\tau=1}^{\infty} \alpha_\tau y_{t+1-\tau}.$$

However, the restriction to one-state models constrains coefficients $\{\alpha_\tau\}_{\tau=1}^{\infty}$ to be given by $\alpha_\tau = (1-\eta)a^\tau \eta^{\tau-1}$ for some $a \in [-1,1]$, some $\eta \in [0,1]$, and all τ. Therefore, the pseudo-true one-state model is the model that picks $\{\alpha_\tau\}_{\tau=1}^{\infty}$ to minimize the KLDR subject to the constraint that $\alpha_\tau = (1-\eta)a^\tau \eta^{\tau-1}$ for all τ. The agent wants to set α_τ to a value that is related to the correlation of y_{t+1} and $y_{t-\tau}$, but the constraint prevents her from fine-tuning the $\{\alpha_\tau\}_{\tau=1}^{\infty}$ coefficients. When the true process is exponentially ergodic, y_{t+1} is much more correlated with y_t than it is with lags of y_t. Then, the best such a constrained agent can do is to fine-tune the coefficient of y_t and entirely disregard its lags. In other words, the constrained minimizer of the KLDR is Markovian even though the unconstrained minimizer is not.

The next example illustrates the use of Theorem 4 in the context of a commonly-used process:

Example 2. Suppose the true process $\mathbb{P}$ has the following representation:

$$\begin{aligned} f_t &= F f_{t-1} + \epsilon_t \\ y_t &= H' f_t, \end{aligned} \tag{6}$$

where $\epsilon_t \sim \mathcal{N}(0, \Sigma)$, $F = \text{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n)$, $\Sigma = \text{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2\right)$, $H \in \mathbb{R}^{n\times n}$ is an invertible square matrix, and $1 > |\alpha_1| > |\alpha_2| > \cdots > |\alpha_n| > 0$. It is easy to verify that $\rho(C_l) = |\alpha_1|^l = \rho(C_1)^l$; that is, the true process is exponentially ergodic. Therefore, Theorem 4 can be used to characterize the pseudo-true one-state forecasts. The persistence, noise, relative attention, and relative sensitivity are, respectively, given by $a = \alpha_1$, $\eta = 0$, $p = (H'VH)^{\frac{1}{2}} H^{-1} V^{-1} e_1$, and $q = (H'VH)^{\frac{-1}{2}} H' V e_1$, where $V \equiv (I - F^2)^{-1}\Sigma$ is the variance-covariance matrix of f_t and e_1 denotes the first coordinate vector.[21]

The agent's forecasts take a particularly simple form when H is the identity matrix, i.e., $y_{it} = f_{it}$ for $i = 1, \ldots, n$. Then, p and q are both multiples of the first coordinate vector e_1, and the agent's forecasts simplify to

$$\begin{aligned} E_t^{\theta}[y_{1,t+s}] &= \alpha_1^s y_{1t} = \mathbb{E}_t[y_{1,t+s}], \\ E_t^{\theta}[y_{i,t+s}] &= 0, \qquad \forall i \neq 1. \end{aligned}$$

The agent's forecast of the most persistent element of the vector of observables coincides with its rational-expectations counterpart, but she forecasts every other element of the observable as if it were i.i.d.

A noteworthy feature of the pseudo-true model in Example 2 is that the persistence parameter a does not depend on the volatilities of the underlying AR(1) processes. The agent uses the

[21] See the proof of Lemma G.5 for a derivation.

subjective latent state to track the most persistent component of y_t, even if the most persistent component has a small variance. However, this result should not come as a surprise given the linear-invariance result: One can always equalize the volatilities of different components of y_t by an appropriate linear transformation of the observable without altering the persistence of the subjective latent state in the agent's pseudo-true model. Therefore, the persistence parameter cannot depend on the volatilities.

The example also illustrates that the agent exhibits a form of *persistence bias*. She forecasts the most persistent component of the vector of observables as accurately as under rational expectations but misses the dynamics of the other components. The intuition for the result is easiest to see when the most persistent component is close to being unit root. In that case, poorly tracking the most persistent component would lead to persistent mistakes in the agent's forecasts. The persistence of those mistakes would make them costly from the point of view of KLDR minimization. Therefore, any pseudo-true model tracks the component close to unit root as best possible, even if doing so results in errors in forecasting the other components. In Section 4.1, I generalize the insight of this example by formally establishing persistence bias in a more general context.[22]

Example 2 can be generalized by relaxing the assumption that matrices F and Σ are diagonal and allowing for non-Gaussian innovations. The key requirement for the process to be exponentially ergodic is that matrix H in representation (6) is full rank. This assumption can be seen as a full-information (or spanning) assumption. If the agent observes an observable of the form (6) with a full-rank matrix H, then she has enough information to forecast the observable as well as in the full-information rational-expectations benchmark—even if she fails to do so due to the constraint on her set of models. See Online Appendix E for a detailed discussion of this generalization.

3.4 Pseudo-True d-State Models

I end this section by discussing how the insights from the $d = 1$ case generalize when $d > 1$. To characterize the pseudo-true d-state models, one needs to find models $\theta = (A, B, Q, R)$ that minimize the KLDR from the true process. Doing so requires minimizing a non-convex function over a non-compact set, consisting of all the matrices A, B, Q, and R of appropriate dimensions. This problem does not lend itself to an analytical solution without further restrictions.

To address this issue, I restrict the models the agent considers to be Markovian. A d-state model θ is *Markovian* if P^θ satisfies the Markov property, i.e., $P^\theta(y_{t+1}|y_t, y_{t-1}, \dots) = P^\theta(y_{t+1}|y_t)$. An agent who believes the observable follows a Markovian d-state model believes that (a) the current realization of the observable contains all the information required for forecasting, and (b) all the

[22] Bidder and Dew-Becker (2016) and Dew-Becker and Nathanson (2019) propose an alternative reason agents might focus on tracking the most persistent components of a payoff-relevant variable. Bidder and Dew-Becker (2016) show that long-run risk is the worst case scenario for ambiguity-averse agents. Dew-Becker and Nathanson (2019) show that, as a result, ambiguity-averse agents will learn most about dynamics at the lowest frequencies.

relevant information contained in y_t can be summarized by a d-dimensional state variable. The following proposition provides a necessary and sufficient condition for a model to be Markovian:

Proposition 2. *Let $Var_t^\theta(y_{t+1})$ denote the variance-covariance matrix of y_{t+1} given model θ and conditional on the history $\{y_\tau\}_{\tau\leq t}$ of the observable, and let $Var^\theta(y_{t+1}|z_t)$ denote the corresponding variance-covariance matrix conditional on the time-t realization of the subjective latent state. $Var_t^\theta(y_{t+1}) \geq Var^\theta(y_{t+1}|z_t) = B'QB + R$ for any d-state model θ, with equality if and only if θ is Markovian.*[23]

The proposition highlights an intuitive property of Markovian models. Note that the agent can observe the history $\{y_\tau\}_{\tau\leq t}$ of the observable but not the subjective latent state z_t. The first part of the proposition shows that the agent cannot forecast any better than if she knew the realization of the latent subjective state. In other words, the forecast error given z_t provides a lower bound on the forecast error given $\{y_\tau\}_{\tau\leq t}$. The second part of the proposition shows that the agent can achieve this lower bound when her model of the world is Markovian. She can then forecast as well as an agent who knows the latent subjective state, because all the relevant information in the latent state can be extracted from the realized history of the observable. Markovian models can thus be seen as models that feature *full information.*

Markovian models constitute only a subset of the class of all state-space models of a given dimension. However, the pseudo-true one-state models happen to be Markovian when the true process is exponentially ergodic, as shown by the following corollary of Theorem 4:

Corollary 1. *If the true process is exponentially ergodic, then any pseudo-true one-state model is Markovian.*

The result shows that constraining the agent to Markovian models is without loss when $d = 1$ and the true process is exponentially ergodic. Even with the flexibility to choose non-Markovian models, an agent who is attempting to minimize the KLDR from an exponentially ergodic process settles on a Markovian model. Whether this result continues to hold for d-state models with $d > 1$ remains an open question. However, I can still make progress by taking the restriction to Markovian models as an assumption and characterizing the resulting pseudo-true models. A Markovian d-state model θ is a *pseudo-true Markovian d-state model* if $\text{KLDR}(\theta) \leq \text{KLDR}(\tilde{\theta})$ for any Markovian d-state model $\tilde{\theta}$.

The pseudo-true Markovian models have a number of appealing properties. They satisfy a version of the linear-invariance result of Theorem 1. They share Bayesian and quasi-maximum-likelihood learning foundations with other pseudo-true models. Perhaps most importantly, they can be fully characterized in closed-form in some useful cases:

Theorem 5. *Suppose either $d = 1$ or the lag-one autocovariance matrix is symmetric. Then the following statements hold:*

[23] I use the usual convention that $X \geq Y$ for symmetric positive semidefinite matrices X and Y if $X - Y$ is positive semidefinite.

(a) *Under any pseudo-true Markovian d-state model θ, the agent's s-period-ahead forecast is given by*

$$E_t^\theta[y_{t+s}] = \sum_{i=1}^{d} a_i{}^s q_i p_i' y_t, \tag{7}$$

where $a_1, \ldots, a_d$ are d eigenvalues of C_1 largest in magnitude (with the possibility that some of the a_i are equal), u_i denotes an eigenvector corresponding to a_i normalized such that $u_i' u_k = \mathbb{1}_{\{i=k\}}$ for all i and k, $p_i \equiv \Gamma_0^{\frac{-1}{2}} u_i$, and $q_i \equiv \Gamma_0^{\frac{1}{2}} u_i$.

(b) *Under any pseudo-true Markovian d-state model θ, the subjective variance-covariance of the vector of observables, $E^\theta[y_t y_t']$, coincides with the true variance-covariance matrix, $\Gamma_0 \equiv \mathbb{E}[y_t y_t']$.*

The result shows that the insights from the analysis of one-state simple models broadly carry over to d-state ones. In particular, agents who are restricted to Markovian d-state models exhibit a form of persistence bias. They focus on perfectly forecasting the d most persistent components of the vector of observables at the expense of the other components. Moreover, agents who are constrained to use Markovian d-state models uncover the true variance-covariance matrix of the observable.

The theorem also suggests that state-space models can be estimated consistently by principal component analysis (PCA). This conclusion is reminiscent of a central result in the theory of dynamic factor models on the consistency of the principal components estimator for the common components.[24] However, Theorem 5 is different along several dimensions. First, it concerns state-space models, not dynamic factor models. Second, the estimator suggested by the theorem uses the principal components of the lag-one autocorrelation matrix, while the PCA estimator of dynamic factor models is constructed from the principal components of the variance-covariance matrix. Lastly, Theorem 5 suggests that the PCA estimator is consistent (at least under the theorem's assumptions) even if the number of states is misspecified.[25] I am aware of no similar result on the consistency of the PCA estimator for dynamic factor models when the number of common factors is misspecified.

4 Behavioral Implications

In this section, I apply the characterization results from the previous section to develop the behavioral implications of the simple models framework. Throughout the section, I maintain the assumption that at least one of the following is satisfied for every agent: (a) the agent is constrained to use one-state models and the true process is exponentially ergodic; (b) the agent is

[24] See, for instance, Stock and Watson (2002).

[25] An estimator for a misspecified model is consistent if the estimate converges to a pseudo-true model as the sample size goes to infinity.

constrained to use Markovian one-state models; or (c) the agent is constrained to use Markovian d-state models and the lag-one autocovariance matrix, Γ_1, is symmetric.

To elaborate on the behavioral implications of the framework, I embed it in a reduced-form economy. Consider a finite set of agents, indexed by $j = 1, \ldots, J$. In every period t, each agent j takes a purely forward-looking decision x_{jt}, which depends on her forecasts via the best-response function

$$x_{jt} = E_{jt}\left[\sum_{s=1}^{\infty} c_{js}' y_{t+s}\right], \tag{8}$$

where $y_t \in \mathbb{R}^n$ is as before the vector of observables, $E_{jt}[\cdot]$ denotes agent j's subjective forecasts, and $c_{js} \in \mathbb{R}^n$ are preference parameters satisfying $\sum_{s=1}^{\infty} \|c_{js}\|_2 < \infty$ for all j.[26] I continue to take the true process $\mathbb{P}$ as a primitive of the economy and assume that agent j can entertain state-space models with no more than d_j states. In Online Appendix F, I provide an analysis suggesting that the partial equilibrium insights would generalize to a general equilibrium economy, in which $\mathbb{P}$ itself is an endogenous outcome of agents' choices.

The reduced-form specification in (8) allows the derivation of sharp theoretical results, which highlight the role of simple models and biased forecasts and are independent of the specifics of agents' decision problems. These results are valid up to first order for purely forward-looking decisions that depend non-linearly on the forecasts of the observable. They also hold arbitrarily well when decisions are sufficiently forward-looking (e.g., when the discount factor is close to one). In the next section, I further develop the implications of the general framework in the context of a microfounded general equilibrium macro model.

4.1 Persistence Bias

Decomposing the observable into its more and less persistent components will be useful for the subsequent discussions:

Proposition 3. *Let a_i denote the ith largest eigenvalue of the first autocorrelation matrix, C_1, in magnitude, and let u_i denote the corresponding eigenvector, normalized such that $u_i' u_k = \mathbb{1}_{\{i=k\}}$ for all i and k. The observable can be decomposed as follows:*

$$y_t = \sum_{i=1}^{n} y_t^{(i)} q_i, \tag{9}$$

where $y_t^{(i)} \equiv p_i' y_t$, $p_i \equiv \Gamma_0^{\frac{-1}{2}} u_i$, $q_i \equiv \Gamma_0^{\frac{1}{2}} u_i$, u_i is as in Theorem 5, and scalars $y_t^{(i)}$ all have unit variance. If ρ_i denotes the lag-one autocorrelation of $y_t^{(i)}$, then $|\rho_1| \geq |\rho_2| \geq \cdots \geq |\rho_n|$.

This proposition represents the observable in terms of the basis vectors $\{q_i\}_{i=1}^n$, with $y_t^{(i)}$ denoting the components (or coordinates) of y_t with respect to this basis. The components

[26] The assumption that each agent takes a single action is without loss of generality. The analysis would be identical if one instead assumed that agent j makes multiple choices in each period, with the kth action of agent j given by $x_{jkt} = E_{jt}\left[\sum_{s=1}^{\infty} c_{jks}' y_{t+s}\right]$.

of y_t are sorted by their persistence, with $y_t^{(1)}$ representing the *most persistent component* and $y_t^{(n)}$ the *least persistent component* of the observable. This decomposition is valid for arbitrary stationary stochastic processes and is independent of agents' forecasting and decision problems. However, the way agents' choices respond to changes in the observable neatly aligns with the decomposition in (9). This is shown in the following corollary of Theorems 4 and 5:

Corollary 2 (persistence bias)**.** *Agent j's time-t forecasts and forward-looking actions only respond to changes in the d_j most persistent components of y_t.*

Agents who use pseudo-true d-state models treat the most persistent and least persistent components of y_t in qualitatively different ways. A change in the current value of the observable can be decomposed into changes in the components $y_t^{(i)}$ of y_t. Agents do not change their forecasts in response to changes in the least persistent components of y_t. Consequently, their forward-looking actions also remain unresponsive to current changes in these less persistent components.

It is worth noting that agents' forecasts and actions are unresponsive to changes in the less persistent components of the observable only on impact. In general, different components of y_t do not evolve independently. Therefore, a change in the current value of $y_t^{(i)}$ could lead to changes in the values of $y_{t+s}^{(j)}$ for some $j \neq i$ and $s > 0$. This can result in a delayed response of agents' forecasts and actions to changes in the observable's less persistent components.

4.2 Increased Comovement

Constraining agents to use simple models increases the comovement between their forward-looking choices. The argument for this prediction is best seen by considering agents j and k, both of whom are constrained to use one-state models. Because of persistence bias, the agents' time-t actions can be written as time-invariant linear functions of the observable's most persistent component. More specifically, $x_{jt} = g_j^{(1)} y_t^{(1)}$ and $x_{kt} = g_k^{(1)} y_t^{(1)}$, where $g_j^{(1)}$ and $g_k^{(1)}$ are constants that depend on the true process and the agents' preferences. Thus, one agent's actions can be expressed as a constant multiple of the other agent's actions. In other words, the agents' actions comove perfectly. The following proposition formalizes and extends this conclusion:

Proposition 4. *Let $D \equiv \max_j d_j$ denote the largest value of d_j among agents and $x_t \equiv (x_{jt})_j \in \mathbb{R}^J$ denote the vector containing agents' time-t actions. Given generic true processes, x_t has the factor structure*

$$x_t = G y_t^{(1:D)},$$

where G is a $J \times D$ matrix of loadings and $y_t^{(1:D)}$ is the D-dimensional vector consisting of the D most persistent components of the observable.[27]

[27] The result requires all pseudo-true models of a given dimension to be observationally equivalent, a condition that holds for generic true processes.

The proposition establishes that the J-dimensional vector of all the forward-looking actions of all the agents in the economy moves with the D factors collected in $y_t^{(1:D)}$. The number of factors depends solely on the complexity of agents' models, while the composition of these factors depends on the properties of the true process. The loadings of actions on different factors depend on the preference parameters c_{js}. If D is much smaller than J (which is often a reasonable assumption), then a large number of actions comove with movements in a small number of factors.

Agents' actions exhibit comovement not only among those using models with the same value of d but also among those using models of different dimensions. To see the intuition for this result, consider agents j and k who use models of dimensions d_j and $d_k > d_j$, respectively. While the agents disagree on the number of state variables needed to forecast the observable, they agree on what d_j of those state variables ought to be. The d_j states used by agent j are a subset of the d_k states used by agent k (up to linear transformations). This strong form of comovement is a unique prediction of the framework of simple models. This result relies on the fact that pseudo-true d-state models rank the components of y_t consistently across d: As d increases, pseudo-true forecasts condition on additional components of y_t but without altering the components already being used to forecast.

A low-dimensional factor structure is one natural expression of comovement. Another commonly used comovement measure is the Pearson correlation coefficient between two variables. The following corollary of Proposition 4 shows that constraining any two agents to one-state models increases the correlation between their actions: [28]

Corollary 3. *Consider actions j and k, both of the form (8), taken by agents j and k with $d_j = d_k = 1$. Generically,*

$$1 = \left|Corr\left(x_{jt}^{1d}, x_{kt}^{1d}\right)\right| > \left|Corr\left(x_{jt}^{RE}, x_{kt}^{RE}\right)\right|,$$

where x_{jt}^{1d} and x_{jt}^{RE} denote agent j's time-t action when using a pseudo-true one-state model and the true model, respectively.

The time-t actions of agents using a pseudo-true one-state model depend solely on the current realization of $y_t^{(1)}$, the most persistent component of y_t. Consequently, an econometrician who analyzes those actions will conclude that the actions are driven by a single shock to the economy. This conclusion holds regardless of the specifics of preferences, technology, or market structure. It holds both in partial equilibrium and in general equilibrium, as suggested by the analysis in Online Appendix F. However, the single shock recovered by the econometrician is not a true shock. It is an endogenous index whose statistical properties depend on the primitives of the economy, the stochastic properties of the shocks that hit it, and the parameters of policy rules.

[28] Note that this corollary does not generalize beyond the one-state case; constraining agents to models with $d_j, d_k > 1$ states might actually *decrease* the correlation between their actions compared to the rational-expectations benchmark.

4.3 Under- and Over-Extrapolation

The framework proposed in this paper is neither a model of under-extrapolation nor of over-extrapolation. Instead, agents who use simple models forecast using a parsimonious model that provides an approximation to the true process and balances forecast errors across different horizons and variables. Consequently, simple models do not lead to mistakes that invariably go in the same direction. In fact, agents who use a pseudo-true d-state model under-extrapolate some variables and over-extrapolate others.

Proposition 5. *Let $y_t^{(1)}$ denote the most persistent component of y_t and $y_t^{(n)}$ denote its least persistent component. If the true process is exponentially ergodic and $d_j < n$, then:*

(a) Agent j overestimates the magnitude of $y_t^{(1)}$'s autocorrelation at all lags.

(b) Agent j underestimates the magnitude of $y_t^{(n)}$'s autocorrelation at all lags.

The following example illustrates the proposition:

Example 3. Suppose the vector of observables is given by $y_t = (y_{1t}, y_{2t})' \in \mathbb{R}^2$, and each element of y_t follows an independent ARMA$(1,1)$ process

$$y_{1t} = \phi_1 y_{1,t-1} + \epsilon_{1t} + \vartheta_1 \epsilon_{1,t-1},$$
$$y_{2t} = \phi_2 y_{2,t-1} + \epsilon_{2t} + \vartheta_2 \epsilon_{2,t-1},$$

where $\phi_1, \phi_2, \vartheta_1, \vartheta_2 \in (0,1)$ are constants, and ϵ_{1t} and ϵ_{2t} are i.i.d. mean-zero random variables with finite variances. Additionally, assume that $\phi_1 > \phi_2$ and $\frac{(\phi_1+\vartheta_1)(1+\phi_1\vartheta_1)}{1+2\phi_1\vartheta_1+\vartheta_1^2} > \frac{(\phi_2+\vartheta_2)(1+\phi_2\vartheta_2)}{1+2\phi_2\vartheta_2+\vartheta_2^2}$. This assumption ensures that y_{1t} has a higher autocorrelation than y_{2t} at all lags.

The lag-l autocorrelation matrix is given by

$$C_l = \begin{pmatrix} \frac{(\phi_1+\vartheta_1)(1+\phi_1\vartheta_1)}{1+2\phi_1\vartheta_1+\vartheta_1^2}\phi_1^{l-1} & 0 \\ 0 & \frac{(\phi_2+\vartheta_2)(1+\phi_2\vartheta_2)}{1+2\phi_2\vartheta_2+\vartheta_2^2}\phi_2^{l-1} \end{pmatrix}.$$

The ith largest eigenvalue of C_1 in magnitude is $\frac{(\phi_i+\vartheta_i)(1+\phi_i\vartheta_i)}{1+2\phi_i\vartheta_i+\vartheta_i^2}$, and the corresponding eigenvector is $u_i = e_i$, where e_i denotes the ith standard coordinate vector. Since the two elements of y_t are independent, the variance-covariance matrix, Γ_0, is diagonal. Therefore, $y_t^{(1)} q_1 = y_{1t} e_1$ and $y_t^{(2)} q_2 = y_{2t} e_2$, i.e., the most persistent component of y_t is its first component in the standard coordinates and its least persistent component is its second component in the standard coordinates. The spectral radius of the lag-l autocorrelation matrix satisfies

$$\rho(C_l) = \frac{(\phi_1+\vartheta_1)(1+\phi_1\vartheta_1)}{1+2\phi_1\vartheta_1+\vartheta_1^2}\phi_1^{l-1} \geq \left(\frac{(\phi_1+\vartheta_1)(1+\phi_1\vartheta_1)}{1+2\phi_1\vartheta_1+\vartheta_1^2}\right)^l = \rho(C_1)^l,$$

with the inequality strict for $l > 1$. That is, the true process is exponentially ergodic.

The pseudo-true one-state model is described by Theorem 4. Under any such model, y_{1t} follows an AR(1) process with persistence parameter $a = \frac{(\phi_1+\vartheta_1)(1+\phi_1\vartheta_1)}{1+2\phi_1\vartheta_1+\vartheta_1^2}$ and y_{2t} is i.i.d. over time. The pseudo-true lag-l autocorrelation of y_{1t} is equal to a^l, while the pseudo-true lag-l autocorrelation of y_{2t} is zero for any $l \geq 1$. On the other hand, the true lag-l autocorrelation of y_{it} is given by $\frac{(\phi_i+\vartheta_i)(1+\phi_i\vartheta_i)}{1+2\phi_i\vartheta_i+\vartheta_i^2}\phi_i^{l-1}$ for $i = 1, 2$. Therefore, an agent who uses a pseudo-true one-state model overestimates the autocorrelation of y_{1t} at all lags (strictly so for lags $l > 1$) while strictly underestimating the autocorrelation of y_{2t} at all lags.

Agents who use simple models over-extrapolate the most persistent components of the observable and under-extrapolate the least persistent ones. For observables with intermediate persistence, the pattern could be under- or over-extrapolation depending on the variable and the horizon being considered. These predictions set this paper's framework apart from models that hardwire under- or over-extrapolation.

5 A Business-Cycle Application

In this section, I use the framework of simple models to study how bounded rationality changes the response of an economy to supply, demand, and policy shocks and the propagation of those shocks to endogenous variables. I do so in the context of a standard business-cycle model economy, which combines elements of the New Keynesian and real business cycle models. This exercise demonstrates that the macroeconomic model's empirical fit is improved when bounded rationality is introduced in the form of dimensionality reduction. Moreover, simple models can serve as a parsimonious substitute for add-ons, such as external habit formation, investment-adjustment costs, and endogenous capital utilization, which are used to increase the persistence and comovement of endogenous variables.

5.1 The Model Economy

The model economy is a New Keynesian economy with price and wage rigidities and endogenous capital formation. Alternatively, the model can be viewed as a DSGE model à la Christiano, Eichenbaum, and Evans (2005), Smets and Wouters (2007), and Justiniano, Primiceri, and Tambalotti (2010) but without the following add-ons: (i) external habit formation in consumption, (ii) investment-adjustment costs, (iii) price and wage indexation, (iv) endogenous capital utilization, (v) a monetary policy that responds to the level and growth rate of the output gap. The first three add-ons increase the persistence of consumption, investment, inflation, and wages. The last two enhance the comovement properties of the economy.

To focus on the core mechanisms, I initially consider a version of the model with only three shocks: a total-factor productivity (TFP) shock, an investment shock, and a monetary-policy shock. The TFP shock is the main supply shock in DSGE models, while the investment shock is the demand shock that explains the largest variance shares of real variables at business-cycle frequencies (Justiniano et al., 2010). Monetary-policy shocks contribute little to the variance of nominal and real variables at business-cycle frequencies. However, they have clear empirical counterparts that can be identified using vector autoregression (VAR), narrative, and high-frequency approaches. Comparing model-implied impulse-response functions (IRFs) to monetary-policy shocks with their empirical counterparts provides a powerful test of the model economy's internal propagation mechanism (Christiano et al., 2005). I focus on these three shocks to clearly illustrate how bounded rationality alters the behavior of the model economy. Later, I enrich the economy with a full suite of standard DSGE shocks, estimate it using Bayesian techniques—both given simple models and under rational expectations—and perform Bayesian model selection.

The economy is populated by seven groups of agents: final-good producers, intermediate-goods producers, investment firms, employment agencies, households, labor unions, and the government. In what follows, I describe each group's problem in detail.

Final-good producers

The final good Y_t is produced by competitive firms by combining a continuum of intermediate goods, indexed by i, according to the CES production function

$$Y_t = \left[\int_0^1 Y_t(i)^{\frac{1}{1+\lambda_p}}\, di\right]^{1+\lambda_p},$$

where λ_p denotes the elasticity of substitution. Profit maximization and the zero-profit condition imply that the price of the final good is given by the price index

$$P_t = \left[\int_0^1 P_t(i)^{\frac{1}{\lambda_p}}\, di\right]^{\lambda_p},$$

where $P_t(i)$ denotes the price of intermediate good i. The demand for good i is given by the isoelastic demand schedule

$$Y_t(i) = \left(\frac{P_t(i)}{P_t}\right)^{-\frac{1+\lambda_p}{\lambda_p}} Y_t.$$

Intermediate-goods producers

A monopolist produces intermediate good i according to the production function

$$Y_t(i) = \max\left\{a_t K_t(i)^{\alpha}\left(\gamma^t L_t(i)\right)^{1-\alpha} - \gamma^t F, 0\right\},$$

where $K_t(i)$ and $L_t(i)$ denote the capital and labor inputs of the monopolist, respectively, F is a fixed cost of production, chosen so that profits are zero along the balanced-growth path, γ denotes the exogenous rate of labor-augmenting technological progress, and a_t is a stationary TFP shock, which follows the AR(1) process $\log a_t = \rho_a \log a_{t-1} + \varepsilon_{at}$ with ε_{at} i.i.d. $\mathcal{N}(0, \sigma_a^2)$.

Intermediate-goods producers are subject to nominal frictions à la Calvo. Each period the price of a randomly-selected fraction ξ_p of intermediate goods grows at rate π, where π denotes the value of gross inflation rate along the balanced-growth path. The remaining intermediate-goods producers choose their prices $P_t(i)$ optimally by maximizing the present-discounted value of future profits,

$$E_{pt}\left[\sum_{s=0}^{\infty}\xi_p^s\beta^s\Lambda_{t+s}\Big(\pi^s P_t(i)Y_{t+s}(i)-W_{t+s}L_{t+s}(i)-r_{t+s}K_{t+s}(i)\Big)\right],$$

subject to the demand curve

$$Y_{t+s}(i)=\left(\frac{\pi^s P_t(i)}{P_{t+s}}\right)^{-\frac{1+\lambda_p}{\lambda_p}}Y_{t+s},$$

where Λ_t is the marginal utility of nominal income, W_t is the nominal wage, r_t is the rental rate of capital, and E_{pt} denotes the time-t forecasts of intermediate-goods producers about the path $\{\Lambda_{t+s}, W_{t+s}, r_{t+s}, P_{t+s}, Y_{t+s}, a_{t+s}\}_{s\geq 1}$ of variables they take as given.

Investment firms

The capital stock of the economy is owned by competitive investment firms. They take the rental rate of capital and the price of the final good as given and maximize the present-discounted value of profits

$$E_{it}\left[\sum_{s=0}^{\infty}\beta^s\Lambda_{t+s}\left(r_{t+s}K_{t+s}-P_{t+s}I_{t+s}\right)\right],$$

subject to the capital accumulation equation

$$K_{t+1}=(1-\delta)K_t+\mu_t\left(I_t-S_k\left(\frac{I_t}{K_t}\right)K_t\right),$$

where I_t is investment, K_t denotes the physical capital, E_{it} denotes the time-t forecasts of investment firms, $S_k(\cdot)$ represents the adjustment cost function, and μ_t is the investment shock, which follows the AR(1) process $\log\mu_t = \rho_\mu\log\mu_{t-1}+\varepsilon_{\mu t}$ with $\varepsilon_{\mu t}$ is i.i.d. $\mathcal{N}(0,\sigma_\mu^2)$. I assume that the adjustment cost satisfies $S_k = S_k' = 0$ and $S_k'' = \varsigma_k > 0$ along the balanced-growth path.[29] I also assume there is no spot market for installed capital.[30]

Employment agencies

There is a continuum of households, indexed by j, each of which is a monopolistic supplier of a specialized type of labor. A competitive employment agency combines specialized labor into a

[29] Note that the adjustment cost is a neoclassical cost à la Hayashi (1982), not the investment-adjustment cost common in the DSGE literature. The investment-adjustment cost specification leads to an investment Euler equation with a backward-looking term, whereas investment will have no backward-looking term in the current specification.

[30] This assumption is immaterial under rational expectations. However, this may no longer be the case away from rational expectations: When there is no spot market for capital, investment depends on agents' expectations about the infinite future path of returns to capital; when a spot market exists, investment only depends on agents' expectations of the rental rate of capital and its price in the next period.

homogeneous labor input using the CES function

$$L_t = \left[\int_0^1 L_t(j)^{\frac{1}{1+\lambda_w}} dj \right]^{1+\lambda_w},$$

where λ_w denotes the elasticity of substitution among differentiated types of labor. Profit maximization by employment agencies and the zero-profit condition imply that the price of the homogeneous labor input is given by the wage index

$$W_t = \left[\int_0^1 W_t(j)^{\frac{1}{\lambda_w}} \right]^{\lambda_w},$$

and the demand for the labor of type j is given by the isoelastic labor-demand curve

$$L_t(j) = \left(\frac{W_t(j)}{W_t} \right)^{-\frac{1+\lambda_w}{\lambda_w}} L_t.$$

Households

Households supply labor, consume the final good, and save in a short-term nominal government bond. Their wages are subjective to nominal rigidities à la Calvo. However, as is common in the literature, I assume that a competitive insurance agency fully insures households against fluctuations in their labor income resulting from nominal frictions. Consequently, the equilibrium labor income of each household is equal to $W_t L_t$, the average labor income in the economy.

Each household takes the labor income and the stream of profits from the ownership of firms as given and chooses consumption and saving in government bonds to maximize the utility function

$$E_{ct} \left[\sum_{s=0}^{\infty} \beta^s \left(\log(C_{t+s}) - \varphi \frac{L_{t+s}(j)^{1+\nu}}{1+\nu} \right) \right],$$

subject to a no-Ponzi condition and the nominal budget constraint

$$P_t C_t + T_t + B_t \leq R_{t-1} B_{t-1} + W_t L_t + \Pi_t,$$

where C_t is consumption, T_t denotes lump-sum taxes, B_t is the holding of one-period government bonds, R_t is the gross nominal interest rate, Π_t denotes profits from the ownership of firms, ν is the inverse Frisch elasticity of labor supply, and φ is a constant that determines the steady-state working hours. The operator E_{ct} denotes the time-t forecasts of households about the path $\{L_{t+s}, W_{t+s}, P_{t+s}, T_{t+s}, R_{t+s}, \Pi_{t+s}\}_{s\geq 1}$ of aggregate and idiosyncratic observables that enter their decision problem.

Labor unions

Wages are set by a continuum of labor unions, also indexed by j, each representing a household. Each period, a randomly-selected fraction ξ_w of unions cannot freely set the wage of the household they represent. The nominal wages of those households grow at the rate $\gamma\pi$.[31] The remaining fraction of labor unions sets the optimal wage $W_t(j)$ by maximizing

$$E_{wt}\left[\sum_{s=0}^{\infty}\xi_w^s\beta^s\left(-\varphi\frac{L_{t+s}(j)^{1+\nu}}{1+\nu}+\Lambda_{t+s}(\gamma\pi)^s W_t(j)L_{t+s}(j)\right)\right]$$

subject to the labor demand curve

$$L_{t+s}(j)=\left(\frac{(\gamma\pi)^s W_t(j)}{W_{t+s}}\right)^{-\frac{1+\lambda_w}{\lambda_w}}L_{t+s},$$

where E_{wt} denotes the time-t forecasts of labor unions about the variables they take as given.

The government

The monetary policy sets the nominal interest rate following a Taylor rule

$$\frac{R_t}{R}=\left(\frac{R_{t-1}}{R}\right)^{\rho_R}\left(\frac{\pi_t}{\pi}\right)^{(1-\rho_R)\phi_\pi}m_t,$$

where $\pi_t\equiv P_t/P_{t-1}$, and R and π are the steady-state gross nominal interest rate and inflation rate, respectively.[32] [33] m_t is a monetary policy shock that follows the AR(1) process $\log m_t = \rho_m\log m_{t-1}+\varepsilon_{mt}$ with ε_{mt} is i.i.d. $\mathcal{N}(0,\sigma_m^2)$.

Government spending G_t is exogenous. In the baseline specification, I assume that government spending grows at the same rate as GDP, that is, $G_t = g\gamma^t$ for some g. The government finances spending by issuing short-term nominal bonds and levying lump-sum taxes on households. The nominal government budget constraint is given by

$$R_{t-1}B_{t-1}+P_tG_t-T_t=B_t,$$

where T_t denotes nominal taxes. Taxes follow a tax rule that ensures that the real value of public debt B_t/P_t grows at rate γ, the deterministic growth rate of the economy.[34]

5.2 Equilibrium

The analysis proceeds in two steps. The first is to characterize the *temporary equilibrium*, which imposes individual optimality and market clearing but not rational expectations.[35] The second

[31] Since there is technological progress, absent this assumption, there would be no balanced-growth path without wage dispersion. Note that this is different than the assumption of wage indexation common in the DSGE literature: Wages are not indexed to the current inflation rate but to its steady-state value.

[32] In the New Keynesian literature, it is often assumed that the monetary authority responds both to changes in the inflation rate and to changes in the output gap. However, the right notion of the output gap is not clear here: It can be defined relative to the flexible price allocation in which agents re-estimate their models, the one in which agents' models are unchanged, or the rational-expectations flexible-price allocation. I bypass the question of how the output gap ought to be defined by instead assuming that the monetary authority only responds to deviations in the inflation rate.

[33] The steady-state gross nominal interest rate π can also be seen as the central bank's inflation target.

[34] Ricardian equivalence does not necessarily hold when agents use simple models. Therefore, both the timing of taxes and the value of the outstanding public debt might affect the response of the economy to shocks. See also Eusepi and Preston (2018), where the authors use an adaptive learning framework to study the effects of the level of public debt on the transmission of monetary policy.

[35] The notion of temporary equilibrium goes back to Grandmont (1977). See Woodford (2013) for a discussion of temporary equilibria in the context of modern monetary models and Farhi and Werning (2019) for an application to heterogeneous-agent New Keynesian economies.

step is to supplement the temporary equilibrium with the model of expectation formation and characterize the resulting (full) equilibrium.

The first step of the analysis is relatively straightforward. I start by deriving agents' first-order optimality conditions. I then characterize the balanced-growth path along which inflation is constant and equal to the monetary authority's inflation target, and output, consumption, investment, government spending, capital stock, real wages, and the public debt all grow at rate γ, the deterministic growth rate of labor productivity. Finally, I log-linearize the optimality conditions around the balanced-growth path.

Extra care is necessary when working with the optimality conditions without imposing rational expectations. Away from rational expectations, one agent's optimality conditions cannot be simplified using equilibrium conditions that are not necessarily respected by the agent's expectation operator—conditions such as other agents' optimality conditions.[36] For instance, the optimality conditions of firms resetting their prices cannot be combined with the law of motion of the price index to obtain the usual recursive version of the Phillips curve. Likewise, households' optimality conditions cannot be combined with the government budget constraint to obtain the usual recursive consumption-Euler equation.

I instead simplify each agent's first-order optimality conditions using only equations that are fully understood by the agent. Combining households' first-order optimality conditions with their budget constraints and no-Ponzi conditions yields the following version of the permanent-income hypothesis:

$$\hat{c}_t = -\hat{R}_t + E_{ct}\left[\sum_{s=1}^{\infty}\beta^s\left(\frac{1-\beta}{\beta}\frac{x}{c}\hat{x}_{t+s} - \frac{1-\beta}{\beta}\frac{\tau}{c}\hat{\tau}_{t+s} - \frac{x-\tau}{c}\hat{R}_{t+s} + \frac{1}{\beta}\frac{x-\tau}{c}\hat{\pi}_{t+s}\right)\right], \tag{10}$$

where lowercase letters with hats denote log-deviations from the balanced-growth path and lowercase letters without hats denote steady-state values, π refers to inflation rate, x refers to households' total pre-tax real income, and τ is their real tax burden. Likewise, the solution to investment firms' optimality conditions yields the following investment equation:

$$\hat{\imath}_t = \hat{k}_t + \frac{1}{\varsigma_k}\left(\hat{\mu}_t - \hat{\lambda}_t\right) + E_{it}\left[\sum_{s=1}^{\infty}\beta^s\left(\frac{1-\beta}{\varsigma_k\beta}\hat{\lambda}_{t+s} + \frac{1}{\varsigma_k}\left(\frac{1}{\beta} - \frac{1-\delta}{\gamma}\right)\hat{\rho}_{t+s} + \frac{1}{\varsigma_k}\left(1 - \frac{1-\delta}{\gamma}\right)\hat{\mu}_{t+s}\right)\right], \tag{11}$$

where $\hat{\rho}$ and $\hat{\lambda}$ denote log-deviations of the rental rate of capital and the stochastic discount factor, respectively, from their values along the balanced-growth path. Meanwhile, intermediate-goods firms' optimality conditions yield the following price Phillips curve:

$$\hat{\pi}_t = \kappa\left(\alpha\hat{\rho}_t + (1-\alpha)\hat{w}_t - \hat{a}_t\right) + E_{pt}\left[\sum_{s=1}^{\infty}\xi_p^s\beta^s\left(\frac{1-\xi_p}{\xi_p}\hat{\pi}_{t+s} + \kappa\left(\alpha\hat{\rho}_{t+s} + (1-\alpha)\hat{w}_{t+s} - \hat{a}_{t+s}\right)\right)\right], \tag{12}$$

[36] Preston (2005) is the first to make this point in the context of adaptive-learning models. See, also, Woodford (2003b, p. 272) for a discussion.

where κ is a constant. Finally, the solution to labor unions' wage-setting problem yields the following wage Phillips curve:

$$\hat{w}_t = \frac{(1+\beta)\kappa_w}{1+\beta-\xi_w\beta}\hat{\ell}_t + \frac{\hat{w}_{t-1}-\hat{\pi}_t}{1+\beta-\xi_w\beta} + \frac{1+\beta}{1+\beta-\xi_w\beta}E_{wt}\left[\sum_{s=1}^{\infty}\xi_w^s\beta^s\left(\frac{\nu_w\kappa_w}{1-\xi_w\beta}\hat{\pi}_{t+s} + \kappa_w\hat{\ell}_{t+s} + \nu_w\kappa_w\hat{w}_{t+s}\right)\right], \tag{13}$$

where ν_w and κ_w are constants and $\hat{\ell}_t \equiv \nu\hat{L}_t - \hat{\lambda}_t - \hat{w}_t$. Equations (10)–(13) are the model economy's only forward-looking temporary-equilibrium conditions. The remaining conditions are static and do not feature agents' subjective expectation operator. The full list of temporary-equilibrium conditions can be found in Appendix A.

Equations (10)–(13) determine consumption, investment, inflation, and the real wage as functions of current and past observables as well as agents' expectations of future values of observables. They are valid under arbitrary expectations—as long as agents all understand their own problems and their expectations satisfy the law of iterated expectations. Together with the remaining temporary-equilibrium conditions and the specification of agents' expectations, they fully determine the equilibrium of the economy.

I next describe agents' subjective expectations. For simplicity, I assume that households, investment firms, intermediate-goods producers, and labor unions all face identical constraints on the models they can entertain, thus ending up with identical subjective expectations. Every agent has perfect foresight about the balanced-growth path of the economy. Agents also all have full information about the log-deviations of all endogenous and exogenous variables from the balanced-growth path. In particular, agents' time-t information set is given by the history $\{\omega_s\}_{s\le t}$ of vector

$$\omega_s \equiv \left(\hat{a}_s, \hat{m}_s, \hat{\mu}_s, \hat{k}_s, \hat{y}_s, \hat{x}_s, \hat{c}_s, \hat{i}_s, \hat{L}_s, \hat{\rho}_s, \hat{\pi}_s, \hat{R}_s, \hat{w}_s, \hat{\tau}_s\right)',$$

consisting of the time-s realizations of TFP, monetary-policy, and investment shocks as well as log-deviations of capital stock, GDP, pre-tax income, consumption, investment, hours, rental rate of capital, inflation, nominal interest rate, wages, and taxes.[37] Instead of imposing rational expectations, I assume that agents are constrained to use one-dimensional state-space models of the form (1) to forecast future values of ω.

The equilibrium definition is straightforward. An equilibrium consists of a stochastic process $\mathbb{P}^*$ for $\{\omega_t\}_t$ and a model θ^* for agents such that (i) $\mathbb{P}^*$ is derived from market-clearing conditions and optimal behavior by all agents in the economy given subjective model θ^*, and (ii) θ^* is a pseudo-true one-state model given the stochastic process $\mathbb{P}^*$.[38]

[37] In equilibrium, some elements of ω_t are linear combinations of other elements of ω_t. By the linear invariance result, dropping the redundant variables from ω_t does not change any of the equilibrium outcomes. Similarly, I can add additional variables to ω_t that, in equilibrium, are linearly dependent on variables already included in ω_t without changing anything.

[38] In earlier work (Molavi, 2019), I referred to this equilibrium notion as the constrained rational-expectations equilibrium.

5.3 Impulse-Response Functions

I begin my analysis of the business-cycle economy by studying its impulse-response functions (IRFs) to TFP, investment, and monetary-policy shocks—both under rational expectations and under the assumption that agents use pseudo-true one-state models. I use the same calibration of primitive parameters and shock processes in both variants. This makes it possible to transparently see how bounded rationality changes the internal propagation mechanism of the economy.[39]

The model parameters are calibrated as follows: A period represents a quarter. I set $\beta = 0.99$, $\gamma = 1.005$, $\delta = 0.025$, $g/y = b/y = 0$, $\nu = 1$, $\alpha = 1/3$, $\lambda_p = 0.5$, $\rho_R = 0.8$, and $\phi_\pi = 1.5$. The persistence parameters of the shocks are set to $\rho_a = 0.95$ for TFP, $\rho_m = 0.4$ for monetary policy, and $\rho_\mu = 0.7$ for investment shocks. The standard deviations of shocks are set to $\sigma_a = 1$ for TFP, $\sigma_m = 0.5$ for monetary policy, and $\sigma_\mu = 2$ for investment shocks. These values are in the ballpark of both the values chosen in the literature and those obtained from Bayesian estimation of the model. The values of three parameters—price rigidity ξ_p, wage rigidity ξ_w, and capital-adjustment cost ς_k—have a great influence on the persistence of endogenous variables and the propagation of shocks. I assume flexible wages and set $\xi_p = 0.6$ and $\varsigma_k = 0.5$. These conservative values are picked to highlight the fact that the model economy can generate realistic IRFs without relying on counterfactually large degrees of nominal and real frictions often assumed in business-cycle models.

There are no free parameters for agents' expectations (other than d, which I have set equal to one). Agents' models, beliefs, and forecasts are all pinned down by structural parameters of the economy and the stochastic processes of the shocks. In equilibrium, agents' forecasts of elements of vector ω_s are given by

$$E_t[\omega_{t+s}] = a^{*s} q^* p^{*\prime} \omega_t,$$

where the perceived persistence is given by

$$a^* = 0.997,$$

and the relative-attention vector p^* and the relative-sensitivity vector q^* are given by[40]

$$\begin{array}{rcccccccccccccc}
 & & \hat{a} & \hat{m} & \hat{\mu} & \hat{k} & \hat{y} & \hat{x} & \hat{c} & \hat{\imath} & \hat{L} & \hat{\rho} & \hat{\pi} & \hat{R} & \hat{w} \\
p^{*\prime} & = & \big(0.01 & 0.02 & 0.01 & 0.92 & 0.07\big), & & & & & & & & \\
q^{*\prime} & = & \big(0.47 & -0.00 & 0.19 & 1.00 & 1.02 & 1.08 & 1.08 & 0.85 & -0.19 & -0.31 & 0.14 & 0.13 & 0.88\big).
\end{array}$$

Agents forecast following a three-step procedure. First, they project the vector of observables on the relative-attention vector p^* to form their estimate $\hat{z}_t \equiv p^{*\prime}\omega_t$ of the current value of the

[39] Figure B.2 plots the posterior on IRFs from a Bayesian estimation of the fully flexible version of the model (enriched with a full suite of DSGE shocks). The posterior concentrates its mass on parameter configurations with implied IRFs that are qualitatively similar to those obtained in this subsection.

[40] Vector p^* is identified only up to a set of linear transformations. Since ω_s contains redundant variables, $\tilde{p}'\omega_s = p^{*\prime}\omega_s$ for all s and a set of vectors $\tilde{p}$ belonging to a subspace. By the linear invariance result, all such vectors lead to identical forecasts and actions for agents at all times.

latent subjective state—I refer to $\hat{z}_t$ as agents' "nowcast." Then, they form their forecasts of future values of the subjective state given its perceived persistence: $E_t[z_{t+s}] = a^{*s}\hat{z}_t$. Finally, they multiply their forecasts of the subjective state by the relative-sensitivity vector q^* to form their forecasts of observables: $E_t[\omega_{t+s}] = q^* E_t[z_{t+s}]$.

Agents perceive the subjective state as highly persistent but not unit root ($a^* = 0.997$). The nowcast is much more sensitive to changes in the capital stock ($p_k^* = 0.92$) than to changes in GDP ($p_y^* = 0.07$), and it barely responds to innovations in the three exogenous shocks ($|p^*| \leq 0.02$). This is a manifestation of persistence bias: In equilibrium, the capital stock is more persistent than output and TFP, monetary-policy, and investment shocks. Agents' forecasts of capital stock, output, gross income, consumption, investment, and the real wage move almost one-for-one with changes in agents' nowcast ($|q^*| \geq 0.88$), whereas their forecasts of TFP, investment shock, hours, and the rental rate of capital exhibit much less sensitivity. The observables whose forecasts are least sensitive to new information ($|q^*| \leq 0.14$) are the nominal variables: agents' expectations of monetary-policy shock, inflation, and nominal interest rate are somewhat "anchored" to their steady-state values.

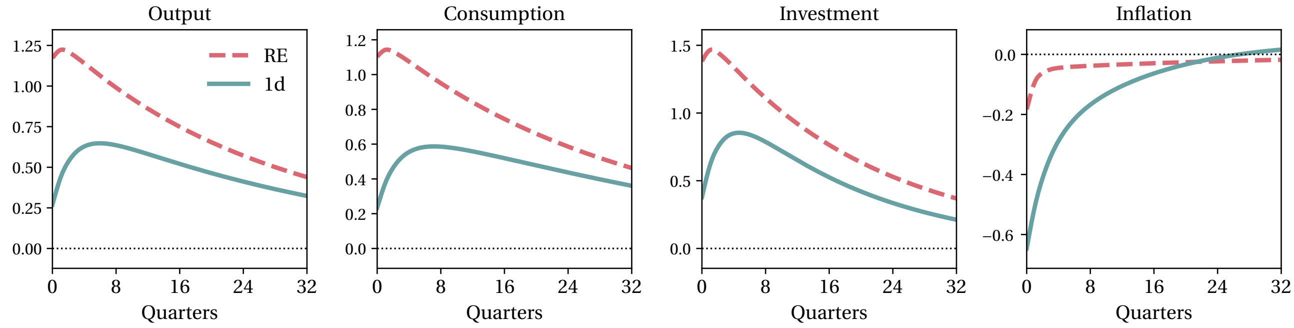

Figure 1. Impulse-response functions to a positive TFP shock.

Notes: Baseline calibration. One-dimensional simple models in solid blue. Rational expectations in dashed red. Responses to a one-percent increase in TFP. Output, consumption, and investment in percents; inflation in percentage points.

Figure 1 plots the IRFs to a positive TFP shock. With one-dimensional simple models, the IRFs of real variables mimic the hump-shaped responses found in models with features that serve to increase the sluggishness of aggregate variables, such as consumption-habit formation and investment-adjustment costs. The response of output on impact is 77% smaller with simple models than under rational expectations (RE). The corresponding figures for consumption and investment are 79% and 73%, respectively. The responses of real variables peak after one quarter under RE. With simple models, the peak ranges from six quarters after impact for consumption to eight quarters after impact for investment. Simple models thus provide a novel account of the hump-shaped responses of aggregate variables to TFP shocks in empirical studies, which does not rely on auxiliary frictions.[41] [42] The fact that quantities respond less with simple models requires

[41] For a meta-analysis of the responses of aggregate variables to technology shocks, see Ramey (2016, pp. 135–151).

[42] With simple models, the IRFs to TFP shocks are hump-shaped even when there are no nominal rigidities, no adjustment costs, and TFP

more of the increase in TFP to be absorbed by a fall in prices. The result is a larger decrease in inflation in response to the increase in TFP. Nevertheless, the response of inflation is more muted and more transitory than those of real variables (both under RE and with simple models).

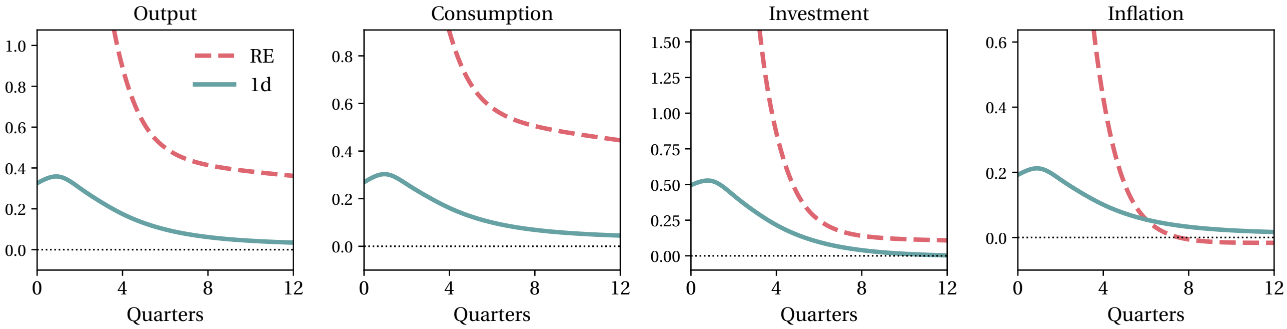

Figure 2. Impulse-response functions to an expansionary monetary-policy shock.

Notes: Baseline calibration. One-dimensional simple models in solid blue. Rational expectations in dashed red. The shock is normalized to reduce the nominal rate by 25 basis points on impact in each variant. Output, consumption, and investment in percents; inflation in percentage points.

Figure 2 plots the IRFs to an expansionary monetary-policy shock. Agents' use of simple models dampens the responses of real variables by about 98% on impact. However, here, simple models also dampen the response of inflation on impact by 98.6%. Bounded rationality also increases the persistence of responses to monetary-policy shocks. The response of all aggregate variables decrease monotonically and rapidly under rational expectations. In contrast, the responses are hump-shaped and significantly more persistent when agents use simple models.

The responses of aggregate variables to monetary-policy shocks are counterfactually strong under rational expectations. This is because changes in future real interest rates have the same effect on current output as changes in current real interest rates. With nominal rigidities and rational expectations, monetary-policy shocks change agents' forecasts of future real interest rates. These expected changes pass through almost one-for-one to aggregate output and consumption.[43] The additional persistence resulting from agents' use of simple models allows the model to generate realistic IRFs to monetary-policy shocks—despite the fact that the economy has no wage rigidity, wage or price indexation, habit formation, or investment-adjustment costs, and it only has moderate degrees of price rigidity and capital-adjustment costs.

Figure 3 plots the IRFs to an investment shock. Under rational expectations, the economy produces a negative comovement between consumption and investment in response to the shock. This is due to Barro and King (1984)'s observation that investment shocks increase the marginal productivity of investment and the rate of return, incentivizing households to save

is i.i.d. over time, suggesting a resolution to Cogley and Nason (1993)'s observation that the RBC model has a weak propagation mechanism. See the earlier version of this paper (Molavi, 2023) for the IRFs in the standard RBC model, which has no nominal or real frictions.

[43] This is the forward guidance puzzle of Del Negro, Giannoni, and Patterson (2023). An earlier version of this paper (Molavi, 2023) showed that replacing rational expectations with simple models greatly reduces the power of forward guidance in an estimated three-equation New Keynesian model and offers a quantitatively plausible resolution to the forward-guidance puzzle. See Angeletos and Lian (2018), García-Schmidt and Woodford (2019), and Farhi and Werning (2019) for other resolutions based on relaxations of the rational-expectations assumption.

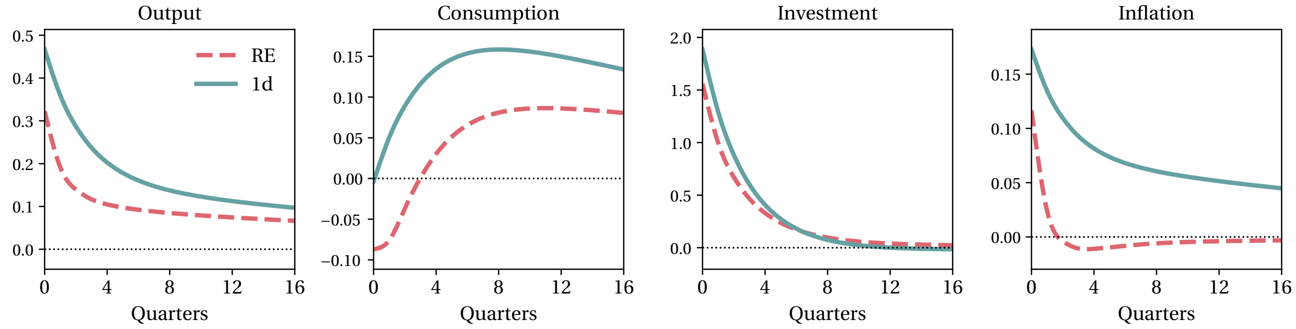

Figure 3. Impulse-response functions to a positive investment shock.

Notes: Baseline calibration. One-dimensional simple models in solid blue. Rational expectations in dashed red. Responses to a one-percent increase in the demand for investment. Output, consumption, and investment in percents; inflation in percentage points.

more and postpone consumption. DSGE models overturn this prediction through frictions such as nominal-wage rigidity, non-time-separable preferences, endogenous capital utilization, and investment-adjustment costs. The framework of simple models offers an alternative solution that relies on the anchoring of expectations. Since p_μ^* is close to zero in equilibrium, agents' expectations do not move much in response to the positive investment shocks. This dampens the initial response of forward-looking variables such as consumption to the shock. As time passes, the investment boom increases the capital stock and aggregate output, leading to an increase in the value of agents' nowcast. This improvement in agents' nowcast makes them optimistic about their future income, thus leading to an increase in consumption through the permanent-income equation (10).

These impulse-response functions offer suggestive evidence that simple models could improve the empirical fit of the business-cycle model economy. However, to establish this conclusively requires the use of a statistical model-selection criterion. I do so by estimating the parameters of the economic model separately under rational expectations and under simple models and performing Bayesian model selection.

5.4 Bayesian Inference

The model is estimated with Bayesian estimation techniques using seven key macroeconomic quarterly US time series as observable variables: the log differences of real GDP, real consumption, real investment, and the real wage; log hours worked; the log difference of the GDP deflator; and the federal funds rate. The construction of the time series closely follows Justiniano et al. (2010). In particular, consumption includes services and non-durables but excludes consumer durables, whereas investment includes consumer durables. The sample period is 1954:III–2007:IV. Online Appendix C includes more details on the data used to construct the likelihood function as well as the prior densities and posterior estimates of model parameters.

Since the likelihood uses seven macroeconomic series, the theoretical economic model needs seven exogenous shocks to avoid issues with stochastic singularity. I enrich the model with four additional shocks: an intertemporal preference shock, price- and wage-markup shocks, and a government spending shock, ending up with the seven shocks commonly assumed in the DSGE literature. The intertemporal preference and government spending shocks follow AR(1) processes with Gaussian innovations. Following Smets and Wouters (2007) and Justiniano et al. (2010), I assume that markup shocks follow ARMA(1,1) processes with Gaussian innovations. The moving average component of these shocks help capture high-frequency fluctuations in price and wage inflation.

I partition the model parameters into two groups. The first group consists of β, γ, δ, φ, F, g/y, and b/y. These parameters are set using level information not used in the Bayesian estimation step. I set $\beta = 0.99$, which implies a steady-state annualized real interest rate of about four percent. I set $\gamma = 1.005$, which implies an annual real GDP growth rate of two percent. I set $\delta = 0.025$, implying an annual rate of depreciation on capital equal to 10 percent. No value is picked for φ because the value of φ does not affect anything other than the steady-state working hours. The fixed cost of production F is set to guarantee that profits are zero along the balanced-growth path. I set the steady-state ratio of government spending to GDP g/y to 0.21 and the steady-state ratio of public debt to GDP b/y to 0.39. These values correspond to the average ratios of government spending to GDP and public debt to GDP, respectively, in the sample used in Bayesian estimation.

The priors on the remaining parameters are fairly diffuse and in line with those adopted in Smets and Wouters (2007) and Justiniano et al. (2010). Following those papers, the intertemporal preference, the price-markup, and the wage-markup shocks are normalized to enter with a unit coefficient in the consumption, inflation, and wage equations, respectively. The prior distributions of all the persistence parameters are beta, with mean 0.6 and standard deviation 0.15. The priors on the standard deviations of innovations are disperse and chosen to generate volatilities for the endogenous variables broadly in line with the data. Specifically, the priors for the standard deviations of innovations are inverse gamma, with mean 0.5 and standard deviation 1.0.

Table 1 reports the contribution of each shock to the variance of each macroeconomic variable at business cycle-frequencies. The first three columns make clear that the three aggregate demand shocks account for the largest share of the fluctuations in aggregate quantities: 94% for output, 82% for consumption, more than 99% for investment, and 52% for hours. The only other shock with significant contribution to fluctuations in these variables is the technology shock, which explains 45% of fluctuations in hours but almost no part of the fluctuations in the other three. The aggregate demand shocks are non-inflationary: together they explain less than 4% of fluctuations in inflation. This does not rely on the monetary policy aggressively leaning against them: the three aggregate demand shocks explain a negligible part of the fluctuations in the nominal interest rate.[44]

[44] These findings also hold in a (time-domain) forecast-error variance decomposition, as seen in Figure B.1.

Table 1. Posterior variance decomposition at business-cycle frequencies.

Series\shock	Government	Preference	Investment	Technology	Price markup	Wage markup	Monetary
GDP	14.2 [12.0, 15.7]	14.5 [12.3, 15.9]	65.5 [62.5, 69.6]	0.0 [0.0, 0.0]	0.2 [0.1, 0.3]	0.1 [0.0, 0.1]	5.5 [4.2, 6.3]
Consumption	0.0 [0.0, 0.0]	70.2 [61.9, 73.6]	12.3 [8.0, 22.1]	0.1 [0.0, 0.1]	1.0 [0.5, 1.3]	0.1 [0.1, 0.2]	16.2 [11.6, 19.1]
Investment	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	99.5 [99.3, 99.7]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.1 [0.0, 0.1]	0.4 [0.2, 0.6]
Hours	7.4 [6.5, 8.2]	7.5 [6.6, 8.4]	36.9 [33.2, 39.9]	45.2 [41.8, 48.2]	0.1 [0.1, 0.2]	0.0 [0.0, 0.0]	2.9 [2.2, 3.4]
Inflation	0.1 [0.1, 0.1]	0.1 [0.1, 0.1]	3.6 [0.1, 6.6]	7.6 [4.0, 9.8]	76.6 [66.3, 85.1]	12.0 [6.4, 14.8]	0.0 [0.0, 0.0]
Real wage	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	4.8 [3.1, 5.7]	1.2 [0.4, 1.8]	13.8 [9.2, 16.9]	80.2 [76.1, 85.5]	0.0 [0.0, 0.0]
Interest rate	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	0.3 [0.0, 0.5]	0.5 [0.1, 0.6]	4.1 [2.1, 5.2]	0.9 [0.3, 1.2]	94.3 [92.3, 96.8]

Notes: One-dimensional simple models. Business-cycle frequencies correspond to periodic components with cycles between 6 and 32 quarters, as in Stock and Watson (1999). Variance decomposition is performed at the posterior mode. 68 percent HPDIs computed using Laplace's approximation in brackets. HPDI bounds need not add up to one.

These findings are consistent with Angeletos, Collard, and Dellas (2020)'s anatomy of the US business cycles in the post-war period. They identify non-inflationary aggregate demand shocks as the main drivers of business cycles. These are shocks that lead to increases in output, consumption, investment, and hours with essentially no effect on inflation and no movement in TFP. Standard DSGE shocks cannot simultaneously meet all these requirements under rational expectations: Technology shocks move TFP while expansionary aggregate demand shocks generate inflation through the New Keynesian Phillips curve. Simple models increase the persistence of subjective expectations, thus significantly weakening the impact of the feedback embedded in the New Keynesian Phillips curve from expectations to current inflation. This allows demand-driven fluctuations in aggregate quantities with essentially no movement in productivity, inflation, or the nominal interest rate.

The observations made in this section give some credence to the idea that models featuring boundedly rational agents can better account for various aspects of business-cycle fluctuations. I end this section by comparing the fit of the estimated model to the data under rational expectations and simple models. I do so by comparing the marginal likelihoods of the Bayesian posteriors, reported in columns two and three of Table 2. The marginal likelihood is more than 150 log points higher with simple models than rational expectations, implying overwhelming posterior odds in favor of the former. Columns four to seven report the marginal likelihood given alternative specifications of the economic model, which feature additional add-ons but maintain rational expectations. While many of these add-ons can improve the fit, none of them do so as much as bounded rationality—although consumption-habit formation comes close. This finding suggests that the framework introduced in this paper is not only favored by data over rational expectations but also can act as a parsimonious substitute for the ad hoc frictions commonly assumed in DSGE models.

Table 2. Marginal likelihoods.

	Simple models	Rational expectations					
Add-on	—	—	Indexation	Utilization	Investment adjustment	Taylor	Habit
Log marginal likelihood	−1319.5	−1470.4	−1470.8	−1468.5	−1457.1	−1400.0	−1321.8

Notes: Log marginal likelihoods are computed using Laplace's approximation. The specifications in columns two and three only feature price and wage rigidities. The specifications in columns four to eight each feature a single addition: price and wage indexation, endogenous capital utilization, investment-adjustment costs (instead of neoclassical capital-adjustment costs), a Taylor rule that responds to the level and growth rate of the output gap, and external habit formation in consumption, respectively.

6 Conclusion

This paper suggests a novel approach to modeling bounded rationality. This approach is portable across different applications. I illustrated the use of the framework in a medium-scale new neoclassical synthesis economy. The additional persistence and comovement arising from agents' use of simple models simultaneously resolves several puzzles (the weak propagation of TFP shocks, the inability of demand shocks to generate the right comovement between investment and consumption, and the forward guidance puzzle) while improving the fit of the model to data.

An earlier version of this paper (Molavi, 2023) integrated simple models into three other workhorse models in macroeconomics, including the Diamond–Mortensen–Pissarides (DMP) model. Replacing rational expectations with simple models in the DMP model adds persistence to the unemployment rate, number of vacancies, and job-finding rate in response to both productivity and separation shocks, thus improving the internal propagation mechanism of the DMP model. Moreover, it enables the DMP model to generate negative comovement between the unemployment rate and vacancies in response to separation shocks—reversing a counterfactual prediction of the model under rational expectations.

This paper focuses on representative-agent macro models to allow for a transparent discussion of how simple models work. However, one can easily incorporate simple models into modern heterogeneous-agent macro models. This allows one to examine how bounded rationality in the face of intertemporal complexity affects predictions of such models. This can be done because neither the additional degrees of freedom nor the computational burden of finding the equilibrium with simple models scale with the size of the macro model. This paper's approach can also be extended to allow for heterogeneity in d without having to contend with the complications associated with heterogeneous-belief macro models (such as the "infinite regress" problem).

Throughout the paper, I took dimension d of agents' models as a primitive parameter. This parameter can be identified using expectations data.[45] I leave the problem of estimating d to future research.

[45] See Molavi, Tahbaz-Salehi, and Vedolin (2024) for a discussion of how this can be done in a closely related framework.

Omitted Details and Additional Results for the Business-Cycle Application

A Temporary Equilibrium

In this appendix, I list the equations that characterize the log-linearized temporary equilibrium of the business-cycle economy studied in Section 5. The derivations are straightforward, and so, are omitted for brevity. These temporary-equilibrium conditions impose individual optimality and market clearing conditions but not rational expectations. They are valid under arbitrary specifications of expectations—as long as agents all understand their own problems, and their expectations satisfy linear-invariance and the law of iterated expectations. I provide the conditions that characterize the fully flexible model used in Bayesian estimation. The baseline specification can be obtained by setting the values of parameters ξ_w, g/y, b/y, σ_ψ, σ_g, σ_p, and σ_w equal to zero.

The steady-state values are given by

$$\begin{aligned}
\rho &= \frac{\gamma}{\beta} - (1-\delta), \\
w &= \left[\frac{1}{1+\lambda_p}\alpha^\alpha(1-\alpha)^{1-\alpha}\frac{1}{\rho^\alpha}\right]^{\frac{1}{1-\alpha}}, \\
\frac{k}{L} &= \frac{w}{\rho}\frac{\alpha}{1-\alpha}, \\
\frac{F}{L} &= \left(\frac{k}{L}\right)^\alpha - \rho\frac{k}{L} - w, \\
\frac{y}{L} &= \left(\frac{k}{L}\right)^\alpha - \frac{F}{L}, \\
\frac{i}{L} &= (\gamma - (1-\delta))\,\frac{k}{L}, \\
\frac{c}{L} &= \frac{y}{L} - \frac{i}{L} - \frac{g}{y}\frac{y}{L}, \\
\frac{x}{L} &= \frac{y}{L} - \frac{i}{L}, \\
\frac{\tau}{L} &= \left(\frac{g}{y} + \frac{1-\beta}{\beta}\frac{b}{y}\right)\frac{y}{L}.
\end{aligned}$$

The log-linear permanent-income equation is given by

$$\hat{c}_t = \hat{\psi}_t - \hat{R}_t + E_{ct}\left[\sum_{s=1}^{\infty}\beta^s\left(\frac{1-\beta}{\beta}\frac{x}{c}\hat{x}_{t+s} - \frac{1-\beta}{\beta}\frac{\tau}{c}\hat{\tau}_{t+s} - \frac{1-\beta}{\beta}\hat{\psi}_{t+s} - \frac{x-\tau}{c}\hat{R}_{t+s} + \frac{1}{\beta}\frac{x-\tau}{c}\hat{\pi}_{t+s}\right)\right]. \tag{A.1}$$

Households' pre-tax income is given by

$$\hat{x}_t = \frac{y}{x}\hat{y}_t - \frac{i}{x}\hat{i}_t. \tag{A.2}$$

The intertemporal preference shock follows the exogenous process

$$\hat{\psi}_t = \rho_\psi\hat{\psi}_{t-1} + \varepsilon_{\psi t}, \qquad \varepsilon_{\psi t} \sim \mathcal{N}(0, \sigma_\psi^2). \tag{A.3}$$

Investment is given by

$$\hat{i}_t = \hat{k}_t + \frac{1}{\varsigma_k}\left(\hat{\mu}_t - \hat{\psi}_t + \hat{c}_t\right) + E_{it}\left[\sum_{s=1}^{\infty}\beta^s\left(\frac{1-\beta}{\varsigma_k\beta}\hat{\psi}_{t+s} - \frac{1-\beta}{\varsigma_k\beta}\hat{c}_{t+s} + \frac{1}{\varsigma_k}\left(\frac{1}{\beta} - \frac{1-\delta}{\gamma}\right)\hat{\rho}_{t+s} + \frac{1}{\varsigma_k}\left(1 - \frac{1-\delta}{\gamma}\right)\hat{\mu}_{t+s}\right)\right], \tag{A.4}$$

where the investment shock follows the AR(1) process

$$\hat{\mu}_t = \rho_\mu \hat{\mu}_{t-1} + \varepsilon_{\mu t}, \qquad \varepsilon_{\mu t} \sim \mathcal{N}(0, \sigma_\mu^2). \tag{A.5}$$

Capital stock evolves according to

$$\hat{k}_t = \frac{1-\delta}{\gamma}\hat{k}_{t-1} + \left(1 - \frac{1-\delta}{\gamma}\right)\left(\hat{i}_{t-1} + \hat{\mu}_{t-1}\right). \tag{A.6}$$

Government spending follows the exogenous process

$$\hat{g}_t = \rho_g \hat{g}_{t-1} + \varepsilon_{gt}, \qquad \varepsilon_{gt} \sim \mathcal{N}(0, \sigma_g^2), \tag{A.7}$$

and GDP is given by

$$\hat{y}_t = \frac{c}{y}\hat{c}_t + \frac{i}{y}\hat{i}_t + \frac{g}{y}\hat{g}_t. \tag{A.8}$$

Inflation is given by

$$\hat{\pi}_t = \hat{\lambda}_{pt} + \kappa\left(\alpha\hat{\rho}_t + (1-\alpha)\hat{w}_t - \hat{a}_t\right) + E_{pt}\left[\sum_{s=1}^{\infty}\xi_p^s\beta^s\left(\frac{1-\xi_p}{\xi_p}\hat{\pi}_{t+s} + \hat{\lambda}_{p,t+s} + \kappa\left(\alpha\hat{\rho}_{t+s} + (1-\alpha)\hat{w}_{t+s} - \hat{a}_{t+s}\right)\right)\right], \tag{A.9}$$

where $\kappa \equiv \frac{(1-\xi_p)(1-\xi_p\beta)}{\xi_p}$ is a constant, TFP follows the exogenous process

$$\hat{a}_t = \rho_a \hat{a}_{t-1} + \varepsilon_{at}, \qquad \varepsilon_{at} \sim \mathcal{N}(0, \sigma_a^2), \tag{A.10}$$

and the price markup shock follows the exogenous process

$$\hat{\lambda}_{pt} = \rho_p \hat{\lambda}_{p,t-1} + \varepsilon_{pt}, \qquad \varepsilon_{pt} \sim \mathcal{N}(0, \sigma_p^2). \tag{A.11}$$

The real wage is given by

$$\begin{aligned}\hat{w}_t =& \frac{1+\beta}{1+\beta-\xi_w\beta}\left(\hat{\lambda}_{wt} + \kappa_w\hat{\ell}_t\right) + \frac{1}{1+\beta-\xi_w\beta}\left(\hat{w}_{t-1} - \hat{\pi}_t\right) \\ &+ \frac{1+\beta}{1+\beta-\xi_w\beta}E_{wt}\left[\sum_{s=1}^{\infty}\xi_w^s\beta^s\left(\frac{\nu_w\kappa_w}{1-\xi_w\beta}\hat{\pi}_{t+s} + \hat{\lambda}_{w,t+s} + \kappa_w\hat{\ell}_{t+s} + \nu_w\kappa_w\hat{w}_{t+s}\right)\right],\end{aligned} \tag{A.12}$$

where $\kappa_w \equiv \frac{(1-\xi_w)(1-\xi_w\beta)}{\xi_w\nu_w(1+\beta)}$ is a constant,

$$\hat{\ell}_t = \nu\hat{L}_t + \hat{c}_t - \hat{w}_t, \tag{A.13}$$

and the wage markup shock $\hat{\lambda}_{wt}$ follows the exogenous process

$$\hat{\lambda}_{wt} = \rho_w \hat{\lambda}_{w,t-1} + \varepsilon_{wt}, \qquad \varepsilon_{wt} \sim \mathcal{N}(0, \sigma_w^2). \tag{A.14}$$

Hours are given by

$$\hat{L}_t = \frac{1}{1-\alpha}\left(\frac{y}{y+F}\hat{y}_t - \alpha\hat{k}_t - \hat{a}_t + \left(\frac{\rho k}{y+F} - \alpha\right)\hat{\rho}_t\right), \tag{A.15}$$

and the rental rate of capital by

$$\hat{\rho}_t = \hat{w}_t + \hat{L}_t - \hat{k}_t. \tag{A.16}$$

The nominal interest rate follows the interest rate rule

$$\hat{R}_t = \rho_R \hat{R}_{t-1} + (1-\rho_R)\phi_\pi \hat{\pi}_t + \hat{m}_t, \tag{A.17}$$

where the monetary-policy shock follows the exogenous process

$$\hat{m}_t = \rho_m \hat{m}_{t-1} + \varepsilon_{mt}, \qquad \varepsilon_{mt} \sim \mathcal{N}(0, \sigma_m^2). \tag{A.18}$$

Finally, taxes follow the tax rule

$$\hat{\tau}_t = \frac{g}{\tau}\hat{g}_t + \frac{b}{\beta\tau}\left(\hat{R}_{t-1} - \hat{\pi}_t\right). \tag{A.19}$$

B Additional Results

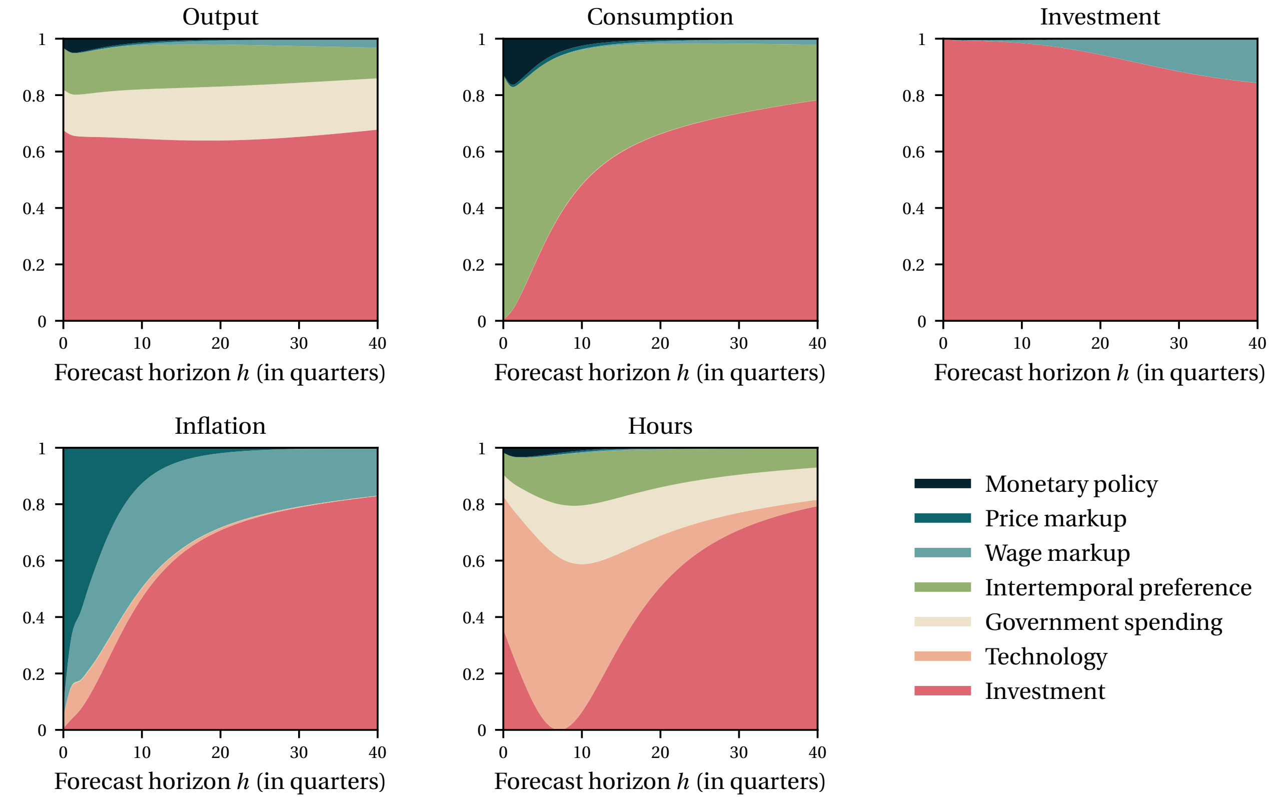

Figure B.1. Forecast-error variance decomposition.

Notes: One-dimensional simple models. Variance decomposition is performed at the posterior mode.

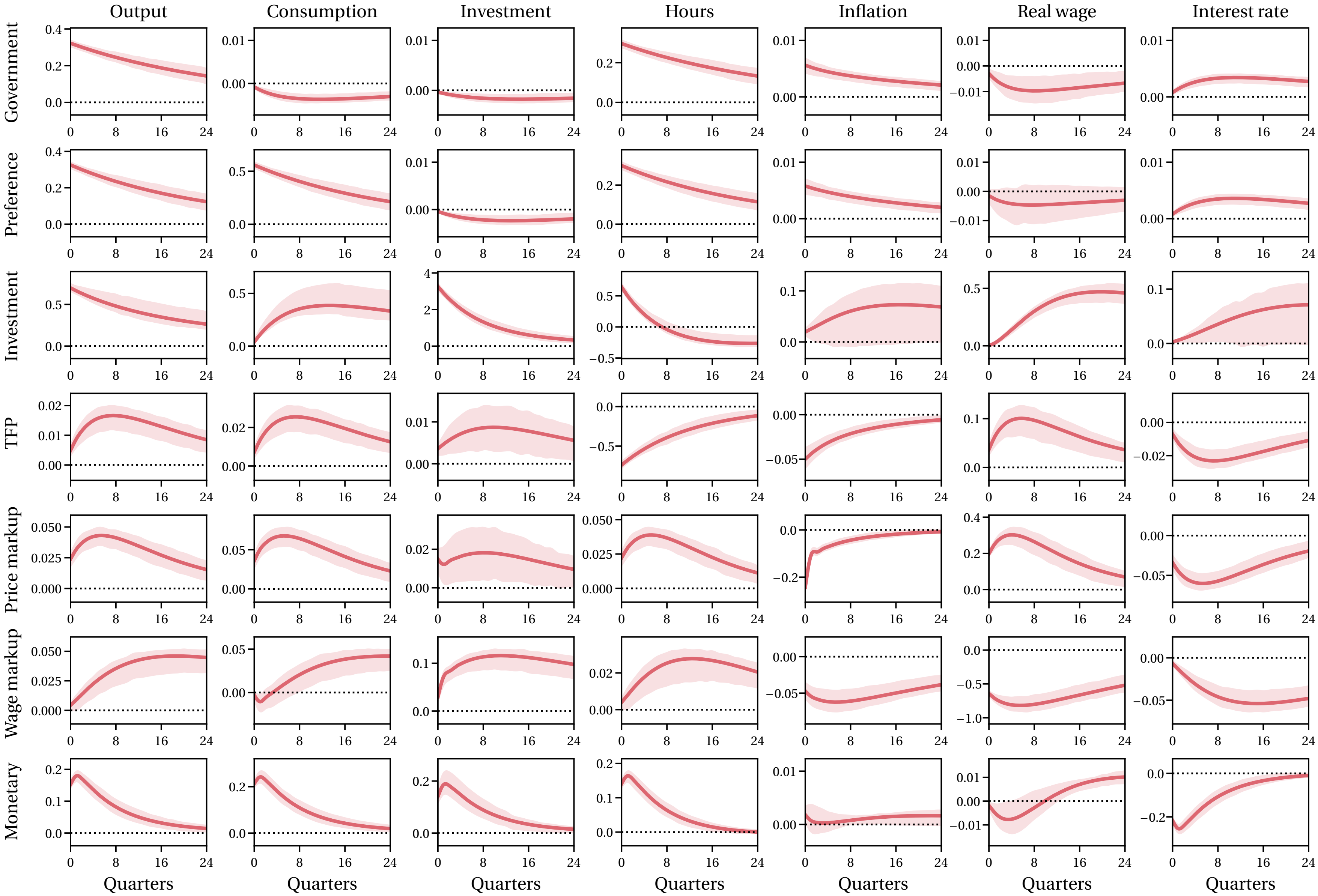

Figure B.2. Impulse-response functions.

Notes: Responses of endogenous variables (columns) to one-standard-deviation shocks (rows). One-dimensional simple models. The solid line represents the posterior mode. Shaded areas are 68 percent HPDIs computed using Laplace's approximation. Output, consumption, investment, hours, and real wage measured in percents; inflation and nominal interest rates measured in percentage points. Shocks are normalized to increase output on impact at the posterior mode.

C Bayesian Estimation

The likelihood is based on the measurement equation

$$\begin{pmatrix}\Delta Y_t & \Delta C_t & \Delta I_t & L_t & \pi_t & \Delta w_t & R_t\end{pmatrix}' = \begin{pmatrix}\hat{y}_t & \hat{c}_t & \hat{i}_t & \hat{L}_t & \hat{\pi}_t & \hat{w}_t & \hat{R}_t\end{pmatrix}' + \overline{\zeta},$$

where Δ denotes the temporal difference operator, Y denotes real GDP per capita, C denotes real consumption per capita, I denotes real investment per capita, L denotes hours worked per capita, π denotes the inflation rate, w denotes the real wage index, R denotes the nominal interest rate, and $\overline{\zeta}$ is the vector containing the sample mean of the vector on the left side of above equation. The vector of means $\overline{\zeta}$ is only informative about level variables that are fixed in the estimation step. Therefore, the likelihood can be constructed using the demeaned values of ΔY_t, ΔC_t, ΔI_t, L_t, π_t, Δw_t, and R_t.

The data are from the Federal Reserve Economic Database (FRED). Tables C.1 and C.2 describe the original data and the transformations used in Bayesian estimation.

Table C.1. Description of data.

Data	Mnemonic	Frequency	Transform
Real gross domestic product per capita	A939RX0Q048SBEA	Q	—
Share of GDP: personal consumption expenditures: nondurable goods	DNDGREI1Q156NBEA	Q	—
Share of GDP: personal consumption expenditures: services	DSERREI1Q156NBEA	Q	—
Share of GDP: personal consumption expenditures: durable goods	DDURREI1Q156NBEA	Q	—
Share of GDP: gross private domestic investment	A006REI1Q156NBEA	Q	—
Nonfarm business sector: average weekly hours	PRS85006023	Q	—
Civilian employment level	CE16OV	M	EoP
Civilian non-institutional population	CNP16OV	M	EoP
Gross domestic product: implicit price deflator	GDPDEF	Q	—
Non-farm business sector: real hourly compensation for all workers	COMPRNFB	Q	—
Effective federal funds rate	FEDFUNDS	M	Ave

Note: Q: quarterly, M: monthly, EoP: end of period, Ave: quarterly average.

Table C.2. Variables used in Bayesian estimation.

Variable	Definition
Real GDP per capita	$Y = 100 \times \log(\text{A939RX0Q048SBEA})$
Real consumption per capita	$C = 100 \times \log((\text{DNDGREI1Q156NBEA} + \text{DSERREI1Q156NBEA}) \times \text{A939RX0Q048SBEA})$
Real investment per capita	$I = 100 \times \log((\text{DDURREI1Q156NBEA} + \text{A006REI1Q156NBEA}) \times \text{A939RX0Q048SBEA})$
Hours worked	$L = 100 \times \log(\text{PRS85006023} \times \text{CE16OV}/\text{CNP16OV})$
Inflation rate	$\pi = 100 \times \log(\text{GDPDEF}/\text{GDPDEF}(-1))$
Real wage	$w = 100 \times \log(\text{COMPRNFB})$
Interest rate	$R = \text{FEDFUNDS}/4$

Table C.3. Prior densities and posterior estimates.

		Prior distribution			Posterior mode	
Coeff.	Description	Distr.	Mean	Std. Dev.	1d	RE
ν	Inverse Frisch elasticity	G	2.00	0.50	1.91 [1.48, 2.46]	0.56 [0.44, 0.71]
α	Capital share	N	0.30	0.05	0.28 [0.27, 0.29]	0.28 [0.27, 0.29]
λ_p	Steady-state price markup	B	0.15	0.05	0.50 [0.45, 0.55]	0.41 [0.36, 0.47]
λ_w	Steady-state wage markup	B	0.15	0.05	0.11 [0.07, 0.16]	0.21 [0.16, 0.26]
ξ_p	Calvo, prices	B	0.50	0.10	0.77 [0.74, 0.79]	0.65 [0.61, 0.68]
ξ_w	Calvo, wages	B	0.50	0.10	0.77 [0.72, 0.81]	0.15 [0.11, 0.19]
ρ_R	Taylor-rule smoothing	B	0.60	0.20	0.87 [0.83, 0.90]	0.54 [0.48, 0.61]
ϕ_π	Taylor rule, inflation	N	1.50	0.20	1.07 [1.03, 1.10]	1.67 [1.57, 1.77]
ς_k	Capital-adjustment cost	G	4.00	1.00	1.70 [1.47, 1.98]	1.06 [0.87, 1.29]
ρ_a	Technology shock, AR	B	0.60	0.15	0.93 [0.90, 0.95]	0.91 [0.90, 0.93]
ρ_m	Monetary-policy shock, AR	B	0.60	0.15	0.30 [0.24, 0.37]	0.26 [0.20, 0.33]
ρ_g	Government-spending shock, AR	B	0.60	0.15	0.97 [0.95, 0.98]	0.97 [0.96, 0.98]
ρ_p	Price-markup shock, AR	B	0.60	0.15	0.91 [0.87, 0.93]	0.96 [0.95, 0.97]
ρ_w	Wage-markup shock, AR	B	0.60	0.15	0.97 [0.96, 0.98]	0.97 [0.96, 0.98]
ρ_ψ	Preference shock, AR	B	0.60	0.15	0.96 [0.94, 0.97]	0.92 [0.90, 0.94]
ρ_μ	Investment shock, AR	B	0.60	0.15	0.87 [0.85, 0.89]	0.90 [0.88, 0.92]
θ_p	Price-markup shock, MA	B	0.50	0.20	0.41 [0.35, 0.48]	0.22 [0.14, 0.34]
θ_w	Wage-markup shock, MA	B	0.50	0.20	0.64 [0.57, 0.70]	0.32 [0.24, 0.41]
σ_a	Technology shock, SD	IG	0.50	1.00	0.54 [0.51, 0.57]	0.56 [0.53, 0.59]
σ_m	Monetary-policy shock, SD	IG	0.50	1.00	0.22 [0.21, 0.23]	0.31 [0.29, 0.35]
σ_g	Government-spending shock, SD	IG	0.50	1.00	1.53 [1.46, 1.61]	1.52 [1.45, 1.60]
σ_p	Price-markup shock, SD	IG	0.50	1.00	0.26 [0.24, 0.27]	0.23 [0.20, 0.27]
σ_w	Wage-markup shock, SD	IG	0.50	1.00	0.42 [0.39, 0.45]	0.91 [0.72, 1.15]
σ_ψ	Preference shock, SD	IG	0.50	1.00	0.56 [0.53, 0.59]	1.45 [1.21, 1.72]
σ_μ	Investment shock, SD	IG	0.50	1.00	5.74 [4.92, 6.70]	4.23 [3.54, 5.05]

Notes: B: beta, G: gamma, IG: inverse gamma, N: normal. 68 percent HPDIs computed using Laplace's approximation in brackets.

References

Afrouzi, Hassan, Spencer Yongwook Kwon, Augustin Landier, Yueran Ma, and David Thesmar (2021), “Overreaction in Expectations: Evidence and Theory.” Working paper.

Akaike, Hirotugu (1975), “Markovian Representation of Stochastic Processes by Canonical Variables.” *SIAM Journal on Control*, 13, 162–173.

Alvarez, Fernando E., Francesco Lippi, and Luigi Paciello (2015), “Monetary Shocks in Models with Inattentive Producers.” *The Review of Economic Studies*, 83, 421–459.

Anderson, Brian D.O. and John B. Moore (2005), *Optimal Filtering*. Dover Publications, Mineola, N.Y.

Angeletos, George-Marios, Fabrice Collard, and Harris Dellas (2018), “Quantifying Confidence.” *Econometrica*, 86, 1689–1726.

Angeletos, George-Marios, Fabrice Collard, and Harris Dellas (2020), “Business-Cycle Anatomy.” *American Economic Review*, 110, 3030–70.

Angeletos, George-Marios and Zhen Huo (2021), “Myopia and Anchoring.” *American Economic Review*, 111, 1166–1200.

Angeletos, George-Marios, Zhen Huo, and Karthik A. Sastry (2021), “Imperfect Macroeconomic Expectations: Evidence and Theory.” *NBER Macroeconomics Annual*, 35, 1–86.

Angeletos, George-Marios and Jennifer La’O (2009), “Incomplete Information, Higher-Order Beliefs and Price Inertia.” *Journal of Monetary Economics*, 56, S19–S37.

Angeletos, George-Marios and Chen Lian (2018), “Forward Guidance without Common Knowledge.” *American Economic Review*, 108, 2477–2512.

Barro, Robert J. and Robert G. King (1984), “Time-Separable Preferences and Intertemporal-Substitution Models of Business Cycles.” *The Quarterly Journal of Economics*, 99, 817–839.

Berk, Robert H. (1966), “Limiting Behavior of Posterior Distributions When the Model is Incorrect.” *Annals of Mathematical Statistics*, 37, 51–58.

Bianchi, Francesco, Cosmin Ilut, and Hikaru Saijo (2023), “Diagnostic Business Cycles.” *The Review of Economic Studies*, 91, 129–162.

Bidder, Rhys and Ian Dew-Becker (2016), “Long-Run Risk Is the Worst-Case Scenario.” *American Economic Review*, 106, 2494–2527.

Bordalo, Pedro, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer (2020), “Overreaction in Macroeconomic Expectations.” *American Economic Review*, 110, 2748–82.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer (2018), “Diagnostic Expectations and Credit Cycles.” *Journal of Finance*, 73, 199–227.

Bray, M. M. and N. E. Savin (1986), “Rational Expectations Equilibria, Learning, and Model Specification.” *Econometrica*, 54, 1129–1160.

Bray, Margaret (1982), “Learning, Estimation, and the Stability of Rational Expectations.” *Journal of Economic Theory*, 26, 318–339.

Broer, Tobias and Alexandre N. Kohlhas (2024), “Forecaster (Mis-)Behavior.” *The Review of Economics and Statistics*, 106, 1334–1351.

Bunke, Olaf and Xavier Milhaud (1998), “Asymptotic Behavior of Bayes Estimates Under Possibly Incorrect Models.” *Annals of Statistics*, 26, 617–644.

Chahrour, Ryan, Kristoffer Nimark, and Stefan Pitschner (2021), “Sectoral Media Focus and Aggregate Fluctuations.” *American Economic Review*, 111, 3872–3922.

Christiano, Lawrence J., Martin Eichenbaum, and Charles L. Evans (2005), “Nominal Rigidities and the Dynamic Effects of a Shock to Monetary Policy.” *Journal of Political Economy*, 113, 1–45.

Cogley, Timothy and James M. Nason (1993), “Impulse Dynamics and Propagation Mechanisms in a Real Business Cycle Model.” *Economics Letters*, 43, 77–81.

Coibion, Olivier and Yuriy Gorodnichenko (2015), "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts." *American Economic Review*, 105, 2644–78.

Del Negro, Marco, Marc P. Giannoni, and Christina Patterson (2023), "The Forward Guidance Puzzle." *Journal of Political Economy Macroeconomics*, 1, 43–79.

Dew-Becker, Ian and Charles G. Nathanson (2019), "Directed Attention and Nonparametric Learning." *Journal of Economic Theory*, 181, 461–496.

Douc, Randal and Eric Moulines (2012), "Asymptotic Properties of the Maximum Likelihood Estimation in Misspecified Hidden Markov Models." *Annals of Statistics*, 40, 2697–2732.

Esponda, Ignacio and Demian Pouzo (2016), "Berk-Nash Equilibrium: A Framework for Modeling Agents with Misspecified Models." *Econometrica*, 84, 1093–1130.

Esponda, Ignacio and Demian Pouzo (2021), "Equilibrium in Misspecified Markov Decision Processes." *Theoretical Economics*, 16, 717–757.

Eusepi, Stefano and Bruce Preston (2018), "Fiscal Foundations of Inflation: Imperfect Knowledge." *American Economic Review*, 108, 2551–89.

Farhi, Emmanuel and Iván Werning (2019), "Monetary Policy, Bounded Rationality, and Incomplete Markets." *American Economic Review*, 109, 3887–3928.

Faurre, Pierre L. (1976), "Stochastic Realization Algorithms." In *Mathematics in Science and Engineering*, volume 126, 1–25, Elsevier.

Forni, Mario and Marco Lippi (2001), "The Generalized Dynamic Factor Model: Representation Theory." *Econometric Theory*, 17, 1113–1141.

Fuster, Andreas, Benjamin Hebert, and David Laibson (2012), "Investment Dynamics with Natural Expectations." *International Journal of Central Banking*, 8, 243–265.

Fuster, Andreas, David Laibson, and Brock Mendel (2010), "Natural Expectations and Macroeconomic Fluctuations." *Journal of Economic Perspectives*, 24, 67–84.

Gabaix, Xavier (2014), "A Sparsity-Based Model of Bounded Rationality." *Quarterly Journal of Economics*, 129, 1661–1710.

Gabaix, Xavier (2020), "A Behavioral New Keynesian Model." *American Economic Review*, 110, 2271–2327. Working paper.

García-Schmidt, Mariana and Michael Woodford (2019), "Are Low Interest Rates Deflationary? A Paradox of Perfect-Foresight Analysis." *American Economic Review*, 109, 86–120.

Gevers, Michel and Vincent Wertz (1984), "Uniquely Identifiable State-Space and ARMA Parametrizations for Multivariable Linear Systems." *Automatica*, 20, 333–347.

Grandmont, Jean-Michel (1977), "Temporary General Equilibrium Theory." *Econometrica*, 45, 535–572.

Hansen, Lars Peter and Thomas J. Sargent (2008), *Robustness*. Princeton University Press.

Hayashi, Fumio (1982), "Tobin's Marginal Q and Average Q: A Neoclassical Interpretation." *Econometrica*, 213–224.

Ho, B. L. and R. E. Kálmán (1966), "Effective Construction of Linear State-Variable Models from Input/Output Functions." *at-Automatisierungstechnik*, 14, 545–548.

Justiniano, Alejandro, Giorgio E. Primiceri, and Andrea Tambalotti (2010), "Investment Shocks and Business Cycles." *Journal of Monetary Economics*, 57, 132–145.

Katayama, Tohru (2005), *Subspace Methods for System Identification*, volume 1. Springer.

Kleijn, B. J. K. and A. W. Van Der Vaart (2006), "Misspecification in Infinite-Dimensional Bayesian Statistics." *Annals of Statistics*, 34, 837–877.

Krusell, Per and Anthony A. Smith, Jr. (1998), "Income and Wealth Heterogeneity in the Macroeconomy." *Journal of Political Economy*, 106, 867–896.

Lorenzoni, Guido (2009), "A Theory of Demand Shocks." *American Economic Review*, 99, 2050–84.

Lucas, Robert E. (1972), "Expectations and the Neutrality of Money." *Journal of Economic Theory*, 4, 103–124.

Maćkowiak, Bartosz and Mirko Wiederholt (2009), "Optimal Sticky Prices under Rational Inattention." *American Economic Review*, 99, 769–803.

Maćkowiak, Bartosz and Mirko Wiederholt (2015), "Business Cycle Dynamics under Rational Inattention." *Review of Economic Studies*, 82, 1502–1532.

Mankiw, N. Gregory and Ricardo Reis (2002), "Sticky Information versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." *Quarterly Journal of Economics*, 117, 1295–1328.

Molavi, Pooya (2019), "Macroeconomics with Learning and Misspecification: A General Theory and Applications." Working paper.

Molavi, Pooya (2023), "Simple Models and Biased Forecasts." Working paper, arXiv:2202.06921.

Molavi, Pooya, Alireza Tahbaz-Salehi, and Andrea Vedolin (2024), "Model Complexity, Expectations, and Asset Prices." *The Review of Economic Studies*, 91, 2462–2507.

Nimark, Kristoffer (2008), "Dynamic Pricing and Imperfect Common Knowledge." *Journal of Monetary Economics*, 55, 365–382.

Orphanides, Athanasios (2003), "Monetary Policy Evaluation With Noisy Information." *Journal of Monetary Economics*, 50, 605–631.

Preston, Bruce (2005), "Learning about Monetary Policy Rules when Long-Horizon Expectations Matter." *International Journal of Central Banking*.

Rabin, Matthew and Dimitri Vayanos (2010), "The Gambler's and Hot-Hand Fallacies: Theory and Applications." *Review of Economic Studies*, 77, 730–778.

Ramey, Valerie A. (2016), "Macroeconomic Shocks and Their Propagation." *Handbook of Macroeconomics*, 2, 71–162.

Sawa, Takamitsu (1978), "Information Criteria for Discriminating Among Alternative Regression Models." *Econometrica*, 46, 1273–1291.

Shalizi, Cosma Rohilla (2009), "Dynamics of Bayesian Updating with Dependent Data and Misspecified Models." *Electronic Journal of Statistics*, 3, 1039–1074.

Silverman, Leonard M. (1976), "Discrete Riccati Equations: Alternative Algorithms, Asymptotic Properties, and System Theory Interpretations." In *Control and Dynamic Systems*, volume 12, 313–386, Elsevier.

Sims, Christopher A. (2003), "Implications of Rational Inattention." *Journal of Monetary Economics*, 50, 665–690.

Slobodyan, Sergey and Raf Wouters (2012), "Learning in a medium-scale dsge model with expectations based on small forecasting models." *American Economic Journal: Macroeconomics*, 4, 65–101.

Smets, Frank and Rafael Wouters (2007), "Shocks and Frictions in US Business Cycles: A Bayesian DSGE Approach." *American Economic Review*, 97, 586–606.

Stock, James H. and Mark Watson (2011), "Dynamic Factor Models." *Oxford Handbooks Online*.

Stock, James H. and Mark W. Watson (1999), "Business Cycle Fluctuations in US Macroeconomic Time Series." volume 1 of *Handbook of Macroeconomics*, 3–64, Elsevier.

Stock, James H. and Mark W. Watson (2002), "Forecasting Using Principal Components From a Large Number of Predictors." *Journal of the American Statistical Association*, 97, 1167–1179.

Stock, James H. and Mark W. Watson (2016), *Dynamic Factor Models, Factor-Augmented Vector Autoregressions, and Structural Vector Autoregressions in Macroeconomics*, first edition, volume 2. Elsevier B.V.

Woodford, Michael (2003a), "Imperfect Common Knowledge and the Effects of Monetary Policy." *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, 25.

Woodford, Michael (2003b), *Interest and Prices: Foundations of a Theory of Monetary Policy*. Princeton University Press.

Woodford, Michael (2013), "Macroeconomic Analysis Without the Rational Expectations Hypothesis." *Annual Review of Economics*, 5, 303–346.

Online Appendices

D Weighted Mean-Squared Forecast Error

The agent's time-t one-step-ahead forecast error given model θ is defined as

$$e_t(\theta) \equiv y_{t+1} - E_t^\theta[y_{t+1}],$$

where E_t^θ denotes the agent's subjective expectation conditional on her information at time t and given model θ. The weighted average of mean-squared forecast errors given a symmetric weight matrix $W \in \mathbb{R}^{n\times n}$ is defined as

$$\text{MSE}_W(\theta) = \mathbb{E}\left[e_t'(\theta) W e_t(\theta)\right].$$

Instead of assuming that the agent uses a model that minimizes the KLDR, one can assume that she makes her forecasts using a model θ that minimizes $\text{MSE}_W(\theta)$ for some matrix W. Using the mean-squared forecast error as the notion of fit has two disadvantages relative to the KLDR. First, the choice of matrix W introduces additional degrees of freedom when the observable is not a scalar. Second, the minimizer of the weighted mean-squared error is in general not invariant to linear transformations of the vector of observable (unless if the weight matrix W is transformed accordingly). However, the following proposition establishes that mean-squared forecast-error minimization coincides with KLDR minimization under the appropriate choice of the weighting matrix W:

Proposition D.1. *Let θ denote a pseudo-true d-state model, and let $\hat{\Sigma}_y^\theta$ denote the implied subjective variance of y_{t+1} conditional on the agent's information at time t. If W is equal to the inverse of $\hat{\Sigma}_y^\theta$, then $\theta \in \arg\min_{\theta\in\Theta_d} MSE_W(\theta)$.*

The proof of the proposition is standard, and so, is omitted.

E Exponential Ergodicity

This appendix provides a set of sufficient conditions for a process to be exponentially ergodic and discusses the relationship between those conditions and the notion of full information.

Proposition E.1. *Consider a process $\mathbb{P}$ that can be represented as*

$$\begin{aligned} f_t &= F f_{t-1} + \epsilon_t \\ y_t &= H' f_t, \end{aligned} \tag{E.1}$$

where $f_t \in \mathbb{R}^m$, ϵ_t is a zero mean i.i.d. shock with a finite variance-covariance matrix, $F \in \mathbb{R}^{m\times m}$ is a convergent matrix, $H \in \mathbb{R}^{m\times n}$, and the variance-covariance of f_t is normalized to be the identity matrix. If H is a rank-m matrix and $\left\|\frac{F+F'}{2}\right\|_2 = \|F\|_2$, where $\|\cdot\|_2$ denotes the spectral norm, then $\mathbb{P}$ is exponentially ergodic.

The assumption that the process has a representation of the form (E.1) is without loss of generality. The Wold representation theorem implies that any mean zero, covariance stationary, and purely non-deterministic process has a representation of this form (possibly with $m = \infty$). The assumption that the variance-covariance of f_t equals identity is also without loss of generality. It can always be arranged to hold by an appropriate normalization of f_t.[46] The assumption on matrix F rules out a severe form of defectiveness by guaranteeing that the largest eigenvalue of the symmetric part of F coincides with the largest singular value of F. It is satisfied, for example, if F is diagonal or symmetric, but it is much weaker than symmetry.

The most substantial assumption of the proposition is the requirement that H is a rank-m matrix. This assumption can be seen as a full-information (or spanning) assumption: If the agent observes an observable of the form (E.1) with a full-rank matrix H, then she has enough information to forecast the observable as well as in the full-information rational-expectations benchmark—even if she fails to do so due to the constraint on her set of models. The following proposition shows that this assumption, in general, cannot be dispensed with:

Proposition E.2. *Suppose the observable is one-dimensional, and the true process $\mathbb{P}$ can be represented as in* (E.1) *for some $f_t \in \mathbb{R}^m$, $\epsilon_t \sim \mathcal{N}(0, \Sigma)$, diagonal divergent matrix $F \in \mathbb{R}^{m\times m}$, diagonal matrix $\Sigma \in \mathbb{R}^{m\times m}$, and matrix $H \in \mathbb{R}^{m\times n}$. If the representation in* (E.1) *is minimal and $m > 1$, then the s-period-ahead forecast of an agent who uses a pseudo-true one-state model θ is given by*

$$E_t^{\theta}[y_{t+s}] = a^s(1-\eta)\sum_{\tau=0}^{\infty} a^{\tau}\eta^{\tau} y_{t-\tau}$$

for some $a \in (-1, 1)$ and $\eta \in (0, 1)$.[47]

F Partial Equilibrium and General Equilibrium

In this appendix, I argue that the implications of the general framework are largely unchanged in a general equilibrium setting where the observable's law of motion depends on agents' choices. I consider a stylized general equilibrium (GE) economy in which observables are linear functions of exogenous shocks and agents' actions. Specifically, I assume that, in equilibrium, the vector of observables $y_t \in \mathbb{R}^n$ can be written as

$$y_t^{\text{GE}} = \tilde{H}' f_t + g x_t^{\text{GE}}, \tag{F.1}$$

where $x_t \in \mathbb{R}$ is agents' time-t action, $f_t \in \mathbb{R}^m$ is the vector of exogenous shocks, $\tilde{H} \in \mathbb{R}^{m\times n}$ is a rank-m matrix, and $g \in \mathbb{R}^n$ is a vector that parameterizes the strength of the GE feedback from agents' actions to the aggregate observable. Agents' best-response functions are given by

$$x_t^{\text{GE}} = b' y_t^{\text{GE}} + E_t\left[\sum_{s=1}^{\infty} \beta^s c' y_{t+s}^{\text{GE}}\right]. \tag{F.2}$$

[46] See Lemma G.5 of the Online Appendix and its proof for how this can be done.

[47] The representation in (E.1) of a process is *minimal* if there exists no representation for the process of the same form in which the dimension of f_t is strictly smaller.

For simplicity, I assume that the shocks follow m independent AR(1) processes:

$$f_t = F f_{t-1} + \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, \Sigma), \tag{F.3}$$

where $F = \operatorname{diag}(\alpha_1, \ldots, \alpha_m)$ and $\Sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_m^2)$. Equations (F.1)–(F.3) together with the specification of agents' subjective expectations fully characterize the (general) equilibrium of the economy.

I contrast this economy with a partial equilibrium (PE) economy in which

$$y_t^{\mathrm{PE}} = H' f_t, \tag{F.4}$$

$$x_t^{\mathrm{PE}} = b' y_t^{\mathrm{PE}} + E_t \left[\sum_{s=1}^{\infty} \beta^s c' y_{t+s}^{\mathrm{PE}} \right], \tag{F.5}$$

and f_t follows (F.3). The term "partial equilibrium" is inspired by the following hypothetical scenario: Suppose we considered the economy described by equations (F.1)–(F.3) but ignored the fact that agents' actions affect the observable, which in turn affects agents' actions, and so on. Then the response of the GE economy to shocks would be described by equations (F.4)–(F.5). The following result establishes an observational equivalence between the GE and PE economies:

Proposition F.1. *Consider the general equilibrium economy* (F.1)–(F.3) *and the partial equilibrium economy* (F.3)–(F.5)*, and suppose that, in each economy, agents use pseudo-true Markovian d-state models to forecast the observable. If*

$$\tilde{H} = H \left(I - \left(b + \sum_{k=1}^{d} \frac{\alpha_k \beta}{1 - \alpha_k \beta} H^{\dagger} e_k e_k' H c \right) g' \right),$$

then the linear equilibria of the two economies are observationally equivalent.

Several remarks are in order. First, the result is a corollary of the linear-invariance result (Theorem 1) and the fact that agents' actions are linear in the observable. Second, the proposition covers the rational-expectations case by setting $d = m$. Third, when $\beta = 0$, the effect of going from PE to GE is to amplify the response of observables to shocks, as measured by matrix H', by the GE multiplier $(I - gb')^{-1}$. When $\beta > 0$, the multiplier has an additional term, which captures the general-equilibrium effect of the updating of expectations by agents.

Last but not least, the distinctions between exogenous and endogenous variables, on one hand, and PE and GE, on the other, are largely inconsequential in this framework. Agents' expectations of endogenous variables are consistent with their expectations of exogenous variables and the structural equations of the economy, the GE economy is just the PE economy with a linearly transformed H matrix, and agents' expectations in the GE economy are just linear transformations of their expectations in the PE economy.

G Proofs

Proof of Theorem 1

As a preliminary step, I fix an arbitrary d-state model $\theta = (A, B, Q, R)$ for the agent and compute her forecasts and the KLDR of her model from the true process. If the support of P^θ does not coincide with $\mathcal{W}$, the support of the true process, then $\text{KLDR}(\theta) = +\infty$. In what follows, I assume that P^θ is supported on $\mathcal{W}$.

Note that minimizing the KLDR over the set Θ_d of d-state models is equivalent to minimizing the KLDR over the set $\Theta_0^m \cup \Theta_1^m \cup \cdots \cup \Theta_d^m$, where Θ_k^m denotes the set of models whose minimal realization requires k state variables. Therefore, in the proofs, I assume without loss of generality that the d-state model θ is minimal, i.e., that there exists no d'-state model with $d' < d$ that is observationally equivalent to θ.

The Kullback–Leibler divergence rate. Since the entropy rate of the true process is finite, the KLDR of θ from the true process is given by

$$\text{KLDR}(\theta) = \lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left[-\log f^\theta(y_1, \ldots, y_t)\right] + \text{constant}.$$

Furthermore, by the chain rule,

$$\lim_{t\to\infty} \frac{1}{t}\mathbb{E}\left[-\log f^\theta(y_1, \ldots, y_t)\right] = \lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}\left[-\log f^\theta(y_\tau | y_{\tau-1}, \ldots, y_1)\right].$$

Since P^θ and $\mathbb{P}$ are both stationary,

$$\mathbb{E}\left[-\log f^\theta(y_\tau | y_{\tau-1}, \ldots, y_1)\right] = \mathbb{E}\left[-\log f^\theta(y_0 | y_{-1}, \ldots, y_{1-\tau})\right].$$

On the other hand, since P^θ is a stationary ergodic Gaussian process and $\mathbb{E}[\|y_t\|^2] < \infty$, the sequence $\{-\log f^\theta(y_0|y_{-1}, \ldots, y_{1-\tau})\}_\tau$ is uniformly bounded by an integrable function for any θ. Thus, by the dominated convergence theorem,

$$\lim_{\tau\to\infty} \mathbb{E}\left[-\log f^\theta(y_\tau | y_{\tau-1}, \ldots, y_1)\right] = \mathbb{E}\left[-\log f^\theta(y_0 | y_{-1}, \ldots)\right] = \mathbb{E}\left[-\log f^\theta(y_{t+1} | y_t, \ldots)\right],$$

where the second equality uses the stationarity of P^θ, and the fact that $\log f^\theta(y_{t+1}|y_t, \ldots)$ is well defined is a consequence of the assumption that A is convergent and Q is positive definite for any $\theta = (A, B, Q, R)$. The above display implies that the Cesàro sum also converges:

$$\lim_{t\to\infty} \frac{1}{t}\sum_{\tau=1}^{t} \mathbb{E}\left[-\log f^\theta(y_\tau | y_{\tau-1}, \ldots, y_1)\right] \to \mathbb{E}\left[-\log f^\theta(y_{t+1} | y_t, \ldots)\right].$$

Therefore, to compute the KLDR, I only need to compute the subjective distribution of y_{t+1} under model θ conditional on the history of observations $\{y_t, y_{t-1}, \ldots\}$.

Let $E_t^\theta[\cdot]$ denote the agent's subjective expectation given model θ and conditional on history $\{y_\tau\}_{\tau=-\infty}^{t}$, and let $\text{Var}_t^\theta(\cdot)$ denote the corresponding variance-covariance matrix. Let $\hat{z}_t \equiv E_t^\theta[z_{t+1}]$

denote the agent's conditional expectation of the subjective state. I can express $\hat{z}_t$ recursively using the Kalman filter:

$$\hat{z}_t = (A - KB')\hat{z}_{t-1} + Ky_t, \tag{G.1}$$

where $K \in \mathbb{R}^{d\times n}$ is the Kalman gain defined as

$$K \equiv A\hat{\Sigma}_z B \left(B'\hat{\Sigma}_z B + R\right)^{\dagger}, \tag{G.2}$$

the dagger denotes the Moore–Penrose pseudo-inverse, and $\hat{\Sigma}_z \equiv \mathrm{Var}_t^{\theta}(z_{t+1})$ is the subjective conditional variance of z_{t+1}, which solves the following (generalized) algebraic Riccati equation:[48] [49]

$$\hat{\Sigma}_z = A\left(\hat{\Sigma}_z - \hat{\Sigma}_z B \left(B'\hat{\Sigma}_z B + R\right)^{\dagger} B'\hat{\Sigma}_z\right) A' + Q. \tag{G.3}$$

Solving equation (G.1) backward, I get

$$\hat{z}_t = \sum_{\tau=0}^{\infty} (A - KB')^{\tau} K y_{t-\tau}.$$

The agent's subjective conditional expectation of y_{t+1} can be written in terms of her conditional expectation of z_{t+1}:

$$E_t^{\theta}[y_{t+1}] = B' E_t^{\theta}[z_{t+1}] = B' \sum_{\tau=0}^{\infty} (A - KB')^{\tau} K y_{t-\tau}.$$

Likewise, the subjective conditional variance of y_{t+1} can be expressed in terms of the subjective conditional variance of z_{t+1}:

$$\hat{\Sigma}_y \equiv \mathrm{Var}_t^{\theta}(y_{t+1}) = B'\hat{\Sigma}_z B + R. \tag{G.4}$$

More generally, the agent's s-period-ahead forecast of the vector of observables is given by

$$E_t^{\theta}[y_{t+s}] = B' A^{s-1} E_t^{\theta}[z_{t+1}] = B' A^{s-1} \sum_{\tau=0}^{\infty} (A - KB')^{\tau} K y_{t-\tau}. \tag{G.5}$$

The Kullback–Leibler divergence rate is thus equal to

$$\begin{aligned} \mathrm{KLDR}(\theta) = & -\frac{1}{2}\log\det{}^{*}\left(\hat{\Sigma}_y^{\dagger}\right) + \frac{n}{2}\log(2\pi) + \frac{1}{2}\mathrm{tr}\left(\hat{\Sigma}_y^{\dagger}\Gamma_0\right) \\ & -\frac{1}{2}\sum_{\tau=1}^{\infty}\mathrm{tr}\left(\hat{\Sigma}_y^{\dagger}\Phi_{\tau}\Gamma_{\tau}'\right) - \frac{1}{2}\sum_{\tau=1}^{\infty}\mathrm{tr}\left(\hat{\Sigma}_y^{\dagger}\Gamma_{\tau}\Phi_{\tau}'\right) \\ & +\frac{1}{2}\sum_{s=1}^{\infty}\sum_{\tau=1}^{\infty}\mathrm{tr}\left(\hat{\Sigma}_y^{\dagger}\Phi_s\Gamma_{\tau-s}\Phi_{\tau}'\right) + \text{constant}, \end{aligned} \tag{G.6}$$

where $\Gamma_l \equiv \mathbb{E}[y_t y_{t-l}']$ denotes the lag-l autocovariance matrix for the vector of observables under the true process, $\Phi_{\tau} \equiv B'(A - KB')^{\tau-1}K$, and the constant contains terms that do not depend on

[48] Note that I allow for the possibility that P^{θ} is supported on some proper subspace $\mathcal{W}$ of $\mathbb{R}^n$, in which case $B'\hat{\Sigma}_z B + R$ might not be invertible. The Moore–Penrose pseudo-inverse is then the appropriate generalization of matrix inverse in the expression for the Kalman gain. See Chapter 4 of Anderson and Moore (2005) for a treatment in the non-singular case and Silverman (1976) for the case where $B'\hat{\Sigma}_z B + R$ may be singular.

[49] The assumptions that the d-state model θ is minimal and Q is positive definite imply that the Riccati equation has a unique positive semidefinite solution and that $A - KB'$ is a convergent matrix.

θ. Matrix $\hat{\Sigma}_y^\dagger$ denotes the Moore–Penrose pseudo-inverse of $\hat{\Sigma}_y$ and $\det^*(\hat{\Sigma}_y^\dagger)$ denotes its pseudo-determinant.[50] These objects are the appropriate counterparts of the matrix inverse and the determinant for the case where $\mathcal{W}$ does not equal $\mathbb{R}^n$, and so, the subjective model θ is degenerate.

Proof of Theorem 1. Let $\tilde{n}$ denote the dimension of vector $\tilde{y}_t = Ty_t$, let $\widetilde{\mathcal{W}}$ denote the linear subspace of $\mathbb{R}^{\tilde{n}}$ defined as $\widetilde{\mathcal{W}} \equiv \{\tilde{y} \in \mathbb{R}^{\tilde{n}} : \tilde{y} = Ty \text{ for some } y \in \mathcal{W}\}$, let $\widetilde{\Theta}_d$ denote the set of d-state models when the vector of observable is $\tilde{y}_t \in \mathbb{R}^{\tilde{n}}$, and let $\widetilde{\mathrm{KLDR}}(\tilde{\theta})$ denote the KLDR of model $\tilde{\theta} \in \widetilde{\Theta}_d$ from the true process $\widetilde{\mathbb{P}} \equiv T(\mathbb{P})$.

Let $\theta \in \Theta_d$ denote an arbitrary pseudo-true d-state model when the true process is $\mathbb{P}$ and $\tilde{\theta} \in \widetilde{\Theta}_d$ denote an arbitrary pseudo-true d-state model when the true process is $\widetilde{\mathbb{P}}$. I first show that $T(P^\theta)$ and $P^{\tilde{\theta}}$ are both supported on $\widetilde{\mathcal{W}}$. Note that there always exists a d-state model for which the KLDR is finite—one such model is the one according to which y_t is i.i.d. over time and has a variance-covariance matrix that coincides with the true variance-covariance matrix Γ_0. Therefore, for any pseudo-true d-state model, the KLDR is finite. Thus, P^θ is supported on $\mathcal{W}$, and so, $T(P^\theta)$ is supported on $\widetilde{\mathcal{W}}$. On the other hand, since the true distribution $\mathbb{P}$ is supported on $\mathcal{W}$, the transformed distribution $\widetilde{\mathbb{P}}$ is supported on $\widetilde{\mathcal{W}}$. Consequently, by the above argument, $P^{\tilde{\theta}}$ is also supported on $\widetilde{\mathcal{W}}$. Therefore, I can restrict my attention to models $\theta \in \Theta_d$ such that P^θ is supported on $\mathcal{W}$ and models $\tilde{\theta} \in \widetilde{\Theta}_d$ such that $P^{\tilde{\theta}}$ is supported on $\widetilde{\mathcal{W}}$.

For any model $\theta = (A, B, Q, R) \in \Theta_d$, define model $T(\theta) \in \widetilde{\Theta}_d$ as $T(\theta) \equiv (A, BT', Q, TRT')$. I next show that $\widetilde{\mathrm{KLDR}}(T(\theta)) = \mathrm{KLDR}(\theta)$, up to an additive constant that does not depend on θ. Fix some model $\theta \in \Theta_d$. Let $\hat{\Sigma}_z \equiv \mathrm{Var}_t^\theta(z_{t+1})$ denote the subjective conditional variance of the subjective state under model θ, and let $\widetilde{\hat{\Sigma}}_z \equiv \mathrm{Var}_t^{T(\theta)}(z_{t+1})$ denote the corresponding conditional variance under model $T(\theta)$. Matrices $\hat{\Sigma}_z$ and $\widetilde{\hat{\Sigma}}_z$ solve the following Riccati equations:

$$\hat{\Sigma}_z = A\left(\hat{\Sigma}_z - \hat{\Sigma}_z B\left(B'\hat{\Sigma}_z B + R\right)^\dagger B'\hat{\Sigma}_z\right)A' + Q, \tag{G.7}$$

$$\widetilde{\hat{\Sigma}}_z = A\left(\widetilde{\hat{\Sigma}}_z - \widetilde{\hat{\Sigma}}_z BT'\left(TB'\widetilde{\hat{\Sigma}}_z BT' + TRT'\right)^\dagger TB'\widetilde{\hat{\Sigma}}_z\right)A' + Q. \tag{G.8}$$

Since matrix T has full rank, $T^\dagger = (T'T)^{-1}T$ and $T^\dagger T = I$. Therefore, $\widetilde{\hat{\Sigma}}_z = \hat{\Sigma}_z$. Next, let K denote the Kalman gain given model θ, and let denote $\tilde{K}$ denote the Kalman gain given model $T(\theta)$. Note that

$$\tilde{K} = A\widetilde{\hat{\Sigma}}_z BT'\left(TB'\widetilde{\hat{\Sigma}}_z BT' + TRT'\right)^\dagger = KT^\dagger.$$

Let $\Phi_\tau \equiv B'(A - KB')^{\tau-1}K$, and let $\widetilde{\Phi}_\tau$ denote the corresponding matrix given model $T(\theta)$. Note that

$$\widetilde{\Phi}_\tau \equiv TB'(A - KT^\dagger TB')^{\tau-1}KT^\dagger = T\Phi_\tau T^\dagger.$$

Finally, let $\hat{\Sigma}_y \equiv \mathrm{Var}_t^\theta(y_{t+1})$ denote the subjective conditional variance of y_{t+1} given model θ, and let $\widetilde{\hat{\Sigma}}_y \equiv \mathrm{Var}_t^{T(\theta)}(\tilde{y}_{t+1})$ denote the corresponding conditional variance given model $T(\theta)$. Note that

$$\widetilde{\hat{\Sigma}}_y = TB'\widetilde{\hat{\Sigma}}_z BT' + TRT' = T\hat{\Sigma}_y T'.$$

[50] The pseudo-determinant is the product of all non-zero eigenvalues of a square matrix.

One the other hand, $\widetilde{\Gamma}_l \equiv \widetilde{\mathbb{E}}[\tilde{y}_t \tilde{y}'_{t-l}] = T\mathbb{E}[y_t y_{t-l}]T' = T\Gamma_l T'$. Therefore, by equation (G.6),

$$\begin{aligned}\widetilde{\text{KLDR}}(T(\theta)) = & -\frac{1}{2}\log\det{}^*\left(T^{\dagger\prime}\hat{\Sigma}_y^\dagger T^\dagger\right) + \frac{n}{2}\log(2\pi) + \frac{1}{2}\text{tr}\left(T^{\dagger\prime}\hat{\Sigma}_y^\dagger T^\dagger T\Gamma_0 T'\right) \\ & - \frac{1}{2}\sum_{\tau=1}^{\infty}\text{tr}\left(T^{\dagger\prime}\hat{\Sigma}_y^\dagger T^\dagger T\Phi_\tau T^\dagger T\Gamma'_\tau T'\right) - \frac{1}{2}\sum_{\tau=1}^{\infty}\text{tr}\left(T^{\dagger\prime}\hat{\Sigma}_y^\dagger T^\dagger T\Gamma_\tau T' T^{\dagger\prime}\Phi'_\tau T'\right) \\ & + \frac{1}{2}\sum_{s=1}^{\infty}\sum_{\tau=1}^{\infty}\text{tr}\left(T^{\dagger\prime}\hat{\Sigma}_y^\dagger T^\dagger T\Phi_s T^\dagger T\Gamma_{\tau-s} T' T^{\dagger\prime}\Phi'_\tau T'\right) + \text{constant.}\end{aligned}$$

The fact that $T^\dagger T = I$ implies that the above expression is equal to KLDR(θ), up to an additive constant that does not depend on θ.

Likewise, for any model $\tilde{\theta} = (\tilde{A}, \tilde{B}, \tilde{Q}, \tilde{R}) \in \widetilde{\Theta}_d$, define $T^{-1}(\tilde{\theta}) \equiv (\tilde{A}, \tilde{B}T^{\dagger\prime}, \tilde{Q}, T^\dagger\tilde{R}T^{\dagger\prime}) \in \Theta_d$. By an argument similar to the one in the previous paragraph, KLDR$(T^{-1}(\tilde{\theta})) = \widetilde{\text{KLDR}}(\tilde{\theta})$, up to an additive constant that does not depend on $\tilde{\theta}$.

Therefore, the mapping T defines an isomorphism between the set of models Θ_d and the set of models $\widetilde{\Theta}_d$: Any model $\theta \in \Theta_d$ can be identified with a model $T(\theta) \in \widetilde{\Theta}_d$ such that the KLDR of P^θ from the process $\mathbb{P}$ is equal to the KLDR of $P^{T(\theta)}$ from $T(\mathbb{P})$, and any model $\tilde{\theta} \in \widetilde{\Theta}_d$ can be identified with a model $T^{-1}(\tilde{\theta}) \in \Theta_d$ such that the KLDR of $P^{T^{-1}(\tilde{\theta})}$ from the process $\mathbb{P}$ is equal to the KLDR of $P^{\tilde{\theta}}$ from the process $T(\mathbb{P})$. This conclusion immediately implies that the set of pseudo-true d-state models under true process $\mathbb{P}$ is identified with the set of pseudo-true d-state models under true process $T(\mathbb{P})$.

It only remains to show that $P^{T(\theta)} = T(P^\theta)$ for any model $\theta \in \Theta_d$. Since $P^{T(\theta)}$ and $T(P^\theta)$ are both zero mean, stationary, and normal distributions over $\{\tilde{y}_t\}_{t=-\infty}^{\infty}$, it is sufficient to show that the autocovariance matrices of $\tilde{y}_t$ are identical at all lags under the two distributions. But this follows the definitions of distributions $P^{T(\theta)}$ and $T(P^\theta)$. □

Proof of Theorem 2

Before establishing the theorem, I state and prove a lemma that underpins all the characterization results of the paper:

Lemma G.1. *Model $\theta = (A, B, Q, R)$ is a pseudo-true d-state model given true autocovariance matrices $\{\Gamma_l\}_l$ with Γ_0 invertible if and only if $A = M$, $B = D'N^{-1}$, $Q = I - M(I - D'D)M'$, and $R = N^{-1\prime}(I - DD')N^{-1}$, where (M, D, N) is a tuple that minimizes*

$$\begin{aligned}KLDR(\tilde{M}, \tilde{D}, \tilde{N}) \equiv & -\frac{1}{2}\log det\left(\tilde{N}\tilde{N}'\right) + \frac{1}{2}\text{tr}\left(\tilde{N}'\Gamma_0\tilde{N}\right) - \sum_{\tau=1}^{\infty}\text{tr}\left(\left(\tilde{M}\left(I - \tilde{D}'\tilde{D}\right)\right)^{\tau-1}\tilde{M}\tilde{D}'\tilde{N}'\Gamma'_\tau\tilde{N}\tilde{D}\right) \\ & + \frac{1}{2}\sum_{s=1}^{\infty}\sum_{\tau=1}^{\infty}\text{tr}\left(\tilde{D}\left(\tilde{M}\left(I - \tilde{D}'\tilde{D}\right)\right)^{s-1}\tilde{M}\tilde{D}'\tilde{N}'\Gamma_{\tau-s}\tilde{N}\tilde{D}\tilde{M}'\left(\left(I - \tilde{D}'\tilde{D}\right)\tilde{M}'\right)^{\tau-1}\tilde{D}'\right)\end{aligned} \tag{G.9}$$

subject to the constraints that $\tilde{M}$ is a $d \times d$ convergent matrix, $\tilde{D}$ is an $n \times d$ diagonal matrix with elements in the $[0, 1]$ interval, $\tilde{N}$ is an $n \times n$ invertible matrix, and $\|\tilde{M}(I - \tilde{D}\tilde{D}')\tilde{M}'\|_2 < 1$.

Proof. The assumption that Q is positive definite implies that the solution $\hat{\Sigma}_z$ to the Riccati equation (G.3) is invertible. On the other hand, since Γ_0 is invertible, I can restrict attention to subjective models for which $\hat{\Sigma}_y$ is non-singular.[51] The pseudo-inverses and pseudo-determinants in equations (G.3) and (G.6) thus reduce to matrix inverses and determinants.

I start by expressing $\hat{\Sigma}_y^{\frac{-1}{2}} B' \hat{\Sigma}_z^{\frac{1}{2}}$ as its singular value decomposition:

$$\hat{\Sigma}_y^{\frac{-1}{2}} B' \hat{\Sigma}_z^{\frac{1}{2}} = UDV', \tag{G.10}$$

where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal matrices, and $D \in \mathbb{R}^{n\times d}$ is a rectangular diagonal matrix with singular values of $\hat{\Sigma}_z^{\frac{1}{2}} B \hat{\Sigma}_y^{\frac{-1}{2}}$ on the diagonal. Note that

$$VD'DV' = \hat{\Sigma}_z^{\frac{1}{2}} B \left(B'\hat{\Sigma}_z B + R\right)^{-1} B' \hat{\Sigma}_z^{\frac{1}{2}}. \tag{G.11}$$

Since R is a symmetric positive semidefinite matrix and V is orthogonal, diagonal elements of D are weakly smaller than 1 (strictly so if R is positive definite). Next, define $M \equiv V^{-1}\hat{\Sigma}_z^{\frac{-1}{2}} A \hat{\Sigma}_z^{\frac{1}{2}} V$. Then,

$$A = \hat{\Sigma}_z^{\frac{1}{2}} VMV^{-1} \hat{\Sigma}_z^{\frac{-1}{2}}, \tag{G.12}$$

$$B = \hat{\Sigma}_z^{\frac{-1}{2}} VD'U' \hat{\Sigma}_y^{\frac{1}{2}}, \tag{G.13}$$

$$K = \hat{\Sigma}_z^{\frac{1}{2}} VMD'U' \hat{\Sigma}_y^{\frac{-1}{2}}, \tag{G.14}$$

and so

$$\begin{aligned} KB' &= \hat{\Sigma}_z^{\frac{1}{2}} VMD'DV' \hat{\Sigma}_z^{\frac{-1}{2}}, \\ \Phi_\tau &= \hat{\Sigma}_y^{\frac{1}{2}} UD \left(M\left(I - D'D\right)\right)^{\tau-1} MD'U' \hat{\Sigma}_y^{\frac{-1}{2}}. \end{aligned}$$

Note that since A is a convergent matrix, so is M. Substituting in (G.3) for A from equation (G.12) and for B from (G.13), I get

$$\begin{aligned} Q &= \hat{\Sigma}_z - A\left(\hat{\Sigma}_z - \hat{\Sigma}_z B \left(B'\hat{\Sigma}_z B + R\right)^{-1} B'\hat{\Sigma}_z\right) A' \\ &= \hat{\Sigma}_z - \hat{\Sigma}_z^{\frac{1}{2}} VMV^{-1} \hat{\Sigma}_z^{\frac{-1}{2}} \left(\hat{\Sigma}_z - \hat{\Sigma}_z^{\frac{1}{2}} VD'DV \hat{\Sigma}_z^{\frac{1}{2}}\right) \hat{\Sigma}_z^{\frac{-1}{2}} VM'V^{-1} \hat{\Sigma}_z^{\frac{1}{2}} \\ &= \hat{\Sigma}_z - \hat{\Sigma}_z^{\frac{1}{2}} VM\left(I - D'D\right) M'V^{-1} \hat{\Sigma}_z^{\frac{1}{2}}. \end{aligned} \tag{G.15}$$

Therefore, since Q is positive definite, the eigenvalues of $VM\left(I - D'D\right)M'V^{-1}$ must all lie inside the unit circle. This implies that $\rho(M\left(I - D'D\right)M') = \|M\left(I - D'D\right)M'\|_2 < 1$, where $\rho(\cdot)$ denotes the spectral radius, and I am using the facts that the spectral radius is invariant to similarity transformations and equal to the spectral norm for symmetric matrices.

I can further reduce the number of parameters in the agent's model by transforming $\hat{\Sigma}_y^{\frac{-1}{2}}$ using the orthogonal matrix U. Define $N \equiv \hat{\Sigma}_y^{\frac{-1}{2}} U$. Since $\hat{\Sigma}_y^{\frac{-1}{2}}$ and U are invertible matrices, so is N.

[51] Since the variance-covariance matrix Γ_0 of the true process is invertible, $\mathrm{KLDR}(\theta) = +\infty$ for any subjective model θ with a singular $\hat{\Sigma}_y$. Note that, in light of Theorem 1, the restriction to true processes with invertible variance-covariance matrices is without loss of generality.

Because $\hat{\Sigma}_y^{\frac{-1}{2}}$ is symmetric, $UN' = NU' = \hat{\Sigma}_y^{\frac{-1}{2}}$, so $\hat{\Sigma}_y^{-1} = NU'UN' = NN'$, and

$$\operatorname{tr}\left(\hat{\Sigma}_y^{-1}\Gamma_0\right) = \operatorname{tr}\left(\hat{\Sigma}_y^{\frac{-1}{2}}\Gamma_0\hat{\Sigma}_y^{\frac{-1}{2}}\right) = \operatorname{tr}\left(UN'\Gamma_0 NU'\right) = \operatorname{tr}\left(N'\Gamma_0 N\right).$$

On the other hand,

$$\operatorname{tr}\left(\hat{\Sigma}_y^{-1}\Phi_\tau\Gamma_\tau'\right) = \operatorname{tr}\left(\hat{\Sigma}_y^{\frac{-1}{2}} UD\left(M\left(I-D'D\right)\right)^{\tau-1} MD'U'\hat{\Sigma}_y^{\frac{-1}{2}}\Gamma_\tau'\right) = \operatorname{tr}\left(\left(M\left(I-D'D\right)\right)^{\tau-1} MD'N'\Gamma_\tau' ND\right),$$

and

$$\operatorname{tr}\left(\hat{\Sigma}_y^{-1}\Phi_s\Gamma_{\tau-s}\Phi_\tau'\right) = \operatorname{tr}\left(D\left(M\left(I-D'D\right)\right)^{s-1} MD'N'\Gamma_{\tau-s}NDM'\left(\left(I-D'D\right)M'\right)^{\tau-1}D'\right).$$

Therefore, the KLDR can be expressed in terms of matrices M, D, and N as

$$\mathrm{KLDR}(\theta) = \mathrm{KLDR}(M, D, N) + \text{constant},$$

where $\mathrm{KLDR}(M, D, N)$ is as in the statement of the lemma.

It only remains to show that, for any $(\hat{M}, \hat{D}, \hat{N})$ such that $\hat{M}$ is a $d \times d$ convergent matrix, $\hat{D}$ is an $n \times d$ diagonal matrix with elements in the $[0, 1]$ interval, $\hat{N}$ is an $n \times n$ invertible matrix, and $\|\hat{M}(I - \hat{D}\hat{D}')\hat{M}'\|_2 < 1$, one can construct a corresponding (A, B, Q, R) such that A is convergent, Q is positive definite, and R is positive semidefinite. Given such a tuple $(\hat{M}, \hat{D}, \hat{N})$, let $A = \hat{M}$, $B = \hat{D}'\hat{N}^{-1}$, $Q = I - \hat{M}\left(I - \hat{D}'\hat{D}\right)\hat{M}'$, and $R = \hat{N}^{-1'}\left(I - \hat{D}\hat{D}'\right)\hat{N}^{-1}$. Since $\hat{M}$ is convergent, so is A. Since $\|\hat{M}(I - \hat{D}\hat{D}')\hat{M}'\|_2 < 1$, matrix Q is positive definite. And since $\hat{D}$ is a diagonal matrix with elements in the $[0, 1]$ interval, R is positive semidefinite. It is easy to verify that then $\hat{\Sigma}_z = I$ is then the solution to the Riccati equation (G.3), and so, $\hat{\Sigma}_y = (\hat{N}\hat{N}')^{-1}$. Therefore, I can choose $U = (\hat{N}\hat{N}')^{\frac{-1}{2}}\hat{N}$, $D = \hat{D}$, and $V = I$ in equation (G.10). Substituting in the expressions for M and N, I get $M = \hat{M}$ and $N = \hat{N}$. This completes the proof of the lemma.

For future reference, I also compute several other objects in terms of the M, D, and N matrices. The matrix of Kalman gain is given by

$$K = MD'N'. \tag{G.16}$$

The subjective forecasts can then be found by substituting for A, B, and K in (G.5):

$$E_t^\theta\left[y_{t+s}\right] = N'^{-1}DM^{s-1}\sum_{\tau=0}^{\infty}\left(M\left(I-D'D\right)\right)^\tau MD'N'y_{t-\tau}. \tag{G.17}$$

The subjective variance of y_{t+1} conditional on the information available to the agent at time t is given by $\hat{\Sigma}_y = (NN')^{-1}$. The unconditional subjective variance of y is given by

$$\mathrm{Var}^\theta(y) = B'\mathrm{Var}^\theta(z)B + R,$$

where $\mathrm{Var}^\theta(z)$ solves the discrete Lyapunov equation

$$\mathrm{Var}^\theta(z) = A\mathrm{Var}^\theta(z)A' + Q.$$

Solving the above equation forward, I get

$$\mathrm{Var}^{\theta}(z) = I + \sum_{\tau=1}^{\infty} M^{\tau} D' D M'^{\tau}.$$

Therefore,

$$\mathrm{Var}^{\theta}(y) = B' \sum_{\tau=0}^{\infty} A^{\tau} Q A'^{\tau} B + R = N^{-1\prime} \left(I + \sum_{\tau=1}^{\infty} D M^{\tau} D' D M'^{\tau} D' \right) N^{-1}. \tag{G.18}$$

□

I can now establish Theorem 2.

Proof of Theorem 2. Let M, D, and N be as in Lemma G.1. When $d = 1$, then $M = a$ for some $a \in [-1, 1]$ and $D = d_1 e_1$ for some $d_1 \in [0, 1]$, where e_1 denotes the first coordinate vector. Define $\eta \equiv 1 - d_1^2$ and $S \equiv \Gamma_0^{\frac{1}{2}} N$. Then KLDR, defined in (G.9), can be written (with slight abuse of notation) as a function of a, η, and S:

$$\mathrm{KLDR}(a, \eta, S) = -\frac{1}{2} \log\det(SS') + \frac{1}{2} \mathrm{tr}(S'S) - \frac{1}{2} e_1' S' \Omega(a, \eta) S e_1 + \text{constant},$$

where

$$\Omega(a, \eta) \equiv a(1-\eta) \sum_{\tau=1}^{\infty} (a\eta)^{\tau-1} \Gamma_0^{\frac{-1}{2}} (\Gamma_\tau + \Gamma_\tau') \Gamma_0^{\frac{-1}{2}} - a^2 (1-\eta)^2 \sum_{s=1}^{\infty} \sum_{\tau=1}^{\infty} (a\eta)^{s+\tau-2} \Gamma_0^{\frac{-1}{2}} \Gamma_{\tau-s} \Gamma_0^{\frac{-1}{2}}.$$

I can simplify the second term of $\Omega(a, \eta)$ further:

$$\begin{aligned}
\sum_{s=1}^{\infty} \sum_{\tau=1}^{\infty} (a\eta)^{s+\tau-2} \Gamma_0^{\frac{-1}{2}} \Gamma_{\tau-s} \Gamma_0^{\frac{-1}{2}} &= \sum_{s=1}^{\infty} \sum_{\tau=s+1}^{\infty} (a\eta)^{s+\tau-2} \Gamma_0^{\frac{-1}{2}} \left(\Gamma_{\tau-s} + \Gamma_{\tau-s}' \right) \Gamma_0^{\frac{-1}{2}} + \sum_{s=1}^{\infty} (a\eta)^{2(s-1)} I \\
&= \sum_{s=1}^{\infty} \sum_{\tau=1}^{\infty} (a\eta)^{2(s-1)+\tau} \Gamma_0^{\frac{-1}{2}} \left(\Gamma_\tau + \Gamma_\tau' \right) \Gamma_0^{\frac{-1}{2}} + \sum_{s=1}^{\infty} (a\eta)^{2(s-1)} I \\
&= \left(\sum_{s=1}^{\infty} (a\eta)^{2(s-1)} \right) \left(I + \sum_{\tau=1}^{\infty} (a\eta)^{\tau} \Gamma_0^{\frac{-1}{2}} (\Gamma_\tau + \Gamma_\tau') \Gamma_0^{\frac{-1}{2}} \right) \\
&= \frac{1}{1 - a^2\eta^2} \left(I + a\eta \sum_{\tau=1}^{\infty} (a\eta)^{\tau-1} \Gamma_0^{\frac{-1}{2}} (\Gamma_\tau + \Gamma_\tau') \Gamma_0^{\frac{-1}{2}} \right).
\end{aligned}$$

Therefore,

$$\Omega(a, \eta) = -\frac{a^2(1-\eta)^2}{1 - a^2\eta^2} I + \frac{(1-\eta)(1-a^2\eta)}{1 - a^2\eta^2} \sum_{\tau=1}^{\infty} a^{\tau} \eta^{\tau-1} \Gamma_0^{\frac{-1}{2}} (\Gamma_\tau + \Gamma_\tau') \Gamma_0^{\frac{-1}{2}}. \tag{G.19}$$

By Lemma G.1, minimizing the KLDR with respect to A, B, Q, and R is equivalent to minimizing KLDR(M, D, N) with respect to M, D, and N. But for any a, η, and S, one can construct a corresponding M, D, and N, and vice versa. Therefore, I can instead minimize KLDR(a, η, S) with respect to a, η, and S.

I first minimize $\mathrm{KLDR}(a, \eta, S)$ with respect to S taking a and η as given. The first-order optimality condition with respect to S is given by $S^{-1} = S' - e_1 e_1' S' \Omega(a, \eta)$, which implies that

$$S'S - e_1 e_1' S' \Omega(a, \eta) S = I. \tag{G.20}$$

Therefore, for any solution to the problem of minimizing $\mathrm{KLDR}(a, \eta, S)$,

$$n = \mathrm{tr}(I) = \mathrm{tr}\left(S'S\right) - \mathrm{tr}\left(e_1 e_1' S' \Omega(a, \eta) S\right) = \mathrm{tr}\left(S'S\right) - e_1' S' \Omega(a, \eta) S e_1.$$

Thus, minimizing $\mathrm{KLDR}(a, \eta, S)$ with respect to a, η, and S is equivalent to solving the following program:

$$\max_{a,\eta} \quad \det\left(S(a, \eta) S'(a, \eta)\right),$$

where

$$S(a, \eta) \in \arg\min_{S} \; -\frac{1}{2} \log\det\left(SS'\right) + \frac{1}{2}\mathrm{tr}\left(S'S\right) - \frac{1}{2} e_1' S' \Omega(a, \eta) S e_1. \tag{G.21}$$

I proceed by first characterizing $S(a, \eta)$. Note that the necessary first-order optimality conditions for problem (G.21) are given by matrix equation (G.20).

Claim G.1. *For any matrix S that solves equation* (G.20)*, the necessary first-order optimality condition for problem* (G.21)*,*

(i) $Se_1 = \dfrac{1}{\sqrt{1-\lambda}} u,$

(ii) $S'^{-1} e_1 = \sqrt{1-\lambda}\, u,$

(iii) $SS' = I + \frac{\lambda}{1-\lambda} uu',$

where λ is an eigenvalue of the real symmetric matrix $\Omega(a, \eta)$ and u is a corresponding eigenvector normalized such that $u'u = 1$.

I return to proving the claim toward the end of the proof. Equation (G.20) in general has multiple solutions, with each solution corresponding to a local extremum of problem (G.21). The global optimum of problem (G.21) is given by the solution to equation (G.20) that results in the largest value for $\det(SS')$. But by part (iii) of Claim G.1, $\det(SS') = (1-\lambda)^{-1}$. Thus, for any pseudo-true one-state model, a and η maximize $\lambda_{\max}(\Omega(a, \eta))$ and S satisfies parts (i)–(iii) of Claim G.1, with $\lambda = \lambda_{\max}(\Omega)$ and $u = u_{\max}(\Omega)$ the corresponding eigenvector.

I next find parameters A, B, Q, and R representing the a, η, and S that minimize $\mathrm{KLDR}(a, \eta, S)$. First, note that $M = a$, $D = \sqrt{1-\eta}\, e_1$, and $N = \Gamma_0^{\frac{-1}{2}} S$. The representation in Lemma G.1 is thus given by $A = a$, $B = \sqrt{1-\eta}\, e_1' S^{-1} \Gamma_0^{\frac{1}{2}}$, $Q = 1 - a^2 \eta$, and $R = \Gamma_0^{\frac{1}{2}} S^{-1\prime} \left(I - (1-\eta) e_1 e_1'\right) S^{-1} \Gamma_0^{\frac{1}{2}}$.[52] By Claim G.1 and the argument above,

$$e_1' S^{-1} = \sqrt{1 - \lambda_{\max}(\Omega)}\, u_{\max}'(\Omega),$$

[52] For this (A, B, Q, R) tuple to represent a one-state model, I need A to be convergent, Q to be positive definite, and R to be positive semidefinite. That R is always positive semidefinite is immediate. Showing that A is convergent and Q is positive definite takes more work. I do so in Lemma G.3.

$$S^{-1\prime}S^{-1} = (SS')^{-1} = I - \lambda_{\max}(\Omega)u_{\max}(\Omega)u'_{\max}(\Omega).$$

Thus,

$$B = \sqrt{(1-\eta)\,(1-\lambda_{\max}(\Omega))}\,u'_{\max}(\Omega)\Gamma_0^{\frac{1}{2}},$$

and

$$\begin{aligned} R &= \Gamma_0^{\frac{1}{2}}\left(I - \lambda_{\max}(\Omega)u_{\max}(\Omega)u'_{\max}(\Omega)\right)\Gamma_0^{\frac{1}{2}} - (1-\eta)\,(1-\lambda_{\max}(\Omega))\,\Gamma_0^{\frac{1}{2}}u_{\max}(\Omega)u'_{\max}(\Omega)\Gamma_0^{\frac{1}{2}} \\ &= \Gamma_0^{\frac{1}{2}}\left[I - (1-\eta+\eta\lambda_{\max}(\Omega))\,u_{\max}(\Omega)u'_{\max}(\Omega)\right]\Gamma_0^{\frac{1}{2}}. \end{aligned}$$

Finally, note that $M = a$, $D = \sqrt{1-\eta}e_1$, and $N = \Gamma_0^{\frac{-1}{2}}S$. Therefore, by equation (G.17), the subjective forecasts are given by

$$E_t^{\theta}[y_{t+s}] = a^s(1-\eta)\Gamma_0^{\frac{1}{2}}S'^{-1}e_1e_1'S'\Gamma_0^{\frac{-1}{2}}\sum_{\tau=0}^{\infty}a^{\tau}\eta^{\tau}y_{t-\tau}. \tag{G.22}$$

Using Claim G.1 to substitute for the optimal S, I get

$$E_t^{\theta}[y_{t+s}] = a^s(1-\eta)\Gamma_0^{\frac{1}{2}}u_{\max}(\Omega)u'_{\max}(\Omega)\Gamma_0^{\frac{-1}{2}}\sum_{\tau=0}^{\infty}a^{\tau}\eta^{\tau}y_{t-\tau},$$

where $u_{\max}(\Omega)$ is a unit-norm eigenvector of Ω with eigenvalue $\lambda_{\max}(\Omega)$. The theorem then follows by the definitions of p and q. □

Proof of Claim G.1. The first-order optimality condition with respect to S is given by

$$S'S - e_1e_1'S'\Omega S = I. \tag{G.23}$$

Multiplying the transpose of the above equation from right by e_1 and from left by S'^{-1}, I get

$$Se_1 - \Omega Se_1 = S'^{-1}e_1. \tag{G.24}$$

On the other hand, multiplying equation (G.23) from left by S and from right by S^{-1}, I get

$$SS' = I + Se_1e_1'S'\Omega. \tag{G.25}$$

By the Sherman–Morrison formula,

$$S'^{-1}S^{-1} = I - \frac{Se_1e_1'S'\Omega}{1+e_1'S'\Omega Se_1}.$$

Multiplying the above equation from right by Se_1, I get

$$S'^{-1}e_1 = \frac{1}{1+e_1'S'\Omega Se_1}Se_1. \tag{G.26}$$

Substituting for $S'^{-1}e_1$ from the above equation in (G.24) and rearranging the terms, I get

$$\Omega Se_1 = \frac{e_1'S'\Omega Se_1}{1+e_1'S'\Omega Se_1}Se_1. \tag{G.27}$$

That is, Se_1 is an eigenvector of Ω. Let λ denote the corresponding eigenvalue and let $u = Se_1/\sqrt{e_1'S'Se_1}$. Then equation (G.27) implies

$$\lambda = \frac{\lambda e_1'S'Se_1}{1+\lambda e_1'S'Se_1}.$$

I separately consider the cases $\lambda \neq 0$ and $\lambda = 0$. If $\lambda \neq 0$, then $e_1'S'Se_1 = 1/(1-\lambda)$ and $Se_1 = u/\sqrt{1-\lambda}$. Equation (G.26) then implies that $S'^{-1}e_1 = \sqrt{1-\lambda}u$, and equation (G.25) implies that

$$SS' = I + \frac{\lambda}{1-\lambda}uu'.$$

If $\lambda = 0$, then equation (G.24) implies that $Se_1 = S'^{-1}e_1$, and so, Se_1 and $S'^{-1}e_1$ are both multiples of u. Furthermore, $e_1'S^{-1}Se_1 = e_1'e_1 = 1$. Therefore, $Se_1 = S'^{-1}e_1 = u$. On the other hand, equation (G.25) implies that $SS' = I$. This completes the proof of the claim. □

Proof of Theorem 3

I first prove two useful lemmas:

Lemma G.2. *For any purely non-deterministic, stationary ergodic, and non-degenerate process with autocorrelation matrices $\{C_l\}_l$, the spectral radii of autocorrelation matrices satisfy $\rho(C_l) \leq 1$ for any l with the inequality strict for $l = 1$.*

Proof. Let λ_l denote an eigenvalue of C_l largest in magnitude and let u_l denote the corresponding eigenvector normalized such that $u_l'u_l = 1$. Define the process $\omega_t^{(l)} \equiv u_l'\Gamma_0^{\frac{-1}{2}}y_t \in \mathbb{R}$. Since y_t is a purely non-deterministic, stationary ergodic, and non-degenerate process, so is $\omega_t^{(l)}$ for any l. I first show that λ_l is the autocorrelation of process $\omega_t^{(l)}$ at lag l. Note that

$$\mathbb{E}[\omega_t^{(l)}\omega_{t-l}^{(l)}] = u_l'\Gamma_0^{\frac{-1}{2}}\mathbb{E}[y_ty_{t-l}']\Gamma_0^{\frac{-1}{2}}u_l = u_l'\Gamma_0^{\frac{-1}{2}}\Gamma_l\Gamma_0^{\frac{-1}{2}}u_l = u_l'\Gamma_0^{\frac{-1}{2}}\left(\frac{\Gamma_l+\Gamma_l'}{2}\right)\Gamma_0^{\frac{-1}{2}}u_l = u_l'C_lu_l = \lambda_l.$$

Furthermore,

$$\mathbb{E}[\omega_t^{(l)}\omega_t^{(l)}] = u_l'\Gamma_0^{\frac{-1}{2}}\mathbb{E}[y_ty_t']\Gamma_0^{\frac{-1}{2}}u_l = u_l'\Gamma_0^{\frac{-1}{2}}\Gamma_0\Gamma_0^{\frac{-1}{2}}u_l = u_l'u_l = 1.$$

Therefore, since $\omega_t^{(l)}$ is purely non-deterministic, stationary ergodic, and non-degenerate,

$$\rho(C_l) = |\lambda_l| = \frac{\mathbb{E}[\omega_t^{(l)}\omega_{t-l}^{(l)}]}{\mathbb{E}[\omega_t^{(l)}\omega_t^{(l)}]} \leq 1.$$

Next, toward a contradiction suppose that $\rho(C_1) = 1$. Then $\omega_t^{(1)}$ is perfectly correlated with $\omega_{t-1}^{(1)}$, and so, with $\omega_{t-l}^{(1)}$ for every l, contradicting the assumption that $\omega_t^{(1)}$ is purely non-deterministic, stationary ergodic, and non-degenerate. □

Lemma G.3. *If $\mathbb{P}$ is purely non-deterministic and stationary ergodic, then so is P^θ for any pseudo-true one-state model θ.*

Proof. Define

$$C(a,\eta) \equiv \sum_{\tau=1}^{\infty} a^{\tau}\eta^{\tau-1}C_{\tau}. \tag{G.28}$$

Then

$$\lambda_{\max}(\Omega(a,\eta)) = -\frac{a^2(1-\eta)^2}{1-a^2\eta^2} + \frac{2(1-\eta)(1-a^2\eta)}{1-a^2\eta^2}\lambda_{\max}(C(a,\eta)), \tag{G.29}$$

where $\lambda_{\max}(C(a,\eta))$ denotes the largest eigenvalue of $C(a,\eta)$. To simplify the exposition, I prove the result under the assumption that the largest eigenvalue of $C(a,\eta)$ is simple at the point (a^*,η^*) that maximizes $\lambda_{\max}(\Omega(a,\eta))$.[53] The partial derivatives of $\lambda_{\max}(\Omega(a,\eta))$ with respect to a and η are given by

$$\begin{aligned}
\frac{\partial\lambda_{\max}(\Omega(a,\eta))}{\partial a} &= \frac{-2a(1-\eta)^2}{\left(1-a^2\eta^2\right)^2} - \frac{4a\eta(1-\eta)^2}{\left(1-a^2\eta^2\right)^2}\lambda_{\max}(C)\\
&\quad + \frac{2(1-\eta)(1-a^2\eta)}{1-a^2\eta^2}u'_{\max}(C)\frac{\partial C}{\partial a}u_{\max}(C), && \text{(G.30)}\\
\frac{\partial\lambda_{\max}(\Omega(a,\eta))}{\partial \eta} &= \frac{2a^2(1-\eta)(1-a^2\eta)}{\left(1-a^2\eta^2\right)^2} - \frac{2\left(1+a^4\eta^2+a^2(1-4\eta+\eta^2)\right)}{\left(1-a^2\eta^2\right)^2}\lambda_{\max}(C)\\
&\quad + \frac{2(1-\eta)(1-a^2\eta)}{1-a^2\eta^2}u'_{\max}(C)\frac{\partial C}{\partial \eta}u_{\max}(C), && \text{(G.31)}
\end{aligned}$$

where $u_{\max}(C)$ denotes the eigenvector of C with eigenvalue $\lambda_{\max}(C)$, normalized such that $u'_{\max}(C)u_{\max}(C) = 1$, and

$$\begin{aligned}
\frac{\partial C}{\partial a} &= \sum_{\tau=1}^{\infty} \tau a^{\tau-1}\eta^{\tau-1}C_{\tau},\\
\frac{\partial C}{\partial \eta} &= \sum_{\tau=1}^{\infty} (\tau-1)a^{\tau}\eta^{\tau-2}C_{\tau}.
\end{aligned}$$

Note that

$$\eta u'_{\max}(C)\frac{\partial C}{\partial \eta}u_{\max}(C) + \lambda_{\max}(C) = a u'_{\max}(C)\frac{\partial C}{\partial a}u_{\max}(C). \tag{G.32}$$

for any a and η.

Let a^* and η^* be scalars in the $[-1,1]$ and $[0,1]$ intervals, respectively, that maximize $\lambda_{\max}(\Omega(a, \eta))$. I separately consider the cases $\eta^* = 1$ and $\eta^* < 1$. If $\eta^* = 1$, then $B = 0$ in the representation in the proof of Theorem 2, the pseudo-true one-state model is i.i.d., and $A = a^*$ can be chosen arbitrarily to satisfy $|a^*| < 1$.[54]

In the rest of the proof, I assume that $\eta^* < 1$ and show that this implies $a^* \neq 1$—by a similar argument $a^* \neq -1$. Toward a contradiction, suppose $a^* = 1$. Setting $a = 1$ in the partial derivatives

[53] The argument can easily be adapted to the case where the largest eigenvalue of $C(a^*,\eta^*)$ is not necessarily simple by replacing the gradient of $\lambda_{\max}(C(a,\eta))$ with its subdifferential and replacing the usual first-order optimality condition with the condition that the zero vector belongs to the subdifferential.

[54] The pseudo-true one-state model then has a zero-state minimal representation.

of $\lambda_{\max}(\Omega(a,\eta))$, I get

$$\left.\frac{\partial \lambda_{\max}(\Omega(a,\eta))}{\partial a}\right|_{a=1} = \frac{2(1-\eta)^2}{\left(1-\eta^2\right)^2}\left[-1-2\eta\lambda_{\max}(C)+(1-\eta^2)u'_{\max}(C)\frac{\partial C}{\partial a}u_{\max}(C)\right],$$
$$\left.\frac{\partial \lambda_{\max}(\Omega(a,\eta))}{\partial \eta}\right|_{a=1} = \frac{2(1-\eta)^2}{\left(1-\eta^2\right)^2}\left[1-2\lambda_{\max}(C)+(1-\eta^2)u'_{\max}(C)\frac{\partial C}{\partial \eta}u_{\max}(C)\right],$$

where $C = C(1,\eta)$ and its partial derivatives are computed at $a = 1$. Multiplying the second equation above by η and subtracting from it the first equation, I get

$$\begin{aligned}
&\eta\left.\frac{\partial \lambda_{\max}(\Omega(a,\eta))}{\partial \eta}\right|_{a=1} - \left.\frac{\partial \lambda_{\max}(\Omega(a,\eta))}{\partial a}\right|_{a=1}\\
&= \frac{2(1-\eta)^2}{\left(1-\eta^2\right)^2}\left[1+\eta+(1-\eta^2)\left(\eta u'_{\max}(C)\frac{\partial C}{\partial \eta}u_{\max}(C)-u'_{\max}(C)\frac{\partial C}{\partial a}u_{\max}(C)\right)\right]\\
&= \frac{2(1-\eta)^2}{\left(1-\eta^2\right)^2}\left[1+\eta-(1-\eta^2)\lambda_{\max}(C)\right],
\end{aligned}$$

where in the second equality I am using identity (G.32). Therefore,

$$\left.\frac{\partial \lambda_{\max}(\Omega(a,\eta))}{\partial a}\right|_{a=1} = \eta\left.\frac{\partial \lambda_{\max}(\Omega(a,\eta))}{\partial \eta}\right|_{a=1} - \frac{2(1-\eta)^2}{\left(1-\eta^2\right)^2}\left(1+\eta-(1-\eta^2)\lambda_{\max}(C(1,\eta))\right).$$

Note that

$$\lambda_{\max}(C(1,\eta)) \le \sum_{\tau=1}^{\infty}\eta^{\tau-1}\lambda_{\max}(C_\tau) < \sum_{\tau=1}^{\infty}\eta^{\tau-1} = \frac{1}{1-\eta},$$

where the second inequality is by Lemma G.2. Therefore,

$$-\frac{2(1-\eta)^2}{\left(1-\eta^2\right)^2}\left(1+\eta-(1-\eta^2)\lambda_{\max}(C(1,\eta))\right) < \frac{2(1-\eta)^2}{\left(1-\eta^2\right)^2}(1+\eta-1-\eta) = 0.$$

On the other hand, by the optimality of $a^* = 1$ and $\eta^* < 1$,

$$\left.\frac{\partial \lambda_{\max}(\Omega(a,\eta))}{\partial \eta}\right|_{a^*=1,\eta=\eta^*} \le 0.$$

Thus,

$$\left.\frac{\partial \lambda_{\max}(\Omega(a,\eta))}{\partial a}\right|_{a^*=1,\eta=\eta^*} < 0,$$

a contradiction to the assumption of optimality of $a^* = 1$ and $\eta^* < 1$. This proves that $a^* < 1$ and establishes that the one-state model with $a = a^*$ and $\eta = \eta^*$ is purely non-deterministic and stationary ergodic. □

I can now prove the theorem.

Proof of Theorem 3. Setting $M = a$, $D = \sqrt{1-\eta}e_1$, and $N = \Gamma_0^{\frac{-1}{2}} S$ in equation (G.18), I get

$$\mathrm{Var}^{\theta}(y) = \Gamma_0^{\frac{1}{2}} \left[I + \frac{1}{1-a^2} \left[a^2(1-\eta)^2 - \left(1 - 2a^2\eta + a^2\eta^2\right)\lambda \right] uu' \right] \Gamma_0^{\frac{1}{2}},$$

where a, η, $\lambda = \lambda_{\max}(\Omega(a,\eta))$, and u are as in Theorem 2. Substituting for $\lambda_{\max}(\Omega(a,\eta))$ from equation (G.29) in the above equation, I get

$$\mathrm{Var}^{\theta}(y_t) = \Gamma_0^{\frac{1}{2}} \left[I + \frac{2(1-\eta)(1-a^2\eta)}{(1-a^2)(1-a^2\eta^2)} \left(a^2(1-\eta) - (1 - 2a^2\eta + a^2\eta^2)\lambda_{\max}(C) \right) uu' \right] \Gamma_0^{\frac{1}{2}}. \tag{G.33}$$

Let a^* and η^* be scalars in the $[-1,1]$ and $[0,1]$ intervals, respectively, that maximize $\lambda_{\max}(\Omega(a,\eta))$. I separately consider the cases $\eta^* = 1$ and $\eta^* < 1$. If $\eta^* = 1$, then the right-hand side of equation (G.33) is equal to Γ_0.

Next suppose $\eta^* < 1$. By the argument in the proof of Lemma G.3, the first-order optimality condition with respect to a must hold with equality at $a = a^*$ and $\eta = \eta^* < 1$. Setting $\partial\lambda_{\max}(\Omega(a, \eta))/\partial a = 0$ in (G.30) and multiplying both sides of the equation by a^*, I get, using (G.32),

$$\begin{aligned} &\frac{2a^{*2}(1-\eta^*)^2}{\left(1-a^{*2}\eta^{*2}\right)^2} + \frac{4a^{*2}\eta^*(1-\eta^*)^2}{\left(1-a^{*2}\eta^{*2}\right)^2}\lambda_{\max}(C) \\ &\quad = \frac{2(1-\eta^*)(1-a^{*2}\eta^*)}{1-a^{*2}\eta^{*2}}\lambda_{\max}(C) + \frac{2(1-\eta^*)(1-a^{*2}\eta^*)}{1-a^{*2}\eta^{*2}}\eta^* u'_{\max}(C)\frac{\partial C}{\partial \eta}u_{\max}(C). \end{aligned} \tag{G.34}$$

Setting $\eta^* = 0$ in the above equation, I get $a^{*2} = \lambda_{\max}(C)$. Setting $a^{*2} = \lambda_{\max}(C)$ in equation (G.33) then establishes that $\mathrm{Var}^{\theta}(y_t) = \Gamma_0$ in the case where $\eta^* = 0$.

Finally, I consider the case where $\eta^* \in (0,1)$. Then additionally the first-order optimality condition with respect to η must hold with equality. Setting $\partial\lambda_{\max}(\Omega(a,\eta))/\partial\eta = 0$ in equation (G.31), multiplying it by η^*, solving for $\eta^* u'_{\max}(C)\frac{\partial C}{\partial \eta}u_{\max}(C)$, and substituting in equation (G.34), I get

$$\begin{aligned} &\frac{2a^{*2}(1-\eta^*)^2}{\left(1-a^{*2}\eta^{*2}\right)^2} + \frac{4a^{*2}\eta^*(1-\eta^*)^2}{\left(1-a^{*2}\eta^{*2}\right)^2}\lambda_{\max}(C) \\ &\quad = \frac{2(1-\eta^*)(1-a^{*2}\eta^*)}{1-a^{*2}\eta^{*2}}\lambda_{\max}(C) - \frac{2a^{*2}\eta^*(1-\eta^*)(1-a^{*2}\eta^*)}{\left(1-a^{*2}\eta^{*2}\right)^2} \\ &\qquad + \frac{2\eta^*\left(1 + a^{*4}\eta^{*2} + a^{*2}(1 - 4\eta^* + \eta^{*2})\right)}{\left(1-a^{*2}\eta^{*2}\right)^2}\lambda_{\max}(C). \end{aligned}$$

Simplifying the above expression leads to

$$a^{*2}(1-\eta^*) = \left(1 - 2a^{*2}\eta^* + a^{*2}\eta^{*2}\right)\lambda_{\max}(C).$$

Combining the above identity with equation (G.33) implies that $\mathrm{Var}^{\theta}(y_t) = \Gamma_0$ and finishes the proof of the theorem. □

Proof of Theorem 4

Let λ denote the eigenvalue of C_1 largest in magnitude.[55] If $\rho(C_1) = 0$, then $\rho(C_\tau) = 0$ for all $\tau \geq 1$. Since C_τ are symmetric matrices, this implies that $C_\tau = 0$ for all $\tau \geq 1$. Therefore,

$$\lambda_{\max}(\Omega(a,\eta)) = -\frac{a^2(1-\eta)^2}{1-a^2\eta^2}.$$

The above expression is maximized by setting $(1-\eta)a = 0$. Therefore, by Theorem 2, for any pseudo-true one-state model, $E_t^\theta[y_{t+s}] = a^s(1-\eta)qp' \sum_{\tau=0}^{\infty} a^\tau \eta^\tau y_{t-\tau} = 0$. On the other hand, if $\rho(C_1) = 0$, then $\lambda = 0$. Therefore, the theorem holds in the case $\rho(C_1) = 0$.

In the rest of the proof, I assume $\rho(C_1) > 0$. Define

$$\begin{aligned} \overline{f}(a,\eta) &\equiv -\frac{a^2(1-\eta)^2}{1-a^2\eta^2} + \frac{2(1-\eta)(1-a^2\eta)}{1-a^2\eta^2} \sum_{\tau=1}^{\infty} |a|^\tau \eta^{\tau-1} \rho(C_1)^\tau \\ &= -\frac{a^2(1-\eta)^2}{1-a^2\eta^2} + \frac{2(1-\eta)(1-a^2\eta)}{1-a^2\eta^2} \frac{|a|\rho(C_1)}{1-\eta|a|\rho(C_1)}, \end{aligned}$$

where in the second equality I am using the fact that $\rho(C_\tau) < 1$, established in Lemma G.2. Function $\overline{f}(a,\eta)$ has two maximizers given by $(\overline{a}^*, \overline{\eta}^*) = (-\rho(C_1), 0)$ and $(\overline{a}^*, \overline{\eta}^*) = (\rho(C_1), 0)$ with the maximum given by $\overline{f}^* = \rho(C_1)^2$. I establish the theorem by showing that $\lambda_{\max}(\Omega(a, \eta)) \leq \overline{f}(a,\eta)$ for all a and η, $\lambda_{\max}(\Omega(\lambda,0)) = \overline{f}(\lambda,0) = \overline{f}^*$, and $\lambda_{\max}(\Omega(-\lambda,0)) \leq \overline{f}(-\lambda,0) = \overline{f}^*$ with the inequality strict if $-\lambda$ is not an eigenvalue of C_1. This establishes that $(a^*,\eta^*) = (\lambda,0)$ is the unique maximizer of $\lambda_{\max}(\Omega(a,\eta))$ if $-\lambda$ is not eigenvalue of C_1 and that $(a^*,\eta^*) = (\lambda,0)$ and $(a^*,\eta^*) = (-\lambda,0)$ are the only maximizers of $\lambda_{\max}(\Omega(a,\eta))$ if λ and $-\lambda$ are both eigenvalues of C_1.

As the first step in doing so, I show that for all a and τ,

$$\lambda_{\max}\left(a^\tau C_\tau\right) \leq |a|^\tau \rho(C_1)^\tau,$$

by considering four disjoint cases: If $a \leq 0$ and $\lambda_{\min}(C_\tau) \leq 0$, then

$$\lambda_{\max}\left(a^\tau C_\tau\right) = a^\tau \lambda_{\min}(C_\tau) = |a|^\tau \left|\lambda_{\min}(C_\tau)\right| \leq |a|^\tau \rho(C_1)^\tau.$$

If $a \leq 0$ and $\lambda_{\min}(C_\tau) > 0$, then

$$\lambda_{\max}\left(a^\tau C_\tau\right) = a^\tau \lambda_{\min}(C_\tau) \leq 0 \leq |a|^\tau \rho(C_1)^\tau.$$

If $a > 0$ and $\lambda_{\max}(C_\tau) \leq 0$, then

$$\lambda_{\max}\left(a^\tau C_\tau\right) = a^\tau \lambda_{\max}(C_\tau) \leq 0 \leq |a|^\tau \rho(C_1)^\tau.$$

Finally, if $a > 0$ and $\lambda_{\max}(C_\tau) > 0$, then

$$\lambda_{\max}\left(a^\tau C_\tau\right) = a^\tau \lambda_{\max}(C_\tau) = |a|^\tau \left|\lambda_{\max}(C_\tau)\right| \leq |a|^\tau \rho(C_1)^\tau.$$

[55] The proof does not assume that λ is unique. I allow for the possibility that λ and $\lambda' = -\lambda$ are both eigenvalues of C_1 and $|\lambda| = |\lambda'| = \rho(C_1)$.

Thus, $\lambda_{\max}\left(a^\tau C_\tau\right) \leq |a|^\tau \rho(C_1)^\tau$ regardless of the value of a and the eigenvalues of C_1. Therefore,

$$\lambda_{\max}\left(\sum_{\tau=1}^{\infty} a^\tau \eta^{\tau-1} C_\tau\right) \leq \sum_{\tau=1}^{\infty} \eta^{\tau-1} \lambda_{\max}\left(a^\tau C_\tau\right) \leq \sum_{\tau=1}^{\infty} \eta^{\tau-1} |a|^\tau \rho(C_1)^\tau = \frac{|a|\rho(C_1)}{1-\eta|a|\rho(C_1)},$$

where the first inequality is using the fact that $\eta^{\tau-1} \geq 0$ for all $\tau \geq 1$ and Weyl's inequality. Consequently,

$$\lambda_{\max}(\Omega(a,\eta)) \leq \overline{f}(a,\eta) < \rho(C_1)^2$$

for any a, η such that $(|a|, \eta) \neq (\rho(C_1), 0)$.

I finish the proof by arguing that $\lambda_{\max}(\Omega(\lambda,0)) = \rho(C_1)^2$ and $\lambda_{\max}(\Omega(-\lambda,0)) \leq \overline{f}(-\lambda,0) = \rho(C_1)^2$ with the inequality strict if $-\lambda$ is not an eigenvalue of C_1. To see this, first note that

$$\lambda_{\max}(\Omega(a,0)) = -a^2 + 2\lambda_{\max}(aC_1) = \begin{cases} -a^2 + 2a\lambda_{\min}(C_1) & \text{if} \quad a < 0, \\ -a^2 + 2a\lambda_{\max}(C_1) & \text{if} \quad a \geq 0. \end{cases}$$

Thus,

$$\max_{a\in[-1,1]} \lambda_{\max}(\Omega(a,0)) = \begin{cases} \lambda_{\min}(C_1)^2 & \text{if} \quad |\lambda_{\min}(C_1)| > \lambda_{\max}(C_1), \\ \lambda_{\max}(C_1)^2 & \text{if} \quad |\lambda_{\min}(C_1)| \leq \lambda_{\max}(C_1), \end{cases}$$

and

$$\arg\max_{a\in[-1,1]} \lambda_{\max}(\Omega(a,0)) = \begin{cases} \{\lambda_{\min}(C_1)\} & \text{if} \quad |\lambda_{\min}(C_1)| > \lambda_{\max}(C_1), \\ \{\lambda_{\min}(C_1), \lambda_{\max}(C_1)\} & \text{if} \quad |\lambda_{\min}(C_1)| = \lambda_{\max}(C_1), \\ \{\lambda_{\max}(C_1)\} & \text{if} \quad |\lambda_{\min}(C_1)| < \lambda_{\max}(C_1). \end{cases}$$

Since C_1 is a symmetric matrix, the eigenvalues of C_1 are all real, and so,

$$\rho(C_1) = \begin{cases} -\lambda_{\min}(C_1) & \text{if} \quad |\lambda_{\min}(C_1)| > \lambda_{\max}(C_1), \\ \lambda_{\max}(C_1) & \text{if} \quad |\lambda_{\min}(C_1)| \leq \lambda_{\max}(C_1). \end{cases}$$

This establishes that, in any pseudo-true one-state model, $\eta = 0$, $a = \lambda$, and

$$\Omega(a,\eta) = -\lambda^2 I + 2\lambda C_1.$$

By Theorem 2, u is an eigenvector of $\Omega(a,\eta)$ with eigenvalue $\lambda_{\max}(\Omega(a,\eta)) = \lambda^2$ and $u'u = 1$. Therefore, u is also an eigenvector of C_1, but with eigenvalue λ. This completes the proof of the theorem. □

Proof of Proposition 2

I first state and prove a lemma, which is used in the proof of the proposition.

Lemma G.4. *Any Markovian model θ has a representation as in Lemma G.1 for which $D'D = I$.*

Proof. Fix a Markovian model θ, and let M, D, and N be as in Lemma G.1. By (G.17), the s-step-ahead forecast under model θ is given by

$$E_t^\theta[y_{t+s}] = N'^{-1} D M^{s-1} \sum_{\tau=0}^{\infty} \left(M\left(I - D'D\right)\right)^\tau M D' N' y_{t-\tau}.$$

Since θ is Markovian and N is invertible, $D\left(M\left(I-D'D\right)\right)^{\tau} MD' = \mathbf{0}$ for all $\tau \geq 1$. As the first step of the proof, I use this identity and an inductive argument to show that $DM^{s}D' = (DMD')^{s}$ for all $s \geq 2$. The following equation establishes the induction base:

$$\mathbf{0} = D\left(M\left(I-D'D\right)\right) MD' = DM^{2}D' - DMD'DMD'.$$

As the induction hypothesis, suppose $DM^{s}D' = (DMD')^{s}$ for some $s \geq 2$. Note that

$$\begin{aligned} D\left(M\left(I-D'D\right)\right)^{s} MD' &= D\left(M\left(I-D'D\right)\right)^{s-1} M(I-D'D)MD' \\ &= D\left(M\left(I-D'D\right)\right)^{s-1} M^{2}D' - D\left(M\left(I-D'D\right)\right)^{s-1} MD'DMD' \\ &= D\left(M\left(I-D'D\right)\right)^{s-1} M^{2}D', \end{aligned}$$

where the last equality follows the fact that $D\left(M\left(I-D'D\right)\right)^{s-1} MD' = \mathbf{0}$ for any $s \geq 2$. By a similar argument,

$$D\left(M\left(I-D'D\right)\right)^{s} MD' = D\left(M\left(I-D'D\right)\right)^{s-2} M^{3}D' = \cdots = D\left(M\left(I-D'D\right)\right) M^{s}D'.$$

Therefore,

$$D\left(M\left(I-D'D\right)\right)^{s} MD' = DM^{s+1}D' - DMD'DM^{s}D' = DM^{s+1}D' - (DMD')^{s+1},$$

where the last equality follows the induction hypothesis. The assumption that $D\left(M\left(I-D'D\right)\right)^{s} MD' = \mathbf{0}$ then proves the induction step.

I next find a model $\tilde{\theta}$, represented by matrices $\tilde{M}$, $\tilde{D}$, and $\tilde{N}$, that is observationally equivalent to θ and for which $\tilde{D}'\tilde{D} = I$. Since $D \in \mathbb{R}^{n\times d}$ is a rectangular diagonal matrix and $d \leq n$,

$$DMD' = \begin{pmatrix} M_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

for some $d \times d$ matrix M_1. Let $\tilde{M} = M_1$, $\tilde{D} \in \mathbb{R}^{n\times d}$ be the rectangular diagonal matrix with its diagonal elements equal to one, and $\tilde{N} = N$. Then $\tilde{D}'\tilde{D} = I$. Furthermore,

$$DM^{s}D' = (DMD')^{s} = \begin{pmatrix} M_1^{s} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} = \tilde{D}\tilde{M}^{s}\tilde{D}'.$$

By equation (G.17), the forecasts are identical under the two models:

$$E_t^{\theta}[y_{t+s}] = N'^{-1}DM^{s}D'N'y_t = \tilde{N}'^{-1}\tilde{D}\tilde{M}^{s}\tilde{D}'\tilde{N}'y_t = E_t^{\tilde{\theta}}[y_{t+s}].$$

By equation (G.18), the unconditional variance of the observable is also identical under the two models:

$$\mathrm{Var}^{\theta}(y) = N'^{-1}\left(I + \sum_{\tau=1}^{\infty} DM^{\tau}D'DM'^{\tau}D'\right)N^{-1} = \tilde{N}'^{-1}\left(I + \sum_{\tau=1}^{\infty} \tilde{D}\tilde{M}^{\tau}\tilde{D}'\tilde{D}\tilde{M}'^{\tau}\tilde{D}'\right)\tilde{N}^{-1} = \mathrm{Var}^{\tilde{\theta}}(y).$$

On the other hand,

$$E^{\theta}[y_{t+s}y_t'] = E^{\theta}[E_t^{\theta}[y_{t+s}]y_t'] = N'^{-1}DM^{s}D'N'E^{\theta}[y_t y_t'] = N'^{-1}DM^{s}D'N'\mathrm{Var}^{\theta}(y),$$

and similarly for $E^{\tilde{\theta}}[y_{t+s}y_t']$. Therefore, $E^{\theta}[y_{t+s}y_t'] = E^{\tilde{\theta}}[y_{t+s}y_t']$ for all s; that is, P^{θ} and $P^{\tilde{\theta}}$ also have identical autocovariance matrices at all lags. This conclusion, together with the fact that P^{θ} and $P^{\tilde{\theta}}$ are both zero-mean Gaussian distributions, implies that they are observationally equivalent. □

I can now prove the proposition.

Proof of Proposition 2. By equation (G.4),

$$\mathrm{Var}_t^{\theta}(y_{t+1}) = B'\hat{\Sigma}_z B + R,$$

where $\hat{\Sigma}_z$ solves the algebraic Riccati equation (G.3). The equation can be written as

$$\hat{\Sigma}_z = A\hat{\Sigma}_z^{\frac{1}{2}}\left(I - \hat{\Sigma}_z^{\frac{1}{2}}B\left(B'\hat{\Sigma}_z B + R\right)^{-1}B'\hat{\Sigma}_z^{\frac{1}{2}}\right)\hat{\Sigma}_z^{\frac{1}{2}}A' + Q. \tag{G.35}$$

Since R is a positive semidefinite matrix, so is $I - \hat{\Sigma}_z^{\frac{1}{2}}B\left(B'\hat{\Sigma}_z B + R\right)^{-1}B'\hat{\Sigma}_z^{\frac{1}{2}}$. Therefore, $\hat{\Sigma}_z \geq Q$, and so, $\mathrm{Var}_t^{\theta}(y_{t+1}) \geq B'QB + R$. On the other hand,

$$\mathrm{Var}^{\theta}(y_{t+1}|z_t) = B'\mathrm{Var}^{\theta}(z_{t+1}|z_t)B + R = B'QB + R,$$

where I am using the assumption that w_t is i.i.d. $\mathcal{N}(0, Q)$, v_t is i.i.d. $\mathcal{N}(0, R)$, and w_t and v_t are independent. This proves the first part of the proposition.

To prove the second part, first assume that $\mathrm{Var}_t^{\theta}(y_{t+1}) = B'QB + R$. Together with equation (G.35), this implies that

$$B'A\hat{\Sigma}_z^{\frac{1}{2}}\left(I - \hat{\Sigma}_z^{\frac{1}{2}}B\left(B'\hat{\Sigma}_z B + R\right)^{-1}B'\hat{\Sigma}_z^{\frac{1}{2}}\right)\hat{\Sigma}_z^{\frac{1}{2}}A'B = \mathbf{0}.$$

Since $\left(I - \hat{\Sigma}_z^{\frac{1}{2}}B\left(B'\hat{\Sigma}_z B + R\right)^{-1}B'\hat{\Sigma}_z^{\frac{1}{2}}\right)$ is a symmetric positive semidefinite matrix, the above equation implies that

$$B'A\hat{\Sigma}_z^{\frac{1}{2}}\left(I - \hat{\Sigma}_z^{\frac{1}{2}}B\left(B'\hat{\Sigma}_z B + R\right)^{-1}B'\hat{\Sigma}_z^{\frac{1}{2}}\right) = \mathbf{0}.$$

On the other hand, by equation (G.5), the one-step-ahead forecast under model θ is given by

$$E_t^{\theta}[y_{t+1}] = B'\sum_{\tau=0}^{\infty}(A - KB')^{\tau}Ky_{t-\tau}.$$

Substituting for K from equation (G.2), I get

$$B'(A - KB') = B'\left(A - A\hat{\Sigma}_z B\left(B'\hat{\Sigma}_z B + R\right)^{-1}B'\right) = B'A\hat{\Sigma}_z^{\frac{1}{2}}\left(I - \hat{\Sigma}_z^{\frac{1}{2}}B\left(B'\hat{\Sigma}_z B + R\right)^{-1}B'\hat{\Sigma}_z^{\frac{1}{2}}\right)\hat{\Sigma}_z^{\frac{-1}{2}} = \mathbf{0}.$$

Therefore,

$$E_t^{\theta}[y_{t+1}] = B'\sum_{\tau=0}^{\infty}(A - KB')^{\tau}Ky_{t-\tau} = B'Ky_t.$$

On the other hand, $\mathrm{Var}_t^{\theta}(y_{t+1}) = B'\hat{\Sigma}_z B + R$. Under model θ, the mean and variance of y_{t+1} conditional on $\{y_\tau\}_{\tau\leq t}$ are both independent of $\{y_\tau\}_{\tau<t}$. Furthermore, P^{θ} is Gaussian. Therefore, it is Markovian.

Next, suppose P^θ is Markovian. Then by Lemma G.4, model θ has a representation as in Lemma G.1 for which $D'D = I$. By equation (G.11), then

$$\hat{\Sigma}_z^{\frac{1}{2}} B \left(B'\hat{\Sigma}_z B + R\right)^{-1} B'\hat{\Sigma}_z^{\frac{1}{2}} = VD'DV' = I,$$

where the second equality follows the facts that $D'D = I$ and V is orthogonal. Substituting in equation (G.35), I get $\hat{\Sigma}_z = Q$. Therefore,

$$\text{Var}_t^\theta(y_{t+1}) = B'\hat{\Sigma}_z B + R = B'QB + R.$$

This completes the proof of the proposition. □

Proof of Theorem 5

By Lemma G.1, the agent's model can be represented in terms of matrices M, D, and N. Since the agent is restricted to the set of Markovian models, by Lemma G.4, I can set $D = (I\ \mathbf{0})'$. Let $S \equiv \Gamma_0^{\frac{1}{2}} N$ and $\Gamma \equiv \Gamma_0^{\frac{-1}{2}} \Gamma_1 \Gamma_0^{\frac{-1}{2}}$. The expression for the KLDR in (G.9) then simplifies to

$$\text{KLDR}(M,S,D) = -\frac{1}{2}\log\det(SS') + \frac{1}{2}\text{tr}(S'S) - \text{tr}(MD'S'\Gamma SD) + \frac{1}{2}\text{tr}(MD'S'SDM') + \text{constant}.$$

Write $S = (S_1\ S_2)$, where $S_1 \in \mathbb{R}^{n\times d}$ and $S_2 \in \mathbb{R}^{n\times(n-d)}$. The above expression can then be written as

$$-\frac{1}{2}\log\det\left(S_1S_1' + S_2S_2'\right) + \frac{1}{2}\text{tr}\left(S_1'S_1\right) + \frac{1}{2}\text{tr}\left(S_2'S_2\right) - \text{tr}\left(MS_1'\Gamma S_1\right) + \frac{1}{2}\text{tr}\left(MS_1'S_1M'\right) + \text{constant}.$$

I next optimize the above expression with respect to M, S_1, and S_2. The first-order optimality condition with respect to S_2 is given by $\left(S_1S_1' + S_2S_2'\right)^{-1} S_2 = S_2$, which can be rewritten as $S_1S_1'S_2 + S_2(S_2'S_2 - I) = \mathbf{0}$. Let b_0 be an arbitrary vector in $\mathbb{R}^{n-d}$, $b_1 \equiv S_1'S_2b_0 \in \mathbb{R}^d$, and $b_2 \equiv (S_2'S_2 - I)b_0 \in \mathbb{R}^{n-d}$. The above equation then implies that

$$0 = \left(S_1S_1'S_2 + S_2(S_2'S_2 - I)\right)b_0 = S_1b_1 + S_2b_2 = Sb,$$

where $b \equiv \begin{pmatrix} b_1' & b_2' \end{pmatrix}' \in \mathbb{R}^n$. Since S is an invertible matrix, it must be that $b = 0$. Therefore, $b_1 = 0$ and $b_2 = 0$. Since b_0 was arbitrary, $S_2'S_2 = I$ and $S_1'S_2 = \mathbf{0}$. On the other hand,

$$\log\det\left(S_1S_1' + S_2S_2'\right) = \log\det(SS') = \log\det(S'S) = \log\det\begin{pmatrix} S_1'S_1 & S_1'S_2 \\ S_2'S_1 & S_2'S_2 \end{pmatrix}.$$

Therefore,

$$\log\det\left(S_1S_1' + S_2S_2'\right) = \log\det\begin{pmatrix} S_1'S_1 & \mathbf{0} \\ \mathbf{0} & I \end{pmatrix} = \log\det(S_1'S_1).$$

The KLDR can thus be written only as a function of M and S_1 as

$$\text{KLDR}(M,S_1) = -\frac{1}{2}\log\det\left(S_1'S_1\right) + \frac{1}{2}\text{tr}\left(S_1'S_1\right) - \text{tr}\left(MS_1'\Gamma S_1\right) + \frac{1}{2}\text{tr}\left(MS_1'S_1M'\right) + \text{constant}. \quad (G.36)$$

The first-order optimality conditions with respect to M and S_1 are then given by

$$- S_1'\Gamma' S_1 + M S_1' S_1 = 0, \tag{G.37}$$

$$- S_1^{\dagger\prime} + S_1 - \Gamma S_1 M - \Gamma' S_1 M' + S_1 M' M = 0. \tag{G.38}$$

Since $S_1'S_1$ is invertible, (G.37) can be solve for M to get $M = S_1'\Gamma' S_1 (S_1'S_1)^{-1}$. Substituting in (G.38), I get

$$S_1(S_1'S_1)^{-1} = S_1 - \Gamma S_1 S_1'\Gamma' S_1(S_1'S_1)^{-1} - \Gamma' S_1(S_1'S_1)^{-1}S_1'\Gamma S_1 + S_1(S_1'S_1)^{-1}S_1'\Gamma S_1 S_1'\Gamma' S_1(S_1'S_1)^{-1}, \tag{G.39}$$

where I am using the fact that $S_1^{\dagger} = (S_1'S_1)^{-1}S_1'$. Next consider the singular-value decomposition of S_1:

$$S_1 = U\Sigma V', \tag{G.40}$$

where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{n\times d}$ is a rectangular diagonal matrix. Substituting for S_1 in (G.39) from (G.40) and multiplying the result from left and right by U' and $V\Sigma'$, respectively, I get

$$\Sigma(\Sigma'\Sigma)^{-1}\Sigma' = \Sigma\Sigma' - X\Sigma\Sigma'X'\Sigma(\Sigma'\Sigma)^{-1}\Sigma' - X'\Sigma(\Sigma'\Sigma)^{-1}\Sigma'X\Sigma\Sigma' + \Sigma(\Sigma'\Sigma)^{-1}\Sigma'X\Sigma\Sigma'X'\Sigma(\Sigma'\Sigma)^{-1}\Sigma', \tag{G.41}$$

where $X \equiv U'\Gamma U$. Note that $\Sigma = \begin{pmatrix} \Sigma_1 \\ \mathbf{0} \end{pmatrix}$ for some diagonal matrix $\Sigma_1 \in \mathbb{R}^{n\times d}$. Moreover, since $S_1'S_1$ is invertible, so is Σ_1. Therefore,

$$\Sigma(\Sigma'\Sigma)^{-1}\Sigma' = \begin{pmatrix} I & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

$$\Sigma\Sigma' = \begin{pmatrix} \Sigma_1^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Write $X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}$, where $X_{11} \in \mathbb{R}^{d\times d}$, $X_{12} \in \mathbb{R}^{d\times(n-d)}$, $X_{21} \in \mathbb{R}^{(n-d)\times d}$, and $X_{22} \in \mathbb{R}^{(n-d)\times(n-d)}$. Equation (G.41) then implies

$$X_{11}'X_{11} = I - \Sigma_1^{-2}, \tag{G.42}$$

$$X_{21}\Sigma_1^2 X_{11}' + X_{12}'X_{11}\Sigma_1^2 = 0. \tag{G.43}$$

These equations fully characterize the set of all (local) extrema of the KLDR.

I next use these equations to show that, as long as either d is equal to one or Γ_1 is symmetric, and for any $i = 1, \dots, d$, the ith coordinate vector $e_i \in \mathbb{R}^n$ is an eigenvector of $(X + X')/2$ with eigenvalue $e_i'Xe_i$.[56] If $e_i'Xe_i = 0$, then trivially e_i is an eigenvector of $(X + X')/2$ with eigenvalue $e_i'Xe_i = 0$. So in the rest of the proof, I consider the case where $e_i'Xe_i \neq 0$. First, suppose $d = 1$.

[56] With slight abuse of notation, I use e_i to denote the ith coordinate vector both in $\mathbb{R}^n$ and in $\mathbb{R}^d$. Whether $e_i \in \mathbb{R}^n$ or $e_i \in \mathbb{R}^d$ will be clear from the context.

Then $i = 1$ and $X_{11}' = X_{11} = e_1'Xe_1 \neq 0$. On the other hand, Σ_1 is a non-zero scalar. Equation (G.43) then implies that $X_{21} + X_{12}' = 0$. Therefore,

$$\left(\frac{X+X'}{2}\right)e_1 = \frac{1}{2}\begin{pmatrix} 2X_{11} & X_{12}+X_{21}' \\ X_{21}+X_{12}' & X_{22}+X_{22}' \end{pmatrix}\begin{pmatrix} 1 \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} X_{11} \\ \mathbf{0} \end{pmatrix} = e_1'Xe_1e_1,$$

proving that e_1 is an eigenvector of $(X+X')/2$ with eigenvalue e_1Xe_1. Next, suppose Γ_1 is symmetric. This implies that Γ, and by extension, X are symmetric matrices. Equation (G.42) then implies that X_{11} is a diagonal matrix. Since Σ_1 is also diagonal, it commutes with X_{11}. Equation (G.43) then implies that

$$(X_{21} + X_{12}')X_{11} = 2X_{21}X_{11} = \mathbf{0}, \tag{G.44}$$

where I am using the fact that Σ_1 is non-singular and X is symmetric. But since X_{11} is a diagonal matrix, it can be written as

$$X_{11} = \sum_{k=1}^{d} e_k'X_{11}e_ke_ke_k'.$$

Substituting in (G.44), I get

$$\sum_{k=1}^{d} X_{21}e_k'X_{11}e_ke_ke_k' = \mathbf{0}.$$

In particular, it must be the case that $X_{21}e_i'X_{11}e_ie_i = 0$. But since $e_i'X_{11}e_i = e_i'Xe_i \neq 0$, it must be that $X_{21}e_i = 0$. Therefore,

$$\left(\frac{X+X'}{2}\right)e_i = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}\begin{pmatrix} e_i \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} X_{11}e_i \\ X_{21}e_i \end{pmatrix} = \begin{pmatrix} e_i'X_{11}e_ie_i \\ 0 \end{pmatrix} = e_i'Xe_ie_i,$$

where the third equality relies on the fact that X_{11} is diagonal. This proves that e_i is an eigenvector of $(X+X')/2$ with eigenvalue $e_i'Xe_ie_i$.

I next show that any matrices M and S_1 that satisfy the first-order optimality conditions (G.37) and (G.38) must be of the form

$$M = \sum_{i=1}^{d} a_iv_iv_i', \tag{G.45}$$

$$S_1 = \sum_{i=1}^{d} \frac{1}{\sqrt{1-a_i^2}}u_iv_i', \tag{G.46}$$

where $\{a_i\}_{i=1}^d$ are eigenvalues of C_1, $u_i \in \mathbb{R}^n$ denotes an eigenvector with eigenvalue a_i normalized such that $u_i'u_k = \mathbb{1}_{\{i=k\}}$ for all $i, k \in \{1, \dots, d\}$, and $\{v_i\}_{i=1}^d$ is an orthonormal basis for $\mathbb{R}^d$. To see this, first note that equation (G.40) can be written as

$$S_1 = U\Sigma V' = U\sum_{i=1^d} \sigma_ie_ie_i'V',$$

where σ_i denotes the ith diagonal element of $\Sigma \in \mathbb{R}^{n\times d}$. I let $u_i \equiv Ue_i$ and $v_i \equiv Ve_i$. Since U and V are orthogonal matrices, $\{u_i\}_{i=1}^d$ is a set of orthonormal vectors and $\{v_i\}_{i=1}^d$ is an orthonormal basis for $\mathbb{R}^d$. Therefore, to show that S_1 takes the form given in (G.46), I only need to show that u_i is an eigenvector of C_1 with eigenvalue a_i and $\sigma_i = 1/\sqrt{1-a_i^2}$. Note that

$$C_1 u_i = \frac{1}{2}\left(\Gamma + \Gamma'\right) U e_i = \frac{1}{2} U U' \left(\Gamma + \Gamma'\right) U e_i = U \left(\frac{X + X'}{2}\right) e_i = U e_i' X e_i e_i = e_i' X e_i u_i,$$

where the fourth equality uses the fact that e_i is an eigenvector of $(X + X')/2$. Therefore, u_i is an eigenvector of C_1. On the other hand, multiplying equation (G.42) from left and right by e_i' and e_i, respectively, for $i \in \{1, \ldots, d\}$ and using the fact that X_{11} is diagonal, I get

$$\left(e_i' X_{11} e_i\right)^2 = 1 - \sigma_i^{-2}.$$

But

$$e_i' X_{11} e_i = e_i' X e_i = e_i' \left(\frac{X + X'}{2}\right) e_i = e_i' U \left(\frac{\Gamma + \Gamma'}{2}\right) U e_i = u_i' C_1 u_i = u_i' a_i u_i = a_i,$$

where a_i denotes the eigenvalue of C_1 with eigenvector u_i. Therefore, $\sigma_i = 1/\sqrt{1-a_i^2}$. Finally, recall that $M = S_1'\Gamma' S_1 (S_1' S_1)^{-1}$. By assumption, either $d = 1$, and so, S_1 is a vector in $\mathbb{R}^n$ or Γ is symmetric. Either way $S_1'\Gamma' S_1 = S_1'(\Gamma + \Gamma')S_1/2 = S_1' C_1 S_1$. Therefore,

$$\begin{aligned} M = S_1' C_1 S_1 (S_1' S_1)^{-1} &= \left(\sum_{i,k=1}^{d} \frac{1}{\sqrt{1-a_i^2}} v_i u_i' C_1 \frac{1}{\sqrt{1-a_k^2}} u_k v_k'\right)\left(\sum_{i,k=1}^{d} \frac{1}{\sqrt{1-a_i^2}} v_i u_i' \frac{1}{\sqrt{1-a_k^2}} u_k v_k'\right)^{-1} \\ &= \left(\sum_{i=1}^{d} \frac{1}{1-a_i^2} v_i a_i v_i'\right)\left(\sum_{i=1}^{d} \frac{1}{1-a_i^2} v_i v_i'\right)^{-1} = \sum_{i=1}^{d} a_i v_i v_i', \end{aligned}$$

where I am using the facts that u_i is an eigenvector of C_1 with eigenvalue a_i and that $\{u_i\}_{i=1}^d$ and $\{v_i\}_{i=1}^d$ are orthonormal sets of vectors.

Although any M and S_1 of the forms (G.45) and (G.46) satisfy the necessary optimality condition, not all such candidates are global minimizers of the KLDR. To find the global optima, I substitute the solutions to the first-order optimality conditions in the KLDR and select the solutions that minimize the KLDR. Multiplying equation (G.38) from left by S_1', I get

$$I = S_1' S_1 - S_1' \Gamma S_1 M - S_1' \Gamma' S_1 M' + S_1' S_1 M' M.$$

Computing the trace of the above equation and substituting the result in (G.36), I get

$$\begin{aligned} \text{KLDR}(M, S_1) &= -\frac{1}{2} \log\det\left(S_1' S_1\right) + \frac{1}{2} \operatorname{tr}\left(S_1' S_1\right) - \operatorname{tr}\left(M S_1' \Gamma S_1\right) + \frac{1}{2} \operatorname{tr}\left(M S_1' S_1 M'\right) + \text{constant} \\ &= -\frac{1}{2} \log\det\left(S_1' S_1\right) + \frac{1}{2} \operatorname{tr}(I) + \text{constant}. \end{aligned}$$

Therefore, the M and S_1 pairs that minimize the KLDR are the ones that maximize the determinant of $S_1' S_1$. But since $S_1' S_1$ is a symmetric matrix with eigenvalues $\{1/(1-a_i^2)\}_{i=1}^d$, its determinant is equal to $\prod_{i=1}^d \frac{1}{1-a_i^2}$. Therefore, any M and S_1 pair that minimize the KLDR are of the forms (G.45) and (G.46) with $\{a_i\}_{i=1}^d$ the top d eigenvalues of C_1 in magnitude (with the possibility that some of the a_i are equal).

With the expressions for the pseudo-true M and S_1 in hand, I can prove the theorem.

Part (a). The forecasts given a model parameterized by matrices M, D, and N are given by equation (G.17). Using the definition of $S \equiv \Gamma_0^{\frac{1}{2}} N$ and the fact that $D'D$ can be taken to be identity matrix, I can write equation (G.17) as follows:

$$E_t^\theta[y_{t+s}] = \Gamma_0^{\frac{1}{2}} S'^{-1} D M^s D' S' \Gamma_0^{\frac{-1}{2}} y_{t-\tau}.$$

Note that for any matrix $S = (S_1 \; S_2)$ that satisfies the first-order optimality condition with respect to S_2,

$$S^{-1} = \begin{pmatrix} (S_1' S_1)^{-1} S_1' \\ S_2' \end{pmatrix}.$$

Therefore,

$$S'^{-1} = \begin{pmatrix} S_1 (S_1' S_1)^{-1} & S_2 \end{pmatrix},$$

and so

$$S'^{-1} D = S_1 (S_1' S_1)^{-1}. \tag{G.47}$$

The forecasts can thus be written only in terms of matrices M and S_1 as follows:

$$E_t^\theta[y_{t+s}] = \Gamma_0^{\frac{1}{2}} S_1 (S_1' S_1)^{-1} M^s S_1' \Gamma_0^{\frac{-1}{2}}.$$

Substituting for M and S_1 using (G.45) and (G.46) and simplifying the resulting expression, I get

$$E_t^\theta[y_{t+s}] = \Gamma_0^{\frac{1}{2}} \sum_{i=1}^{d} a_i^s u_i u_i' \Gamma_0^{\frac{-1}{2}}.$$

Letting $p_i \equiv \Gamma_0^{\frac{-1}{2}} u_i$ and $q_i \equiv \Gamma_0^{\frac{1}{2}} u_i$ completes the proof of part (a).

Part (b). Equation (G.18) gives the variance-covariance matrix under a model parameterized by matrices M, D, and N. Using the definition of S and setting $D'D = I$, equation (G.18) can be written as follows:

$$\text{Var}^\theta(y) = \Gamma_0^{\frac{1}{2}} \left(S'^{-1} S^{-1} + S^{-1'} D \sum_{\tau=1}^{\infty} M^\tau M'^\tau D' S^{-1} \right) \Gamma_0^{\frac{1}{2}}.$$

To prove part (b), I need to show that the terms in parentheses add up to the identity matrix. I start with the first term:

$$S'^{-1} S^{-1} = (SS')^{-1} = (S_1 S_1' + S_2 S_2')^{-1}. \tag{G.48}$$

The fact that $S_2' S_2 = I$ implies that S_2 can be written as

$$S_2 = \sum_{i=d+1}^{n} u_i w_i',$$

where $u_i \in \mathbb{R}^n$ and $w_i \in \mathbb{R}^{n-d}$ for $i = d+1, \ldots, n$, $\{u_i\}_{i=d+1}^n$ are orthonormal vectors, and $\{w_i\}_{i=d+1}^n$ constitutes an orthonormal basis for $\mathbb{R}^{n-d}$. On the other hand, the fact that $S_1' S_2 = \mathbf{0}$ implies that

$u_i'u_k = 0$ for any $i \in \{1, \dots, d\}$ and $k \in \{d+1, \dots, n\}$. Therefore, $\{u_i\}_{i=1}^n$ constitutes an orthonormal basis for $\mathbb{R}^n$. Substituting for S_1 and S_2 in (G.48), I get

$$S'^{-1}S^{-1} = \left(\sum_{i=1}^{d} \frac{1}{1-a_i^2} u_i u_i' + \sum_{i=d+1}^{n} u_i u_i' \right)^{-1} = \sum_{i=1}^{d} (1-a_i^2) u_i u_i' + \sum_{i=d+1}^{n} u_i u_i',$$

where the second equality uses the fact that $\{u_i\}_{i=1}^n$ are orthonormal. Next consider the second term:

$$\begin{aligned} S^{-1\prime} D \sum_{\tau=1}^{\infty} M^{\tau} M'^{\tau} D' S^{-1} &= S_1 (S_1'S_1)^{-1} \sum_{\tau=1}^{\infty} M^{\tau} M'^{\tau} (S_1'S_1)^{-1} S_1' \\ &= \sum_{i=1}^{d} \sqrt{1-a_i^2} u_i v_i' \sum_{\tau=1}^{\infty} \sum_{k=1}^{d} a_k^{2\tau} v_k v_k' \sum_{l=1}^{d} \sqrt{1-a_l^2} v_l u_l' \\ &= \sum_{i=1}^{d} (1-a_i^2) u_i u_i' \sum_{\tau=1}^{\infty} a_i^{2\tau} \\ &= \sum_{i=1}^{d} a_i^2 u_i u_i', \end{aligned}$$

where the first equality uses (G.47) and the second equality is by (G.45) and (G.46). Putting everything together,

$$S'^{-1}S^{-1} + S^{-1\prime} D \sum_{\tau=1}^{\infty} M^{\tau} M'^{\tau} D' S^{-1} = \sum_{i=1}^{d} (1-a_i^2) u_i u_i' + \sum_{i=d+1}^{n} u_i u_i' + \sum_{i=1}^{d} a_i^2 u_i u_i' = \sum_{i=1}^{n} u_i u_i' = I,$$

where the last equality follows the fact that $\{u_i\}_{i=1}^n$ is an orthonormal basis for $\mathbb{R}^n$. □

Proof of Proposition 3

Recall that I have assumed (without loss of generality) that Γ_0 is non-singular. Since C_1 is symmetric, $\{u_i\}_{i=1}^d$ constitutes an orthonormal basis for $\mathbb{R}^n$, and so, $\Gamma_0^{\frac{-1}{2}} y_t$ can be expressed as

$$\Gamma_0^{\frac{-1}{2}} y_t = \sum_{i=1}^{n} \omega_{it} u_i,$$

where $\omega_{it} \equiv u_i' \Gamma_0^{\frac{-1}{2}} y_t$. Therefore,

$$y_t = \Gamma_0^{\frac{1}{2}} \sum_{i=1}^{n} \omega_{it} u_i = \sum_{i=1}^{n} \Gamma_0^{\frac{1}{2}} u_i u_i' \Gamma_0^{\frac{-1}{2}} y_t = \sum_{i=1}^{n} y_t^{(i)} q_i,$$

where the last equality uses the definitions of $y_t^{(i)}$ and q_i.

The lag-one autocovariance of $y_t^{(i)}$ is given by

$$\mathbb{E}\left[y_t^{(i)} y_{t-1}^{(i)} \right] = p_i{}' \mathbb{E}[y_t y_{t-1}] p_i = p_i{}' \Gamma_1 p_i = p_i{}' \left(\frac{\Gamma_1 + \Gamma_1'}{2} \right) p_i = p_i{}' \Gamma_0^{\frac{1}{2}} C_1 \Gamma_0^{\frac{1}{2}} p_i = u_i' C_1 u_i,$$

where the first equality uses the definition $y_t^{(i)}$, and the last equality uses the definition of p_i. The fact that u_i is an eigenvector of C_1 implies $u_i' C_1 u_i = a_i u_i' u_i = a_i$, where a_i is the ith largest (in magnitude) eigenvalue of C_1. Moreover, $\mathbb{E}\left[{y_t^{(i)}}^2\right] = p_i{}' \Gamma_0 p_i = u_i' u_i = 1$. Therefore,

$$\rho_i \equiv \mathbb{E}\left[y_t^{(i)} y_{t-1}^{(i)}\right] / \sqrt{\mathbb{E}\left[{y_t^{(i)}}^2\right]} = a_i.$$

The proposition follows the fact that a_i is the ith largest eigenvalue of C_1 in magnitude. □

Proof of Proposition 4

I prove the result under the assumption that the top D eigenvalues of the first autocorrelation matrix, C_1, are all distinct. This assumption is true for generic true processes. By Theorem 5(a) (or Theorem 4), the forecasts of an agent who uses a pseudo-true d-state model θ are given by

$$E_t^{\theta}\left[y_{t+s}\right] = \sum_{i=1}^{d} a_i{}^s q_i p_i{}' y_t, \tag{G.49}$$

where a_i is the ith largest eigenvalue of C_1, u_i denotes the corresponding eigenvector, normalized to have unit norm, $p_i \equiv \Gamma_0^{\frac{-1}{2}} u_i$, and $q_i \equiv \Gamma_0^{\frac{1}{2}} u_i$. Since the eigenvalues of C_1 are all distinct, the corresponding eigenvectors are unique (up to multiplicative constants). Therefore, all agents use the same values of $\{(a_i, p_i, q_i)\}_i$ to forecast.

Consider agent j who is constrained to models of dimension d_j. The agent's optimal action given her pseudo-true d-state model is given by

$$\begin{aligned} x_{jt} &= E_t^{\theta_j}\left[\sum_{s=1}^{\infty} c_{js}' y_{t+s}\right] = \sum_{s=1}^{\infty} c_{js}' E_t^{\theta_j}\left[y_{t+s}\right] \\ &= \sum_{s=1}^{\infty} c_{js}' \sum_{i=1}^{d_j} a_i{}^s q_i p_i{}' y_t = \sum_{i=1}^{d_j} g_{ji} y_t^{(i)}, \end{aligned}$$

where θ_j denotes agent j's pseudo-true model, $y_t^{(i)} \equiv p_i' y_t$ as before, $g_{ji} \equiv \sum_{s=1}^{\infty} a_i{}^s c_{js}' q_i$ is a constant, which is a finite since $\{c_{js}\}_s$ is absolutely summable. Using vector notation, $x_t \equiv (x_{1t}, \ldots, x_{Jt})' \in \mathbb{R}^J$, I can write the above expression as $x_t = G y_t^{(1:D)}$, where $G \equiv \begin{pmatrix} g_1' & g_2' & \cdots & g_J' \end{pmatrix}' \in \mathbb{R}^{J\times D}$, $g_j \equiv \begin{pmatrix} g_{j1} & g_{j2} & \ldots g_{jd_j} & 0 & \ldots & 0 \end{pmatrix} \in \mathbb{R}^{1\times D}$, and $y_t^{(1:D)} \equiv \begin{pmatrix} y_t^{(1)} & y_t^{(2)} & \cdots & y_t^{(D)} \end{pmatrix}' \in \mathbb{R}^D$. □

Proof of Proposition 5

I start by taking v to be an arbitrary n-dimensional vector and computing the autocovariances of $v' y_t$ under the pseudo-true and true models. Define $w \equiv \Gamma_0^{\frac{1}{2}} v$. Under a pseudo-true d-state model θ,

$$E^{\theta}\left[v' y_t v' y_{t-l}\right] = v' E^{\theta}\left[y_t y_{t-l}'\right] v = v' E^{\theta}\left[E_{t-l}^{\theta}\left[y_t\right] y_{t-l}'\right] v$$

$$= v' \sum_{i=1}^{d} a_i{}^l q_i p_i' E^\theta \left[y_{t-l} y_{t-l}' \right] v = \sum_{i=1}^{d} a_i{}^l v' q_i p_i' \Gamma_0 v$$
$$= \sum_{i=1}^{d} a_i{}^l v' \Gamma_0^{\frac{1}{2}} u_i u_i' \Gamma_0^{\frac{1}{2}} v = \sum_{i=1}^{d} a_i{}^l w' u_i u_i' w,$$

where the first equality is by Theorem 1, the second equality follows the fact that the agent's subjective model satisfies the law of iterated expectations, the third and fifth equalities are by Theorem 5(a) (or Theorem 4 depending on the assumptions), and the fourth equality is by Theorem 5(b) (or Theorem 3). On the other hand, under the true model,

$$\mathbb{E}[v' y_t v' y_{t-l}] = v' \mathbb{E}[y_t y_{t-l}'] v = v' \Gamma_l v = v' \left(\frac{\Gamma_l + \Gamma_l'}{2} \right) v = v' \Gamma_0^{\frac{1}{2}} C_l \Gamma_0^{\frac{1}{2}} v = w' C_l w.$$

To prove part (a), set $v = p_1$, which implies $v' y_t = y_t^{(1)}$ and $w = \Gamma_0^{\frac{1}{2}} p_1 = u_1$. Therefore,

$$\left| E^\theta \left[y_t^{(1)} y_{t-l}^{(1)} \right] \right| = \left| \sum_{i=1}^{d} a_i{}^l u_1' u_i u_i' u_1 \right| = \left| a_1{}^l \right| = |a_1|^l,$$

for any pseudo-true model θ. Furthermore,

$$\left| \mathbb{E} \left[y_t^{(1)} y_{t-l}^{(1)} \right] \right| = \left| u_1' C_l u_1 \right| \leq \rho(C_l) u_1' u_1 = \rho(C_l) \leq \rho(C_1)^l = |a_1|^l,$$

where the second inequality is using the assumption that the true process is exponentially ergodic, and the last equality is due to the fact that a_1 is the eigenvalue of C_1 largest in magnitude. On the other hand, by Theorem 5(b) (or Theorem 3), the variance of $y_t^{(1)}$ is the same under the true and pseudo-true d-state models. Therefore, the agent overestimates the magnitude of $y_t^{(1)}$'s autocorrelation at all lags.

In part (b), set $v = p_n$, which implies $v' y_t = y_t^{(n)}$ and $w = \Gamma_0^{\frac{1}{2}} p_n = u_n$. Thus,

$$\left| E^\theta \left[y_t^{(n)} y_{t-l}^{(n)} \right] \right| = \left| \sum_{i=1}^{d} a_i{}^l u_n' u_i u_i' u_n \right| = 0,$$

for any pseudo-true model θ, where I am using the fact that $\{u_i\}_{i=1}^n$ is an orthonormal basis and the assumption that $d < n$. Hence, the agent underestimates the magnitude of $y_t^{(n)}$'s autocorrelation at all lags, regardless of the true autocorrelation of $y_t^{(n)}$. □

Proof of Proposition E.1

I first prove a useful lemma, which offers a canonical representation of the autocorrelation matrices for stochastic processes that can be represented as in (E.1):[57]

[57] Versions of this result have previously appeared in the control and time-series literatures. For early examples, see Ho and Kálmán (1966) and Akaike (1975).

Lemma G.5. *Suppose $\{C_l\}_l$ are the autocorrelation matrices of a non-degenerate n-dimensional stationary ergodic process that can be represented as in (E.1) with $f_t \in \mathbb{R}^m$. There exists a convergent $m \times m$ matrix $\mathbb{F}$ with $\|\mathbb{F}\|_2 \leq 1$ and a semi-orthogonal $m \times n$ matrix $\mathbb{H}$ such that*

$$C_l = \mathbb{H}' \left(\frac{\mathbb{F}^l + \mathbb{F}'^l}{2} \right) \mathbb{H}. \tag{G.50}$$

Conversely, for any positive integers $m \geq n$, $m \times m$ convergent matrix $\mathbb{F}$ with $\|\mathbb{F}\|_2 \leq 1$, and semi-orthogonal $m \times n$ matrix $\mathbb{H}$, there exists an n-dimensional stationary ergodic process with autocorrelation matrices $\{C_l\}_l$ of the form (G.50), which can be represented as in (E.1).[58]

Proof. The assumption that the process is non-degenerate requires $m \geq n$, an assumption I maintain throughout the first part of the proof. Given representation (E.1), the autocovariance matrices are given by

$$\Gamma_l = \mathbb{E}\left[y_t y_{t-l}'\right] = H'F^l \mathbb{E}\left[f_{t-l} f_{t-l}'\right] H = H'F^l V H,$$

where $V \equiv \mathbb{E}\left[f_t f_t'\right]$ is the unique solution to the following discrete-time Lyapunov equation:

$$V = FVF' + \Sigma, \tag{G.51}$$

and Σ is the variance-covariance matrix of ϵ_t. Therefore,

$$C_l = (H'VH)^{\frac{-1}{2}} \left(\frac{H'F^l VH + H'VF'^l H}{2} \right) (H'VH)^{\frac{-1}{2}}.$$

Matrix V is positive semidefinite; it is positive definite if the representation in (E.1) is minimal.[59] Without loss of generality, I assume that that is the case. Define $\mathbb{H}' \equiv (H'VH)^{\frac{-1}{2}} H'V^{\frac{1}{2}}$ and $\mathbb{F} \equiv V^{\frac{-1}{2}} F V^{\frac{1}{2}}$. Then

$$C_l = \mathbb{H}' \left(\frac{\mathbb{F}^l + \mathbb{F}'^l}{2} \right) \mathbb{H}. \tag{G.52}$$

Note that since F is a convergent matrix, so is $\mathbb{F}$. Substituting $\mathbb{F} = V^{\frac{-1}{2}} F V^{\frac{1}{2}}$ in equation (G.51), I get

$$1 - \mathbb{F}\mathbb{F}' = V^{\frac{-1}{2}} \Sigma V^{\frac{-1}{2}}.$$

Therefore, since Σ is positive semidefinite, the spectral radius of $\mathbb{F}\mathbb{F}'$ is weakly smaller than one. This implies that $\|\mathbb{F}\|_2 \leq 1$. On the other hand,

$$\mathbb{H}'\mathbb{H} = (H'VH)^{\frac{-1}{2}} H'VH \, (H'VH)^{\frac{-1}{2}} = I.$$

That is, $\mathbb{H}$ is a (full-rank) semi-orthogonal matrix. This proves the first part of the lemma.

I next argue that given a convergent matrix $\hat{\mathbb{F}} \in \mathbb{R}^{m \times m}$ with $\|\hat{\mathbb{F}}\|_2 \leq 1$ and a semi-orthogonal matrix $\hat{\mathbb{H}} \in \mathbb{R}^{m \times n}$ with $m \geq n$, there exists a stationary ergodic process such that the corresponding autocorrelation matrices are given by (G.52) with $\mathbb{F} = \hat{\mathbb{F}}$ and $\mathbb{H} = \hat{\mathbb{H}}$. Given any such $\hat{\mathbb{F}}$ and $\hat{\mathbb{H}}$, let $F = \hat{\mathbb{F}}$, $H = \hat{\mathbb{H}}$, and $\Sigma = I - \hat{\mathbb{F}}\hat{\mathbb{F}}'$. The solution to the Lyapunov equation (G.51) is then given by $V = I$. Therefore, $\mathbb{F} = F = \hat{\mathbb{F}}$ and $\mathbb{H} = \hat{\mathbb{H}}(\hat{\mathbb{H}}'\hat{\mathbb{H}})^{\frac{-1}{2}} = \hat{\mathbb{H}}$, where in the last equality I am using the assumption of semi-orthogonality of $\hat{\mathbb{H}}$. By construction, then the autocorrelation matrices of a process of the form (E.1) with matrices F, H, and Σ as above are given by (G.52) with $\mathbb{F} = \hat{\mathbb{F}}$ and $\mathbb{H} = \hat{\mathbb{H}}$. □

[58] Matrix $\mathbb{H} \in \mathbb{R}^{m \times n}$ is semi-orthogonal if $\mathbb{H}'\mathbb{H} = I$, where I denotes the $n \times n$ identity matrix.

[59] See, for instance, Akaike (1975).

Proof of Proposition E.1. By Lemma G.5, $C_l = \frac{1}{2}\mathbb{H}'\left(\mathbb{F}^l + \mathbb{F}'^l\right)\mathbb{H}$, where $\mathbb{H}' \equiv (H'VH)^{\frac{-1}{2}} H'V^{\frac{1}{2}}$, $\mathbb{F} \equiv V^{\frac{-1}{2}}FV^{\frac{1}{2}}$, and $V \equiv \mathbb{E}\left[f_t f_t'\right]$ is the variance-covariance of f_t. Note that since the variance-covariance of f_t is normalized to be the identity matrix, $V = I$, $\mathbb{F} = F$, and $\mathbb{H} = H$. Recall that vector y_t does not contain any redundant observables (which are linear combinations of other observables). This assumption, together with the assumption that H is a rank-m matrix, ensures that H is an invertible $m \times m$ matrix. Therefore, by Lemma G.5, $\mathbb{H} = H$ is an orthogonal matrix. Thus,

$$\rho(C_l) = \rho\left(\mathbb{H}'\left(\frac{\mathbb{F}^l + \mathbb{F}'^l}{2}\right)\mathbb{H}\right) = \rho\left(\frac{\mathbb{F}^l + \mathbb{F}'^l}{2}\right) = \rho\left(\frac{F^l + F'^l}{2}\right) \tag{G.53}$$

for all l. But since the spectral radius of a symmetric matrix equals its spectral norm,

$$\rho\left(\frac{F^l + F'^l}{2}\right) = \left\|\frac{F^l + F'^l}{2}\right\|_2 \leq \frac{1}{2}\left\|F^l\right\|_2 + \frac{1}{2}\left\|F'^l\right\|_2 = \left\|F^l\right\|_2 \leq \|F\|_2^l . \tag{G.54}$$

Therefore, $\rho(C_l) \leq \|F\|_2^l$. On the other hand, by equations (G.53) and (G.54), $\rho(C_1) = \frac{1}{2}\|F + F'\|_2 = \|F\|_2$, where the second equality is by assumption. Thus, $\rho(C_l) \leq \|F\|_2^l = \rho(C_1)^l$, and the process is exponentially ergodic. □

Proof of Proposition E.2

I first state and prove a useful lemma:

Lemma G.6. *Suppose C_1 has a unique and simple eigenvalue λ with $|\lambda| = \rho(C_1) > 0$, and let u denote the corresponding eigenvector normalized to have $u'u = 1$.*[60] *If $u'C_2u > \rho(C_1)^2$, then the agent's forecasts in any pseudo-true one-state model are given by* (4) *with a tuple (a, η, p, q) such that $\eta > 0$.*

Proof. Define $C(a, \eta)$ as in the proof of Lemma G.3. As in the proof of Lemma G.3, I present the argument under the assumption that the largest eigenvalue of $C(a, \eta)$ is simple at the point (a^*, η^*) that maximizes $\lambda_{\max}(C(a, \eta))$.[61] I start by proposing a candidate solution to the problem of maximizing $\lambda_{\max}(\Omega(a, \eta))$ at which $\eta = 0$ and argue that the candidate does not satisfy the necessary first-order optimality conditions. Setting $\eta = 0$ in equations (G.30) and (G.31), I get

$$\left.\frac{\partial\lambda_{\max}(\Omega(a,\eta))}{\partial a}\right|_{\eta=0} = -2a + 2u'_{\max}(aC_1)C_1u_{\max}(aC_1),$$

$$\left.\frac{\partial\lambda_{\max}(\Omega(a,\eta)))}{\partial \eta}\right|_{\eta=0} = 2a^2 - 2(1+a^2)\lambda_{\max}(aC_1) + 2a^2u'_{\max}(aC_1)C_2u_{\max}(aC_1),$$

where I am using the fact that $C = aC_1$ when $\eta = 0$. Any solution to $\partial\lambda_{\max}(\Omega(a,\eta))/\partial a|_{\eta=0} = 0$ satisfies $a = \lambda$, where $\lambda = \lambda_{\min}(C_1)$ if $\lambda_{\max}(C_1) \leq 0$, $\lambda = \lambda_{\max}(C_1)$ if $\lambda_{\min}(C_1) \geq 0$, and $\lambda \in \{\lambda_{\max}(C_1), \lambda_{\min}(C_1)\}$ otherwise. Evaluating $\lambda_{\max}(\Omega(a, \eta))$ at $a = \lambda$ and $\eta = 0$, I get $\lambda_{\max}(\Omega(\lambda, 0)) =$

[60] The assumption that λ is unique and simple is not necessary for the result. The result generalizes to arbitrary matrices C_1 with $\rho(C_1) \neq 0$ by replacing $u'C_2u$ with the maximum of $u'C_2u$ over all unit-norm eigenvectors u of C_1 with eigenvalues λ such that $|\lambda| = \rho(C_1)$.

[61] See footnote 53 for how the argument can be generalized.

λ^2. Therefore, for the solution $(a, \eta) = (\lambda, 0)$ to the first-order condition $\partial\lambda_{\max}(\Omega(a,\eta))/\partial a = 0$ to be a maximizer of $\lambda_{\max}(\Omega(a,\eta))$, it must be the case that λ is the eigenvalue of C_1 largest in magnitude and $u = u_{\max}(aC_1)$ is a corresponding eigenvector normalized such that $u'u = 1$. Substituting in the expression for $\partial\lambda_{\max}(\Omega(a,\eta))/\partial\eta|_{\eta=0}$, I get

$$\left.\frac{\partial\lambda_{\max}(\Omega(a,\eta))}{\partial\eta}\right|_{a=\lambda,\eta=0} = 2\rho(C_1)^2\left(u'C_2u - \rho(C_1)^2\right) > 0,$$

where the inequality follows the assumption that $u'C_2u > \rho(C_1)^2$. This implies that the pair $\eta = 0$ and $a = \lambda$ does not constitute a local maximizer of $\lambda_{\max}(\Omega(a,\eta))$. Since this pair is the only candidate with $\eta = 0$ that may satisfy the first-order conditions, in any pseudo-true one-state model, $\eta > 0$. This establishes the lemma. □

Proof of Proposition E.2. Let σ^2 denote the variance of y_t. By the argument in the proof of Lemma G.5, the lag-l autocorrelation of y_t is given by

$$C_l = \mathbb{H}'\left(\frac{\mathbb{F}^l + \mathbb{F}'^l}{2}\right)\mathbb{H},$$

where $\mathbb{F} \equiv V^{\frac{-1}{2}}FV^{\frac{1}{2}}$, $\mathbb{H}' \equiv (H'VH)^{\frac{-1}{2}}H'V^{\frac{1}{2}}$, and V is the solution to the discrete-time Lyapunov equation (G.51). Since F and Σ are diagonal matrices, so is V. Therefore, $\mathbb{F} = F$. On the other hand, by Lemma G.5, $\mathbb{H}$ is a semi-orthogonal matrix. Therefore, $\mathbb{H}'\mathbb{H} = 1$, and so, $C_l = \sum_{i=1}^m w_i\alpha_i^l$, where $w_i \equiv \mathbb{H}_i^2 \geq 0$, $\sum_{i=1}^m w_i = 1$, and α_i is the ith diagonal element of F. That is, $C_l^{\frac{1}{l}}$ is equal to the weighted l-norm of vector $(\alpha_1, \ldots, \alpha_m)$ with weights $w = (w_1, \ldots, w_m)$.

Since the representation in (E.1) is minimal, $w_i > 0$ for all i, and all α_i are distinct. If that were not the case, there would exist some $\tilde{m} < m$ such that $C_l = \sum_{i=1}^{\tilde{m}}\tilde{w}_i\tilde{\alpha}_i^l$ for some non-negative weights $\tilde{w}_i$ that sum up to one and some $\tilde{\alpha}_i \in (-1,1)$. Consider the process $\widetilde{\mathbb{P}}$ represented as in (E.1) with $F = \text{diag}(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{\tilde{m}})$, $\epsilon_t \sim \mathcal{N}(0,\Sigma)$, $\Sigma = I - FF'$, and $H = \sigma\,\text{diag}(\sqrt{\tilde{w}_1}, \ldots, \sqrt{\tilde{w}_{\tilde{m}}})$. By the argument in the proof of Lemma G.5, $\widetilde{\mathbb{P}}$ has the same autocorrelation matrices as $\mathbb{P}$. Moreover, both $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ are mean-zero and normal and both have variance σ^2. Therefore, $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ are observationally equivalent, a contradiction to the assumption that the representation I started with was minimal.

Next, note that, by the generalized mean inequality, $C_l^{\frac{1}{l}} > C_1$ for all $l \geq 2$, where the strictness of the inequality follows the facts that $w_i > 0$ for all i and all α_i are distinct. In particular, $u'C_2u = C_2 > C_1^2 = \rho(C_1)^2$, where I am using the fact that y_t is a scalar. Thus, by Lemma G.6, $\eta > 0$. To see why $\eta < 1$, recall that by Theorem 2, the (a,η) pair maximizes

$$\Omega(\tilde{a},\tilde{\eta}) = -\frac{\tilde{a}^2(1-\tilde{\eta})^2}{1-\tilde{a}^2\tilde{\eta}^2} + \frac{2(1-\tilde{\eta})(1-\tilde{a}^2\tilde{\eta})}{1-\tilde{a}^2\tilde{\eta}^2}\sum_{\tau=1}^{\infty}\tilde{a}^\tau\tilde{\eta}^{\tau-1}C_\tau.$$

But $\Omega(\tilde{a},1) = 0 < C_1^2 = \Omega(C_1,0)$ for any $\tilde{a}$. Therefore, $\eta = 1$ cannot be part of the description of a pseudo-true one-state model. Finally, $a \in (1,1)$ by Lemma G.3. The proposition then follows Theorem 2 by noting that $qp' = 1$ whenever y_t is a scalar. □