"Small Data": Inference with Occasionally Observed States*

Alexandros Gilch† Andreas Lanz‡ Philipp Müller§ Gregor Reich¶
Ole Wilms‖**

January 2025

Abstract

We study the estimation of dynamic economic models for which some of the state variables are observed only *occasionally* by the econometrician—a common problem in many fields, ranging from marketing to finance to industrial organization. If those occasional state observations are serially correlated, the likelihood function of the model becomes a high-dimensional integral over a nonstandard domain. We generalize the recursive likelihood function integration procedure (RLI; Reich, 2018) to incorporate the occasional observations, enabling likelihood-based inference in such estimation problems. In extensive Monte Carlo studies, we demonstrate the favorable properties of the proposed method for identifying all model parameters and compare it to alternative methods.

Keywords: maximum likelihood estimation, occasional state observations, recursive likelihood function integration, interpolation, numerical quadrature, Markov models, dynamic discrete choice models, long-run risk models.

*We thank Jaap Abbring, Einar Breivik, Max Diegel, Frank de Jong, Philipp Eisenhauer, Robert Erbe, Joachim Freyberger, Joris Gillis, Bo Honoré, Fedor Iskhakov, Kenneth Judd, Tobias Klein, Felix Kübler, Jasmin Maag, John Rust, Karl Schmedders, Dominik Wied, Bas Werker, Tom Zimmermann, seminar participants at the University of Cologne, NHH Bergen, the University of Bonn, Tilburg University, WHU Vallendar, and the University of Zurich, and participants in the 2020 and 2024 Econometric Society Winter Meetings and the 2020 EEA Annual Congress for helpful comments and discussions. Philipp Müller and Gregor Reich gratefully acknowledge the financial support of Kenneth Judd, Senior Fellow at the Hoover Institution. Alexandros Gilch gratefully acknoledges financial support by the Collaborative Research Center Transregio 224, funded by the German Research Foundation (DFG).

†Institute of Finance and Statistics, University of Bonn, Adenauerallee 24-26, 53113 Bonn, Germany. Email: alexandros.gilch@uni-bonn.de

‡Faculty of Business and Economics, Peter Merian-Weg 6, 4002 Basel, Switzerland. Email: andreas.lanz@unibas.ch

§Department of Business Administration, University of Zurich, Plattenstrasse 14, 8032 Zurich, Switzerland. Email: philipp.mueller@business.uzh.ch

¶Tsumcor Research AG, Sonnenbergstrasse 74, 8603 Schwerzenbach, Switzerland. Email: gregor.reich@tsumcor.ch

‖Department of Economics, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany. Email: ole.wilms@uni-hamburg.de

**Department of Finance, Tilburg University, PO Box 90153, 5000 LE Tilburg, the Netherlands.

1 Introduction

A ubiquitous problem in the estimation of dynamic economic models is that the econometrician does not have access to data for all the state variables of the model. In such a regime, the likelihood function of the parameters forms an integral over the unobserved states; this integral has—if those states are *serially correlated*—a dimensionality that is proportional to the time horizon of the dataset. While some estimation approaches exist to cope with this large integral, they are usually limited to the special case in which the state is either fully observed or fully unobserved. In this paper, we study regimes where the state is observed infrequently, which we refer to as an *occasionally observed state*. This type of observability can take many forms: states can be observed randomly, endogenously (i.e., tied to a particular realization of some other model state or action), or follow a regular time pattern; moreover, states that are observed in a time-aggregated fashion can be modeled as occasional state observations too. All these imperfect observability forms occur across fields, ranging from marketing to finance to industrial organization. And their causes are diverse, and include confidentiality or privacy issues and immeasurability.

Since such observability problems occur in many fields, there exists an emerging—but still relatively recent—body of literature covering the inherent unobservability of serially correlated states and its implications for estimation (e.g., Cosslett and Lee, 1985; Kitagawa, 1987; Keane, 1994; Norets, 2009; Arcidiacono and Miller, 2011; Blevins, 2016; Reich, 2018).[1] All these approaches assume a particular state to be either fully observed over time, or fully unobserved over time—but nothing in between. Often, however, managerial problems are characterized by occasional observations of such serially correlated states, including in marketing (e.g., prices in scanner data; Erdem et al., 1999), labor and health economics (e.g., health status in retirement choices; Iskhakov, 2010), finance (e.g., transaction data; Engle and Russell, 1998), and industrial organization (e.g., steel trading and inventory optimization; Hall and Rust, 2021).[2]

Erdem et al. (1999) demonstrate the importance of appropriately incorporating such occasional observations in the estimation procedure itself—as opposed to just imputing the missing values by, say, the mean of their occasional observations—and how conclusions might be misleading otherwise: in a model of product purchases, they estimate consumers' price elasticity based on scanner data, which contains daily product prices only if actual transactions have taken place. However, since the decision to buy likely depends on the price faced by the consumer, imputing prices from their average observed values creates a selection bias in these estimates because, in fact, a *conditional* mean price was used for the imputation (conditional on having made at least one purchase). The authors use maximum simulated likelihood to integrate out the missing observations, obtaining an unbiased estimator for the price elasticity. Their approach, however, does not incorporate any serial correlation in prices, which is quite a limiting

[1] More recently, Connault (2016) and Farmer (2021) have addressed similar issues in discrete state contexts as in Cosslett and Lee (1985), though in different domains.

[2] A related problem appears in the domain of *partially observable Markov decision processes* (POMDP), where some states are fully or partially unobservable to the decision maker. Whether or not the econometrician observes the same set of states (up to the error terms) as the agent or less, is, however, an independent question. See, for example, Chang et al. (2020) for a recent treatment of POMDP estimation in a similar context as our second application.

assumption.

Our contribution is to enable likelihood-based inference even under occasional state observations by generalizing the recursive likelihood function integration (RLI) method of Reich (2018)—an approach originally developed to estimate models with one or more states being completely unobserved. The RLI method provides a recursive approach to integrating out the serially correlated unobserved state variables in the likelihood function. In the presence of occasionally observed states, this integration cannot, however, be carried out directly because its domain would be nonstandard[3]—a lower-dimensional sub-manifold in particular, and thus a set of measure zero in the original state space. Therefore, we derive a generalized recursive formulation of the integral, which yields a series of interconnected, lower-dimensional integration problems over standard domains. The latter can then be approximated by alternating highly efficient quadrature and interpolation methods, such as Gauss quadrature rules and splines, respectively. The proposed method nests the full observation and the no-observation regimes as special cases, and applies to various dynamic state-space models, including instances where occasional state observations occur due to time aggregation of the data.

Two issues arise when performing likelihood-based inference under occasional state observations: Firstly, the observation pattern of the occasionally observed state might be endogenous, i.e., the probability to observe a state in some period t is not independent of the realization of the state in period t. Constructing the likelihood purely from the model ignores this possible dependence in the available sample, hence making the likelihood invalid and the estimator possibly biased, even before any approximation of the likelihood. To avoid this issue, we show how to construct a generally correct likelihood, including the case of endogeneity, if the conditional probability of observation is known. Furthermore, we provide conditions under which a simplified likelihood purely based on the model is also valid. Secondly, for both, the true and the approximate estimator, the validity of large sample properties is not ex ante innate. However, the companion paper Gilch et al. (2025) derives conditions under which both the true and approximate estimators examined in this paper are consistent and asymptotically normal.

Recently, an alternative approach to estimate dynamic models with occasionally observed states has been developed by Hall and Rust (2021), one that proposes estimating partially observed Markov processes using the simulated method of moments (SMM), explicitly allowing for serial correlation of the occasionally observed variables:[4] First, the authors forward simulate the full process for a particular set of parameter values; second, they censor the simulated data using the same pattern as in the original dataset; finally, they compare the moments of the so obtained simulated data with the empirical ones. Based on the (appropriately weighted) difference in moments, a new set of parameter values is chosen, and the procedure is repeated until the difference is minimal. The authors formally derive conditions for the consistency and asymptotic normality of their estimator. Given the estimator's roots in the moment-based methodology, it is broadly applicable (in fact, more broadly than likelihood-based approaches),

[3]We refer to an integration domain as "nonstandard" if neither a change of variables exists to map it to the unit hypercube nor a specialized quadrature rule is available for that domain.

[4]A note on nomenclature: We refer to the phenomenon in question as *occasionally* observed—in contrast to Hall and Rust's 2021 notion of *endogenously* sampled—to explicitly include cases in which states are periodically or randomly observed, but not necessarily tied to a particular decision or any other endogenous outcome of the model. We formalize the allowable observation patterns below.

but we expect—and confirm—it to be less efficient than likelihood function-based estimation.

We compare this approach to the performance of our method, which we assess, demonstrating its broad applicability by applying it to three relevant problems from the literature that cover all forms of unobservability: (i) a long-run risk model as in Bansal and Yaron (2004); (ii) a model with stochastic volatility as in Schorfheide et al. (2018)—where both applications feature one completely unobserved state and one that is defined quarterly in the model but observed only aggregated over a whole year—and (iii) a hypothetical setup of the bus engine replacement problem of Rust (1987), a dynamic discrete choice model where the econometrician only observes the state variable in the rare case of an engine replacement. Such a setup can occur, for example, if the decision-maker outsources the actual replacement operation and the third-party contractor wants to infer on the decision-maker's implied demand for replacement as a function of replacement costs, using a structural model based on his own incomplete data.

We analyze these examples in extensive Monte Carlo studies: First, we simulate many complete datasets for various sample lengths, while discarding some observations to obtain the corresponding datasets with occasional observations. Then, we estimate the models based on both types of datasets, and compare the distributions for the different estimators, and for the different sample lengths. Our study yields two insights: (i) the approach is computationally highly efficient and thus makes likelihood-based inference even in the presence of serially correlated occasionally observed states feasible; and (ii) we provide simulation-based evidence for the statistical efficiency of the likelihood-based estimator even in small samples. Finally, we compare our approach to the SMM estimator for endogenously sampled time series of Hall and Rust (2021), showing that our method can yield significant efficiency gains, particularly for small samples and non-linear models. In fact, we confirm Grammig and Küchlin (2018) in finding that it is very hard to identify the parameters of a highly persistent stochastic volatility process using SMM, and we show that this is not an issue with likelihood-based approaches like ours.

The remainder of this paper is organized as follows: In Section 2, we show how to compute the likelihood for (discrete-time) Markov processes with occasional state observations. The section begins with a motivating example highlighting three challenges, endogeneity of the observation process, high-dimensional integrals and validity of large sample properties, which we each address in the subsequent subsections. In Section 3.1 we apply our method to estimate the long-run risk and stochastic volatility models; in Section 3.2 we apply our method to estimate the dynamic discrete choice model of Rust (1987). Section 4 concludes.

2 Maximum likelihood estimation with occasionally observed states

In this section, we show how to estimate structural models with states that are observed infrequently—i.e., only at particular values of the states or decisions, periodically, or completely at random—using recursive likelihood function integration (RLI). We refer to this setting as *occasionally observed states*. While there are many empirical applications involving occasionally observed states, common estimation methods assume that states are either fully observed or

completely unobserved. We propose a new methodology to estimate such models and show that fully observed states and unobserved states are nested as special cases.

We proceed as follows: We begin with an introductory example in Section 2.1, where we demonstrate the main challenges that arise when estimating dynamic models with occasionally observed states using maximum likelihood estimation. This includes (i) potential endogeneity of the observation process that needs to be taken into account when formulating the likelihood function, (ii) computational challenges arising from the high dimensional integrals that are induced by the marginalization of the unobserved states, and (iii) the validity of large sample properties of the approximates likelihood estimator. The subsequent subsections then each cover one of the problems encountered in the introductory example: In Section 2.2, we outline how to appropriately treat generic observation patterns in the likelihood function to ensure validity of our approach; we demonstrate that observability is not limited to fixed, regular time intervals, but can depend on the realization of certain states or controls of the Markov process itself. Section 2.3 contains our main methodology, and we show how to estimate structural models with occasionally observed states based on an efficient recursive approximation of the likelihood function. Section 2.4 briefly discusses large sample properties of our estimator which are formally treated in the companion paper Gilch et al. (2025).

For the sake of argument, we limit ourselves to one observed and one occasionally observed state in the following. It is important to highlight that this is by no means a limitation of our method but rather lets us use very instructive notation. We provide a fully general description of our methodology for an arbitrary number of states in Appendix A.2.

2.1 Introductory example

Consider a discrete-time Markov process $\{z_t, x_t\}$—possibly controlled, like in dynamic discrete choice models—with two one-dimensional state variables, $z_t, x_t \in \mathbb{R}$, and a parametric family of transition probability functions, $P(z_t, x_t | z_{t-1}, x_{t-1}; \theta)$. We want to estimate the model parameter θ using a maximum likelihood approach. In particular, we are interested in a case of limited data availability, where the variable z_t is observed for all periods $t \in \mathcal{T} \equiv \{1, \dots, T\}$ of the sample, whereas x_t is observed only at the times $t \in \bar{\mathcal{T}}$ with $\bar{\mathcal{T}} \subseteq \mathcal{T}$. In this subsection, we analyze the simple case with only one observation, $\bar{\mathcal{T}} = \{\bar{t}\}$ and illustrate three challenges for our estimation: endogeneity of the observation process, computation of high-dimensional integrals, and the large sample properties of the resulting estimator. The subsequent sections then each deal with one of these issues in more detail, and cover the general observation set $\bar{\mathcal{T}} \subseteq \mathcal{T}$.

To introduce basic notation and the fundamental treatment of unobserved states, let us first consider two counterfactual cases: Under full observability for both states x_t and y_t—i.e., $\bar{\mathcal{T}} = \mathcal{T}$—the (unconditional) likelihood function of the parameter vector θ reads

$$\begin{aligned} L(\theta) &= P(\{z_t, x_t\}_{t \in \mathcal{T}}; \theta) \\ &= P(z_1, x_1; \theta) \prod_{t=2}^{T} P(z_t, x_t | z_{t-1}, x_{t-1}; \theta), \end{aligned} \tag{1}$$

where $P(z_1, x_1; \theta)$ is the stationary distribution of x_t (if available). Conversely, if no state observations on x_t are available—i.e., $\bar{\mathcal{T}} = \emptyset$—the likelihood function forms an integral with respect to the unobserved state,

$$\begin{aligned} L(\theta) &= P(\{z_t\}_{t\in\mathcal{T}}; \theta) \\ &= \int \cdots \int_{\mathcal{S}_x^T} P(z_1, \tilde{x}_1; \theta) \prod_{t=2}^{T} P(z_t, \tilde{x}_t | z_{t-1}, \tilde{x}_{t-1}; \theta) \, \mathrm{d}(\tilde{x}_1, \ldots, \tilde{x}_T) \,. \end{aligned} \tag{2}$$

Here and in the following, we decorate any integration variable with a tilde; in (2), we write $\tilde{x}_t$ to clearly distinguish them from any data set element or state variable, x_t. Note that the overall dimensionality of the integral in (2) is proportional to the time horizon of the data, T. Thus, computing this integral constitutes a delicate task.

Suppose we have a single observation $x_{\bar{t}}$ at $\bar{t}$ that lies in the "interior" of $\mathcal{T}$—i.e., $1 < \bar{t} < T$ and $\bar{\mathcal{T}} = \{\bar{t}\}$. If we were to integrate the likelihood as in (2), the domain of integration in the likelihood function would read $\{(\tilde{x}_1, \ldots, \tilde{x}_T) \in \mathcal{S}_x^T : \tilde{x}_{\bar{t}} = x_{\bar{t}}\}$, which is no longer a full-dimensional subset of $\mathcal{S}_x^T$ (for general state spaces $\mathcal{S}_x$), and thus potentially creates ill-defined integrals. Therefore, we rewrite the integral to explicitly exclude the integration variable $\tilde{x}_{\bar{t}}$ and only integrate w.r.t. the unobserved states $\tilde{x}_t$ for $t \in \mathcal{T} \setminus \bar{\mathcal{T}}$:

$$\begin{aligned} L(\theta) = \int \cdots \int & P(z_1, \tilde{x}_1; \theta) \left(\prod_{t=2}^{\bar{t}-1} P(z_t, \tilde{x}_t | z_{t-1}, \tilde{x}_{t-1}; \theta) \right) P(z_{\bar{t}}, x_{\bar{t}} | z_{\bar{t}-1}, \tilde{x}_{\bar{t}-1}; \theta) \\ & \cdot P(z_{\bar{t}+1}, \tilde{x}_{\bar{t}+1} | z_{\bar{t}}, x_{\bar{t}}; \theta) \left(\prod_{t=\bar{t}+2}^{T} P(z_t, \tilde{x}_t | z_{t-1}, \tilde{x}_{t-1}; \theta) \right) \mathrm{d}(\tilde{x}_1, \ldots, \tilde{x}_{\bar{t}-1}, \tilde{x}_{\bar{t}+1}, \ldots, \tilde{x}_T) \end{aligned} \tag{3}$$

In the following, we use the likelihood (3) to illustrate three key challenges that arise when estimating dynamic models with occasional state observations.[5]

(i) Observation process The fact that we observe x_t in some periods, but we don't in others, is not innocuous, in particular if the (un-)observability of a variable inherently depends on the value of the variable itself. This, in turn, implies that the observation pattern carries information about realization of the underlying variable—even if it is unobserved. Ignoring this information generally leads to endogeneity, which needs to be accounted for by adapting the integration domain and the (conditional) distribution for the unobserved x_t accordingly.[6]

To maintain the simplicity of this example, suppose there is an endogenous observation mechanism such that x_t is observed if $x_t \in \tilde{\mathcal{S}}$, for some subset of the state space $\tilde{\mathcal{S}} \subset \mathcal{S}_x$, but not otherwise. A prominent example is *censoring*: here, x_t is only observed if it exceeds—or falls below—some threshold. To further simply the outline, consider a situation where $T = 2$ and $\bar{t} = 1$, i.e., we have a realization of this mechanism for which we know that $x_1 \in \tilde{\mathcal{S}}$ and

[5] The likelihood (3) bares a close resemblance to the Chapman-Kolmogorov equation. For $\bar{\mathcal{T}} = \{1, T\}$ the function $L(\theta)$ is defined as $P(x_1, x_T, \{z_t\}_{t\in\mathcal{T}}; \theta)$, hence dividing by $P(z_1, x_1; \theta)$ returns exactly the Chapman-Kolmogorov equation, $P(x_T, \{z_t\}_{t=2}^T | x_1, z_1) = \int_{\mathcal{S}_x^{T-2}} \prod_{t=2}^{T-1} P(z_t, \tilde{x}_t | z_{t-1}, \tilde{x}_{t-1}; \theta) \, \mathrm{d}(\tilde{x}_2, \ldots, \tilde{x}_{T-1})$, which marginalizes the unobserved states $\tilde{x}_2, ..., \tilde{x}_{T-1}$ between the unobserved ones.

[6] We thank the referee for pointing out this issue, which was only informally discussed in previous versions of the paper.

$x_2 \in \mathcal{S}_x \setminus \tilde{\mathcal{S}}$. Note that if we were to integrate out the unobserved variable over the full state space $\mathcal{S}_x$, we would include values from $\tilde{\mathcal{S}}$ which are impossible to have happened, conditional on non-observation of x_2. To account for the endogeneity in the likelihood function, we include the probability to observe x_t conditional on x_t itself:

$$L(\theta) = \int_{\mathcal{S}_x} P(z_1, x_1; \theta) P(x_1 \text{ is observed} | x_1) P(z_2, \tilde{x}_2 | z_1, x_1; \theta) P(\tilde{x}_2 \text{ is observed} | \tilde{x}_2) \mathrm{d}\tilde{x}_2.$$

In our example, the observation probability turns out to be an indicator function,

$$P(x_t \text{ is observed} | x_t) = \mathbf{1}_{\{x_t \in \tilde{\mathcal{S}}\}}. \tag{4}$$

Thus, we can simplify the likelihood by transforming the integral over the domain $\mathcal{S}_x \setminus \tilde{\mathcal{S}}$:

$$L(\theta) = \int_{\mathcal{S}_x \setminus \tilde{\mathcal{S}}} P(z_1, x_1; \theta) P(z_2, \tilde{x}_2 | z_1, x_1; \theta) \mathrm{d}\tilde{x}_2. \tag{5}$$

Note that the likelihood function (5) neither that for the model under full observations (which would feature no integral at all), nor that for the model with no observation of the state variable x at all (which would integrate over $\mathcal{S}_x$).

Importantly, given our sample and the knowledge that observation is endogenous in our example, only this adjusted likelihood is correct for inference. In Section 2.2, we return to the case with any $T \in \mathbb{N}$ and provide a general likelihood formulation, which accounts for multiple observations from a (potentially) endogenous observation process, covering different levels of endo- or exogeneity of the observation pattern $\bar{T}$. The only requirement is that we can express it through a parametric family of distributions; moreover, we show under which circumstances the likelihood original (3) can be used"as-is".

(ii) High-dimensional integrals Computing the integral in Equation (3) is numerically challenging, as its dimension grows proportional with the time horizon T. In the following we show how to rewrite the integral in (3) recursively. As we demonstrate in the methodology sections below, we can then form an approximation of that recursive form, which effectively breaks the curse of dimensionality otherwise induced by the time horizon.

Due to the integrability and boundedness assumption on all Ps, the Fubini–Tonelli theorem and the Markov structure of the model allow us to express the second integral in Equation (3) (i.e., the integral from $\bar{t}+2$ up to T) recursively as

$$f_t^\theta(x) = \begin{cases} 1 & t > T \\ \int P(z_t, \tilde{x} | z_{t-1}, x; \theta) f_{t+1}^\theta(\tilde{x}) \mathrm{d}\tilde{x} & \bar{t}+2 \leq t \leq T, \end{cases}$$

where z_t and z_{t-1} come from the dataset, but x and $\tilde{x}$ are function arguments and integration variables, respectively. Using the result of this recursion, $f_{\bar{t}+2}^\theta$, the integration over $\bar{t}+1$ yields the following *constant* function, which depends only on the state observation $x_{\bar{t}}$, but not on the function argument x:

$$f_{\bar{t}+1}(x) = \int P(z_{\bar{t}+1}, \tilde{x} | z_{\bar{t}}, x_{\bar{t}}; \theta) f_{\bar{t}+2}^\theta(\tilde{x}) \mathrm{d}\tilde{x}.$$

Defining

$$f_{\bar{t}}(x) = P(z_{\bar{t}}, x_{\bar{t}} | z_{\bar{t}-1}, x; \theta) f_{\bar{t}+1}(x_{\bar{t}}),$$

we can express the remaining dimensions of the integral through recursion, as

$$f_t^\theta(x) = \int P(z_t, \tilde{x} | z_{t-1}, x; \theta) f_{t+1}^\theta(\tilde{x}) \mathrm{d}\tilde{x} \quad 2 \le t \le \bar{t} - 1,$$

and evaluate the final likelihood based on f_2^θ.

However, evaluating f_2^θ is still computationally challenging, as an evaluation of f_2^θ triggers an evaluation of f_3^θ, which again triggers an evaluation of f_4^θ, and so on. As each evaluation level t requires multiple evaluations at $t+1$ to compute the integrals, the resulting number of function evaluations grows exponentially with T. In Section 2.3 we show how to use an approximation of the recursion f_t^θ so that the computational complexity only grows linearly in T and hence becomes computationally feasible; together with our results on convergence of the resulting estimator in Section 2.4, this formulation can be shown to truly break the curse of dimensionality.

(iii) Asymptotic properties In the standard setup with fully observed x_t, asymptotic properties of the likelihood estimator are obtained by taking the logarithm of the likelihood (1),

$$\log L(\theta) = \log P(z_1, x_1; \theta) + \sum_{t=2}^{T} \log P(z_t, x_t | z_{t-1}, x_{t-1}; \theta).$$

Taking the sample size T to infinity, consistency is derived using a law of large numbers and asymptotic normality follows from a central limit theorem.

In the case of occasional state observations, this approach is not applicable: Taking the logarithm of the likelihood (3) does not yield a sum over t summands, but rather a sum of two summands only, as the logarithm and the integral cannot be interchanged:

$$\begin{aligned} \log L(\theta) = \log \int \cdots \int P(z_1, \tilde{x}_1; \theta) \left(\prod_{t=2}^{\bar{t}-1} P(z_t, \tilde{x}_t | z_{t-1}, \tilde{x}_{t-1}; \theta) \right) P(z_{\bar{t}}, x_{\bar{t}} | z_{\bar{t}-1}, \tilde{x}_{\bar{t}-1}; \theta) \, \mathrm{d}(\tilde{x}_1, \ldots, \tilde{x}_{\bar{t}-1}) \\ + \log \int \cdots \int P(z_{\bar{t}+1}, \tilde{x}_{\bar{t}+1} | z_{\bar{t}}, x_{\bar{t}}; \theta) \left(\prod_{t=\bar{t}+2}^{T} P(z_t, \tilde{x}_t | z_{t-1}, \tilde{x}_{t-1}; \theta) \right) \mathrm{d}(\tilde{x}_{\bar{t}+1}, \ldots, \tilde{x}_T) \,. \end{aligned} \tag{6}$$

Of course, with a fixed number of observations (here: one), we do not obtain the infinite sum required to derive the desired properties of the estimator when taking T to infinity. Instead, it is only the dimension of the integral that becomes larger, and convergence is in our general framework—to the best of our knowledge—unclear. This makes asymptotic statements impossible, even if we were able to compute these integrals exactly.

However, as shown by Gilch et al. (2025) and further discussed in Section 2.4, it is possible to recover the asymptotical results known from many other log-likelihood-based estimators, if the number of occasional observations, $N \equiv |\bar{\mathcal{T}}|$, also tends to infinity as T grows. In particular, these asymptotics can be derived based on the joint probability of all states between

two observation periods, $P\left(\{z_t, x_t\}_{t=\bar{t}_i+1}^{\bar{t}_{i+1}} | z_{\bar{t}_i}, x_{\bar{t}_i}; \theta\right)$.

Based on this observation, Gilch et al. (2025) continue to show how these asymptotics can be derived if the likelihood has to be approximated, as it is typically the case in realistically-sized applications: As mentioned in Paragraph (ii), our likelihood cannot be evaluated analytically and is therefore approximated numerically. Hence, the estimator we are actually interested in is not the maximizer of (6) but the maximizer of this approximated likelihood. However, approximation of the likelihood introduces an additional deterministic error to our estimator on top of the stochastic estimation error and thus also affects its asymptotic properties. We further explain this issue and the proposed solution by Gilch et al. (2025) in Section 2.4.

2.2 Endogeneity of the observation process

We start by formalizing the assumptions on the formation of the observation pattern $\bar{\mathcal{T}}$ which are necessary to ensure validity of our approach. As we have indicated above, observability is not limited to fixed, regular time intervals, but can, in general, depend on the realization of certain states or controls of the Markov process itself. However, such dependence can lead to endogeneity if one fails to incorporate the information about the conditional distribution of the unobserved variables inherent in the observation pattern realization where necessary. As a consequence, if maximum likelihood estimation does not account for this endogeneity, it is potentially biased. In this section, we incorporate the observation pattern into the sample, and we describe different levels of susceptibility to endogeneity that it can induce. Subsequently, we derive a general likelihood function based on both the economic model and the observation pattern, and we show under which circumstances the original likelihood (i.e., the likelihood based purely on the economic model) is sufficient for consistent inference. For the ease of notation, we keep the assumption that x_t is one-dimensional. At the end of this section, we describe how the results generalize in a setting with multi-dimensional $x_t \in \mathbb{R}^{d_x}$, $d_x \geq 1$.

Let us define the *observation variable* as

$$m_t = \begin{cases} 1, & \text{if } x_t \text{ is not observed,} \\ 0, & \text{otherwise.} \end{cases}$$

If we allow the observability of x_t to be random, m_t is in fact a random variable with probability distribution P^m. This distribution is typically called the *missing data mechanism* as it specifies the probability for the value of x_t to be (un-) observed, and it may condition on past realizations of x_s, $s < t$, as well as current and past realizations of z_s and m_s, $s \leq t$.[7] In this paper, we only consider cases where P^m is Markov. Formally, we write

$$P^m\left(m_t | \{m_s\}_{s=1}^{t-1}, \{z_s, x_s\}_{s=1}^{t}; \eta\right) = P^m\left(m_t | z_t, x_t, z_{t-1}, x_{t-1}, m_{t-1}; \eta\right), \tag{7}$$

where $\eta \in \mathcal{H}$ is a nuisance parameter. To distinguish the *model process* for the model variables, x_t and z_t, from the missing data mechanism P^m for the observation variable, we denote the

[7] See, for example, Little and Rubin (2002). Note that while the missing data literature calls x_t (potentially) "missing", we call it "unobserved", in order to distinguish it from truncation and related concepts (also see our comment at the end of the section). Moreover, our definition of P^m is a slight extension of the typical definition since it allows for dependence on the past observations.

former by

$$P^{zx}\left(z_t, x_t | z_{t-1}, x_{t-1}; \theta\right), \tag{8}$$

which is specified by the underlying economic model and parametrized by $\theta \in \Theta$. In line with many applications, and to allow the application of RLI, we also impose the Markov property on P^{zx}. Note that the observation variable m_t is only relevant to the econometrician; the agent whose behavior is explained by the model observes all model variables at all times.[8] Therefore, m_t does not appear in P^{zx}.

In this paper, we address the estimation of the parameter of interest θ using maximum likelihood estimation under occasional observations of some model variables. Recall that the likelihood function is the joint probability of the sample, seen as function of the parameter of the underlying model. If the econometrician observes all variables at all times (by construction, not by chance), the model process P^{zx} together with the observed model variables, $\{z_t, x_t\}_{t=1}^T$ is sufficient for the determination of the likelihood and its maximizer. However, under occasional observations, the sample is deprived of some x_t, but their absence is itself informative about their realized (but unobserved) values—and thus about θ—as specified through P^m. Hence, the relevant sample is in fact the *information set*, $\left\{\{z_t, m_t\}_{t=1}^T, \{x_t | 1 \leq t \leq T, m_t = 0\}\right\}$, and the likelihood function needs to be augmented by P^m. To formalize this augmentation, we first introduce some helpful concepts and notation:

Definition 1 (Observation pattern)**.**

1. *The set of periods in which x_t is observed is called the* ***observation pattern,*** $\bar{\mathcal{T}}$***,*** *and is denoted by*

$$\bar{\mathcal{T}} \equiv \{t \in \{1, ..., T\} | m_t = 0\}. \tag{9}$$

   It is a random variable with support $\{0,1\}^T$.

2. *The* ***number of observations,*** N***,*** *is defined by $N \equiv |\bar{\mathcal{T}}|$ and is a random variable with support $\{0, 1, ..., T\}$.*

3. *Given $N \geq 1$, the* ***periods of observations,*** $\bar{t}_i$***,*** *are the elements of $\bar{\mathcal{T}}$ and numbered ascendingly by $i = 1, ..., N$, s.t. $\bar{t}_i < \bar{t}_{i+1}$ for all i. Each $\bar{t}_i$ is a random variable with recursively defined support conditional on N and, if $i > 1$, $\bar{t}_{i-1}$,*

$$supp(\bar{t}_i) = \begin{cases} \{1, ..., T - (N-1)\} & \text{if } i = 1, \\ \{\bar{t}_{i-1} + 1, ..., T - (N-i)\} & \text{if } i > 1. \end{cases} \tag{10}$$

   We further define $\bar{t}_0 \equiv 0$ and $\bar{t}_{N+1} \equiv T + 1$ for notational purposes and without correspondence to any observation.

4. *The* ***length of the non-observation segment between the*** $(i-1)$***-th and the*** i***-th***

[8] This is in contrast to partially observable Markov decision processes (POMDP), where the agent observes relevant states only imperfectly; also see Footnote 2.

observation, *τ_i, is defined by $\tau_i \equiv \bar{t}_i - \bar{t}_{i-1} - 1$ for $i = 1, ..., N+1$. Each τ_i is a random variable as it is the function of the two random variables $\bar{t}_i$ and $\bar{t}_{i-1}$.*

Note that Definition 1 allows for three equivalent representations of the observation pattern; we continue with the representation through m_t, but still use $N, \bar{t}_i$, and τ_i to simplify the notation of the likelihood function below.

Next, we characterize the potential endogeneity of the observation pattern, $\bar{\mathcal{T}}$, w.r.t. the model variables, x_t and z_t, in terms of P^m. Generally, the functional form (7) allows for any probabilistic Markov dependence of the observability of x_t on z_t, x_t, x_{t-1}, and z_{t-1}. The following definition restricts the functional form of P^m selectively to obtain three forms of endogeneity which are common in the missing data literature:

Definition 2 (Properties of the observation pattern)**.** *The missing data mechanism P^m is* ***missing at random*** *(MAR) if for all t*

$$P^m(m_t|z_t, x_t, m_{t-1}, z_{t-1}, x_{t-1}; \eta) = P^m(m_t|z_t, z_{t-1}, m_{t-1}; \eta), \tag{11}$$

i.e., if observation of x_t only depends on the fully observed variables. The missing data mechanism is called ***missing completely at random*** *(MCAR) if for all t*

$$P^m(m_t|z_t, x_t, m_{t-1}, z_{t-1}, x_{t-1}; \eta) = P^m(m_t|\eta), \tag{12}$$

i.e., if observation of x_t is independent of all other variables.
If P^m is neither missing at random nor missing completely at random, we call P^m ***missing not at random*** *(MNAR).*[9]

Note that MCAR implies MAR but not vice versa. Also, while an MNAR mechanism corresponds to an endogenous observation pattern, MAR and MCAR mechanisms are associated with observation patterns that are exogenous (conditional on z_{t-1}, z_t in the former case). We provide examples for all forms of endogeneity at the end of the section, and relate them to the following assumption and proposition:

Assumption A1 (Admissible scenarios for the observation pattern)**.** *One of the following two cases holds:*

(a) The missing data mechanism is MNAR and the functional form of $P^m(m_t|z_t, x_t, m_{t-1}, z_{t-1}, x_{t-1}; \eta)$ is known.

[9]Two alternative definitions of MAR and MCAR are also admissible for our setup—one that relates closer to the definition known from the missing data literature, and one that extends ours even further: For MAR, we could also keep the dependence on x_{t-1} and x_t if they are actually observed:

$$P^m(m_t|z_t, x_t, m_{t-1}, z_{t-1}, x_{t-1}; \eta) = P^m(m_t|z_t, z_{t-1}, m_{t-1}, \{x_s\}_{s \in \{t-1,t\} \cap \bar{\mathcal{T}}}; \eta).$$

For MCAR, we could also keep the Markov dependence on the previous observation variable, m_{t-1}:

$$P^m(m_t|z_t, x_t, m_{t-1}, z_{t-1}, x_{t-1}; \eta) = P^m(m_t|m_{t-1}; \eta).$$

(b) The missing data mechanism is MAR and the parameters η and θ specifying P^m and P^{zx} are distinct, i.e the joint parameter space of η and θ is the product of the parameter spaces $\mathcal{H}$ and Θ.

Based on Assumption A1, and using a set of common notational conventions to achieve a compact notation,[10] the following proposition proposes a maximum likelihood estimator for each of the two cases:

Proposition 1 (Likelihood function under occasional observations)**.** *Let m_t, z_t, x_t for $t = 1, ..., T$, as well as $\bar{t}_i$ for $i = 0, ..., N+1$, τ_i for $i = 1, ..., N$ and P^{zx}, P^m, and N be defined as above. Let $y_t \equiv (m_t, z_t)$ be the vector of variables which are observed in all periods, i.e., it is composed of both model and observation variables.*

1. *If A1(a) holds, then let $\psi \equiv (\theta, \eta)$ and $\Psi \equiv \Theta \times \mathcal{H}$ and define the transition probability P by*

$$P(y_t, x_t | y_{t-1}, x_{t-1}; \psi) \equiv P^m(m_t | z_t, x_t, m_{t-1}, z_{t-1}, x_{t-1}; \eta) P^{zx}\left(z_t, x_t | z_{t-1}, x_{t-1}; \theta\right). \quad (13)$$

2. *If A1(b) holds, then let $\psi \equiv \theta$ and $\Psi \equiv \Theta$ and define the transition probability P by*

$$P(y_t, x_t | y_{t-1}, x_{t-1}; \psi) \equiv P^{zx}\left(z_t, x_t | z_{t-1}, x_{t-1}; \theta\right). \quad (14)$$

For each of the cases and the according definitions of ψ and P, the maximum likelihood estimator for ψ is given by

$$\hat{\psi} = \underset{\psi \in \Psi}{\operatorname{argmax}} L(\psi),$$

[10] To achieve a compact formulation for the likelihoods, we adhere to the following notational convention for the product sign:

$$\prod_{s=t}^{t-1} f(s) = \prod_{s \in \emptyset} f(s) = 1,$$

i.e., the empty product is equal to 1. Furthermore, we write

$$P^{zx}\left(z_1, x_1 | z_0, x_0; \theta\right) \equiv P^{zx}\left(z_1, x_1; \theta\right),$$
$$P^m\left(m_1 | z_1, x_1, m_0, z_0, x_0; \theta\right) \equiv P^m\left(m_1 | z_1, x_1; \theta\right)$$

for the respective initial distributions, and

$$P^{zx}\left(z_{T+1}, x_{T+1} | z_T, x_T; \theta\right) \equiv 1,$$
$$P^m\left(m_{T+1} | z_{T+1}, x_{T+1}, m_T, z_T, x_T; \theta\right) \equiv 1,$$

since period $(T+1)$ is outside of the sample period $\{1, ..., T\}$ and hence $m_{T+1}, z_{T+1}, x_{T+1}$ are not defined as random variables.

where

$$L(\psi) = L(\psi|\{y_t\}_{t=1}^T, \{x_t\}_{t\in\bar{\mathcal{T}}}) \tag{15a}$$

$$\equiv \prod_{i\in\{1,\dots,N+1\}\bigcap\{i|\tau_i>0\}} \int\cdots\int_{\mathcal{S}_x^{\tau_i}} P\left(y_{\bar{t}_{i-1}+1}, \tilde{x}_{\bar{t}_{i-1}+1}|y_{\bar{t}_{i-1}}, x_{\bar{t}_{i-1}}; \psi\right) \tag{15b}$$

$$\cdot \prod_{t=\bar{t}_{i-1}+2}^{\bar{t}_i-1} P\left(y_t, \tilde{x}_t|y_{t-1}, \tilde{x}_{t-1}; \psi\right) \tag{15c}$$

$$\cdot P\left(y_{\bar{t}_i}, x_{\bar{t}_i}|y_{\bar{t}_i-1}, \tilde{x}_{\bar{t}_i-1}; \psi\right) \mathrm{d}\left(\tilde{x}_{\bar{t}_{i-1}+1}, ..., \tilde{x}_{\bar{t}_i-1}\right) \tag{15d}$$

$$\cdot \prod_{i\in\{j=1,\dots,N+1|\tau_j=0\}} P\left(y_{\bar{t}_i}, x_{\bar{t}_i}|y_{\bar{t}_i-1}, x_{\bar{t}_i-1}; \psi\right). \tag{15e}$$

The proof of Proposition 1 can be found in Appendix A.1. The idea of the proposition is as follows: In the full information case, the likelihood of a sample would simply be the product over each observation's conditional probability. In our case of non-observation of some x_t, we derive the likelihood function (15) by marginalizing the unobserved x_t, i.e., integrating them w.r.t. their associated conditional probability and their domain $\mathcal{S}_x$.

In contrast to the case of RLI examined in Reich (2018), the occasional observations of x_t allow us to split up the joint integral over x_t, $t \notin \bar{\mathcal{T}}$, into $N+1$ separate factors which are each associated with one of the segments $(\bar{t}_{i-1}, ..., \bar{t}_i)$ for $i = 1, ..., N+1$. These $N+1$ factors are in turn split up into two sets, according to the length of their non-observation segment, τ_i: If $\tau_i = 0$ (Equation (15e)), i.e., the respective segment is $(\bar{t}_i - 1, \bar{t}_i)$, then the i-th factor is simply the probability of $y_{\bar{t}_i}, x_{\bar{t}_i}$ conditional on $y_{\bar{t}_i-1}, x_{\bar{t}_i-1}$. If $\tau_i > 0$ (equations (15b)–(15d)), there is at least one period of non-observation between two periods of observation, i.e., the i-th segment $(\bar{t}_{i-1}, ..., \bar{t}_i)$ contains at least 3 time periods. The factor associated with this segment is the integral of the product of conditional probabilities over all successive unobserved x_t with $\bar{t}_{i-1} < t < \bar{t}_i$: the p.d.f. in (15b) is the probability of the first unobserved variable of this segment conditional on the observations in period $\bar{t}_{i-1}$; the p.d.f. in (15d) is the probability of the observations in period $\bar{t}_i$ conditional on the last unobserved variable of this segment; the product in (15c) is empty if there is only one period of unobservation in this segment and otherwise is the product of all probabilities of an unobserved variable conditional on the previous unobserved variable.

If the missing data mechanism is MNAR, i.e., if the observation pattern is endogenous, x_t and x_{t-1} are part of P^m. Therefore, marginalization of the unobserved x_t involves the product of P^m and P^{zx} as the integrand. Consequently, we require (the knowledge of) a specific functional form for P^m. Moreover, we need to estimate η as well, since it is now also a nuisance parameter of L (the likelihood function "inherits" the nuisance parameter from P^m). In particular, η cannot be estimated separately by conditional maximum likelihood, i.e., by only using the process for m_t. The conditional likelihood of η is again an integral over x_t for $t \notin \bar{\mathcal{T}}$ and is in fact the same as (15).

If the missing data mechanism is MAR, i.e., if the observation pattern is exogenous, x_t is not a part of P^m for any t, allowing us to "pull" P^m out of the integral. By Assumption A1(b), P^m is parametrized independently of θ and thus becomes merely a scaling factor to

the likelihood of θ. In other words, the likelihood given in (15) with $\psi = \theta$ and Definition (14) is equivalent to the likelihood with $\psi = (\theta, \eta)$ and Definition (13) for any $\eta \in \mathcal{H}$. Thus, maximum likelihood estimation with an M(C)AR missing data mechanism does neither require any explicit specification of P^m, nor the estimation of an additional nuisance parameter. Note that if the second part of Assumption A1(b) regarding the support of η fails to be satisfied, the resulting estimator of θ is still valid, but less efficient.

Let us briefly comment on the dimensionality of the model variable x_t.[11] For simplicity, we have assumed in this section that x_t is one-dimensional, and hence m_t is also one-dimensional. We can generalize this setting to a setting with multi-dimensional $x_t \in \mathbb{R}^{d_x}$, $d_x > 1$, in two ways: First, consider a situation in which all elements of the vector x_t are observed at the same time, i.e., if one element of x_t is observed, all other elements are also observed. Then, m_t and all of the auxiliary variables $N, \bar{t}_i$, and τ_i, $i = 1, ..., N$, are still one-dimensional variables. In particular, all definitions and derivations in this section only depend on the entire vector x_t and never directly on individual elements of x_t. Thus, all of the definitions and derivations and especially Proposition 1 above hold verbatim and can be used without further adjustments for estimation.

The second case further generalizes our setting in that it allows some elements of x_t to be observed while others are unobserved in the same period, i.e., m_t is now also d_x-dimensional. This entails a more intricate definition of the observation pattern $\bar{\mathcal{T}}$ and causes the integration domain $\mathcal{S}_x$ to differ in each period, depending on how many elements of x_t need to be marginalized. We show in the Appendix A.2 how our notation has to be adjusted for this case. However, the general intuition of this section remains valid: If the missing data mechanism is MAR, then the likelihood can be constructed only using the model process. If it is MNAR, then we additionally require the observation process. We conclude this section with three examples each corresponding to one of the types of endo- or exogeneity of the observation pattern described in Definition 2:

Example 1 (Discrete choice models with choice-dependent observability)**.** *Consider a data set $\{d_t, s_t\}_{t=1}^T$ which is analyzed using a Discrete Choice model (DCM): It consists of one or several decision variables d_t which take discrete values $\{1, ..., J\}$ and one or several explanatory variables s_t which are inputs to the decision process of the agent in the model. Suppose that the econometrician does not have access to the full data set but observes some of the s_t only occasionally, according to the missing data mechanism P^m. If the decision variables d_t are always observed and the observability of the explanatory variables depends only on the decision variable, then the missing data mechanism is MAR: Although d_t depends on s_t and hence observability of s_t also depends on s_t, the dependence on s_t in P^m simply drops out by conditioning on the always observed d_t. Put in terms of our framework, the occasionally observed parts of s_t map into x_t, while d_t and the remaining parts of s_t map into z_t. This implies that A1(b) is fulfilled and the likelihood is conveniently constructed using only the decision process which is anyway specified by the DCM.*

Prominent examples for such DCM with occasionally observed explanatory variables are

[11] The dimensionality of z_t did not play a role in any of our definitions and derivations, hence z_t can be freely considered to be multi-dimensional.

models of purchase decisions based on scanner data (Erdem et al., 1999) or price data from steel retailers (Hall and Rust, 2021). In these settings, prices are only observed if the agent actually purchases the product, otherwise the price is not observed by the econometrician but only by the agent (who bases his purchase decision on this price). This reflects exactly the setting above; thus, only the decision process is required for estimation of these models.

In Section 3.2, we present a hypothetical version of the bus engine replacement model by Rust (1987) in which the mileage data is only observed upon replacement and the random utility component of the bus manager is serially correlated. We analyze it with the likelihood from Part 2 of Proposition 1 and provide empirical evidence for its asymptotic properties.

Example 2 (Observation of summary statistics of time series data)**.** *Suppose an econometrician wants to estimate a model of two variables, z_t and s_t, which admit a joint transition probability $P^{sz}(z_t, s_t|z_{t-1}s_{t-1};\theta)$ parameterized by θ. However, he or she only has access to reliable data for the z-variable; observations of s_t are not available, e.g., because they are not reported, or because they are diluted by an unknown seasonality pattern or some other form of noise making them unfit for estimation. Instead, the econometrician occasionally observes another variable, say x_t, which is a function of a finite sequence of s_t. For simplicity, suppose x_t is observed every τ periods, i.e. $\bar{t}_i - \bar{t}_{i-1} = \tau$ for observation periods $\{\bar{t}_i\}_{i=1}^N$, and $x_{\bar{t}_i}$ is a function of the τ variables $s_{\bar{t}_i}, ..., s_{\bar{t}_{i-1}+1}$, which we denote by f in this example.*[12]

In this setting, our framework allows us to use the model of z_t and s_t to derive a likelihood of θ based on the sample $\{\{z_t\}_{t=1}^T, \{x_{\bar{t}_i}\}_{i=1}^N\}$. In general, this likelihood has the same form as that in Equation (15)*, but it features an integral over $\{s_t\}_{t=\bar{t}_{i-1}+1}^{\bar{t}_i}$, where the integration domain is defined by $f(s_{\bar{t}_{i-1}+1}, ..., s_{\bar{t}_i}) = x_{\bar{t}_i}$. This formulation makes maximum likelihood estimation possible even if the observed data is only a function of the variables of interest.*[13] *If we think of the f as a statistic of the sample $\{s_t\}_{t=\bar{t}_{i-1}+1}^{\bar{t}_i}$, then we can still estimate model parameters even if only a sample quantile or moment of the relevant model variables are observed.*

A common example is time aggregation, where $x_{\bar{t}_i} = s_{\bar{t}_{i-1}+1} + s_{\bar{t}_i-\tau+2} + ... + s_{\bar{t}_i}$, i.e., where x_t is proportional to the empirical mean of a sample of length τ of the s-variable. In Section 3.1, we illustrate how our approach can be used to estimate two long-run risk models by Bansal and Yaron (2004) and Schorfheide et al. (2018); while those models are defined on the quarterly level, reliable dividend data necessary to estimate them is only available at the annual level. We show how the aggregated observed variable corresponds with the disaggregated model variable, and demonstrate the empirical efficiency of our approach in comparison to the case with full observability of the model variables.

Example 3 (Censoring with deterministic threshold)**.** *Let us consider the well known case of censored data: Suppose x_t is a variable with support $\mathbb{R}$ which is only observed if $x_t < \bar{x}$ for some threshold $\bar{x} \in \mathbb{R}$; otherwise, we only know that $x_t \geq \bar{x}$. Importantly, non-observation of x_t is still informative as it corresponds to knowing that x_t lies above the threshold. Hence, the number of*

[12] The general formulation allows for varying length of the non-observation segment, i.e. observations occur in periods $t = \bar{t}_1, ..., \bar{t}_N$, and a time-dependent functional relation, $x_{\bar{t}_i} \equiv f_i(s_{\bar{t}_{i-1}+1}, ..., s_{\bar{t}_i})$.

[13] The integrand over $\{s_t\}_{t=\bar{t}_{i-1}+1}^{\bar{t}_i}$ can be interpreted as conditional probability of $\{s_t\}_{t=\bar{t}_{i-1}+1}^{\bar{t}_i}$ conditional on $x_{\bar{t}_i}$. Note that maximum likelihood estimation is possible using the characterization of the likelihood above; however, asymptotic properties can only be derived if the conditional probability of $(z_{\bar{t}_{i-1}+1}, s_{\bar{t}_{i-1}+1})$ conditioned on $(z_{\bar{t}_{i-1}}, x_{\bar{t}_{i-1}})$ is specified by the model.

non-observations provides an estimator for how much mass of x_t's probability distribution lies on the interval $[\bar{x}, \infty)$.[14] *This period-t-information (either from observation or non-observation) is often formalized by considering the censored variable x_t^* which is defined by $x_t^* \equiv x_t$ if $x_t < \bar{x}$ and $x_t^* \equiv \bar{x}$ otherwise and which is observed in all periods. In contrast, we capture the same information by the tuple (x_t, m_t) where the observation variable, m_t, associated with x_t is given by*

$$m_t = \begin{cases} 1, & \text{if } x_t > \bar{x}, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

This way, censoring can be formulated in terms of a missing data problem. Notably, we cannot discard the dependence on x_t in the definition of P^m in the case of censoring. As a matter of fact, P^m is an indicator function and only takes the values 0 or 1, directly depending on whether $x_t > \bar{x}$ or not. Therefore, the missing data mechanism P^m is MNAR and we need to construct the likelihood from Proposition 1.1, including P^m.

Our formulation using the tuple (x_t, m_t) generalizes the censoring intuition: it does not deterministically impose observability based on a fixed threshold but instead allows observability to be probabilistic at every value in the support of x_t. In this sense, our notion of generalized censoring is indeed equivalent to the notion of MNAR.

2.3 Recursive formulation of the likelihood function

Proposition 1 provides the general formulation of the likelihood, Equation (15), which accounts for both MAR and MNAR data. In particular, it abstracts from both cases by introducing a general notation for the observed variables, y_t and the parameter of interest, ψ. In this section, we show how to efficiently compute this likelihood using the recursive likelihood integration proposed by Reich (2018).

By a simple induction argument, the example with only one observation from Section 2.1 can be generalized to any observation pattern $\bar{\mathcal{T}} \subseteq \mathcal{T}$ for $t > 1$ by

$$f_t^\psi(x) = \begin{cases} 1 & t > T \\ \int P(y_t, \tilde{x} | y_{t-1}, x_{t-1}; \psi) f_{t+1}^\psi(\tilde{x}) \mathrm{d}\tilde{x} & t-1 \in \bar{\mathcal{T}}, t \notin \bar{\mathcal{T}} \\ P(y_t, x_t | y_{t-1}, x; \psi) f_{t+1}^\psi(x_t) & t-1 \notin \bar{\mathcal{T}}, t \in \bar{\mathcal{T}} \\ P(y_t, x_t | y_{t-1}, x_{t-1}; \psi) f_{t+1}^\psi(x_t) & t-1 \in \bar{\mathcal{T}}, t \in \bar{\mathcal{T}} \\ \int P(y_t, \tilde{x} | y_{t-1}, x; \psi) f_{t+1}^\psi(\tilde{x}) \mathrm{d}\tilde{x} & \text{otherwise,} \end{cases} \tag{17}$$

where all indexed variables—i.e., x_t, x_{t-1}, y_t, and y_{t-1}—denote observations from the dataset, whereas x and $\tilde{x}$ denote function arguments and integration variables, respectively. Finally, the

[14] A different notion of "missing" data is described by *truncation*. However, truncation is fundamentally a sampling problem, i.e., values outside of the truncated support of x_t are inherently never sampled. Opposed to our setting, it is therefore not possible to obtain any information about the distribution of x_t outside the truncated support.

complete likelihood function can be evaluated as one of the following:

$$L(\psi) = \int P(y_1, \tilde{x}; \psi) f_2^{\psi}(\tilde{x}) \mathrm{d}\tilde{x} \tag{18a}$$

$$L(\psi) = P(y_1, x_1; \psi) f_2^{\psi}(x_1), \tag{18b}$$

where we distinguish two cases: In Equation (18a) we assume $1 \notin \bar{\mathcal{T}}$—that is, the very first state realization is unobserved; the likelihood thus forms an integral against the stationary distribution of the unobserved process, $P(x; \psi)$. In Equation (18b) we require $1 \in \bar{\mathcal{T}}$—that is, the initial state realization is either trivially known or observed.[15]

The appeal—but also the main problem—of the recursive representation of the likelihood function in (18) is that it is *exact*. Theoretically, an evaluation of f_2^{ψ} defined according to (17) would trigger evaluations of f_3^{ψ}, which would themselves trigger evaluations of f_4^{ψ}, and so on, up to f_{T+1}^{ψ}. Moreover, each evaluation at level t triggers potentially multiple evaluations at the level $t+1$ due to the integration, resulting in an exponentially growing number of total function evaluations in T. Therefore, we introduce an approximation operator, a mapping between two function spaces $\mathcal{B}$ and $\mathcal{P}$:

$$\mathcal{I} : \mathcal{B} \to \mathcal{P}, f \mapsto \hat{f}, \tag{19}$$

where the elements of $\mathcal{P}$ can be represented by a countable set of parameters, the size of which is independent of T. This allows us to state an approximation of (17) as follows:

$$\hat{f}_t^{\psi}(x) = \begin{cases} 1 & t > T \\ \int P(y_t, \tilde{x} | y_{t-1}, x_{t-1}; \psi) \hat{f}_{t+1}^{\psi}(\tilde{x}) \mathrm{d}\tilde{x} & t-1 \in \bar{\mathcal{T}}, t \notin \bar{\mathcal{T}} \\ P(y_t, x_t | y_{t-1}, x; \psi) \hat{f}_{t+1}^{\psi}(x_t) & t-1 \notin \bar{\mathcal{T}}, t \in \bar{\mathcal{T}} \\ P(y_t, x_t | y_{t-1}, x_{t-1}; \psi) \hat{f}_{t+1}^{\psi}(x_t) & t-1 \in \bar{\mathcal{T}}, t \in \bar{\mathcal{T}} \\ \mathcal{I}\left(\int P(y_t, \tilde{x} | y_{t-1}, x; \psi) \hat{f}_{t+1}^{\psi}(\tilde{x}) \mathrm{d}\tilde{x} \right) & \text{otherwise,} \end{cases} \tag{20}$$

which now has linear complexity in T. The procedure to actually evaluate $\hat{f}_2^{\psi}$ *numerically*—and thus, the likelihood function (18)—critically depends on the nature of the state variable x, in particular whether it is discrete or continuous. This is not only true for the approximation operator $\mathcal{I}$, but also for the computation of the integrals in (20). However, the actual procedures to compute the value of the likelihood function for some parameter value share the same logic, which is therefore summarized in Algorithm 1 in Online Appendix A.3

In the following, we specialize the likelihood recursion (17) for Markov processes with state x being continuous—that is, $x_t \in \mathcal{S}_x \subseteq \mathbb{R}$ (We do not need to specify the nature of the state space for y at this point). In the presence of a continuous occasionally observed state, we first have to compute the integrals by numerical quadrature, and second, we have to approximate the function $f_t^{\psi}(\cdot)$—an infinite-dimensional object—by some form of function approximation, such as interpolation.[16]

[15] Note that by setting $P(y_1, x_1; \psi)$ equal to unity in (18b) we can also calculate the *conditional* likelihood.

[16] For the case where x follows a discrete process, the recursion function, $f_t^{\psi}(\cdot)$, becomes a finite vector and the integral is simply replaced by a sum. Hence, the recursion can be computed exactly (up to the floating-point arithmetic round-off error).

We start with the discussion of the integration problems. In order to make a wide range of quadrature rules applicable, we formulate two assumptions, which are, however, without loss of generality for the method itself, and their relaxation only limits the choice of the numerical integration methods. First, suppose the transition probabilities for the model variables and the observation variables each satisfy a conditional independence relation s.t. we obtain for the joint probability:

$$P(y_t, x_t|y_{t-1}, x_{t-1}) = P(y_t|x_t)P(x_t|y_{t-1}, x_{t-1}).$$

Second, suppose there exists an invertible and differentiable change of variables ϕ, mapping to a new variable Δx_t by

$$x_t = \phi(\Delta x_t, y_{t-1}, x_{t-1}; \psi),$$

such that

$$P(\phi(\Delta x_t, y_{t-1}, x_{t-1}; \psi)|y_{t-1}, x_{t-1}; \psi) = q(\Delta x_t).$$

A simple but very relevant example is the AR(1) process with $\phi : x_t = \rho x_{t-1} + \Delta x_t$ and $\phi' \equiv 1$. Using this change of variables, we can rewrite the integrals in recursion (17) as

$$f_t^\psi(x) = \int_{\phi(\mathcal{D}, y_{t-1}, x; \psi)} P(y_t|\tilde{x}; \psi)P(\tilde{x}|y_{t-1}, x; \psi)f_{t+1}^\psi(\tilde{x})\mathrm{d}\tilde{x} \tag{21}$$

$$\begin{aligned} &= \int_{\mathcal{D}} P(y_t|\phi(\Delta\tilde{x}, y_{t-1}, x; \psi); \psi)\phi'(\Delta\tilde{x}, y_{t-1}, x; \psi) \\ &\quad \cdot q(\Delta\tilde{x})f_{t+1}^\psi(\phi(\Delta\tilde{x}, y_{t-1}, x; \psi))\mathrm{d}\Delta\tilde{x}, \end{aligned} \tag{22}$$

which allows the application of a wide range of quadrature rules of the following form: Consider an integrand $h : \mathcal{D} \to \mathbb{R}$, a non-negative bounded weighting function $q : \mathcal{D} \to [0, a]$, and a set of nodes and weights $\{(c_i, w_i)\}_{i=1}^{N^Q}$ with

$$\int_{\mathcal{D}} h(\tilde{x})q(\tilde{x})\mathrm{d}\tilde{x} = \sum_{j=1}^{N^Q} w_j h(c_j) + \epsilon_Q, \tag{23}$$

such that the approximation error ϵ_Q is minimized in some sense. For example, for Gauss-type weighting functions, Gauss–Hermite quadrature approximates integrals of the form (23) accurately if the corresponding integrand h is sufficiently smooth (and can thus be approximated well by a polynomial). Therefore, we *approximate* the integral in the recursive definition (22) by

$$\sum_{j=1}^{N^Q} w_j P(y_t|\phi(c_j, y_{t-1}, x; \psi); \psi)\phi'(c_j, y_{t-1}, x; \psi)f_{t+1}^\psi(\phi(c_j, y_{t-1}, x; \psi)), \tag{24}$$

where the weighting function $q(\cdot)$ is now captured in the weights w_j according to the specific quadrature rule.

We now turn our attention to the approximation of the function object f^ψ itself. As we have argued above, this is necessary because otherwise the recursion would trigger a tree of function evaluations that grows exponentially in T at worst. Even when using an approximation of the integral as in (24), we are required to evaluate $f_t^\psi(x)$ at the transformed quadrature nodes

$\phi(c_j, y_{t-1}, x_{t-1}; \psi)$, which do, however, depend on t and thus keep changing over the course of the recursion. Consequently, we want to find a representation of f_t^ψ, say $\hat{f}_t^\psi$, that (i) can be obtained through finitely many evaluations of f_t^ψ, and (ii) does not require the evaluation of f_s^ψ, or $\hat{f}_s^\psi$, for $s > t+1$.

Following Reich (2018), we further specialize the approximation operator $\mathcal{I}$ defined in (19) to create an *interpolant* based on finitely many evaluations of f^ψ.[17] Consider a grid of interpolation nodes $\{g_j\}_{j=1}^{N^I} \in \mathcal{S}^{N^I}$. Then,

$$\hat{\mathcal{I}} : \mathbb{R}_+^{N^I} \to \mathcal{P},\ \{f^\psi(g_j)\}_{j=1}^{N^I} \mapsto \hat{f}^\psi, \tag{25}$$

where $\hat{f}^\psi$ is an object from a function space with finite-dimensional representation, $\mathcal{P}$, and is typically obtained by solving a system of equations on the set of interpolation nodes, $\{g_j\}_{j=1}^{N^I}$, such that $\forall g_j : \hat{f}^\psi(g_j) = f^\psi(g_j)$. Moreover, some interpolation schemes impose further restrictions on the derivatives of the interpolant (e.g., splines for continuous higher-order derivatives of the interpolant, or Hermite interpolation to fit derivatives of the original function). Obviously, this approximation step introduces a second source of error: $\epsilon_I = \|f^\psi - \hat{f}^\psi\|$.[18]

By combining the numerical approaches for integration and function approximation defined above, recursion (20) can be implemented for continuous occasionally observed states as

$$\hat{f}_t^\psi(x) = \begin{cases} 1 & t > T \\ \sum_{j=1}^{N^Q} w_j P(y_t | \phi(c_j, y_{t-1}, x_{t-1}; \psi); \psi) & \\ \quad \cdot \phi'(c_j, y_{t-1}, x_{t-1}; \psi) \hat{f}_{t+1}^\psi(\phi(c_j, y_{t-1}, x_{t-1}; \psi)) & t-1 \in \bar{\mathcal{T}}, t \notin \bar{\mathcal{T}} \\ P(y_t | x_t; \psi) P(x_t | y_{t-1}, x; \psi) \hat{f}_{t+1}^\psi(x_t) & t-1 \notin \bar{\mathcal{T}}, t \in \bar{\mathcal{T}} \\ P(y_t | x_t; \psi) P(x_t | y_{t-1}, x_{t-1}; \psi) \hat{f}_{t+1}^\psi(x_t) & t-1 \in \bar{\mathcal{T}}, t \in \bar{\mathcal{T}} \\ \hat{\mathcal{I}}\Big(\Big\{ \sum_{j=1}^{N^Q} w_j P(y_t | \phi(c_j, g_i, y_{t-1}; \psi); \psi) & \\ \quad \cdot \phi'(c_j, g_i, y_{t-1}; \psi) \hat{f}_{t+1}^\psi(\phi(c_j, g_i, y_{t-1}; \psi)) \Big\}_{i=1}^{N^I} \Big) & \text{otherwise.} \end{cases} \tag{26}$$

The final, approximated likelihood for the continuous case, which depends on whether x_1 is observed or not, reads

$$\tilde{L}(\psi) = \sum_{j=1}^{N^Q} \bar{w}_j P(y_1 | \phi_1(\bar{c}_j; \psi); \psi) \phi_1'(\bar{c}_j; \psi) \hat{f}_2^\psi(\phi_1(\bar{c}_j; \psi)) \tag{27a}$$

$$\tilde{L}(\psi) = P(y_1 | x_1; \psi) P(x_1; \psi) \hat{f}_2^\psi(x_1), \tag{27b}$$

where ϕ_1 is the corresponding change of variables for the stationary distribution of x_t, $P(x_1; \psi)$, and $\{(\bar{c}_j, \bar{w}_j)\}_{i=1}^{N^Q}$ are the respective quadrature nodes and weights.

We conclude this subsection with some remarks on practicalities and alternatives: First,

[17] Regression approaches that minimize a loss function are equally applicable, in particular in higher-dimensional contexts.

[18] Note that the individual interpolation and quadrature errors, ϵ_I and ϵ_Q, respectively, are only the one-step approximation errors, which potentially magnify throughout the recursion. Reich (2018), however, provides a rigorous error and convergence analysis. In particular, the author finds that the error grows only linearly in T, and that the overall convergence rate is half as good as the smaller of the respective interpolation and quadrature schemes. Note that this potentially allows for exponential convergence.

the procedure outlined above is not spared from the usual issues related to floating point arithmetic, over- and under-flow in particular. Therefore, we discuss implementation-related aspects in Appendix A.3, together with a pseudo-code description of the algorithm (Algorithm 1). Moreover, we give a very concise—and yet fully functional—MATLAB implementation for x from a continuous state space in Appendix A.4.

Our second remark concerns the application of (forward) simulation to approximate the integral in Equation (3), which is a popular procedure due to its wide applicability to many integration problems. For example, using the GHK sampler (seen as an importance sampling device, as due to Keane, 1994) one can efficiently simulate an unobserved, serially correlated process even in the presence of observations of other, dependent variables, such as realized choices in a discrete choice model (under appropriate distributional assumptions on the process itself). However, while occasionally observed processes can theoretically be forward-simulated, direct implementations tend to be highly inefficient: Consider again the case with exactly one observation at $\bar{t} > 1$ and some (known) initial value x_0. All forward simulation procedures that we are aware of are inherently one-sided, because the simulated value x_t^i for any variable x_t conditions either on the previous observation x_{t-1}, if $t = 1$, or the previous simulated value x_{t-1}^i. Then, sample sequences $\{x_t^i\}_{t=1}^{\bar{t}-1}$ are generated by applying the forward simulation mechanism $\bar{t} - 1$ times. Note that all sequences obtained like this incorporate solely the data at $t = 0$, but not the information available at $t = \bar{t}$. Nevertheless, in order to utilize the simulated sequences for computation of the integral in (3) we do need to incorporate the information inherent in $x_{\bar{t}}$. Therefore, we need to weight the sequences by the conditional probabilities $P(x_{\bar{t}}|x_{\bar{t}-1}^i; \theta)$, which are known from the model. However, these probabilities can become arbitrarily small, as the sample path—and in particular the final element $x_{\bar{t}-1}$—is generated without specifically "targeting" the observed value $x_{\bar{t}}$ in the $\bar{t}$-th period.

2.4 Large sample properties of the approximate likelihood estimator

We now provide a brief discussion of the statistical properties of our likelihood estimator which are formally treated in Gilch et al. (2025). In the previous section, we have defined two likelihood functions: the exact likelihood L in Equation (15), and the approximated likelihood $\tilde{L}$ in equation (27). Although they both admit estimators of the parameter of interest ψ, namely

$$\hat{\psi} = \underset{\psi \in \Psi}{\operatorname{argmax}}\, L(\psi)$$

and

$$\tilde{\psi} = \underset{\psi \in \Psi}{\operatorname{argmax}}\, \tilde{L}(\psi),$$

respectively, only $\tilde{\psi}$ is computationally feasible, since we cannot evaluate the exact likelihood L in general.

The main requirement for proving asymptotic properties of $\tilde{\psi}$ is the consistency and asymptotic normality of the exact maximum likelihood estimator $\hat{\psi}$ itself. In the case of full state observations (as in equation (1)), these properties are verified for $\hat{\psi}$ by taking the logarithm of the likelihood, granting a sum of log-probabilities and allowing the application of the law

of large numbers and the central limit theorem. This approach is not directly applicable for integrated likelihoods such as (2) because the integral over the unobserved states disallows the log-transformation of the likelihood-product into the loglikelihood-sum.

Yet, as Gilch et al. (2025) show, in contrast to the case of never observed states as in Reich (2018), the availability of occasional observations allows to split up the integral. Gilch et al. (2025) use this decomposition to facilitate a loglikelihood approach and prove consistency and asymptotic normality of $\hat{\psi}$. Besides standard assumptions on stationarity and ergodicity of the Markov process, they only require an assumption about the frequency of observations of x_t for their proof.[19]

However, note that we generally have $\hat{\psi} \neq \tilde{\psi}$. This is because numerical approximation of the likelihood is not only a subject of interest for computational methods, but also affects statistical properties of the resulting maximum likelihood estimator. This is because the approximation introduces an error to the objective function of the underlying maximization problem—and thus the corresponding maximizer, too. Therefore, the large sample properties of the estimator, i.e., its consistency and asymptotic normality, which are the basis for inference about ψ, are potentially affected by the approximation scheme.[20] Gilch et al. (2025) provide a full set of proofs to show consistency and asymptotic normality of the approximated estimator, $\tilde{\psi}$ based on similar methods as developed by Griebel et al. (2019).

3 Applications

In this section, we demonstrate the applicability our method as well as its favorable properties in various examples. In Section 3.1, we use it to estimate two prominent finance applications: First, we apply it to the long-run risk model of Bansal and Yaron (2004), which includes persistent changes in consumption and dividend growth. The model is conditionally Gaussian and features linear state dynamics. Second, we use a model with stochastic volatility as in Schorfheide et al. (2018), which adds non-linear dynamics to consumption and dividend growth. Both models feature one fully unobserved and one occasionally observed process. We show that our method can identify all model parameters even in small data samples, and we provide evidence for the asymptotic normality of our RLI estimator. Furthermore, we demonstrate that our approach can be significantly more efficient than conventional approaches such as simulated method of moments, in particular in the presence of non-linearities, and for short data set time windows. In Section 3.2, we consider an application featuring a controlled Markov process. In particular, we estimate a variant of the optimal replacement of GMC bus engines model by Rust (1987), for which we observe only a subset of the state variables over time. We show that our estimator with

[19]Note that the assumption regarding frequency of observations requires regular complete observations, i.e. the entire state vector is observed after finite time. For that for two of the applications we present below, the long-run risk and the stochastic volatility model, this property is not satisfied because they have permanently unobserved states. For the modified bus engine replacement model it is satisfied because the mileage-state is observed every time the engine is replaced.

[20]Note that we consider deterministic and not stochastic algorithms for the approximation of L: While there is a large literature on simulated maximum likelihood, this literature is not applicable for numerical approximation because its approximation nodes are chosen deterministically. Hence, it is necessary to make use of the known approximation error formulas whereas for simulation methods often the Delta-method and a central limit theorem can be applied.

occasionally observed states is barely less efficient for identifying some key model parameters than the estimator with full state observations in this application, despite the fact that it uses significantly less state data.

3.1 Long-run risk and stochastic volatility models with time-aggregated observations

This section demonstrates the favorable properties and the broad applicability of our approach, using two prominent finance applications: the long-run risk models of Bansal and Yaron (2004) and Schorfheide et al. (2018). Bansal and Yaron (2004) propose a model where consumption and dividend growth are driven by a small but persistent, unobservable component, which can help explain a large number of asset pricing puzzles (see, for example, Hansen et al., 2008; Bollerslev et al., 2009; Drechsler and Yaron, 2011; Bansal et al., 2012; Bansal and Shaliastovich, 2013). Further, they add stochastic volatility to the model to generate time variation in risk premia. Estimating long-run risk models has proven difficult due to the unobserved and persistent nature of the state variables. Difficulties arise, for example, due to data scarcity caused by relatively short observation horizons and to different frequencies of the observable data: Aggregate consumption data for the US is available at a quarterly frequency starting in 1947, which yields around 294 observations as of now. Aggregate dividend data is available on a monthly basis and goes back to the 1920s. However, monthly and quarterly dividends show strong seasonality patterns, which explains why researchers have relied on smoothing methods (see, for example, Grammig and Küchlin, 2018; Schorfheide et al., 2018), which can, in turn, bias the estimation. Hence, while we have reliable, unmodified quarterly data for consumption, only annual data of the same quality is available for dividends.[21] This results in a rather small dataset with, as we show below, an occasional observation pattern due to these mixed data frequencies. In the following, we show how the RLI approach can be used to estimate long-run risk models with occasionally observed and unobserved states.

3.1.1 Model dynamics

We consider two model variants to compare the performance of our estimation approach. First, we use the standard model of Bansal and Yaron (2004, Case 1, without stochastic volatility), where log aggregate consumption growth, Δc_t, and log aggregate dividend growth, Δd_t are given by

$$\begin{aligned} \Delta c_t &= \mu_c + x_t + \sigma_c \eta_{c,t} \\ x_t &= \rho_x x_{t-1} + \sigma_x \eta_{x,t} \\ \Delta d_t &= \mu_d + \Phi x_t + \sigma_d \eta_{d,t}, \end{aligned} \tag{28}$$

with $\eta_{\cdot,t} \sim N(0,1)$ i.i.d.[22] The key feature of the long-run risk model is that there are small but highly persistent shifts in the growth rate of consumption and dividends, which are captured by

[21] For consumption, monthly data is available starting in the 1960s. This data is, however, artificially smoothed as the observed data is only available at a quarterly frequency; see Schorfheide et al. (2018).

[22] Note that the original model of Bansal and Yaron (2004) assumes that Δc_t depends on x_{t-1} instead of on x_t. This does not, however, influence our estimation results, but following the original notation would significantly complicate the notation of the likelihood.

x_t. In this model, all shocks are normally distributed and enter the model equations linearly. This clearly allows highly efficient estimation also by moment-based methods.

As a second example, we use a stochastic volatility model as in Bansal and Yaron (2004, Case 2). Bansal and Yaron (2004) model the variance as an AR(1) process, which has the strong disadvantage that the variance can become negative. Therefore, Schorfheide et al. (2018) propose a model where volatility dynamics follow a log-normal distribution. We follow this approach and use for the second model the following consumption and dividend dynamics:

$$\begin{aligned} \Delta c_t &= \mu_c + \sigma_c e^{h_t} \eta_{c,t} \\ h_t &= \rho_h h_{t-1} + \sigma_h \eta_{h,t} \\ \Delta d_t &= \mu_d + \phi_d \sigma_c e^{h_t} \eta_{d,t} \end{aligned} \tag{29}$$

with $\eta_{\cdot,t} \sim N(0,1)$ i.i.d. Note that volatility dynamics enter consumption and dividends non-linearly, which makes the estimation of the model more challenging. In the following, we describe how the models can be estimated using recursive likelihood integration.

3.1.2 Time aggregation as occasional state observations

As already indicated, the data available to estimate the model is limited. While consumption data is observed quarterly, dividend data at the same frequency shows strong seasonalities, which processes (28) and (29) can not account for. Hence, only annual dividend data can be used in its original form for the estimation.

A full observation regime can be characterized by a quarterly time index set $\mathcal{T} \equiv \{1, \dots, T\}$ where the data is given by $\{\Delta c_t, \Delta d_t\}_{t=1}^{T}$. As the econometrician can only rely on annual dividend data, but the model uses dividend growth in quarterly terms, we obtain an occasional observation regime in the following way: Let us define the time index set $\bar{\mathcal{T}} \equiv \{t \in \mathcal{T} : t \bmod 4 = 0\}$ (with $T \in \bar{\mathcal{T}}$ for notational simplicity); for example, if index 1 represents quarter 1 in year 1, $\bar{\mathcal{T}}$ would contain the indices of all 4th quarters over the years. Along these lines, we assume that the sum over quarterly dividend growth (i.e., annual dividend growth) is observed at the end of each year—so, in quarter 4.

Furthermore, we define the time-aggregated dividend state ΔD_t for the long-run risk model (28) by

$$\Delta D_t = \begin{cases} \mu_d + \Phi x_t + \sigma_d \eta_{d,t} & t+3 \in \bar{\mathcal{T}} \\ \Delta D_{t-1} + \mu_d + \Phi x_t + \sigma_d \eta_{d,t} & \text{otherwise.} \end{cases}$$

At the same time, the time-aggregated dividend state ΔD_t for the stochastic volatility model (29) reads

$$\Delta D_t = \begin{cases} \mu_d + \phi_d \sigma_c e^{h_t} \eta_{d,t} & t+3 \in \bar{\mathcal{T}} \\ \Delta D_{t-1} + \mu_d + \phi_d \sigma_c e^{h_t} \eta_{d,t} & \text{otherwise.} \end{cases}$$

The available data in the occasionally observed regime is then given by $\{\{\Delta c_t\}_{t\in\mathcal{T}}, \{\Delta D_t\}_{t\in\bar{\mathcal{T}}}\}$ for each model. Note that the states x_t and h_t are unobserved in both the full and the occasional observation regime in the respective models, while the state ΔD_t, which captures the sum over dividend growth *within* a year, is only observed every fourth

quarter in the occasional observation regime.

3.1.3 The likelihood function

We begin with the likelihood in the "full information" regime—that is, with consumption and dividend observations at the same frequency; note that the states x_t in the long-run risk model (28) and h_t in the stochastic volatility model (29) are still (completely) unobserved. Denote by θ the vector of model parameters, which is given by $\theta \equiv (\mu_c, \sigma_c, \rho_x, \sigma_x, \mu_d, \Phi, \sigma_d)$ for the long-run risk model and $\theta \equiv (\mu_c, \sigma_c, \rho_h, \sigma_h, \mu_d, \phi_d)$ for the stochastic volatility model. Since the likelihoods of the two models have a very similar structure, we denote the underlying completely unobserved states, x_t and h_t, respectively, by s_t in the following, and complement it with a tilde if it constitutes an integration variable. For full consumption and dividend data $\{\Delta c_t, \Delta d_t\}_{t=1}^T$, the likelihood is given by

$$L(\theta) = \int \cdots \int p(\tilde{s}_1;\theta)p(\Delta c_1|\tilde{s}_1;\theta)p(\Delta d_1|\tilde{s}_1;\theta) \cdot \prod_{t=2}^{T} p(\Delta c_t|\tilde{s}_t;\theta)p(\Delta d_t|\tilde{s}_t;\theta)p(\tilde{s}_t|\tilde{s}_{t-1};\theta)\,\mathrm{d}(\tilde{s}_1,\ldots,\tilde{s}_T)\,.$$

Note that we still need to integrate out the fully unobserved state s_t, representing either x_t or h_t, depending on which model is estimated.

In the occasional observation regime, where dividends are only observed annually and in an aggregated form, the available data is given by $\{\{\Delta c_t\}_{t\in\mathcal{T}}, \{\Delta D_t\}_{t\in\bar{\mathcal{T}}}\}$. Hence, $\Delta\tilde{D}_t$ is integrated out for $t \in \mathcal{T}\setminus\bar{\mathcal{T}}$, and ΔD_t is taken from the dataset otherwise ($t \in \bar{\mathcal{T}}$). This creates a nonstandard domain of integration—indeed, a lower-dimensional sub-manifold of $\mathbb{R}^T$—

$$\mathcal{D} \equiv \left\{(\Delta d_1,\ldots,\Delta d_T) \in \mathbb{R}^T : \sum_{i=0}^{3} \Delta d_{t-i} = \Delta D_t, t \in \bar{\mathcal{T}}\right\},$$

for the likelihood function,

$$L(\theta) = \int \cdots \int_{\mathbb{R}^T\times\mathcal{D}} \prod_{t=1}^{T} p(\Delta c_t|\tilde{s}_t;\theta) \cdot p(\Delta\tilde{d}_t|\tilde{s}_t;\theta)p(\tilde{s}_t|\tilde{s}_{t-1};\theta)\mathrm{d}S((\Delta\tilde{d}_t)_{t\in\mathcal{T}})\,\mathrm{d}(\tilde{s}_t)_{t\in\mathcal{T}}\,.$$

Note that we slightly abuse notation and write the stationary distribution of s as $p(s_1|s_0;\theta) \equiv p(s_1;\theta)$ for brevity. Moreover, note that the data on dividend growth enters the likelihood function only through the domain of integration in this formulation. Finally, since the integration is carried out over a $(3T/4)$-dimensional sub-manifold, we need to introduce—for now just symbolically—the $(3T/4)$-dimensional "surface element," $S(x)$, for the integration to be well defined; otherwise, the integration domain would be a set of measure zero.

Due to the special structure of $\mathcal{D}$, we can carry out a simple change of variables, implying

$$\mathrm{d}S((\Delta\tilde{d}_t)_{t\in\mathcal{T}}) = \mathrm{d}^{3T/4}\Big(\Delta\tilde{D}_t\Big)_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}}\,,$$

and write the likelihood function as an integral over a standard domain:

$$\begin{aligned} L(\theta) = \int \cdots \int_{\mathbb{R}^{T+3T/4}} \prod_{t \in \bar{\mathcal{T}}} & p(\Delta c_{t-3} | \tilde{s}_{t-3}; \theta) p(\Delta \tilde{D}_{t-3} | \tilde{s}_{t-3}; \theta) p(\tilde{s}_{t-3} | \tilde{s}_{t-4}; \theta) \\ & \cdot p(\Delta c_{t-2} | \tilde{s}_{t-2}; \theta) p(\Delta \tilde{D}_{t-2} | \Delta \tilde{D}_{t-3}, \tilde{s}_{t-2}; \theta) p(\tilde{s}_{t-2} | \tilde{s}_{t-3}; \theta) \\ & \cdot p(\Delta c_{t-1} | \tilde{s}_{t-1}; \theta) p(\Delta \tilde{D}_{t-1} | \Delta \tilde{D}_{t-2}, \tilde{s}_{t-1}; \theta) p(\tilde{s}_{t-1} | \tilde{s}_{t-2}; \theta) \\ & \cdot p(\Delta c_t | \tilde{s}_t; \theta) p(\Delta D_t | \Delta \tilde{D}_{t-1}, \tilde{s}_t; \theta) p(\tilde{s}_t | \tilde{s}_{t-1}; \theta) \, \mathrm{d}\Big(\Delta \tilde{D}_t\Big)_{t \in \mathcal{T} \setminus \bar{\mathcal{T}}} \mathrm{d}(\tilde{s}_t)_{t \in \mathcal{T}} \,. \end{aligned}$$

To cast it recursively as in Equation (17), we directly incorporate another change of variables to express the distribution of ΔD_t (conditional on ΔD_{t-1}) by the distribution of Δd_t:

$$f_t^\theta(s, \Delta D) = \begin{cases} 1 & t > T \\ \int p(\Delta c_t | \tilde{s}; \theta) p(\Delta D_t | \Delta D, \tilde{s}; \theta) p(\tilde{s} | s; \theta) & \\ \quad \cdot f_{t+1}^\theta(\tilde{s}, \Delta D_t) \mathrm{d}\tilde{s} & t \in \bar{\mathcal{T}} \\ \iint p(\Delta c_t | \tilde{s}; \theta) p(\Delta \tilde{d} | \tilde{s}; \theta) p(\tilde{s} | s; \theta) & \\ \quad \cdot f_{t+1}^\theta(\tilde{s}, \Delta \tilde{d}) \mathrm{d}\Delta \tilde{d} \mathrm{d}\tilde{s} & t+3 \in \bar{\mathcal{T}} \\ \iint p(\Delta c_t | \tilde{s}; \theta) p(\Delta \tilde{d} | \tilde{s}; \theta) p(\tilde{s} | s; \theta) & \\ \quad \cdot f_{t+1}^\theta(\tilde{s}, \Delta D + \Delta \tilde{d}) \mathrm{d}\Delta \tilde{d} \mathrm{d}\tilde{s} & \text{otherwise.} \end{cases} \tag{30}$$

Note that this formulation renders standard Gauss–Hermite quadrature applicable.

Finally, recall that we assumed $t = 4$ to be the first observation of the aggregate state ΔD; therefore, the unconditional likelihood is given by

$$L(\theta) = \iint p(\Delta c_1 | \tilde{s}_1; \theta) p(\Delta \tilde{d}_1 | \tilde{s}_1; \theta) p(\tilde{s}_1; \theta) f_2^\theta(\tilde{s}_1, \Delta \tilde{d}_1) \mathrm{d}\Delta \tilde{d}_1 \mathrm{d}\tilde{s}_1. \tag{31}$$

Depending on the concrete distribution of s and Δd, we can choose adequate changes of variables to make efficient quadrature rules applicable in (30) and (31); we comment on our strategy for the numerical approximation in the next subsection.

3.1.4 Results: Monte Carlo study

We conduct an extensive Monte Carlo estimation study to demonstrate the high efficiency of our RLI approach for estimating the consumption and dividend dynamics (28) and (29), respectively, with quarterly consumption and annual dividend data. We compare the RLI approach to the simulated method of moments (SMM) approach for occasional state observations introduced by Hall and Rust (2021). To compare the asymptotic behavior of the methods in this example, we consider simulated datasets with sample lengths of 50, 100, 200, 400, and 800 years, respectively—so $T \in \{200, 400, 800, 1600, 3200\}$ quarters. Note that quarterly consumption data starts in 1947 so there are only 294 observations available as of now, which makes around 74 years of data. For each T we simulate $N = 400$ datasets using the monthly calibration of Bansal and Yaron (2004) scaled to a quarterly frequency.[23] The parameters used for the

[23] To compute quarterly from monthly parameters, means are multiplied by 3, standard deviations by $\sqrt{3}$, and the persistence is taken to the power of 3. Bansal and Yaron (2004) use a linearized version for the stochastic

simulation are reported in Table 1, and we refer to the vector of all parameters of the model as θ throughout this section.

Long-Run Risk Model (28)						
μ_c	σ_c	ρ_x	σ_x	μ_d	Φ	σ_d
0.0045	0.0135	0.9383	0.0010	0.0045	3	0.1053
Stochastic Volatility Model (29)						
μ_c	σ_c	ρ_h	σ_h	μ_d	ϕ_d	
0.0045	0.0135	0.9615	0.0109	0.0045	4.5	

Table 1: Parameter calibrations for the asset pricing models (28) and (29) used in the Monte Carlo study. The values are based on the monthly calibration of Bansal and Yaron (2004), adjusted to a quarterly frequency.

For the RLI approach, we use cubic spline interpolation with 50 nodes in each dimension, and compute the integrals by Gauss–Hermite quadrature with 11 nodes per dimension. We compare the results to the SMM estimator for endogenously sampled data by Hall and Rust (2021) using the following parametrization: for the HAC weighting matrix we use $\left\lfloor \sqrt[5]{T} \right\rceil$ lags, where $\lfloor \cdot \rceil$ denotes the rounding operator (rounding to the closest integer). We employ the following moments for both consumption and dividend data: for the long-run risk model (28), we use the first and second non-central moments, autocorrelations up to order 10, as well as the cross-correlation between consumption and dividend growth. For the stochastic volatility model, we additionally use the third, fourth, fifth, and sixth non-central moments to account for higher-order effects and for autocorrelations up to order 10.[24] All computations are carried out in MATLAB.[25]

For the assessment and comparison of the different estimation methods, we use the Mahalanobis distance. The Mahalanobis distance D_M is defined as

$$D_M(\hat{\theta}; \theta, \Sigma) \equiv \sqrt{(\hat{\theta} - \theta)\Sigma^{-1}(\hat{\theta} - \theta)^T}, \tag{32}$$

and measures the distance between the vector $\hat{\theta}$ and a distribution with mean θ and covariance matrix Σ. Since the Mahalanobis distance corrects for the variance of the estimators, it not only normalizes the size of the parameters, it also takes into account how well a parameter is identified relative to the other parameters for some reference method used to estimate Σ, and thus allows meaningful comparisons of the properties of the methods in question, and not just

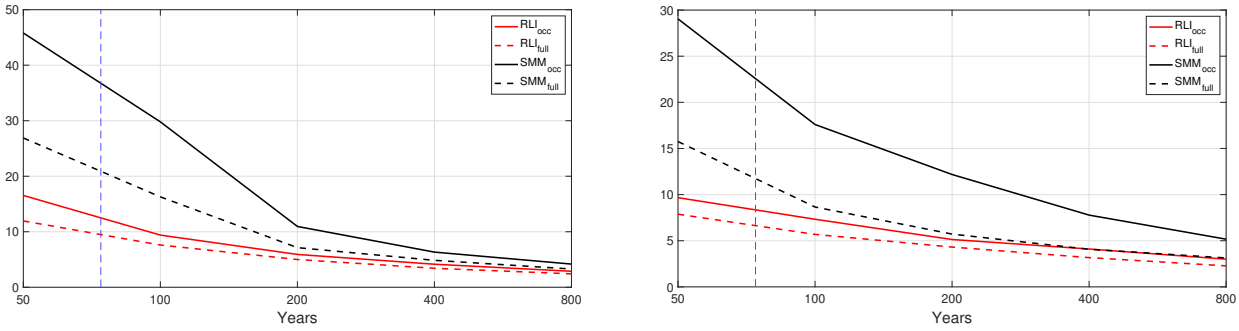
volatility process, which has the disadvantage of allowing negative variances. To obtain the parameters for the exponential process, we use $\sigma_h = \sigma_s/(2\sigma_c^2)$ where σ_s is the volatility of the linearized process; see Schorfheide et al. (2018).

[24] We have tried different numbers of autocorrelations and lags for the HAC weighting matrix; the specification reported here is the one that yields the lowest errors. We also used the Lasso GMM approach by Cheng and Liao (2015) to automatically select the relevant moments. The results are reported in Appendix B.1. We show that our findings are robust with regard to the specific selection of moments and, in particular, using the Lasso GMM approach does not change the conclusions we draw regarding the comparison of the SMM and RLI estimator. We thank an anonymous referee for suggesting this additional robustness check.

[25] We discard all runs that do not converge. For RLI we have that in most cases, all runs or all but one run converge and we get a maximum of 1% non-converged runs, see Table 2 in Appendix B.1. We get similar numbers for SMM with a maximum of 1.25% non-converged runs.

properties of the model or the dataset. Furthermore, we use the Mahalanobis distance in Q–Q plots to analyze the asymptotic normality of our estimators.

Figure 1 plots the median of the Mahalanobis distances over the 400 simulated datasets for different dataset lengths.[26] The Mahalanobis distance (32) for each estimate $\hat{\theta}$ is computed using the true parameter vector θ; as an estimate of Σ, we use the covariance matrix of $\hat{\theta}$ obtained from RLI under the full information regime and the longest dataset, of 800 years; the same Σ is used for both SMM and RLI to make the results comparable. Red lines depict the results for RLI and black lines those for SMM, while dashed lines denote the full observation regime and solid lines the occasional observation regime. Panel (a) shows the results for the long-run risk model (28) and panel (b) those for the stochastic volatility model (29).



(a) Long-Run Risk Model (28)

(b) Stochastic Volatility Model (29)

Figure 1: Median of the Mahalanobis distance over 400 simulated datasets as a function of dataset length in years, using RLI (red lines) and SMM (black lines) under the full observation regime (dashed lines) and the occasional observation regime (solid lines). The dashed blue line shows the number of observations in the real consumption and dividend dataset. Panel (a) shows the results for the long-run risk model (28) and panel (b) for the stochastic volatility model (29).

We observe that for each model the errors for RLI with occasional state observations decrease with the size of the dataset, showing the consistency of our method. Moreover, in each model the errors for RLI in the occasional observation regime are only slightly larger than those in the full information case, although there are four times as many observations of the aggregate dividend process available in the full information case compared to the occasional observation regime.

The errors for SMM in the occasional observation regime also converge for large datasets. However, for small datasets the errors are significantly larger than those of the RLI estimator: approximately 3–4 times as many data points are needed for SMM to achieve the same accuracy as RLI in the occasional observation regime. For the stochastic volatility model, where the state process enters the consumption and dividend dynamics non-linearly, the evidence for the consistency of the SMM estimator in the occasional observation regime is much weaker—a point we investigate in detail below. We observe a large difference for SMM between the full and the occasional observation regime, which shows that the missing state observations significantly

[26] In Figure 11 in Appendix B.1 we plot kernel densities of the Mahalanobis distance which show its full distribution instead of only the median. The kernel density plots show the same patterns as the median and hence we focus on the median in the main text which allows for easier interpretation.

affect the estimation outcomes. The difference is considerably smaller for RLI, which suggests that it can handle the missing state observations more efficiently. Furthermore, the errors for RLI under the occasional observation regime are significantly smaller than the errors for SMM under the occasional information regime, and approximately 5 times as many data points are needed for SMM to achieve the same accuracy as RLI in the occasional observation regime.

In the following, we analyze in detail where the differences between the methods come from. For this, we first analyze the joint normality of the estimators using Q-Q plots; second, we look at kernel density plots to analyze in which dimensions the methods fail to estimate different parameters. Figure 2 shows Q-Q plots of the roots of the χ^2 quantiles with seven degrees of freedom against sample quantiles of the Mahalanobis distance of the estimates for the long-run risk model (28). Note that in contrast to the analysis of Figure 1, which measures the deviation of estimates from the true value of the parameter—projected to a single dimension via the Mahalanobis distance based on a *common* weighting matrix—the analysis of the distribution of the estimators and their comparison to a theoretical benchmark requires us to obtain an estimate of Σ for each sample size T and each method separately. Therefore, we use the empirical covariance matrix $\hat{\Sigma}$ obtained from the respective 400 Monte-Carlo samples with length T for the corresponding method.[27] Our Q-Q plots can be interpreted as follows: Any shift of the distribution away from the 45 degree line corresponds to a bias in the estimate, whereas any rotation centered at zero corresponds to a different covariance matrix. One or more crossings of the 45 degree line (i.e., s-shapes, u-shapes, etc.) implies a different skewness or kurtosis and, hence, non-normality. From visual inspection, we find strong evidence that the RLI estimator asymptotically approaches normality in the occasional observation regime, and that the convergence appears to be almost as fast as in the full observation case. While SMM shows some s-shaping for small samples, we also cannot reject asymptotic normality; this does not come as a surprise given the linear and conditional Gaussian structure of the model.

Figure 3 shows the corresponding Q-Q plots for the stochastic volatility model (29). For RLI in the occasional observation regime, we observe some s-shaping, which most likely comes from the skewness induced by the natural bound of ρ_h at 1, as we argue below using individual kernel density plots. As in the long-run risk model, RLI approaches normality almost as quickly in the occasional observation regime as in the full information case. In contrast, for SMM we observe substantial deviations from the 45 degree line and hence find evidence against the normality of the estimators. Note that as the points are almost linear and close to cutting the origin, the deviations could also suggest a failure in the estimation of the covariance matrix, which can, for example, happen when the distribution is essentially flat in some dimensions. We find evidence for this hypothesis by looking, in the following analysis, at kernel densities.

In the following we analyze kernel density plots for individual parameters. We consider one parameter of the unobserved process and one of the occasionally observed process. Kernel densities for all parameters can be found in Appendix B.1. Figure 4 shows the kernel densities for the persistence ρ_x of the unobserved process in the long-run risk model. The distributions for

[27] To minimize the impact of outliers introduced by the Monte Carlo approach when estimating Σ, we drop the 2% of runs with the largest Mahalanobis distance. Note, however, that we only exclude them from the estimation of the covariance matrix, but not from the samples used to obtain the Q-Q plots. Therefore, the plots are expressive and the conclusions are robust even in the presence of outliers.

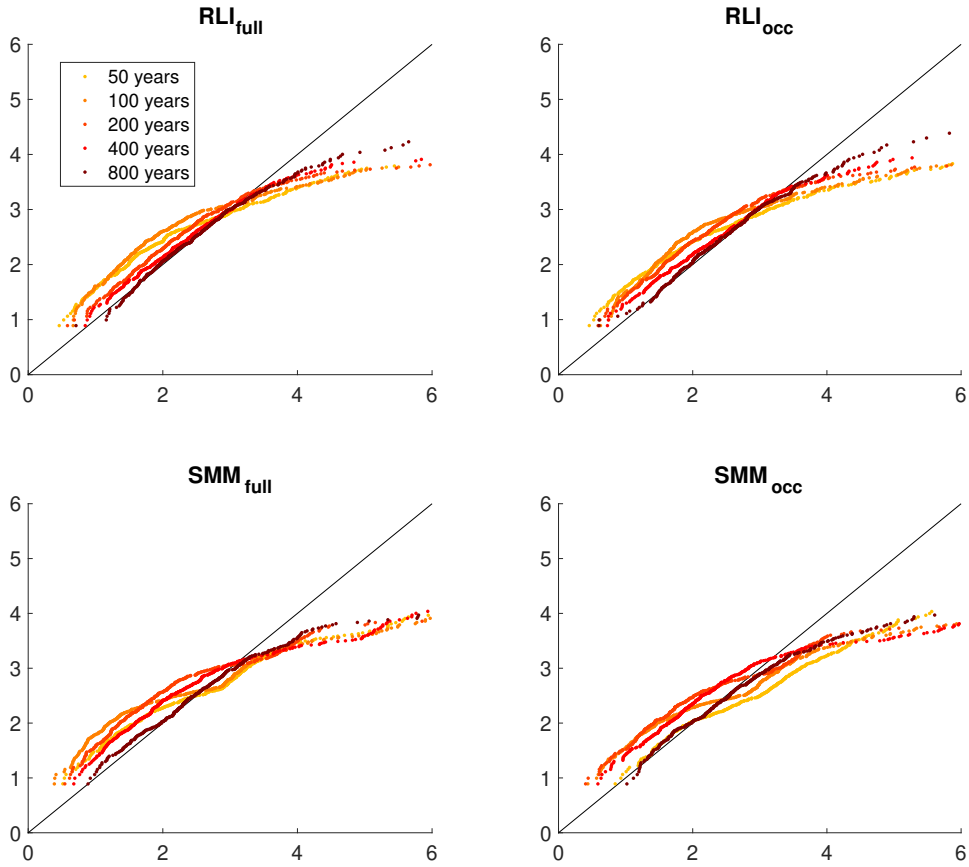


Figure 2: Q-Q plots of the roots of χ^2 quantiles with seven degrees of freedom against sample quantiles of the Mahalanobis distance of the estimates. Results are shown for the long-run risk model (28) using either RLI or SMM in the full or occasional observation regime.

RLI in the occasional observation regime are approaching bell shapes and have modes close to the true parameter value. The distributions are very similar to those in the full information regime. For SMM, the distributions are also bell shaped but with significantly higher variance, which is consistent with the higher errors reported in Figure 1. Figure 5 shows the corresponding kernel densities for the volatility σ_d of the (occasionally) observed dividend process. As expected, the densities for the occasional observation regime show a slightly larger variance compared to those of the full information case. This holds for both RLI and SMM. However, for small datasets SMM shows a large bias especially in the occasional observation regime, which is reflected in the larger errors for small datasets, see Figure 1.

Figure 6 shows kernel densities for the persistence ρ_h of the unobserved process in the stochastic volatility model. Convergence for RLI is slower than for the distributions in the long-run risk model, which reflects the complexity of the non-linear volatility dynamics; but even in the occasional observation regime the distributions approach bell shape with modes close to the true parameter. This departure from normality, partially due to the boundedness of the ρ_h parameter, explains the deviation from the 45 degree line reported in Figure 3.

In contrast, SMM experiences substantial problems in identifying the key model parameters.

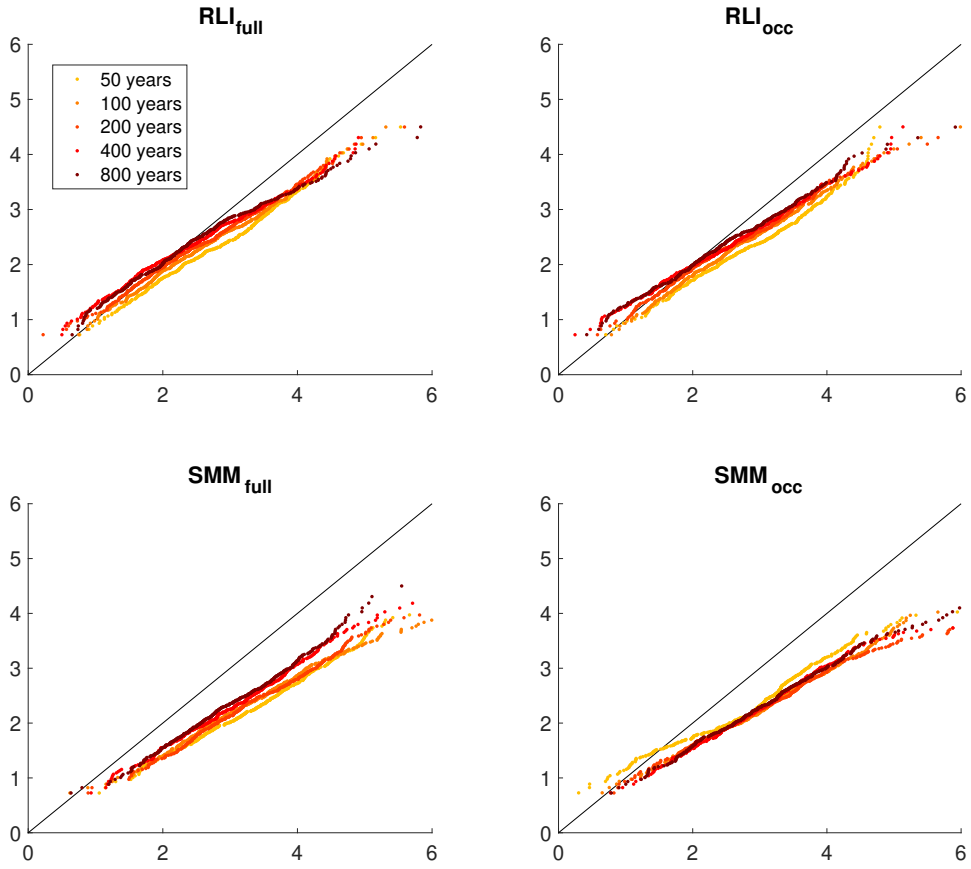


Figure 3: Q-Q plots of the roots of χ^2 quantiles with six degrees of freedom against sample quantiles of the Mahalanobis distance of the estimates. Results are shown for the stochastic volatility model (29) using either RLI or SMM in the full or occasional observation regime.

For both the full information regime and the occasional observation regime the kernel densities for ρ_h are almost flat. SMM is not able to identify the persistence ρ_h; σ_h is not well identified either (see Figure 13 in Appendix B.1). These findings are in line with Grammig and Küchlin (2018), who argue that SMM is not able to identify the parameters of the unobserved stochastic volatility process. We attribute this to a failure to properly estimate the covariance matrix reflected in the flat kernel densities. This explains the deviation in the Q–Q plots (see Figure 3) as well as the larger and maybe even non-converging errors reported in Figure 1. Importantly, this holds for both the full information and the occasional information regime and hence is a feature of the SMM estimator itself rather than its restriction to occasional observations.

Figure 7 shows the kernel densities for the volatility ϕ_d of the (occasionally) observed dividend process. In line with the findings for the long-run risk model, densities for RLI under the occasional observation regime show a slightly larger variance compared to those under the full information case. Furthermore, SMM shows a large bias in the occasional observation regime due to the availability of less data for the dividend process. This bias adds to the deviation from the 45 degree line in the Q–Q plots reported in Figure 3. For completeness, Figures 12 and 13 in Appendix B.1 plot kernel densities for all model parameters for the long-run risk

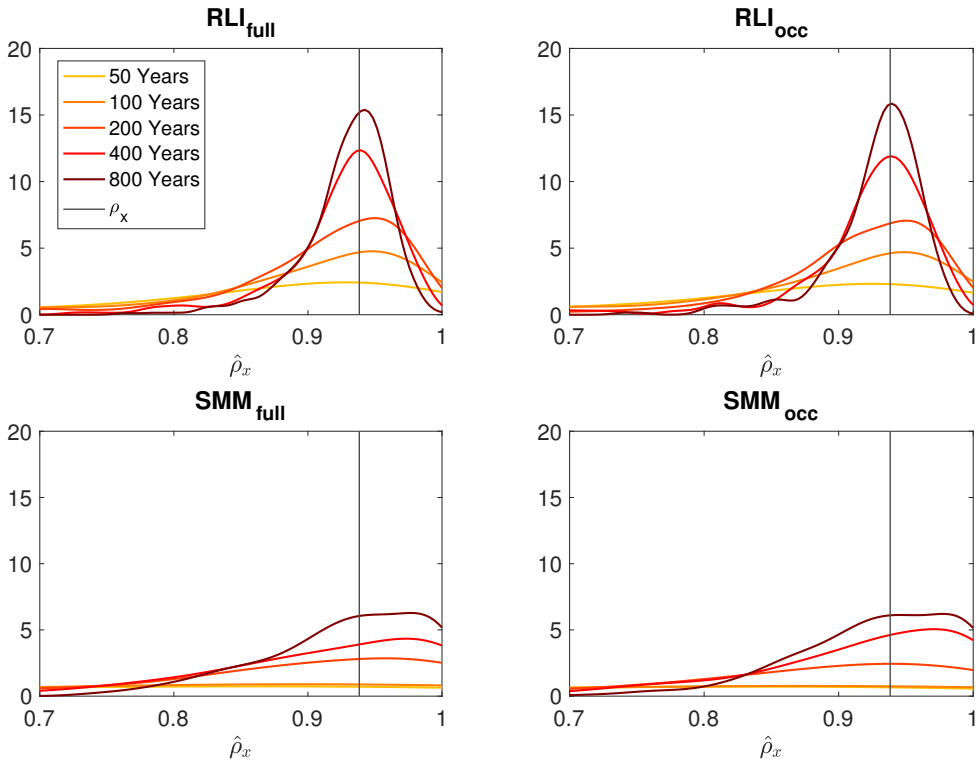


Figure 4: Kernel density estimates for ρ_x for the long-run risk model (28) using RLI (top panels) and SMM (bottom panels) under the full observation regime (left panels) and the occasional observation regime (right panels) for 400 simulated datasets. The color map from yellow to red show results for 50, 100, 200, 400, and 800 years.

and the stochastic volatility model, respectively. We find that RLI yields at least as efficient, and in many cases more efficient, estimates for each parameter individually compared to SMM. Moreover, with few exceptions RLI with occasional state observations is even more efficient than SMM under the full information regime.

To conclude, we find that our RLI estimator for occasionally observed data consistently identifies all model parameters, even for the non-linear and more complex stochastic volatility model. Furthermore, there is no strong evidence against the normality of our estimator even for very small data samples, and the difference in the errors between the full information and occasional information regime decreases with the sample size. The performance of SMM is comparable with RLI for the simple and linear long-run risk model. It fails, however, in estimating the key parameters of the stochastic volatility model. This problem is specific to SMM and does not depend on the observation pattern of the dividend process. This showcases another important aspect of our RLI approach: the broad applicability of a likelihood-based estimation approach.

We conclude this section with a few of remarks regarding other aspects of the comparison of RLI and SMM, aspects that we have not touched on thus far: First and foremost, the RLI approach is only applicable if the likelihood function can be specified in a non-degenerate way. This requires a specification of—and data on—individual behavior as well as rich enough error terms, which are not always available. For example, the steel trading application considered

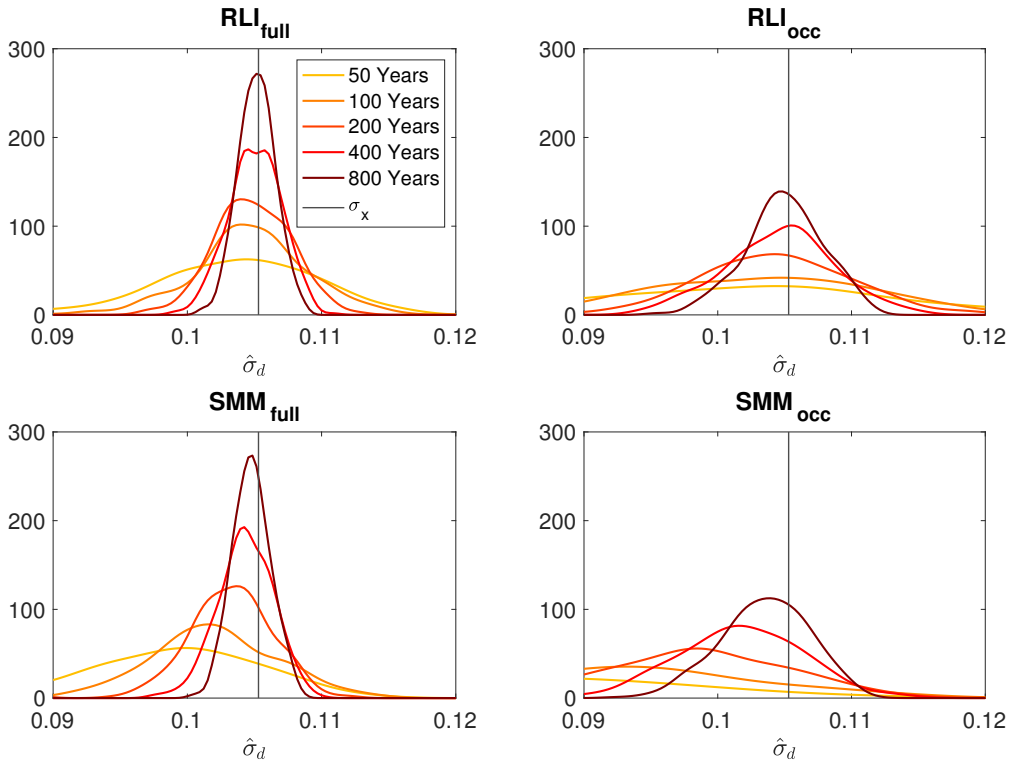


Figure 5: Kernel density estimates for σ_d for the long-run risk model (28) using RLI (top panels) and SMM (bottom panels) under the full observation regime (left panels) and the occasional observation regime (right panels) for 400 simulated datasets. The color map from yellow to red show results for 50, 100, 200, 400, and 800 years.

by Hall and Rust (2021) would induce a degenerate likelihood function, which is non-zero only on a set of measure zero, and thus can not be estimated using RLI; rather, it requires a more generally applicable approach such as SMM. On the other hand, if a (non-degenerate) likelihood function is available, the RLI approach has less "degrees of freedom"—or method-related parameters—and is thus easier to configure from the viewpoint of the researcher: while SMM requires an explicit choice of moments (HAC bandwidth, etc.) and thus potentially significant experimentation and optimization effort, the use of RLI only requires the—mostly straightforward—configuration of the numerical methods for integration and interpolation, for which Reich (2018) provides useful guidelines directly based on the convergence rates of the respective methods. Finally, the fact that the objective function in SMM is subject to simulation noise requires special attention in order to avoid the moment criterion minimization getting stuck in noise-induced local optima; see Hall and Rust (2021), and the references cited therein, for various mitigation strategies.

3.2 The optimal replacement of GMC bus engines (Rust, 1987)

To assess the performance of our method for the estimation of a controlled Markov process with occasional state observations, we apply it to a hypothetical scenario of the well-understood bus engine replacement model of Rust (1987): we assume that the main state variable of the

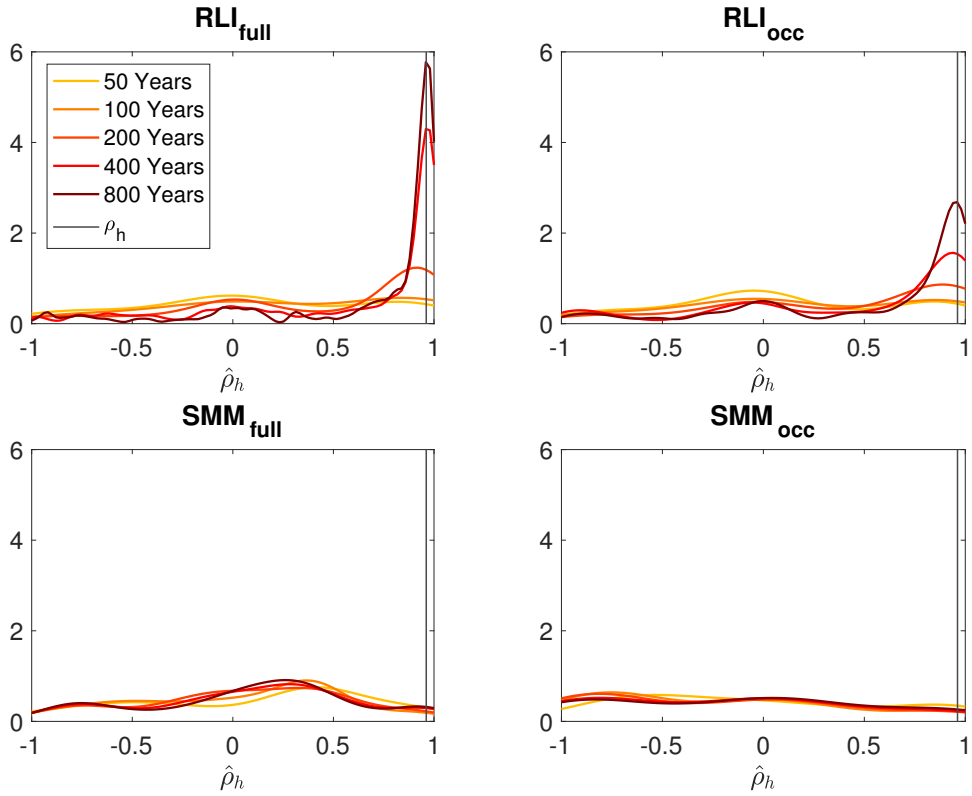


Figure 6: Kernel density estimates for ρ_h for the stochastic volatility model (29) using RLI (top panels) as well as SMM (bottom panels) under the full observation regime (left panels) and the occasional observation regime (right panels) for 400 simulated datasets. The color map from yellow to red show results for 50, 100, 200, 400, and 800 years.

model—the mileage a bus has traveled since its engine was last replaced—is only observed when a particular action takes place; namely engine replacement. To assess the properties of the resulting estimator, we conduct a Monte Carlo simulation study to compare the distribution of the estimators under the full and the occasional observation regimes. To demonstrate the feasibility, the flexibility, and both the statistical and the numerical efficiency of our approach, we use a modified version of the model, featuring a continuous mileage state. We find that the estimator for the parameters of interest under occasionally observed states is, in essence, as efficient as the estimator under full state observations, although it uses only around 2% of the available state data.

3.2.1 Rust (1987) with continuous mileage state

In this canonical dynamic discrete choice model, Harold Zurcher, the manager of a fleet of public transportation buses, faces a dynamic renewal problem: On a regular basis, he inspects all the buses of his fleet. He can decide to fully overhaul a bus, which, most importantly, implies the renewal of its engine; such a bus counts as new and its odometer is reset to zero. Or he can do only the regular maintenance work necessary to keep the bus in service; this option usually comes at lower immediate costs compared to engine replacement, but costs increase with the mileage driven since the vehicle was last fully overhauled. This models a common trade-off—

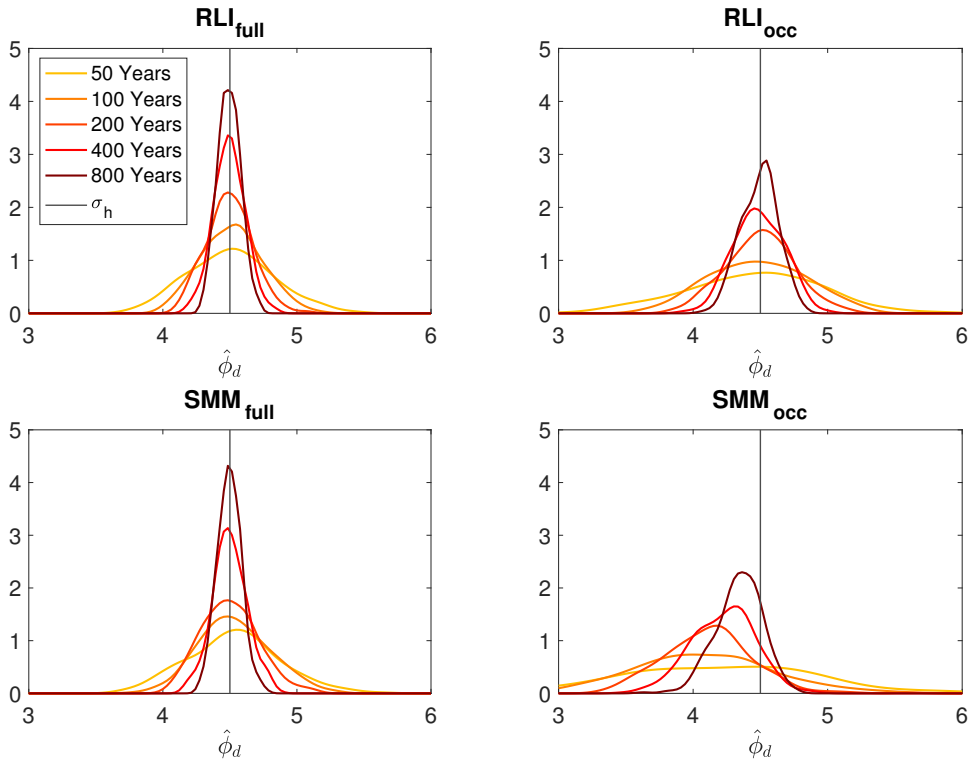


Figure 7: Kernel density estimates for ϕ_d for the stochastic volatility model (29) using RLI (top panel) as well as SMM (bottom panel) under the full observation regime (left panel) and the occasional observation regime (right panel) for 400 simulated datasets. The color map from yellow to red show results for 50, 100, 200, 400 and 800 years.

namely, whether or not to further invest in an old machine to keep it in service, as opposed to replacing it by a new one to reduce future (expected) maintenance costs.

Formally, the agent faces the immediate utility function

$$u(x, i; \theta_1, RC) + \epsilon(i) \equiv \begin{cases} -RC + \epsilon(1) & i = 1 \\ -0.001 \cdot \theta_1 \cdot x + \epsilon(0) & i = 0 \end{cases}$$

for each individual bus, where x is the current mileage (i.e., odometer reading) of the bus, i is the decision of the agent, and θ_1 and RC are two structural parameters of the model. Having decided to replace the engine ($i = 1$), the agent receives a constant, negative utility, $-RC$, plus some random utility shock (with fixed mean), $\epsilon(1)$. When the agent decides to carry out regular maintenance work ($i = 0$), he or she receives a utility $-0.001 \cdot \theta_1 \cdot x$ that is linearly decreasing in mileage, plus some utility shock $\epsilon(0)$. We refer to RC as the replacement costs and θ_1 as the maintenance cost parameter. Both are to be estimated from the observed data. Following Rust (1987), we assume $\epsilon(0)$ and $\epsilon(1)$ to be extreme value type I (EV1) i.i.d.

The model assumes that the agent behaves in a dynamically optimal manner—that is, the agent maximizes the expected sum of discounted future payoffs. A sufficient condition for such

optimality is given by the well-known Bellman equation (Bellman, 1952),

$$V_\theta(x,\epsilon) = \max_{i\in\{0,1\}} \left\{ u(x,i;\theta_1,RC) + \epsilon(i) + \beta \mathbb{E}\left[V_\theta(x',\epsilon') | x,i;\theta_3 \right] \right\}, \tag{33}$$

which has to hold for all possible values of (x,ϵ) in the state space. β denotes the discount factor, and we follow Rust (1987) in fixing it at 0.9999. In order to compute the expectation over the future value in Equation (33), we need to specify a law of motion of the mileage state variable, parametrized by θ_3. We follow Rust's conditional independence assumption and refer to this distribution as $P(x_t|x_{t-1},i_{t-1};\theta_3)$.

Instead of the original Rust (1987) specification where the odometer readings are discretized, we modify the model to feature a continuous mileage state. This specification comes naturally as mileage is an inherently continuous measure, and as the raw data measuring mileage is virtually continuous (i.e., merely rounded). To this end, we assume stationarity of the mileage increments conditional on engine replacement,

$$p(x_t|x_{t-1},i_{t-1};\theta_3) = \begin{cases} q(x_t - x_{t-1};\theta_3) & i_{t-1} = 0 \\ q(x_t;\theta_3) & i_{t-1} = 1, \end{cases}$$

for some parametric density $q(\cdot;\theta_3)$. In other words, the distribution of the increment,

$$\Delta x = x_t - (1 - i_{t-1})x_{t-1},$$

is independent of the decision. Specifically, we assume the increment to follow a log-normal distribution: $\Delta x \sim LN(\mu,\sigma)$, with parameter vector $\theta_3 \equiv (\mu,\sigma)$. We further analyze the fit of this model in Appendix B.2, concluding that the estimated log-normal density provides a reasonable fit.[28] The specifics of the solution of the dynamic program in the presence of the proposed law of motion and how it feeds into the likelihood function can be found ibidum.

3.2.2 Rust (1987) with occasionally observed mileage state

In the remainder of the section, we study the following hypothetical scenario compared to the standard Rust (1987) model: Suppose the fleet manager outsources engine replacement to a third-party company, which can only record the odometer readings when the bus comes to its repair shop for engine replacement; thus, the third-party cannot collect any data on the odometer readings in between replacements. Moreover, we assume that the third-party has access to a document, such as the vehicle's registration certificate, that states when the bus was put into service, giving us the first mileage state observation. In this setup, we ask if it is still possible to estimate the manager's cost trade-off accurately using the limited dataset of only about 2 percent of the state observations.

Formally, we want to compare the estimators for the full observation regime with state observation index set (per bus) $\mathcal{T} = \{1,\dots,T\}$, denoted by $\hat{\theta}^{full}$, and the occasional observation regime with $\bar{\mathcal{T}} = \{t : i_t = 1\} \cup \{1\}$ and estimator $\hat{\theta}^{occ}$. To do so, we compare the estimators'

[28] Rust (1987) and, more recently, Lanz et al. (2022) suggest using exponentially distributed mileage increments; we, however, find the log-normal model to provide a better fit for the original dataset.

distributions based on a Monte Carlo simulation of datasets while keeping their time horizon of comparable length to the original dataset on average with $T = 80$. To assess the estimators' efficiency, we compare kernel density fits of their distributions.

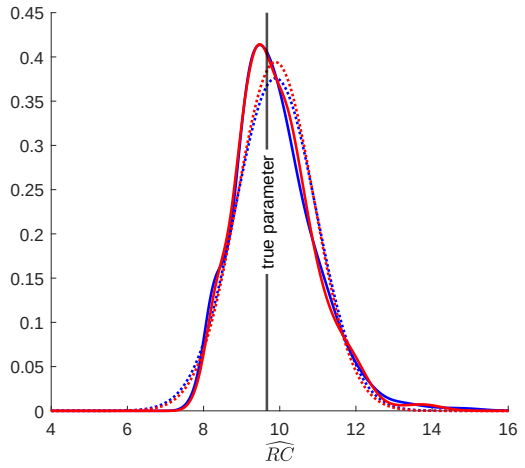
To efficiently and accurately evaluate the likelihood function under occasional state observations, we employ the continuous state variant of the recursive formulation (26). As discussed in Section 2.3, continuous state spaces require efficient numerical approximations for the integrals induced by the likelihood function recursion. Analogously to the computation of the integrals in the expected value function (see Appendix B.2, Eqn. 45), we use Gauss–Hermite quadrature to integrate the log-normally distributed mileage increments. We provide the full likelihood recursion in Appendix B.2, Equations (47) and (48). The maximization of the likelihood function subject to the constraints implied by the model is analogous to that in the full observation case of the previous section. All our computations use MATLAB with CasADi (Andersson et al., 2018) for automatic differentiation as well as the constrained solvers KNITRO and IPOPT (Wächter and Biegler, 2005).

3.2.3 Results: Monte Carlo study

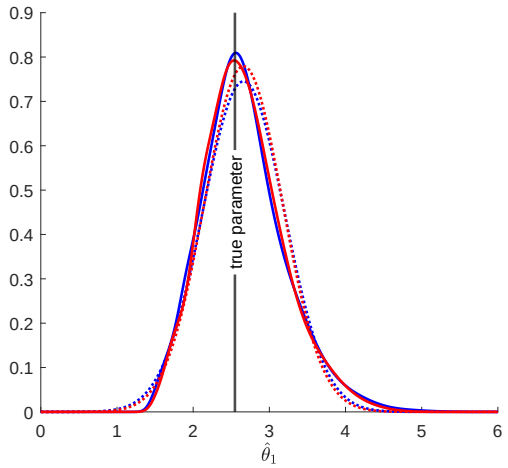
As motivated previously, we study the distribution of estimators of the structural parameters, $\hat{\theta}^{full}$ and $\hat{\theta}^{occ}$, under the full and the occasional observation regimes. These distributions are estimated in a Monte Carlo study with 400 simulated datasets. Figure 8 depicts a kernel fit of the distribution of the two estimators and a normal distribution with sample mean and variance of the estimates from each regime. Apparently, the costs parameters—and thus the objects of interest—are estimated from occasional observations as efficiently as under full mileage state observations; that is to say, $\widehat{RC}^{occ}$ and $\hat{\theta}_1^{occ}$ are unbiased and exhibit little or no additional variance compared to $\widehat{RC}^{full}$ and $\hat{\theta}_1^{full}$, respectively. Thus, with only around 2 percent of the original dataset, we achieve an almost equally good fit as under complete information for the quantity of interest. The estimators for the parameters of the law of motion, $\widehat{\mu}$ and $\hat{\sigma}$, show substantially more variance under occasional observations, which is anything but unexpected. We further elaborate on the properties of the distribution of the estimators in Appendix B.2.

4 Conclusion

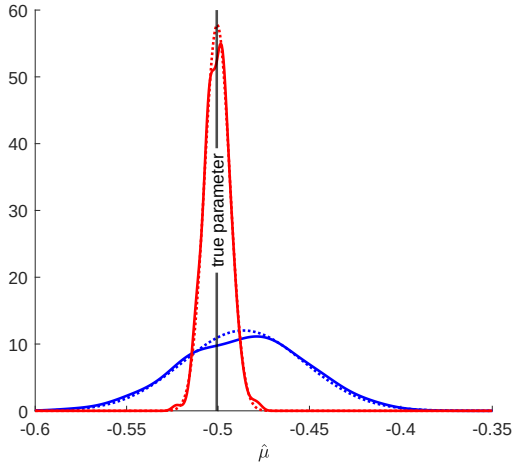
This paper's contribution is to allow for likelihood-based inference under occasional state observations, and harvest its favorable properties in small samples. We propose a method that generalizes the RLI procedure of Reich (2018) to cover various forms of occasional observability (e.g., random observations, endogenous observations, and observations following a time pattern). We provide a general likelihood formulation, which accounts for endogenous observation patterns, and show how it can be simplified if the observation process is exogenous. We demonstrate the high efficiency and broad applicability of our likelihood-based estimator. For this, we apply the proposed method to three relevant problems in finance and industrial organization: (i) a long-run risk model as in Bansal and Yaron (2004), (ii) a model with stochastic volatility as in Schorfheide et al. (2018), and (iii) a counterfactual setup of the famous bus engine replacement problem of Rust (1987).



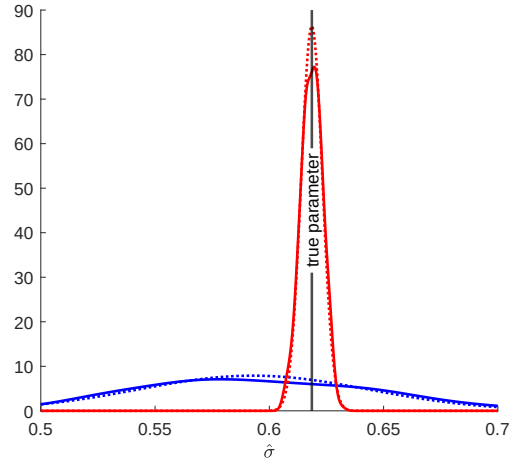
(a) Replacement Parameter RC



(b) Regular Maint. Cost Parameter θ_1



(c) Mileage Trans. Location Parameter μ



(d) Mileage Trans. Shape Parameter σ.

Figure 8: Distributions of the maximum likelihood estimators of the costs and transition parameters of Rust (1987) with continuous mileage states under the full observation regime (red) and the occasional observation regime (blue) for 400 simulated datasets. The solid lines show the kernel fit, the dotted lines the normal distribution using the sample mean and sample variance, the black vertical line the parameter value used for simulation.

We show in extensive Monte Carlo studies that our method can identify all model parameters with high efficiency, and we find that the additional variance of our estimator when going from full to occasional state observations is small for the parameters of interest. This is a valuable finding for the current discussion on optimal data provision as well as privacy considerations and raises the question: How much data do econometricians—and eventually companies—really need to generate satisfactory insights? A better understanding of data requirements, especially with sensitive consumer data, is important in the advent of increasingly strict (self-)regulations.

References

Andersson, J. A. E., Gillis, J., Horn, G., Rawlings, J. B., and Diehl, M. (2018). CasADi: a software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation*, 20(3):1–36.

Arcidiacono, P. and Miller, R. A. (2011). Conditional Choice Probability Estimation of Dynamic Discrete Choice Models with Unobserved Heterogeneity. *Econometrica: Journal of the Econometric Society*, 79(6):1823–1867.

Bansal, R., Kiku, D., and Yaron, A. (2012). An Empirical Evaluation of the Long-Run Risks Model for Asset Prices. *Critical Finance Review*, 1(1):183–221.

Bansal, R. and Shaliastovich, I. (2013). A long-run risks explanation of predictability puzzles in bond and currency markets. *The Review of Financial Studies*, 26(1):1–33.

Bansal, R. and Yaron, A. (2004). Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. *The Journal of Finance*, 59(4):1481–1509.

Bellman, R. (1952). On the Theory of Dynamic Programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716–719.

Blevins, J. R. (2016). Sequential Monte Carlo Methods for Estimating Dynamic Microeconomic Models. *Journal of Applied Econometrics*, 31(5):773–804.

Bollerslev, T., Tauchen, G. E., and Zhou, H. (2009). Expected Stock Returns and Variance Risk Premia. *Review of Financial Studies*, 22(11):4463–4492.

Chang, Y., Garcia, A., and Wang, Z. (2020). Dynamic Discrete Choice Estimation with Partially Observable States and Hidden Dynamics. *arXiv.org*.

Cheng, X. and Liao, Z. (2015). Select the valid and relevant moments: An information-based lasso for gmm with many moments. *Journal of Econometrics*, 186(2):443–464. High Dimensional Problems in Econometrics.

Connault, B. (2016). Hidden Rust Models.

Cosslett, S. R. and Lee, L.-F. (1985). Serial Correlation in Latent Discrete Variable Models. *Journal of Econometrics*, 27(1):79–97.

Drechsler, I. and Yaron, A. (2011). What's Vol Got to Do with It. *Review of Financial Studies*, 24(1):1–45.

Engle, R. F. and Russell, J. R. (1998). Autoregressive Conditional Duration: A New Model for Irregularly Spaced Transaction Data. *Econometrica: Journal of the Econometric Society*, 66(5):1127.

Erdem, T., Keane, M. P., and Sun, B. (1999). Missing price and coupon availability data in scanner panels: Correcting for the self-selection bias in choice model parameters. *Journal of Econometrics*, 89(1-2):177–196.

Farmer, L. E. (2021). The discretization filter: A simple way to estimate nonlinear state space models. *Quantitative Economics*, 12(1):41–76.

Gilch, A., Reich, G., and Wilms, O. (2025). Asymptotic properties of the maximum likelihood estimator under occasionally observed states. Working paper.

Grammig, J. and Küchlin, E.-M. (2018). A two-step indirect inference approach to estimate the long-run risk asset pricing model. *Journal of Econometrics*, 205(1):6–33.

Griebel, M., Heiss, F., Oettershagen, J., and Weiser, C. (2019). Maximum approximated likelihood estimation. Available as University of Bonn INS Preprint No. 1905.

Hall, G. and Rust, J. (2021). Estimation of Endogenously Sampled Time Series: The Case of Commodity Price Speculation in the Steel Market. *Journal of Econometrics*, 222(1):219–243.

Hansen, L. P., Heaton, J. C., and Li, N. (2008). Consumption Strikes Back? Measuring Long Run Risk. *Journal of Political Economy*, 116(2):260–302.

Iskhakov, F. (2010). Structural dynamic model of retirement with latent health indicator. *The Econometrics Journal*, 13(3):126–161.

Judd, K. L. (1992). Projection Methods for Solving Aggregate Growth Models. *Journal of Economic Theory*, 58(2):410–452.

Judd, K. L. (1998). *Numerical Methods in Economics*. The MIT Press, Cambridge, MA.

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica*, 62, No. 1:95–116.

Kitagawa, G. (1987). Non-gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400):1032.

Lanz, A., Reich, G., and Wilms, O. (2022). Adaptive grids for the estimation of dynamic models. *Quantitative Marketing and Economics*, 20(2):179–238.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley.

Norets, A. (2009). Inference in Dynamic Discrete Choice Models with Serially Correlated Unobserved State Variables. *Econometrica: Journal of the Econometric Society*, 77(5):1665–1682.

Reich, G. (2018). Divide and Conquer: Recursive Likelihood Function Integration for Hidden Markov Models with Continuous Latent Variables. *Operations Research*, 66(6):1457–1470.

Rust, J. (1987). Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica: Journal of the Econometric Society*, 55(5):999–1033.

Schorfheide, F., Song, D., and Yaron, A. (2018). Identifying Long-Run Risks: A Bayesian Mixed-Frequency Approach. *Econometrica: Journal of the Econometric Society*, 86(2):617–654.

Su, C.-L. and Judd, K. L. (2012). Constrained Optimization Approaches to Estimation of Structural Models. *Econometrica: Journal of the Econometric Society*, 80(5):2213–2230.

Wächter, A. and Biegler, L. T. (2005). On the Implementation of an Interior-Point Filter Line-Search Algorithm for Large-Scale Nonlinear Programming. *Mathematical Programming*, 106(1):25–57.

A Online Appendix

A.1 Proof of Proposition 1

The likelihood of a parameter vector θ given the sample $\left\{\{y_t\}_{t=1}^T, \{x_t\}_{t\in\bar{\mathcal{T}}}\right\}$ is the joint probability of the sample in the underlying model under parametrization with θ. We use definitions (7) for P^m, (8) for P^{zx}, the notational conventions from Footnote 10, as well as Assumption A1 and the assumptions of the proposition to transform the likelihood to the shape given in (15). We start with Part 1, i.e. Part (a) of Assumption A1 holds and $y_t = (m_t, z_t)$:

$$
\begin{aligned}
L(\theta) &= L(\theta|\{y_t\}_{t=1}^T, \{x_t\}_{t\in\bar{\mathcal{T}}}) \\
&= \int\cdots\int_{\times_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}}\mathcal{S}_x} P\left(\{y_t\}_{t=1}^T, \{x_t\}_{t=1}^T; \psi\right) \mathrm{d}x_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}} \\
&= \int\cdots\int_{\times_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}}\mathcal{S}_x} P\left(\{m_t, z_t\}_{t=1}^T, \{x_t\}_{t=1}^T; \psi\right) \mathrm{d}x_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}} \\
&= \int\cdots\int_{\times_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}}\mathcal{S}_x} \prod_{t=1}^T P^m(m_t|z_t, x_t, m_{t-1}, z_{t-1}, x_{t-1}; \eta) P^{zx}\left(z_t, x_t|z_{t-1}, x_{t-1}\theta\right) \mathrm{d}x_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}} \\
&= \int\cdots\int_{\times_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}}\mathcal{S}_x} \prod_{t=1}^T P(y_t, x_t|y_{t-1}, x_{t-1}; \psi) \mathrm{d}x_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}} \\
&= \int\cdots\int_{\times_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}}\mathcal{S}_x} \prod_{i=1}^{N+1} \prod_{t=\bar{t}_{i-1}+1}^{\bar{t}_i} P(y_t, x_t|y_{t-1}, x_{t-1}; \psi) \mathrm{d}x_{t\in\mathcal{T}\setminus\bar{\mathcal{T}}} \\
&= \prod_{i\in\{1,\ldots,N+1\}\bigcap\{i|\tau_i>0\}} \int\cdots\int_{\mathcal{S}_x^{\tau_i}} P\left(y_{\bar{t}_{i-1}+1}, \tilde{x}_{\bar{t}_{i-1}+1}|y_{\bar{t}_{i-1}}, x_{\bar{t}_{i-1}}; \psi\right) \\
&\quad \cdot \prod_{t=\bar{t}_{i-1}+2}^{\bar{t}_i-1} P\left(y_{t+1}, \tilde{x}_{t+1}|y_t, \tilde{x}_t; \psi\right) \\
&\quad \cdot P\left(y_{\bar{t}_i}, x_{\bar{t}_i}|y_{\bar{t}_i-1}, \tilde{x}_{\bar{t}_i-1}; \psi\right) \mathrm{d}\left(\tilde{x}_{\bar{t}_{i-1}+1}, ..., \tilde{x}_{\bar{t}_i-1}\right) \\
&\quad \cdot \prod_{i\in\{j=1,\ldots,N+1|\tau_j=0\}} P\left(y_{\bar{t}_i}, x_{\bar{t}_i}|y_{\bar{t}_i-1}, x_{\bar{t}_i-1}; \psi\right).
\end{aligned}
\tag{34}
$$

Note that up to the last equation we use x_t for any $t \in \mathcal{T} \setminus \bar{\mathcal{T}}$ as integration variable; we only switch to our previous convention with $\tilde{x}_t$ as integration variable in the last equation. The reason for this is to display the emergence of the product $\prod_{i=1}^N$ s.t. the double product $\prod_{i=1}^{N+1}\prod_{t=\bar{t}_{i-1}+1}^{\bar{t}_i}$ clearly spans the entire sample period $t = 1, ..., T$. Thereafter, in the last line we need to separate the set $\{1, ..., N\}$ into two parts: If $\tau_i = 0$, then x_t is observed in two successive periods and no integration is needed. In particular, the conditional probability of $(y_{\bar{t}_{i+1}}, x_{\bar{t}_{i+1}})$ conditions on $(y_{\bar{t}_i}, x_{\bar{t}_i})$, hence, the factor for this i consists of only one conditional probability. In all other cases, there is at least one period of unobservation between two periods of observation, hence, we need to integrate these unobserved states. Note that the product $\prod_{t=\bar{t}_{i-1}+2}^{\bar{t}_i-1}$ is empty, i.e. equal to 0, if there is exactly one period of unobservation ($\bar{t}_{i+1} = \bar{t}_i + 2$).

In summary, the likelihood in the case of MNAR data follows immediately from our Markov assumptions for P^m and P^{zx} and a suitable decomposition of the observation pattern $\mathcal{T}$. For

the case of MAR data, the conditional independence assumption (11) implies

$$P(y_t, x_t|y_{t-1}, x_{t-1}; \psi) = P^m(m_t|m_{t-1}, z_t, z_{t-1}; \eta) P^{zx}(z_t, x_t|z_{t-1}, x_{t-1}; \theta)$$

Plugging this identity into the likelihood from (34) yields:

$$\begin{aligned}
L(\psi) = & \prod_{i\in\{1,\dots,N+1\}\bigcap\{i|\tau_i>0\}} \int\cdots\int_{\mathcal{S}_x^{\tau_i}} P^m\left(m_{\bar{t}_{i-1}+1}|m_{\bar{t}_{i-1}}, z_{\bar{t}_{i-1}+1}, z_{\bar{t}_{i-1}}; \eta\right) \\
& \cdot P^{zx}\left(z_{\bar{t}_{i-1}+1}, \tilde{x}_{\bar{t}_{i-1}+1}|z_{\bar{t}_{i-1}}, x_{\bar{t}_{i-1}}; \psi\right) \\
& \cdot \prod_{t=\bar{t}_{i-1}+2}^{\bar{t}_i-1} P^m\left(m_t|m_{t-1}, z_t, z_{t-1}; \eta\right) P^{zx}\left(z_t, \tilde{x}_t|z_{t-1}, \tilde{x}_{t-1}; \psi\right) \\
& \cdot P^m\left(m_{\bar{t}_i}|m_{\bar{t}_i-1}, z_{\bar{t}_i}, z_{\bar{t}_i-1}; \eta\right) P^{zx}\left(z_{\bar{t}_i}, x_{\bar{t}_i}|z_{\bar{t}_i-1}, \tilde{x}_{\bar{t}_i-1}; \psi\right) \mathrm{d}\left(\tilde{x}_{\bar{t}_{i-1}+1}, ..., \tilde{x}_{\bar{t}_i-1}\right) \\
& \cdot \prod_{i\in\{j=1,\dots,N+1|\tau_j=0\}} P^m\left(m_{\bar{t}_i}|m_{\bar{t}_i-1}, z_{\bar{t}_i}, z_{\bar{t}_i-1}; \eta\right) P^{zx}\left(z_{\bar{t}_i}, x_{\bar{t}_i}|z_{\bar{t}_i-1}, x_{\bar{t}_i-1}; \psi\right) \\
= & \prod_{i\in\{1,\dots,N+1\}\bigcap\{i|\tau_i>0\}} \int\cdots\int_{\mathcal{S}_x^{\tau_i}} P^{zx}\left(z_{\bar{t}_{i-1}+1}, \tilde{x}_{\bar{t}_{i-1}+1}|z_{\bar{t}_{i-1}}, x_{\bar{t}_{i-1}}; \psi\right) \\
& \cdot \prod_{t=\bar{t}_{i-1}+2}^{\bar{t}_i-1} P^{zx}\left(z_t, \tilde{x}_t|z_{t-1}, \tilde{x}_{t-1}; \psi\right) \\
& \cdot P^{zx}\left(z_{\bar{t}_i}, x_{\bar{t}_i}|z_{\bar{t}_i-1}, \tilde{x}_{\bar{t}_i-1}; \psi\right) \mathrm{d}\left(\tilde{x}_{\bar{t}_{i-1}+1}, ..., \tilde{x}_{\bar{t}_i-1}\right) \\
& \cdot \prod_{i\in\{j=1,\dots,N+1|\tau_j=0\}} P^{zx}\left(z_{\bar{t}_i}, x_{\bar{t}_i}|z_{\bar{t}_i-1}, x_{\bar{t}_i-1}; \psi\right) \\
& \cdot \prod_{t=1}^{T+1} P^m(m_t|m_{t-1}, z_t, z_{t-1}; \eta).
\end{aligned}$$

Due to the data being MAR, the observation variable m_t is conditionally independent on x_t, x_{t-1}, hence its transition probability P^m can be pulled out of the integral for each t. Since under Assumption A1(b) also the model and the nuisance parameter are fully separate, the factor $\prod_{t=1}^{T+1} P^m(m_t|m_{t-1}, z_t, z_{t-1}; \eta)$ is purely scaling the likelihood and does not influence maximization. For this reason the likelihood based solely on the model transition probability P^{zx} is equivalent to the likelihood (34). Setting $\psi = \theta$ and $P = P^{zx}$ and only utilizing the z_t-component of y_t delivers this result for the general result (15).

A.2 General notation

We develop a general notation to formulate the likelihood function of a Markov model with serially correlated states, in particular if some (or all) of the model states are observed only occasionally. In fact, the notation developed below allows for arbitrary observations patterns both w.r.t. time and the state space dimension.

In contrast to previous sections, we consider observed variables Y_t and occasionally observed variables X_t jointly to allow for general observation patterns and simplify notation in this section. Recall that we have introduced Y_t as composite variable of the always observed variables in Proposition 1, i.e., it includes both model variables Z_t and observation variables M_t. We

now join all the model variables into $\bar{X}_t = (Z_t, X_t) \in \mathbb{R}^d$, the vector of all model variables, in particular allowing for any number of occasionally observed variables and no always observed variables. Then, we define $\bar{M}_t \in \{0,1\}^d$ to be the vector of observation variables $m_{it}, i = 1, ..., d$ for each component of the vector $\bar{X}_t$. As before, the conditional joint probability P^m of the vector $\bar{M}_t$ conditional on $\bar{X}_t, \bar{X}_{t-1}$ can account for any (cross) endogeneity of the observation process. However, using the same formalism as in Proposition 1, we can now continue by solely considering the joint vector $W_t = (\bar{X}_t, \bar{M}_t)$. Its joint probability function P is the product of $P^{\bar{x}}$ and P^m if the data is MNAR and only $P^{\bar{x}}$ when the data is MAR. Hence, we consider a stochastic process $\{W_t\}_{t\in\mathbb{N}}$, where the random vector W_t has support $\mathcal{S} \subseteq \mathbb{R}^{2d}$, which we refer to as the "state space".[29] Note that we restrict our attention to continuous state variables here, as all concepts we present below have simple analogues in the discrete case.

We assume the Markov model explaining $\{W_t\}$ to define a parametric family of (conditional) distributions, which can be represented through probability density functions

$$P\left(W_t \mid \{W_s\}_{s<t}; \theta\right) = P\left(W_t \mid W_{t-1}; \theta\right)$$

with $\theta \in \Theta \subset \mathbb{R}^p$.[30] We assume that there exists a unique $\theta_0 \in \Theta$ that perfectly parametrizes the data generating process and aim to estimate this parameter through applying a maximum likelihood approach. In order to precisely express the observation pattern of a dataset, we introduce some more notation: Let $w_{\tau_0} \equiv (w^i)_{i\in\tau_0}$ denote the sub-vector of states for some index set $\tau_0 \subseteq \tau \equiv \{1, \ldots, d\}$. Moreover, we write $\tilde{\tau}_0 \equiv \tau \setminus \tau_0$ for the complement of τ_0 w.r.t. τ, and we express the number of dimensions of w_{τ_0} using the cardinality operator $|\tau_0|$. Finally, note that if we write $(w_{\tau_0}, w_{\tilde{\tau}_0})$, we tacitly assume the elements to be re-ordered appropriately so that $(w_{\tau_0}, w_{\tilde{\tau}_0}) = w$, including the special cases $(w_\tau, w_\emptyset)$ and $(w_\emptyset, w_\tau)$.

This notation allows us to define the observation pattern of a dataset as follows: For an observation horizon $\{0, \ldots, T\}$, the set of index sets $\{\tau_t\}_{t=0}^T$, $\tau_t \subseteq \tau$, specifies which dimensions of the state vector w are observed at each point in time t,[31] and we denote the dataset by $\{w_{t,\tau_t}\}_{t=0}^T$. Note that in order to distinguish entries of the dataset from generic sub-vectors of states such as w_{τ_t}, we have equipped the former with another time subscript besides the index set. This notation also allows us to implicitly distinguish completely, never, and occasionally observed variables and thus ties it back into the context of the previous section: A completely observed variable w_{ti} has $i \in \tau_t$ for all $t \in \mathcal{T}$, an unobserved variable has $i \in \tilde{\tau}_t$ for all t and a variable is occasionally observed if neither holds. At each point in time t, the state realizations are an element of the subset

$$\mathcal{S}_t \equiv \{w \in \mathcal{S} : w_{\tau_t} = w_{t,\tau_t}\},$$

[29] We abstract from the more general case which supposes time-heterogenous dimensionality of the state space in favor of a lighter notation. As the integration dimension will vary over time due to occasional observations of $W_t = w_t$ this extension is straight-forward.

[30] The density P_θ can, of course, be time-dependent, but we spare the additional index here, as our notation encompasses this feature—theoretically—through a deterministic, discrete state.

[31] Note that in the outline of the method in Section 2, we use a single index set $\mathcal{T}$ to denote the points in time where an observation of a single state takes place. Here, each point in time has its own index set τ_t, specifying the dimensions of the state space which are observed at time t.

which "binds" the observed dimensions to the values from the dataset. Note, though, that not necessarily all elements in $\mathcal{S}_t$ have non-zero probability density. For the integration over the unobserved dimensions, we also need the projection of $\mathcal{S}_t$ to the lower-dimensional space where the unobserved dimensions live:

$$\tilde{\mathcal{S}}_t \equiv \left\{ \tilde{w} \in \mathbb{R}^{|\tilde{\tau}_t|} : \tilde{w} = w_{\tilde{\tau}_t}, w \in \mathcal{S}_t \right\}.$$

We write $\tilde{\mathcal{S}}_t = \emptyset$ if $\tau_t = \tau$ and thus $\tilde{\tau}_t = \emptyset$. The (unconditional) likelihood of the model under observation regime $\{\tau_t\}_{t=0}^T$ reads

$$\begin{aligned} L_g^T(\theta) &\equiv L(\theta|\{w_{t,\tau_t}\}_{t=1}^T) \\ &= \int \cdots \int_{\times_{t=1}^T \tilde{\mathcal{S}}_t} \prod_{t=1}^T P\left(\tilde{w}_t, w_{t,\tau_t} | \tilde{w}_{t-1}, w_{t-1,\tau_{t-1}}; \theta\right) \mathrm{d}\tilde{w}_T \ldots \mathrm{d}\tilde{w}_1 \qquad (35) \\ &= \int \cdots \int_{\times_{t=1}^T \tilde{\mathcal{S}}_t} \prod_{t=1}^T g_t\left(\tilde{w}_t, \tilde{w}_{t-1}, \theta\right) \mathrm{d}\tilde{w}_T \cdots \mathrm{d}\tilde{w}_1 \qquad (36) \end{aligned}$$

and thus resembles the definition of L_g^T in Reich (2018). The functions $g_t : \tilde{\mathcal{S}}_t \times \tilde{\mathcal{S}}_{t-1} \times \Theta \to \mathbb{R}$ are defined by

$$g_t(\tilde{w}_t, \tilde{w}_{t-1}, \theta) \equiv \begin{cases} P\left(\tilde{w}_t, w_{t,\tau_t} | \tilde{w}_{t-1}, w_{t-1,\tau_{t-1}}; \theta\right) & \text{if } t > 1 \\ P\left(\tilde{w}_1, w_{1,\tau_1}; \theta\right) & \text{if } t = 1 \end{cases}$$

s.t. the dependence of the integrand on the data is implicitly given in the subscript t of g_t. Note that both $\tilde{\mathcal{S}}_t$ and $\tilde{w}_t$ can be empty if $\tau_t = \tau$, i.e., g_t, g_{t+1} are constant in $\tilde{w}_t$ and no integration w.r.t. $\tilde{w}_t$ takes place. Using the Markov structure of the model and standard regularity conditions for g_t[32], a Fubini–Tonelli theorem (the concrete version of it depending on the nature of $\mathcal{S}$) justifies a recursive formulation of (36),

$$\begin{cases} \varphi_t^\theta \in \mathbb{R}_+ : \begin{cases} 1 & t > T \\ g_t\left(w_t | w_{t-1}; \theta\right) \varphi_{t+1}^\theta & \tau_t = \tau \\ \int_{\tilde{\mathcal{S}}_t} g_t\left((\tilde{w}, w_{t,\tau_t}) | w_{t-1}; \theta\right) \cdot f_{t+1}^\theta(\tilde{w}) \mathrm{d}^{|\overline{\tau}_t|}\tilde{w} & \text{otherwise} \end{cases} & \tau_{t-1} = \tau \\ f_t^\theta : \tilde{\mathcal{S}}_{t-1} \to \mathbb{R}_+, w \mapsto \begin{cases} 1 & t > T \\ g_t\left(w_t | (w, w_{t-1,\tau_{t-1}}); \theta\right) \varphi_{t+1}^\theta & \tau_t = \tau \\ \int_{\tilde{\mathcal{S}}_t} g_t\left((\tilde{w}, w_{t,\tau_t}) | (w, w_{t-1,\tau_{t-1}}); \theta\right) \cdot f_{t+1}^\theta(\tilde{w}) \mathrm{d}^{|\overline{\tau}_t|}\tilde{w} & \text{otherwise} \end{cases} & \text{otherwise,} \end{cases} \qquad (37)$$

[32] These follow from the fact that g_t is derived from a conditional p.d.f. which tend to be continuous and bounded in most economic applications.

and the final likelihood reads

$$L(\theta; \{w_{t,\tau_t}\}_{t=1}^{T}) = \begin{cases} g_t\left(w_1; \theta\right) \varphi_2^\theta & \tau_1 = \tau \\ \int_{\tilde{\mathcal{S}}_1} g_t\left((\tilde{w}, w_{1,\tau_1}); \theta\right) f_2^\theta(\tilde{w}) \mathrm{d}^{|\overline{\tau}_1|}\tilde{w} & \text{otherwise.} \end{cases} \tag{38}$$

While formulation (37) is exact, it is not practical for implementation purposes for the following reasons:

1. Actually evaluating the final likelihood—and thus evaluating either f_2^θ or φ_2^θ—would still require traversing a tree with $T-1$ levels and potentially infinitely many "knots" at each level; thus its computational complexity would explode.

2. No explicit use is made from the knowledge of the observations w_{t,τ_t} to determine the *conditional* distribution of $w_{\overline{\tau}_t}$.

To address issue 1, we introduce a mapping between two function spaces $\mathcal{B}_n$ and $\mathcal{P}_n$, whose elements are real functions of n-dimensional arguments, and with all elements in $\mathcal{P}_n$ having a complete representation through a countable set of parameters:

$$\mathcal{I}_n : \mathcal{B}_n \to \mathcal{P}_n, f \mapsto \hat{f},$$

where

$$f, \hat{f} : \mathbb{R}^n \supseteq \mathcal{D} \to \mathbb{R},$$

and with the norm $\|f - \hat{f}\|$ being "small" in the appropriate sense.

As indicated in issue 2, knowledge of w_{t,τ_t} can be used in many instances to obtain "high density regions" for $w_{\overline{\tau}_t}$ by conditioning its distribution on w_{t,τ_t}. This can often be exploited when numerically approximating the integrals in (37), e.g., by placing the nodes of quadrature rules accordingly. Therefore, we rewrite the relevant cases, conditioning the integrated probability densities on the observed states; note that in practice, this is not always possible.

Consequently, the final likelihood function recursion reads

$$\left\{\begin{array}{l} \hat{\varphi}_t^\theta \in \mathbb{R}_+ : \left\{\begin{array}{ll} 1 & t > T \\ g_t\left(w_t|w_{t-1};\theta\right)\hat{\varphi}_{t+1}^\theta & \tau_t = \tau \\ \int_{\mathcal{S}} g_t\left(\tilde{w}|w_{t-1};\theta\right)\hat{f}_{t+1}^\theta(\tilde{w})\mathrm{d}^n\tilde{w} & \tau_t = \emptyset \\ g_t\left(w_{t,\tau_t}|w_{t-1};\theta\right) & \\ \quad \cdot \int_{\tilde{\mathcal{S}}_t} g_t\left(\tilde{w}|w_{t,\tau_t},w_{t-1};\theta\right)\hat{f}_{t+1}^\theta(\tilde{w})\mathrm{d}^{|\overline{\tau}_t|}\tilde{w} & \text{otherwise} \end{array}\right\} & \tau_{t-1} = \tau \\ \hat{f}_t^\theta : \tilde{\mathcal{S}}_{t-1} \to \mathbb{R}_+, w \mapsto \left\{\begin{array}{ll} 1 & t > T \\ g_t\left(w_t|(w,w_{t-1,\tau_{t-1}});\theta\right)\hat{\varphi}_{t+1}^\theta & \tau_t = \tau \\ \mathcal{I}_{|\overline{\tau}_{t-1}|}\Big(\int_{\mathcal{S}} g_t\left(\tilde{w}|(w,w_{t-1,\tau_{t-1}});\theta\right) & \\ \qquad \cdot \hat{f}_{t+1}^\theta(\tilde{w})\mathrm{d}^n\tilde{w}\Big) & \tau_t = \emptyset \\ \mathcal{I}_{|\overline{\tau}_{t-1}|}\Big(g_t\left(w_{t,\tau_t}|(w,w_{t-1,\tau_{t-1}});\theta\right) & \\ \qquad \cdot \int_{\tilde{\mathcal{S}}_t} g_t\left(\tilde{w}|w_{t,\tau_t},(w,w_{t-1,\tau_{t-1}});\theta\right) & \text{otherwise} \\ \qquad \cdot \hat{f}_{t+1}^\theta(\tilde{w})\mathrm{d}^{|\overline{\tau}_t|}\tilde{w}\Big) & \end{array}\right\} & \text{otherwise,} \end{array}\right. \tag{39}$$

and the actual likelihood can be computed analogously to (38).

A.3 Issues related to floating-point arithmetics and rescaling the likelihood function

In most maximum likelihood applications, it is not the "physical" likelihood function that is maximized, but its logarithm instead. This (smooth) monotonic transformation does not affect the location of the maximum, but has significant numerical advantages: The main motivation for this transformation is the fact that likelihood functions can become very small or—in the case of continuous random variables with densities possibly larger than one—very large, potentially causing severe numerical problems in floating-point arithmetic, most prominently underflows (number can no longer be distinguished from 0) and overflows (absolute value of a number can no longer be represented).[33] However, this transformation cannot be "moved" into the integral in the recursion (20), because generally $\log \int g(\tilde{w})d\tilde{w} \neq \int \log g(\tilde{w})d\tilde{w}$. We therefore employ a simple rescaling scheme based on the following equality (for notational brevity, we restrict ourselves to the case of no observations here):

$$\begin{aligned} &\int \cdots \int p(\tilde{w}_1)\prod_{t=2}^{T} p(\tilde{w}_t|\tilde{w}_{t-1})\,\mathrm{d}(\tilde{w}_1,\ldots,\tilde{w}_T) \\ &= \sum_{t=1}^{T}\log\alpha_t + \log\int\cdots\int p(\tilde{w}_1)\alpha_1^{-1}\prod_{t=2}^{T} p(\tilde{w}_t|\tilde{w}_{t-1})\alpha_t^{-1}\,\mathrm{d}(\tilde{w}_1,\ldots,\tilde{w}_T)\,. \end{aligned} \tag{40}$$

[33] On the other hand, note that subtraction is considered the most accuracy-losing operation. Therefore, the multiplication of physical probabilities would be preferred in cases where the log-likelihood of the individual observations is of mixed sign, which can happen with continuous variables.

We choose the following scaling factors (where the norm is taken over x):

$$\alpha_t \propto \left\| \int \cdots \int p(\tilde{w}_{t+1}|w)\alpha_{t+1}^{-1} \prod_{s=t+2}^{T} p(\tilde{w}_s|\tilde{w}_{s-1})\alpha_s^{-1} \,\mathrm{d}(\tilde{w}_{t+1}, \ldots, \tilde{w}_T) \right\|,$$

because in the context of recursion (20) it can easily be obtained as $\alpha_t \propto \|f_t^\theta(w)\|$. In other words, at every iteration of the recursion we rescale the recursion function f_t^θ to have a unity norm, and separately update the "aggregate" scaling factor to finally obtain the correct log-likelihood function. This procedure is integrated in Algorithm 1, as well as the listings in Appendix A.4.

For the type of the norms itself, we often use L_1 or L_∞, or, to balance better the order of magnitude, $\exp \| \log f_t^\theta \|_1$; the integral over f_t^θ can be easily approximated using the nodes of the interpolant $\mathcal{I}_{f_t^\theta}$ (recall that approximation error for the norm does not affect the accuracy of the likelihood approximation itself by construction). Alternatively, if the point(s) of evaluation of the result of the recursion is known upfront—which is most commonly the case, for example, for quadrature rules to compute (27a), or for a single observation w_1 in (27b)—one can also rescale f_t^θ to unity around that point(s).

Algorithm 1 provides a pseuo-implementation of recursive likelihood function integration under occasional observations, with the rescaling from Equations (40) and (A.3) applied.[34]

A.4 Code listings

The following listing is a MATLAB implementation of our method for continuous x, but which is agnostic about the nature of y, i.e., $(x, y) \in \mathcal{S}_x \times \mathcal{S}_y \subset \mathbb{R}^2$. The numerical integration follows rule (23), and is thus expressed as a set of nodes and weights, `qn` and `qw`, respectively; note that the weights contain the kernel against which to integrate. The nodes and weights to integrate x_1 against the stationary distribution are given by `qn1` and `qw1`. The change of variables to allow for an unconditional kernel, $\phi(\Delta x_t, x_{t-1}, i_{t-1}; \theta)$, and its first derivative w.r.t. x_t is given by `phi` and `phipr`, respectively; those objects are function-valued and take the respective arguments x_t, x_{t-1}, and i_{t-1}. Similarly, the density functions $Pr(x_t|x_{t-1}, y_{t-1}; \theta)$, $Pr(y_t|x_t; \theta)$, and $Pr(x_1; \theta)$ enter as `Px`, `Py`, and `Px1`, respectively, and are all function-valued arguments. Finally, the interpolation grid is passed as `in`.

Note that the parameter to be estimated, θ, enters implicitly through the probability distributions, the change of variables ϕ (and its partial derivative), and the quadrature nodes and weights for the stationary distribution in this implementation (we assume that `qn` and `qw` integrate against a standardized distribution, and thus parameter dependence of the conditional distributions enters solely through the change of variables).

[34] Note that a direct implementation of Algorithm 1 will, in many programming languages, still cause the full recursion tree to be built up, before the actual numerical evaluation is triggered in $t = 1$—in particular if lambda calculus is used (for example, through "anonymous functions" in MATLAB). This is, however, not generally an issue, because only the assignment in line 13 triggers more than one evaluation of the previous $\hat{f}^\psi$ while not being "safeguarded" through $\mathcal{I}$; however, due to the observational pattern, it cannot call itself repeatedly. Alternatively, it is straightforward—and often slightly more efficient—to explicitly force the numerical evaluation of $f^\psi(\cdot)$ on the right-hand side of the assignments in each time step (as we demonstrate in Appendix A.4); moreover, one can wrap each line by $\mathcal{I}$, which is slightly less efficient and introduces a higher numerical error but allows for some simpler implementations, in particular in higher dimensional states with asymmetric observation patters.

Algorithm 1 Recursive likelihood function integration (RLI) for occasionally observed states with adaptive rescaling.

$\hat{f}(x) \leftarrow 1$
$\alpha \leftarrow 0$
$\hat{\alpha} \leftarrow 1$
for $t = T, \ldots, 2$ **do**
    if $t \in \bar{\mathcal{T}}$ **then**
        if $t - 1 \in \bar{\mathcal{T}}$ **then**
            $\hat{f}^{\psi}(x) \leftarrow P(y_t, x_t | y_{t-1}, x_{t-1}; \psi) \hat{f}^{\psi}(x_t) / \hat{\alpha}$
        else
            $\hat{f}^{\psi}(x) \leftarrow P(y_t, x_t | y_{t-1}, x; \psi) \hat{f}^{\psi}(x_t) / \hat{\alpha}$
        end if
    else
        if $t - 1 \in \bar{\mathcal{T}}$ **then**
            $\hat{f}^{\psi}(x) \leftarrow \int P(y_t, \tilde{x} | y_{t-1}, x_{t-1}; \psi) \hat{f}^{\psi}(\tilde{x}) / \hat{\alpha} \mathrm{d}\tilde{x}$
        else
            $\hat{f}^{\psi}(x) \leftarrow \mathcal{I} \left(\int P(y_t, \tilde{x} | y_{t-1}, x; \psi) \hat{f}^{\psi}(\tilde{x}) / \hat{\alpha} \mathrm{d}\tilde{x} \right)$
        end if
    end if
    $\hat{\alpha} \leftarrow \| \hat{f}^{\psi} \|$
    $\alpha \leftarrow \alpha + \log \hat{\alpha}$
end for
if $1 \in \bar{\mathcal{T}}$ **then**
    $L^{\psi} \leftarrow P(y_1, x_1; \psi) \hat{f}^{\psi}(x_1) / \hat{\alpha}$
else
    $L^{\psi} \leftarrow \int P(y_1, \tilde{x}; \psi) \hat{f}^{\psi}(\tilde{x}) / \hat{\alpha} \mathrm{d}\tilde{x}$
end if
return $\log L^{\psi} - \alpha$

In order to keep MATLAB from spanning the full recursion tree first and only evaluate it at $t = 1$, we need to explicitly enforce the numerical evaluation of the recursion function at each time step. This is done through the helper variable `fn_` (instead of directly writing the call to `fn` of the previous iteration into the definition of the new `fn` function object).

```
function l = likelihood_cont(x,y,qn,qw,qn1,qw1,in,Py,Px,Px1,phi,phipr)
% LIKELIHOOD_CONT - Likelihood for occasional continuous observations
%
% Inputs:
%    x     - data for X (MISSING if unobserved)  LENGTH(x):  T
%    y     - data for Y                          LENGTH(y):  T
%    qn    - quadrature nodes                    SIZE(qn):   [nq,1]
%    qw    - quadrature weights                  SIZE(qw):   [nq,1]
%    qn1   - quad. nodes (stat. dist.;theta)     SIZE(qn1):  [nq,1]
%    qw1   - quad. weights (stat. dist.;theta)   SIZE(qw1):  [nq,1]
%    in    - interpolation nodes                 SIZE(in):   [ni,1]
%    Py    - Pr(Y_t|X_t;theta)                   FUNCTION(nargin=2)
%    Px    - Pr(X_t|X_t-1,Y_t-1;theta)           FUNCTION(nargin=3)
%    Px1   - Pr(X_1;theta)                       FUNCTION(nargin=1)
%    phi   - phi(DeltaX,X_t-1,Y_t-1;theta)       FUNCTION(nargin=3)
%    phipr - phi'(DeltaX,X_t-1,Y_t-1;theta)      FUNCTION(nargin=3)
%
% Outputs:
%    l     - value of log-likelihood function


% initialization
ni = length(in);
nq = length(qn);
T = length(x);
qn_ = repmat(qn',ni,1);
in_ = repmat(in,1,nq);

fn = @(x) 1;
a = 0;
a_t = 1;

for t=T:-1:2

    % recursion step
    if ismissing(x(t))
        if ismissing(x(t-1))
            phi_ = phi(qn_,in_,y(t-1));
            phipr_ = phipr(qn_,in_,y(t-1));
            fn_ = fn(phi_) ./ a_t;
            f_vals = (Py(y(t),phi_) .* phipr_ .* fn_) * qw;
            fn = @(x_) max(interp1(in,f_vals,x_,'spline','extrap'),0);
            a_t = max(f_vals);
        else
            phi_ = phi(qn',x(t-1),y(t-1));
            phipr_ = phipr(qn',x(t-1),y(t-1));
```

```
            fn_ = fn(phi_) ./ a_t;
            fn = @(x_) (Py(y(t),phi_) .* phipr_ .* fn_) * qw;
            a_t = fn(NaN);
        end
    else
        fn_ = fn(x(t)) ./ a_t;
        if ismissing(x(t-1))
            fn = @(x_) Py(y(t),x(t)) .* Px(x(t),x_,y(t-1)) .* fn_;
            a_t = max(fn(in));
        else
            fn = @(x_) Py(y(t),x(t)) .* Px(x(t),x(t-1),y(t-1)) .* fn_;
            a_t = fn(NaN);
        end
    end

    a = a + log(a_t);
end

% final likelihood
if ismissing(x(1))
    l = (Py(y(1),qn1') .* Px1(qn1)' .* (fn(qn1') ./ a_t)) * qw1;
else
    l = Py(y(1),x(1)) .* Px1(x(1)) .* (fn(x(1) ./ a_t));
end
l = log(l) + a;
```

A.5 Gauss–Hermite quadrature for Integration against a log-normal Density

As in the main text, we define $m(x, i; \theta) \equiv u(x, i; \theta_1) + \beta EV_\theta(x, i)$ for shorter notation; consider the integrand

$$EV(x,i) = \int_0^\infty \log\left(\sum_{j\in\{0,1\}} \exp(m((1-i)x + \Delta\tilde{x}, j; \theta))\right) q(\Delta\tilde{x})\mathrm{d}\Delta\tilde{x}, \tag{41}$$

where q is the density of the log-normal distribution,

$$\frac{1}{x\sqrt{2\pi}\sigma}\exp\left(-\frac{(\log(x)-\mu)^2}{2\sigma^2}\right). \tag{42}$$

Given that Δx is log-normally distributed, consider the change of variables $\Delta x = \exp(\Delta t)$:

$$\begin{aligned} EV(x,i) = \int_{-\infty}^\infty \log\left(\sum_{j\in\{0,1\}} \exp\big(m((1-i)x + \exp\big(\Delta\tilde{t}\big), j; \theta)\big)\right) \\ \cdot \exp\big(\Delta\tilde{t}\big)\frac{1}{\exp\big(\Delta\tilde{t}\big)}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\Delta\tilde{t}-\mu)^2}{2\sigma^2}\right)\mathrm{d}\Delta\tilde{t}, \end{aligned} \tag{43}$$

which is an integral against the normal density, and thus can be approximated efficiently using Gauss–Hermite quadrature with nodes and weights $\{(c_k, w_k)\}_{k=1}^{N^Q}$; see, for example, Judd (1998).

Thus, if $\{(c_k, w_k)\}_{k=1}^{N^Q}$ denote the nodes and weights of the "standard" degree-n^Q Gauss–

Hermite rule, the nodes and weights to integrate against a log-normal distribution with parameters μ and σ as in (41) read $\{(\exp\big(\mu + \sqrt{2}\sigma c_k\big), \sqrt{\pi}^{-1} w_k)\}_{k=1}^{N^Q}$. Note that this corresponds to the usual transformation of the Gauss–Hermite nodes and weights for integrals against normal kernels with parameters μ and σ, except that the nodes have been exponentiated to capture the $\exp\big(\Delta\tilde{t}\big)$ term in (43).

B Additional material for the applications

B.1 Additional results for LRR and SV models

In Section 3.1.4 we report results for the SMM approach using a specific set of moments. For this we have tried different sets of moments and we report results for the case which provided the lowest errors for our examples. To demonstrate the robustness of our results with regard to the specific selection of moments, we also provide results using the Lasso GMM approach proposed by Cheng and Liao (2015). The idea of the approach is to divide the set of moments into two categories. The *sure* moments, which contain a relatively small subset the moments and are sufficient to identify the model parameters, and the *doubt* moments, which can be a large set of moments which the econometrician is not sure about whether they provide additional value for estimating the model parameters. The moment conditions of the *doubt* moments are then allowed to deviate from zero by introducing the slackness parameters β_l. The Lasso GMM by Cheng and Liao (2015) minimizes the standard GMM objective function plus a Lasso term with penalty parameter λ_n which shrinks the slackness parameters β_l towards 0. For $\lambda_n = 0$, the *doubt* moments can take on any value and hence, only the *sure* moments are relevant for the SMM estimator. For $\lambda_n \to \infty$ the SMM estimator which includes both, the *sure* moments and the *doubt* moments is obtained.

Cheng and Liao (2015) provide theoretical arguments on how to choose the penalty parameter λ_n as well the weights $\omega_{n,l}$ for each slackness parameter β_l (see equation 2.7 of Cheng and Liao (2015)). This *optimal* choice of parameters is based on the information content of each moment condition. For the SMM approach we use in our paper, this information based approach did not yield reasonable results for example due to numerical problems when inverting the variance-covariance matrix of our estimator in the first step estimation with $\lambda_n = 0$. So instead we use a very pragmatic approach which turned out to work very well in practice.

For this, we first run the unconstrained SMM with $\lambda_n = 0$ and an identity weighting matrix as proposed by Cheng and Liao (2015). Then we scale each slackness parameter by a constant term $\omega_{n,l}$ such that the SMM objective (left term in equation 2.7 of Cheng and Liao (2015)) and the Lasso penalty (right term equation 2.7 of Cheng and Liao (2015)) are of equal size for $\lambda_n = 1$. So all slackness parameters obtain the same weight and by increasing for λ_n to for example 100, the Lasso penalty on the objective function will be 100 times as large as the objective from the standard SMM estimator.

For the long-run risk model (28), we use the first and second non-central moments, the cross-correlation between consumption and dividend growth, as well as autocorrelations up to order 5 as *sure* moments. Additionally, we add autocorrelations up to order 10 for both, consumption and dividend growth as *doubt* moments. For the stochastic volatility model (29), we use the

first and second non-central moments, the cross-correlation between consumption and dividend growth, as well as the first-order autocorrelations as *sure* moments. Additionally, we add the third, fourth, fifth, and sixth non-central moments and autocorrelations up to order 7 as *doubt* moments. We report results for $\lambda_n \in [0, 1, 100, 10000]$.

Figure 9 shows the results for the long-run risk model and Figure 10 shows the corresponding results for the stochastic volatility model. Red lines show results for RLI and black lines for the SMM estimator used in the main text in Figure 1. The left panel shows results for the occasional observation regime and the right panel for the full information regime. Colored lines show results for Lasso SMM with different penalty parameters λ_n. We find that for the long-run risk model, the Lasso SMM estimator yields comparable results to the SMM estimator we use in the main text of the paper showing the robustness of our results with regard to the specific selection of moments. For the stochastic volatility model, we find that while increasing the penalty can decrease the errors, the moments selected in the main text of the paper for SMM yield lower errors compared to the Lasso approach. More importantly, for both models, the errors of SMM with any selection of moments are significantly larger compared to our RLI estimator independent of the size of the dataset. Hence, the results show that the conclusions we draw regarding the comparison of the SMM and RLI estimators do not rely on the specific set of moments we selected for the SMM approach.

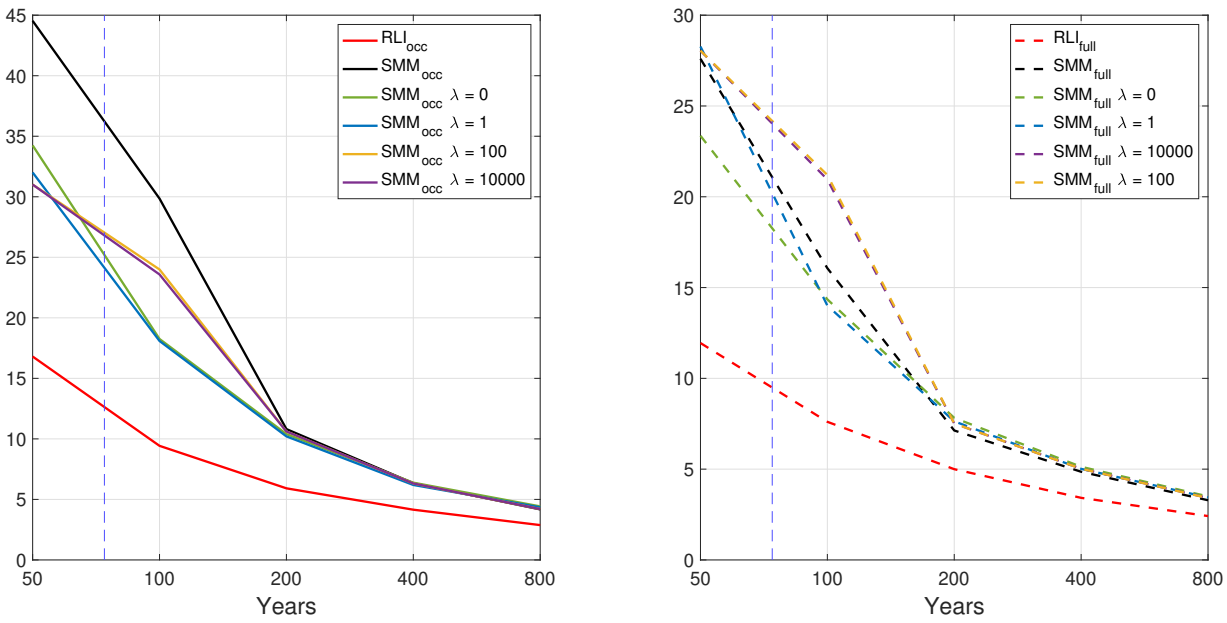


Figure 9: Median of the Mahalanobis distance over 400 simulated datasets as a function of dataset length in years for the long-run risk model (28). The left panel shows results for the occasional observation regime and the right panel for the full information regime. Red lines show results for RLI and black lines for the SMM estimator used in the main text in Figure 1. Colored lines show results for Lasso SMM with different penalty parameters λ_n. The dashed blue line shows the number of observations in the real consumption and dividend dataset.

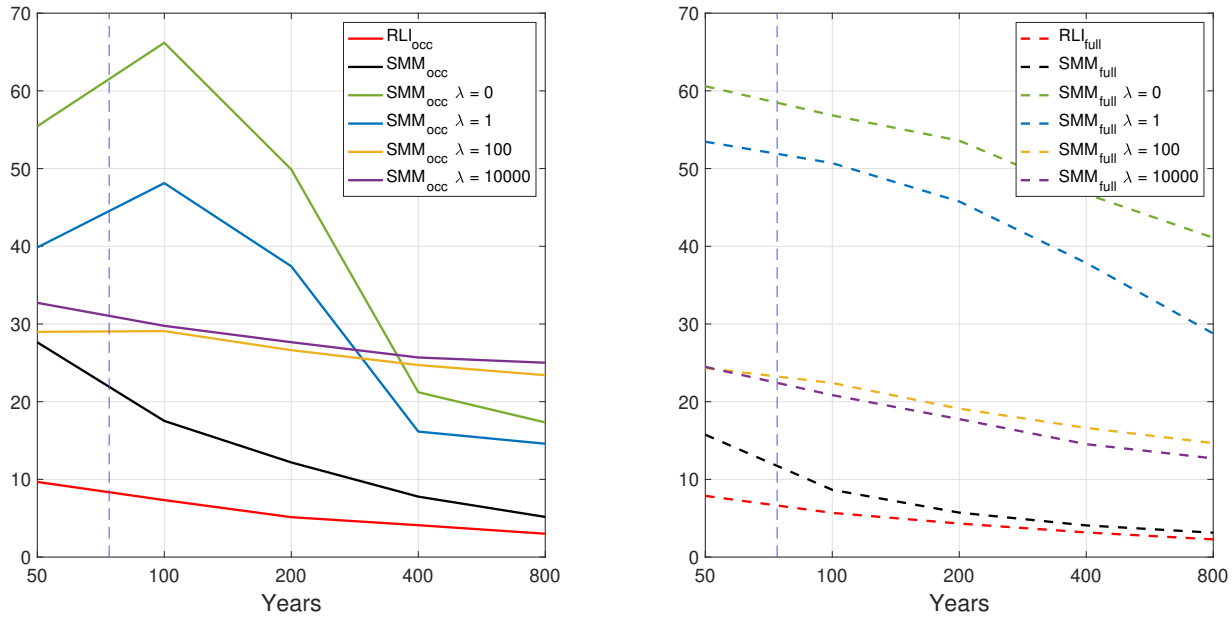
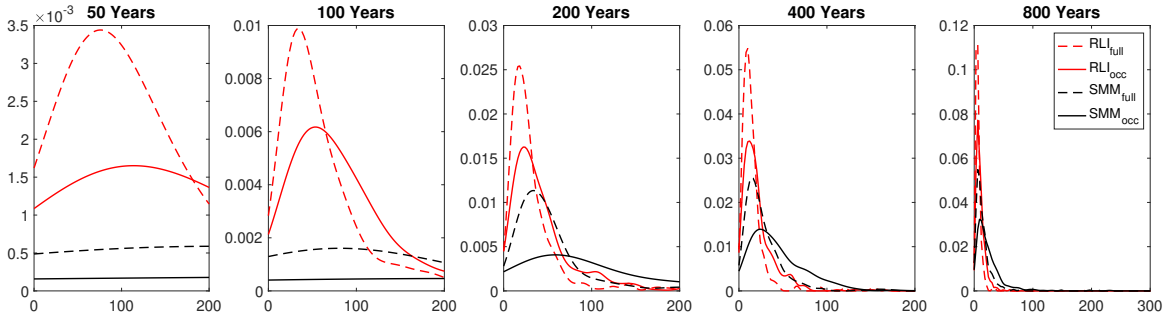


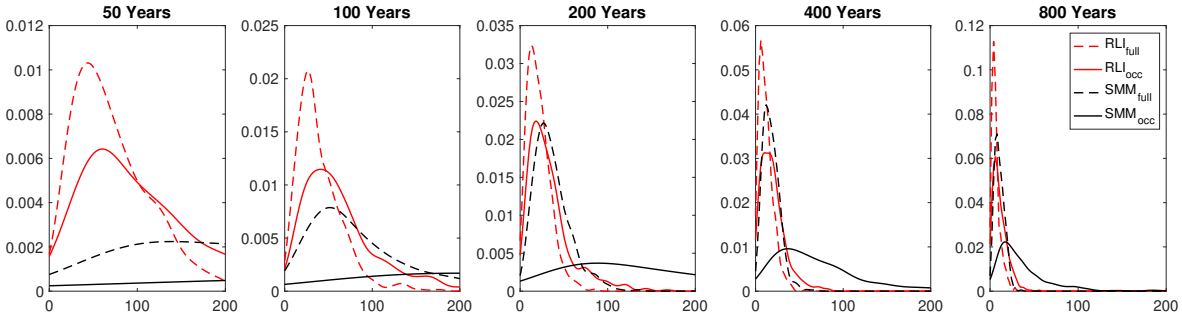
Figure 10: Median of the Mahalanobis distance over 400 simulated datasets as a function of dataset length in years for the stochastic volatility model (29). The left panel shows results for the occasional observation regime and the right panel for the full information regime. Red lines show results for RLI and the black line for the SMM estimator used in the main text in Figure 1. Colored lines show results for Lasso SMM with different penalty parameters λ_n. The dashed blue line shows the number of observations in the real consumption and dividend dataset.

Long-Run Risk Model					
Years	50	100	200	400	800
RLI_{full}	1	0	0	0	0
RLI_{occ}	3	1	1	1	0
SMM_{full}	4	0	2	1	0
SMM_{occ}	4	3	2	1	0
Stochastic Volatility Model					
Years	50	100	200	400	800
RLI_{full}	2	3	3	1	0
RLI_{occ}	3	1	1	0	1
SMM_{full}	0	1	2	4	1
SMM_{occ}	5	4	3	3	1

Table 2: Number of non-converged runs out of the 400 runs in the Monte Carlo study for the long-run risk model and the stochastic volatility model.



(a) Long-Run Risk Model (28)



(b) Stochastic Volatility Model (29)

Figure 11: Kernel density estimates for the Mahalanobis distance over 400 simulated datasets, using RLI (red lines) and SMM (black lines) under the full observation regime (dashed lines) and the occasional observation regime (solid lines). Panel (a) shows the results for the long-run risk model (28) and panel (b) for the stochastic volatility model (29).

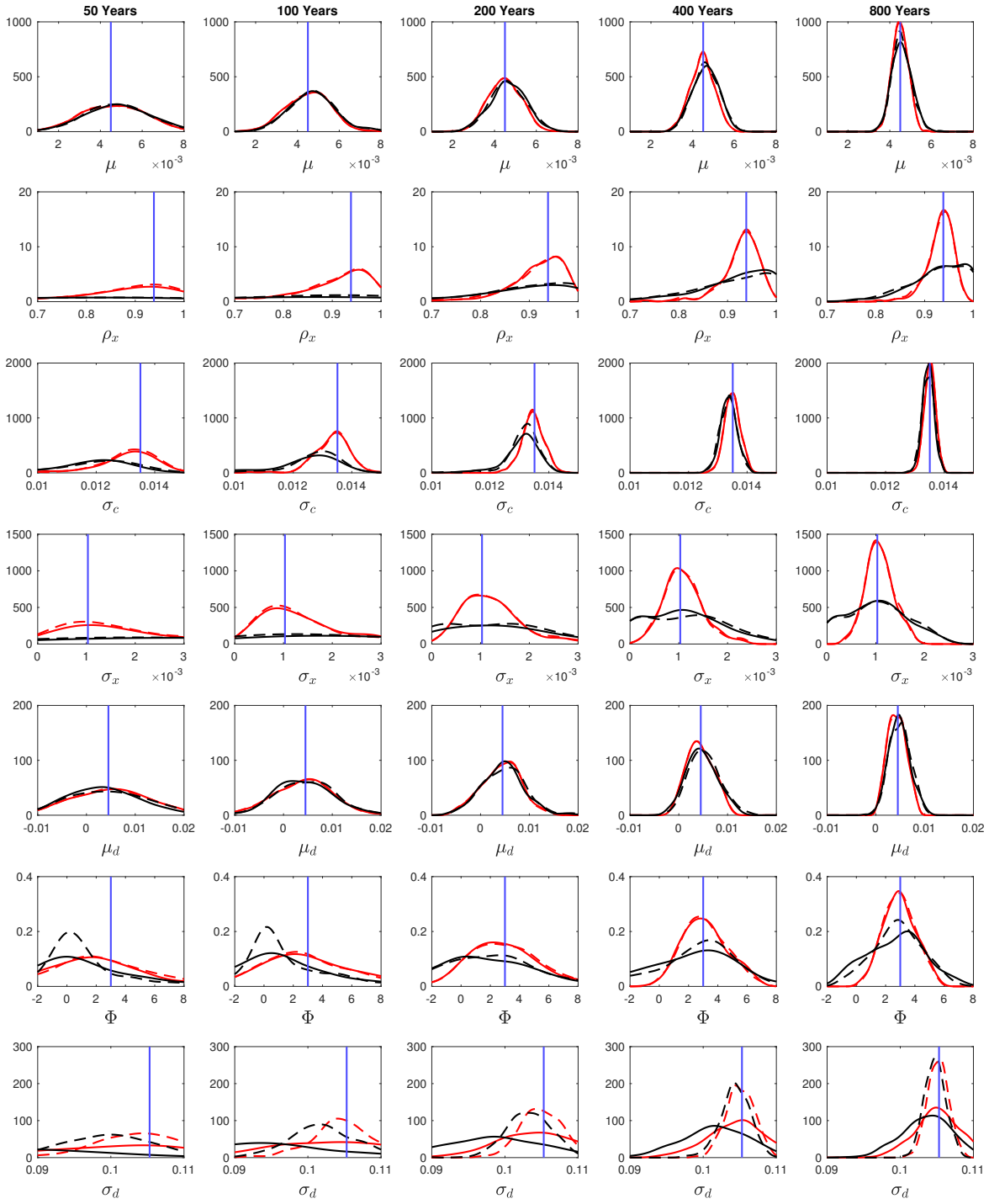


Figure 12: Kernel density estimates for the parameters of the long-run risk model (28) using RLI (red lines) and SMM (black lines) under the full observation regime (dashed lines) and the occasional observation regime (solid lines) for 400 simulated datasets of 50, 100, 200, 400, and 800 years, respectively. The blue vertical lines mark the true population parameter used for the data generation.

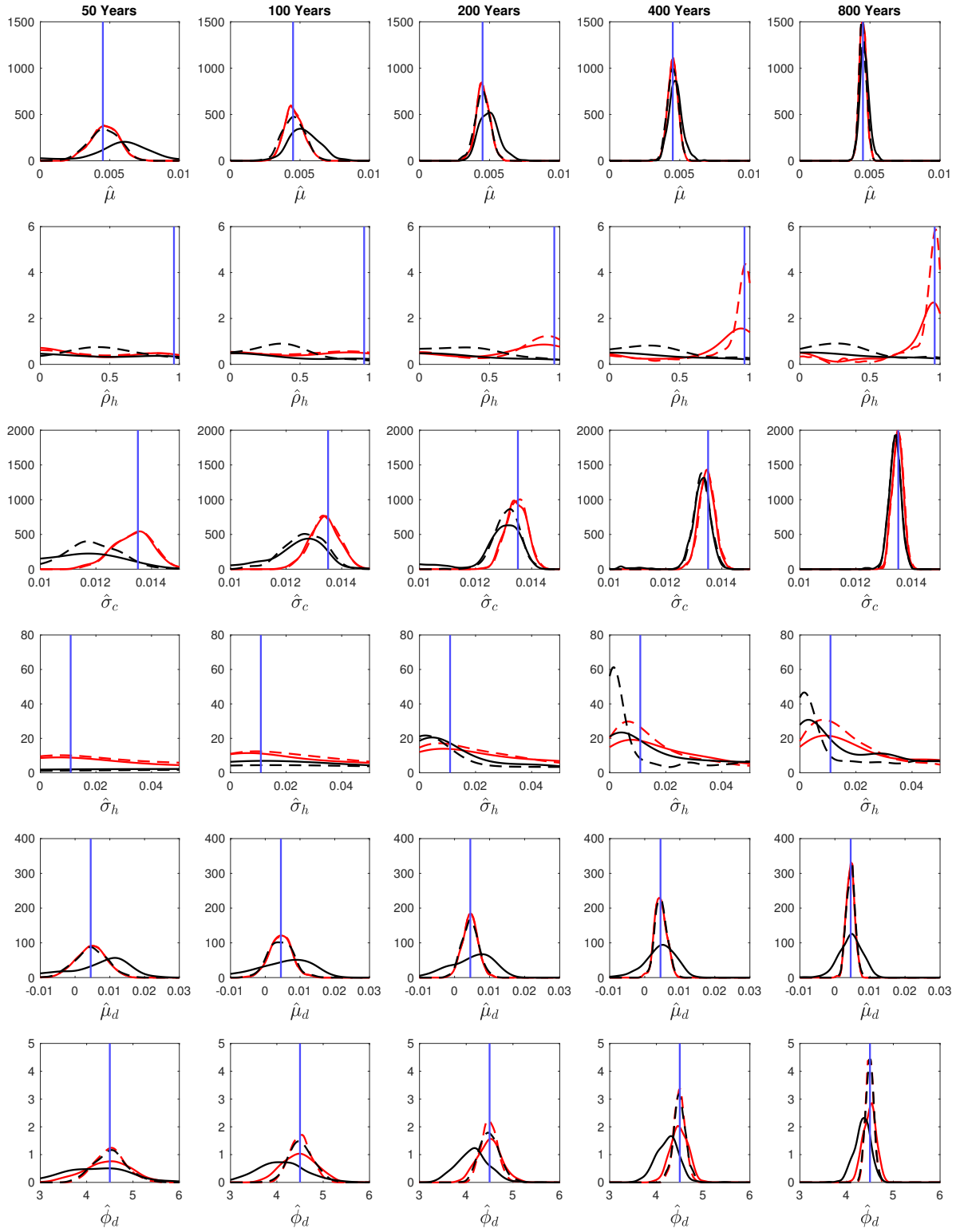


Figure 13: Kernel density estimates for the parameters of the stochastic volatility model (29) using RLI (red lines) and SMM (black lines) under the full observation regime (dashed lines) and the occasional observation regime (solid lines) for 400 simulated datasets of 50, 100, 200, 400, and 800 years, respectively. The blue vertical lines mark the true population parameter used for the data generation.

B.2 Additional material for optimal replacement of GMC bus engines model

In the main body of the paper, we omitted the precise discussion of the solution process for the value function as well as the corresponding likelihood function and approaches to maximize it for brevity.

Under the EV1 assumption on the distribution of ϵ, Rust (1987) derives a (partial) closed-form solution for the conditional expected value as

$$\begin{aligned} EV_\theta(x,i) \equiv \mathbb{E}\left[V_\theta(\tilde{x},\tilde{\epsilon})|x,i;\theta_3\right] &= \int\int V_\theta(\tilde{x},\tilde{\epsilon})P(\tilde{x}|x,i;\theta_3)q(\tilde{\epsilon})\mathrm{d}\tilde{\epsilon}\mathrm{d}\tilde{x} \\ &= \int \log\left(\sum_{j\in\{0,1\}} \exp\left(m(\tilde{x},j;\theta)\right)\right) P(\tilde{x}|x,i;\theta_3)\mathrm{d}\tilde{x} \end{aligned} \tag{44}$$

with q denoting the density function of the (standard) EV1 distribution here, and $m(x,i;\theta) \equiv u(x,i;\theta_1,RC) + \beta EV_\theta(x,i)$ for notational brevity.

The continuous nature of the mileage state requires a numerical approximation of the integral in Equation (44). Gauss–Hermite quadrature rule is designed for computing expectations of (functions of) normally distributed random variables but can be also applied to expectations of log-normal random variables after transforming the integration variable, as we show in Appendix A.5. It preserves its fast convergence properties of polynomial or even exponential order and is thus equally well suited for log-normal random variables if the integrand does not gain a singularity from the transformation. This is usually the case for economic models. Let us denote the Gauss–Hermite nodes and weights by $\{(c_k,w_k)\}_{k=1}^{N^Q}$. Then, the EV function can be approximated by

$$EV_\theta(x,i) \approx \sum_{k=1}^{N^Q} \sqrt{\pi}^{-1} w_k \log\left(\sum_{j\in\{0,1\}} \exp\left(m((1-i)x + \exp\left(\mu+\sqrt{2}\sigma c_k\right),j;\theta)\right)\right). \tag{45}$$

Suppose an econometrician wants to quantify the economic trade-off of the agent, but lacks knowledge of the costs parameter values and parametrization of the state's law of motion. The econometrician does, however, have data on all renewal decisions and observes the mileage state of the buses at each inspection by the manager. Then, the likelihood function of the parameters in question is composed of two sources of contributions: first, the joint probability or density of the state transitions, $P(x_t|x_{t-1},i_{t-1};\theta)$; and second, the conditional choice probabilities, $P(i_t|x_t;\theta)$. In the case of dynamic logit models with EV1 errors, the latter can be expressed as a function of the state variables as follows (Rust, 1987):

$$Pr(i_t|x_t;\theta) = \frac{\exp\left(m(x_t,i_t;\theta)\right)}{\sum_{j\in\{0,1\}}\exp\left(m(x_t,j;\theta)\right)}.$$

Assuming, for notational simplicity, that the sample consists of one bus only—observed over a time horizon $\mathcal{T} = \{1,\ldots,T\}$—the likelihood of the parameter θ given a sample $\{x_t,i_t\}_{t\in\mathcal{T}}$ reads

$$L(\theta;\{x_t,i_t\}_{t\in\mathcal{T}}) = Pr(i_1|x_1;\theta)\prod_{t=2}^{T} Pr(i_t|x_t;\theta)P(x_t|x_{t-1},i_{t-1};\theta). \tag{46}$$

Finally, when maximizing the likelihood function (46) to obtain concrete parameter estimates, the econometrician has to ensure that the model—that is, the Bellman equation (33)—is solved for the likelihood maximizing parameters. This can be done using the nested fixed point algorithm (NFXP; Rust, 1987), which computes the solution to the Bellman equation before each evaluation of the likelihood function. Alternatively, the likelihood can be maximized subject to a set of constraints implied by the Bellman equation, which is known as mathematical programming with equilibrium constraints (MPEC; Su and Judd, 2012)[35].

We now turn to the likelihood recursion for the Rust (1987) model with occasionally observed mileage state presented in Section 3.2.2. Specifically, it is given by:

$$\hat{f}_t^\theta(x) = \begin{cases} 1 & t > T \\ \sum_{k=1}^{N^Q} \sqrt{\pi}^{-1} w_k & \\ \quad \cdot Pr(i_t | \exp\Big(\mu + \sqrt{2}\sigma c_k\Big) + (1 - i_{t-1}) x_{t-1}; \theta) & t-1 \in \bar{\mathcal{T}}, t \notin \bar{\mathcal{T}} \\ \quad \cdot \hat{f}_{t+1}^\theta(\exp\Big(\mu + \sqrt{2}\sigma c_k\Big) + (1 - i_{t-1}) x_{t-1}) & \\ Pr(i_t | x_t; \theta) p(x_t | x, i_{t-1}; \theta_3) \hat{f}_{t+1}^\theta(x_t) & t-1 \notin \bar{\mathcal{T}}, t \in \bar{\mathcal{T}} \\ Pr(i_t | x_t; \theta) p(x_t | x_{t-1}, i_{t-1}; \theta_3) \hat{f}_{t+1}^\theta(x_t) & t-1 \in \bar{\mathcal{T}}, t \in \bar{\mathcal{T}} \\ \hat{\mathcal{I}}\Big(\Big\{ \sum_{k=1}^{N^Q} \sqrt{\pi}^{-1} w_k & \\ \quad \cdot Pr(i_t | \exp\Big(\mu + \sqrt{2}\sigma c_k\Big) + (1 - i_{t-1}) g_j; \theta) & \text{otherwise,} \\ \quad \cdot \hat{f}_{t+1}^\theta(\exp\Big(\mu + \sqrt{2}\sigma c_k\Big) + (1 - i_{t-1}) g_j) \Big\}_{j=1}^{N^I} \Big) & \end{cases} \tag{47}$$

with Gauss–Hermite nodes and weights $\{(c_k, w_k)\}_{k=1}^{N^Q}$. Note that $\{g_j\}_{j=1}^{N^I}$ denotes our interpolation grid, and as for the approximation of the expected value function in (45), we use splines to approximate $\hat{f}_t^\theta$ throughout the recursion. The (conditional) likelihood function based on occasional mileage state observations reads

$$L(\theta; \{\{x_t\}_{t \in \bar{\mathcal{T}}}, \{i_t\}_{t \in \mathcal{T}}\}) = Pr(i_1 | x_1; \theta) \hat{f}_2^\theta(x_1). \tag{48}$$

In the following, we present additional results for the bus engine replacement model of Section 3.2.

Figure 14 illustrates the fit for the original dataset of Rust (1987) by contrasting a kernel density estimate of the empirical distribution, our log-normal model, and a histogram with the same class width as in the discretization of the original model.[36]

We now address why the additional variance of the law of motion estimators does not carry over to the costs parameters estimators. First and foremost, the correlation plot in Figure 15

[35] In this formulation, the expected value EV is a continuous function of the mileage state x and has to be approximated accordingly; we use splines throughout our implementation. Moreover, the implicit nature of the problem requires the application of a projection or collocation method if we want to integrate it with MPEC (as opposed to iterative procedures such as value function iteration, which can be integrated with NFXP, but not with MPEC); see, for example, Judd (1992, 1998).

[36] Note that the histogram depicts the discrete mileage transitions for each period, whereas the transition probabilities in Rust (1987) capture the transitions between mileage bins. For example, a transition from mileage state 4,000 to state 6,000 would fall in the first histogram class in Figure 14, because the difference is only 2,000, but in the second in the original formulation, because the bin changes from first to second.

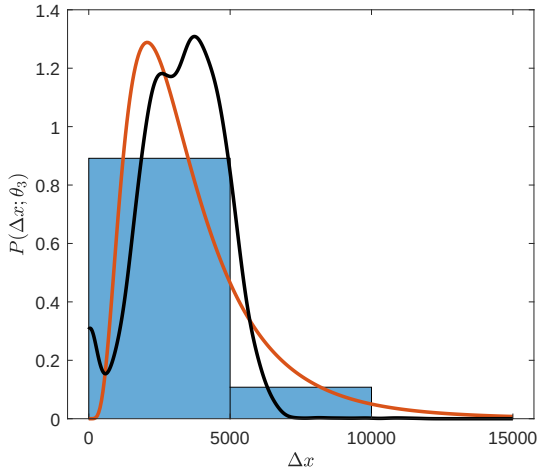


Figure 14: Transition probability density functions for the continuous mileage state space. Kernel fit of the empirical distribution of state transitions (black line); fitted log-normal model (red line); histogram of transitions with class width equal to $5,000$ miles.

delivers insights into the correlation structure of the estimates under occasional observations: while the estimators of the costs parameters are highly correlated among themselves, they are apparently uncorrelated with the estimators of the parameters of the law of motion. Second, it is not the parameters of the law of motion themselves that matter for the decision-making, but rather their implied modes as, e.g., can be seen in Equation (33). Figure 16 depicts the distribution of the mean and variance of the log-normal distribution derived from its parameters estimates; apparently, the mean mileage increment in particular is estimated quite accurately even from occasional state observations. The Q-Q plot in Figure 17 compares the joint distributions of the estimators for full and occasional observation regimes based on their Mahalanobis distance, once for all parameters (left), and once for the two costs parameter estimates, together with the estimated mean mileage increment (right). We conclude that (i) there is barely any evidence, other than their variances, against the two distributions being the same (linearly aligned data points, but rotated around zero), and (ii) the difference in variance becomes substantially smaller if we look at the mean of the mileage increments instead of at their raw distributional parameters.

Finally, Figure 18 plots pairs of estimates for each simulated dataset—the estimate using full state observations on the x-axis against the estimate using occasional state observations for the same dataset on the y-axis. The red line corresponds to the linear regression through all estimates. Note that the fitted linear model intersects with the true parameters and is almost equal to the 45° line. Even though the two types of estimates are not identical for each dataset, they are centered symmetrically around the linear model. In fact, Figure 19 reveals that the distribution of $\hat{\theta}_1^{full} - \hat{\theta}_1^{occ}$ for each dataset is approximately normal, and symmetric around zero with fat tails for $\widehat{RC}^{full} - \widehat{RC}^{occ}$.

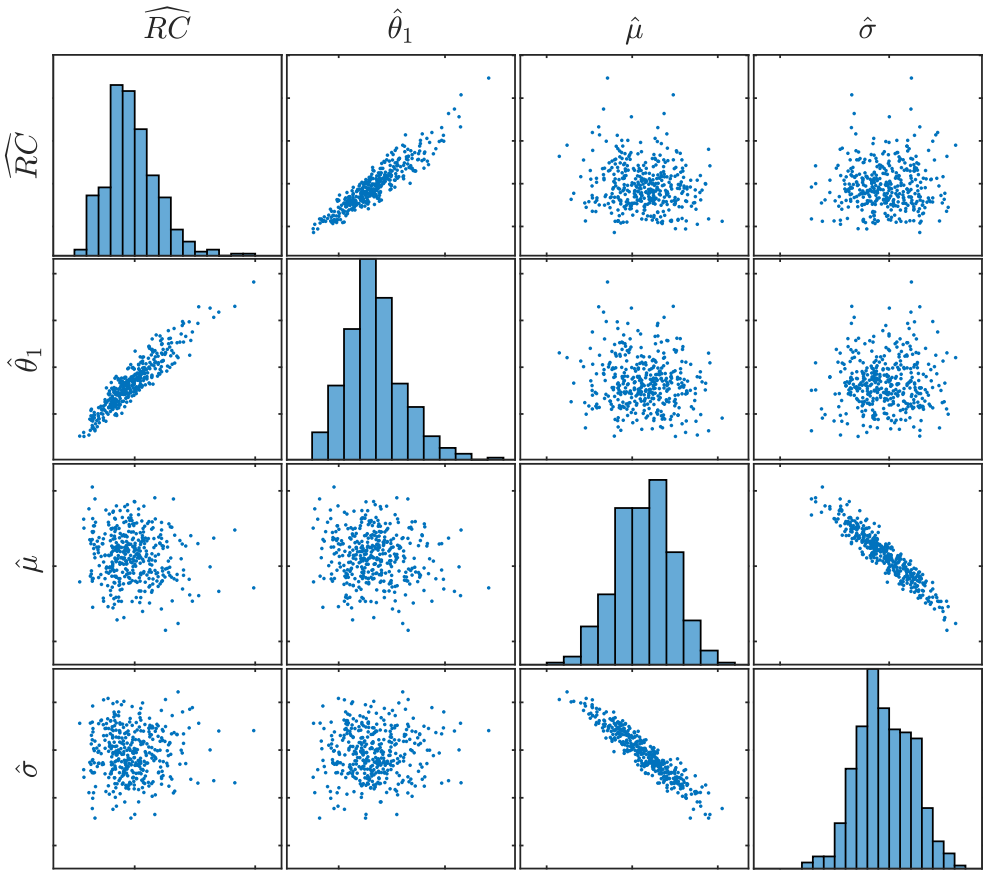
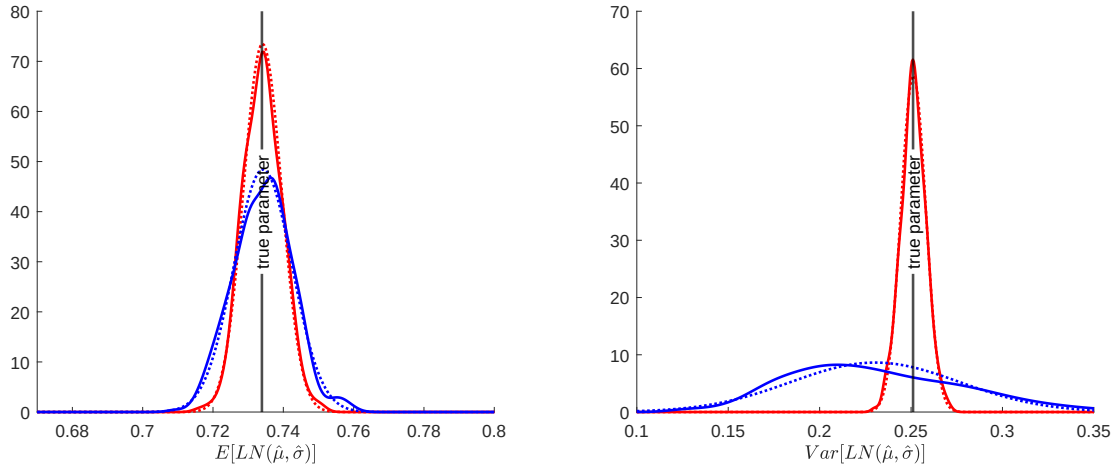
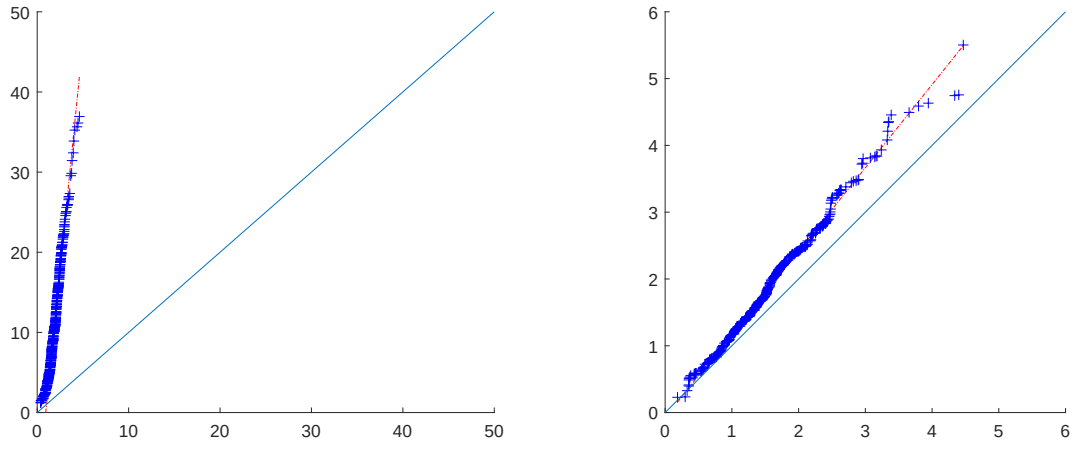


Figure 15: Correlation plot of the maximum likelihood estimates of the costs and transition parameters of Rust (1987) with continuous mileage states under the occasional observation regime for 400 simulated datasets.



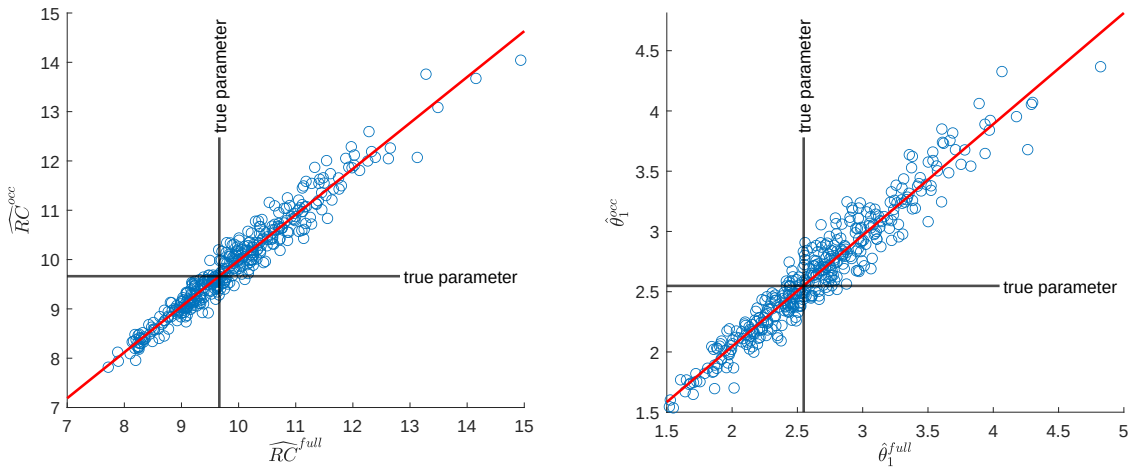
(a) Expectation of $LN(\mu, \sigma)$ (b) Variance of $LN(\mu, \sigma)$

Figure 16: Distributions of the maximum likelihood estimators of the mean and variance of mileage increments in Rust (1987) with continuous mileage states under the full observation regime (red) and the occasional observation regime (blue) for 400 simulated datasets. The solid lines show the kernel fit, the dotted lines the normal distribution using the sample mean and sample variance, the black vertical line the parameter value used for simulation.



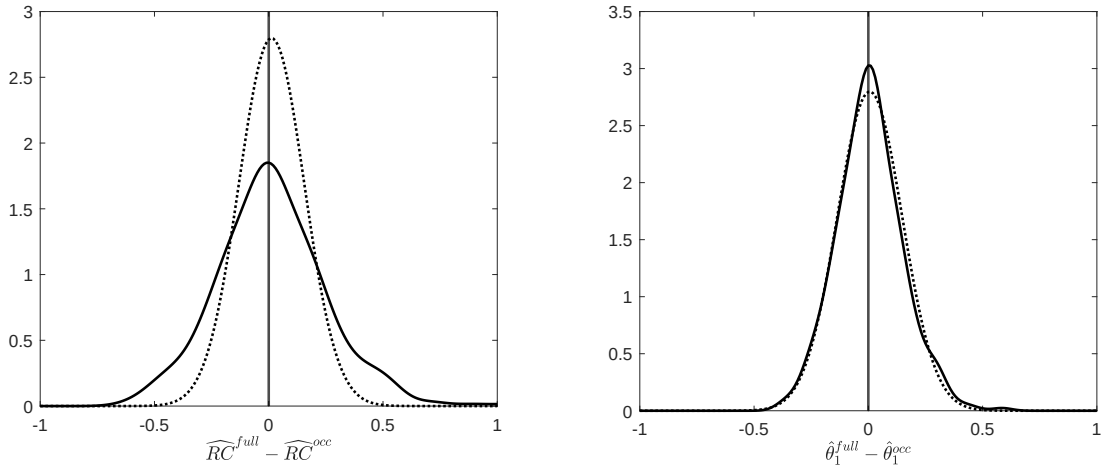
(a) Mahal. Dist. $(\widehat{RC}, \hat{\theta}, \hat{\mu}, \hat{\sigma})$ (b) Mahal. Dist. $(\widehat{RC}, \hat{\theta}, \mathbb{E}[LN(\hat{\mu}, \hat{\sigma})])$

Figure 17: Q-Q plot (Mahalanobis distance) of the maximum likelihood estimates of the costs and transition parameters of Rust (1987) with continuous mileage states under the full observation regime (x-axis) and the occasional observation regime (y-axis) for 400 simulated datasets.



(a) Replacement Parameter RC (b) Regular Maint. Cost Parameter θ_1

Figure 18: Scatter plot of the maximum likelihood estimates (blue circles) under the full observation (x-axis) and the occasional observation regimes (y-axis) for 400 simulated datasets. The red line depicts a fitted linear model; vertical and horizontal black solid lines depict the true parameter values.



(a) Replacement Parameter RC (b) Regular Maint. Cost Parameter θ_1

Figure 19: Distribution of the difference of the maximum likelihood estimates under the full observation and the occasional observation regimes for 400 simulated datasets. The solid line shows the kernel fit, the dotted line the normal distribution using the sample mean and sample variance.