Inference for missing data in state-space models*

Alexandros Gilch† Gregor Reich‡

10th October 2025

PRELIMINARY DRAFT

Abstract

Nonlinear, non-Gaussian state-space models are a standard tool for analyzing time series or panel data with latent state variables, but estimating their parameters and, even more so, the latent states is challenging. We provide a comprehensive methodology to estimate the latent states, particularly addressing two issues: First, because the latent state is serially correlated, accurate point estimators and prediction bands require evaluating high-dimensional integrals arising from marginalizing the latent path. We propose a deterministic recursive quadrature and interpolation (RQI) algorithm to approximate these integrals, exploiting the efficiency of lower-dimensional numerical algorithms. Second, ignoring uncertainty about the model parametrization yields overconfident prediction bands. We develop a framework of prediction-band unions that incorporate parameter uncertainty, which can be computed via a sequence of constrained optimization problems solvable with off-the-shelf packages. We demonstrate the efficiency of RQI in extensive Monte Carlo studies for a Stochastic Volatility model, benchmarking against RQI a popular particle smoothing algorithm. Finally, we conduct full predictive inference for a sequence of endogenously unobserved prices using data from a steel-trading firm and a dynamic profit-maximization model.

Keywords: state-space models, missing data, predictive inference, recursive likelihood integration, smoothing, occasional observations

*We thank Christian Bayer, Einar Breivik, Jörn Frerking, Joachim Freyberger, Björn Höppner, Shayan Hundrieser, Julius Kappenberg, Farzad Saidi, Jan Scherer, Frank Schorfheide, and all seminar participants at the University of Bonn and the University of Pennsylvania for helpful comments. Alexandros Gilch acknowledges funding by the DFG through CRC TR 224 (Project C03).

†University of Bonn, Department of Economics, Adenauerallee 24-26, 53113 Bonn, Germany. E-mail: alexandros.gilch@uni-bonn.de

‡Tsumcor Research AG. E-mail: gregor.reich@tsumcor.ch

1 Introduction

A ubiquitous problem in policy analysis with dynamic economic models is that the econometrician does not observe all state variables at every point in time, resulting in incomplete time series for some states. This lack of information makes it difficult to evaluate how a policy change would propagate through the model, since the reaction of each variable depends on the full joint trajectory. Such observability problems arise across many fields: In marketing or empirical IO, the price of a product is often only available in periods when the product is actually purchased (Erdem, Keane and Sun (1999), Hall and Rust (2021)). If we want to assess how a tax reduction affects demand for this product, we must first know what the true prices might have been. In finance, the unobserved variables often include abstract quantities such as the time-varying volatility of asset returns (Engle and Russell (1998), Engle and Patton (2001)), and a common objective is to compare these volatilities across firms, industries and countries.

These examples point to three overarching motives for imputing the missing observations: (i) visualizing how the realized path supports the empirical narrative in a dataset; (ii) comparing analogous (latent) variables across different datasets on a common time line; and (iii) conducting counterfactuals of the model—i.e., assessing what would have happened under alternative policies, parameters, or model equations. Realistic counterfactuals with correct uncertainty quantification require both accurate point estimators and inference for the missing sequence. For each of these use cases, estimating the specific possible paths including their uncertainty is necessary because the general dynamics implied by the model structure and parametrization alone are insufficient.

However, such inference for the missing sequence is challenging. In general, it is based on a structural state-space model that defines the joint distribution of the unobserved sequence, conditional on the observed data for the other state variables of the model. In this setup, two challenges arise: (i) a computational one—accurate plug-in point estimators and prediction bands are costly to compute; and (ii) a methodological one—inference must account for parameter uncertainty. In this paper, we address both: First, we develop a fast, deterministic algorithm that efficiently approximates point estimators and prediction bands for the missing sequence for fixed structural parameter (i.e., plug-in), providing a more efficient, deterministic alternative to simulation-based methods. Second, we introduce a prediction-band union that explicitly incorporates parameter uncertainty, and we show how to compute its projected envelope by solving a series of constrained optimization problems, which can be solved with standard optimization packages. The two contributions are complementary: plug-in bands can be overconfident in small samples, which the uncertainty-aware construction corrects, while the optimization routine repeatedly calls the plug-in components and is computationally viable only because of we can efficiently approximate them.

High-dimensional integrals. Both the point estimates and the prediction sets are derived from the joint conditional law of the latent path given the data. However, computing them requires marginalizing over the unobserved states—i.e., evaluating a high-dimensional integral whose dimension is proportional to the length of the latent sequence. This integral is usually not available in closed form and must be approximated. Existing approaches such as particle smoothing draw samples from the joint distribution and use them both to approximate the integral and to estimate the latent states (see Chopin and Papaspiliopoulos (2020) for a textbook treatment). These methods face two limitations: (i) sampling-based integration converges slowly, making accurate approximation computationally expensive—especially when estimators must be recomputed many times (e.g., computing volatility paths for a universe of stocks); and (ii) they do not fully exploit occasional observations of the latent states (e.g., in the pricing data case, prices are observed when purchases occur and should be used directly in the estimation process).

We address both limitations with a recursive quadrature and interpolation (RQI) algorithm for computing plug-in point estimators and exact prediction bands when parameters are treated as known. Similar to the recursive likelihood integration (RLI) algorithm (Reich (2018), Gilch et al. (2025)), RQI alternates numerical integration and interpolation within a recursive representation of the latent-state integrals. This leverages the fast convergence of deterministic quadrature/integration and interpolation methods to achieve an approximation error with polynomial convergence rate. Moreover, RQI naturally incorporates occasional observations: whenever a latent state is observed, the corresponding recursion step collapses to a single density evaluation rather than an integral.

Inference under parameter uncertainty. The methodological difficulty is that the structural model is parametrized, and the parameter (vector) θ is typically estimated from the same data used to infer the latent path. Naïve plug-in inference for the missing sequence therefore understates uncertainty if it ignores estimation error in θ. Bayesian inference addresses this issue by imposing distributional assumption through a prior on θ. Parametric bootstrap methods resample the data but face typical challenges when drawing from time series and computing quantiles of high-dimensional objects.

We avoid distributional assumptions and adopt a (fully frequentist) predictive inference perspective. Specifically, we define a prediction-band union as the union of plug-in bands over a confidence set for θ, establish a coverage lower bound for this union, and calibrate it to achieve a target frequentist coverage under parameter uncertainty. Computationally, we show that the (projected) prediction band union can be obtained by solving a series of constrained optimization problems that standard

solvers handle reliably—an approach that is practical precisely because the plug-in components (point path, variance, and coverage) are computed efficiently with RQI.

We demonstrate efficiency and feasibility of the RQI algorithm and our estimators in simulation studies as well as using real data from a steel-trading firm (Hall and Rust, 2021). First, we carry out simulation exercises to verify the theoretical error convergence rates using a standard stochastic volatility model with simulated data. We show that the relative approximation error of the mean integral decays with a polynomial rate, $O(N_{total}^{-4})$, when using the RQI algorithm with cubic splines and Gaussian quadrature, compared to the standard probabilistic Monte Carlo rate, $O_p(N_{total}^{-1/2})$, when using an FFBS particle smoother. The same holds true for the plug-in prediction band, despite its computation involving a root-finding problem. Second, we demonstrate that our proposed prediction band union satisfies the predictive coverage criterion. Simulating data from a linear-Gaussian model, we find that for a target coverage of 95% the prediction band union covers the true sequence in 99% of all simulations, i.e., even overcovers, whereas the plug-in prediction band ignoring parameter uncertainty is overconfident, covering in only 85% of all simulations. Third, we prove applicability of our method in a real-world application, analyzing data from a steel-trading firm that buys on the wholesale market and resells on the retail market. The dataset contains only occasional observations of the wholesale price p_t, specifically in periods when the firm restocks. Using the dynamic profit-maximization model presented in Hall and Rust (2021), and applying the methods developed in this paper, we estimate the wholesale price sequence in non-observed periods; we report the mean path and the projected prediction band union, thereby delivering full predictive inference for p_t under both model-induced randomness and parameter uncertainty.

We contribute to two literatures: First, the approximation of high-dimensional integrals arising from marginalization of the latent state in nonlinear state-space models; second, inference for the latent state under parameter uncertainty. Underlying both is the literature on nonlinear, non-Gaussian state-space models which are a standard tool for time series and panel data with latent states. Influential applications were pioneered in dynamic IO (Pakes (1986), Rust (1987), Rust (1988)), labor economics (Keane and Wolpin (1994), Keane and Wolpin (1997)), marketing (Erdem and Keane (1996), Dubé, Hitsch and Manchanda (2005)), macroeconomics (Fernández-Villaverde and Rubio-Ramírez (2007), Aruoba, Bocola and Schorfheide (2017), Aruoba et al. (2021)), and finance (Shephard (1997), Kim, Shepherd and Chib (1998)).

High-dimensional integrals. For these nonlinear models, inference for the latent state requires evaluating high-dimensional integrals that are not available in closed form

and therefore must be approximated. Linearizing the model yields closed-form expressions for these integrals—obtainable via Kalman filtering and smoothing—but can introduce large errors and often defeats the purpose of a nonlinear specification (Fernandez-Villaverde, Rubio-Ramirez and Santos, 2006). Simulation-based methods retain the full nonlinear model and sample missing states from model-implied conditional distributions—e.g., particle filtering/smoothing (Fernández-Villaverde and Rubio-Ramírez (2007), Herbst and Schorfheide (2014), Blevins (2015), Chopin and Papaspiliopoulos (2020)), Gibbs sampling (Norets (2009)) and the GHK simulator (Keane (1994))—to obtain Monte Carlo approximations of the integrals; however, such approximations achieve only the probabilistic Monte Carlo error rate, thus there is a trade-off between cheap but rough or computationally expensive but accurate point estimators and prediction bands. In contrast, deterministic numerical schemes based on efficient quadrature and interpolation have been developed to approximate the model likelihood, a high-dimensional integral with similar structure (Reich (2018), Gilch et al. (2025)), but not yet to compute the integrals needed for inference on the latent state. Kitagawa (1987) approximates these integrals also with numerical integration and interpolation, however, only for linear non-Gaussian models and without explicitly controlling the coverage of the prediction bands.

Our first contribution fills this gap in the literature by also adopting a scheme of alternating numerical integration and interpolation to approximate these integrals with high numerical efficiency for general nonlinear state-space models. Importantly, the RQI algorithm also incorporates occasional observations of the latent state naturally, allowing us to improve the statistical efficiency of our estimators over approaches that cannot utilize these observations.

Inference under parameter uncertainty. The integrals discussed and approximated in the literature above are derived from a parametrized state-space model. However, proper inference has to address the fact that the parameters of the model are usually estimated themselves and hence are subject to estimation uncertainty. The literature deals with this in different ways: When samples are large, parameter uncertainty may be negligible, which motivates plug-in procedures that condition on estimated parameters and account only for model-implied randomness (Kitagawa (1987), Durbin and Koopman (2012)). However, with too little data parameter uncertainty cannot be ignored. Bayesian approaches integrate over a prior on the parameters (Hamilton (1986), Quenneville and Singh (2000), Durbin (2002)), though the choice of prior is often debated and some researchers prefer not to impose such distributional assumptions. In contrast, frequentist predictive inference (Cox (1975) and Barndorff-Nielsen and Cox (1996), see also Geisser (1993) and Young and Smith (2005) for textbook treatments), addresses prediction under

parameter uncertainty; however, the typical application for predictive inference is the estimation of future observations given past data.

Therefore, our second contribution is to formulate predictive inference for "predicting" latent states and provide frequentist prediction bands under parameter uncertainty for the latent state. To our knowledge, we are the first to construct such simultaneous prediction bands for general nonlinear state-space models that directly satisfy a predictive-inference criterion in the frequentist sense. Related works by Pfeffermann and Tiller (2005) and Rodríguez and Ruiz (2012) use bootstrap methods to compute per-period prediction mean squared errors, but do not deliver (simultaneous) prediction bands for the entire missing sequence and is limited to linear-Gaussian models.

The remainder of this paper is organized as follows: Section 2 introduces the formal setup and states the problem; in particular, Section 2.4 introduces the concepts from predictive inference that are fundamental to our method. Section 3 states our contributions: Section 3.1 develops the recursive quadrature and interpolation (RQI) algorithm for plug-in point estimators and prediction bands and reports its convergence rates in simulation experiments. Section 3.2 defines the prediction band union, derives a lower bound on its coverage and demonstrates its empirical coverage in simulation experiments. Section 4 reports the result from an application of our method to the steel trading model of HR. Section 5 concludes.

2 Framework

This section lays out the setting on which our contribution builds. Section 2.1 formally defines the state-space model that encodes the economic environment; Section 2.2 then discusses data availability and formalizes occasional observations. Next, Section 2.3 details maximum-likelihood estimation via RLI, providing the foundation for the computational methods we develop in Section 3.1. Finally, Section 2.4 presents our framework for inference on the missing sequence, specifying the objects computed in Sections 3.1 and 3.2.

2.1 State-Space Models

In this paper, we consider a discrete-time stochastic process $\{Y_t, X_t\}$, with measurement variable y_t and state variable x_t following a parametric Markov transition density $P_\theta^{X,Y}$. This density may not be available in closed form but rather it is induced by a structural economic model.

The transition density of the state variables, $P_\theta^{X,Y}$, typically arises from a functional relation between x_t and y_t, in our case motivated by a formal economic model:

$$y_t = \Psi_\theta(x_t, \eta_t) \qquad \eta_t \overset{\text{i.i.d.}}{\sim} P_\theta^\eta \tag{1}$$

$$x_t = \Phi_\theta(x_{t-1}, \varepsilon_t) \qquad \varepsilon_t \overset{\text{i.i.d.}}{\sim} P_\theta^\varepsilon \tag{2}$$

where η_t and ε_t are random errors. Their distributions $P_\theta^\eta, P_\theta^\varepsilon$ and the functions Ψ_θ, Φ_θ are known up to a finite-dimensional parameter vector $\theta \in \Theta$. Together, the model equations and the distributions of the random errors induce a probability law for the transition from (y_{t-1}, x_{t-1}) to (y_t, x_t) satisfying a Markov property. We further assume that this law admits a transition density $P_\theta^{X,Y}$, which therefore factorizes as

$$P_\theta^{X,Y}(y_t, x_t | y_{t-1}, x_{t-1}) = P_\theta^Y(y_t|x_t)\, P_\theta^X(x_t|x_{t-1}). \tag{3}$$

Both y_t and x_t may be vectors, so Ψ_θ and Φ_θ are, in general, multivariate functions. However, our analysis focuses on the case of one-dimensional x_t: This is because we are interested in estimating x_t for all t, and then visualizing those estimates, which ultimately amounts to reporting each component of x_t separately. In practice, this means we must marginalize all other components of x_t not under consideration anyway. For the sake of the argument, it is therefore convenient to write the model and all associated densities in terms of the one-dimensional x_t of interest; however, the theory extends naturally to the multivariate case.

Crucially, we allow for non-linear functions and non-Gaussian densities. "Non-linear" may include cases where the functions are defined implicitly as the solution to a (per-period) optimization problem. A large literature addresses how to solve such problems, so we assume that these state-dependent solutions—and hence the functions Φ_θ and Ψ_θ—are either available in closed form or can be approximated to high accuracy.

2.2 Missing Data

In many state-space models, y_t is treated as the measurement variable which is always observed by the researcher, whereas x_t is latent. In this paper, we consider a more general setting in which the researcher observes the full measurement series $\{y_t\}_{t=0}^T$ but, in addition, observes x_t for some periods $t \in \bar{\mathcal{T}} \subset \{0, \dots, T\}$; note that this includes the case of no observations, i.e., $\bar{\mathcal{T}} = \emptyset$. However, the observation process itself may depend on the model variables. Not accounting for such endogeneity can bias inference based on the combined data $\{\{x_t\}_{t\in\bar{\mathcal{T}}}, \{y_t\}_{t=0}^T\}$. Therefore, the structural model must impose assumptions on either exogeneity of the observation process or a functional form governing its endogeneity.

We can formalize occasional observations using a missingness indicator $m_t \in \{0, 1\}$

that records whether x_t is present in the dataset: $m_t = 0$ if $t \in \bar{\mathcal{T}}$ and $m_t = 1$ otherwise. We assume the joint process (y_t, x_t, m_t) is Markov. For simplicity, m_t depends only on current (y_t, x_t). Because missingness is a researcher-side issue, all agents observe the full data, implying

$$P_\theta^{M,X,Y}(m_t, x_t, y_t | m_{t-1}, y_{t-1}, x_{t-1}) = P_\theta^M(m_t | x_t, y_t) P_\theta^{X,Y}(y_t, x_t | y_{t-1}, x_{t-1}), \tag{4}$$

i.e., m_t doesn't affect the economic variables x_t, y_t.

The missingness mechanism $P_\theta^M(m_t|x_t, y_t)$ formalizes when and with what probability the researcher observes x_t. Its key feature is whether it depends on x_t: if it does not, we call the mechanism exogenous and the data missing-at-random (MAR), because observation does not depend on the realized value;[1] if it does depend on x_t, we call $P_\theta^M(m_t|y_t, x_t)$ endogenous and the data not-missing-at-random (NMAR).[2]

For notational simplicity, we adopt two conventions for the remainder of the paper: First, we set $\bar{\mathcal{T}} = \emptyset$, i.e., x_t is never observed, and we estimate the missing sequence $\{x_t\}_{t=0}^T$ given the data $\{y_t\}_{t=0}^T$. In the case with occasional observations, all derivations carry over verbatim by estimating $\{x_t\}_{t\in\bar{\mathcal{T}}}$ given $\{\{x_t\}_{t\in\bar{\mathcal{T}}}, \{y_t\}_{t=0}^T\}$; this does not alter the formal arguments, and our computational methods are designed to accommodate occasional observations. Second, we treat the data as MAR. All results extend to NMAR settings by including m_t as an additional observed variable and treating (m_t, y_t) as the full set of observations. In both cases, see Gilch et al. (2025) for details on modeling the missing mechanism for latent states in state-space models.

Under non-observation of x_t in some or all periods, allowing non-linearity and/or non-Gaussianity in the economics creates two challenges: First, key objects of interest—such as the likelihood—typically lack closed-form expressions. We show how to approximate these objects using recursive likelihood integration (RLI): in Section 2.3 for the likelihood itself, and in Section 3 for inference over $\{x_t\}_{t=0}^T$. Second, the densities P_θ^X and P_θ^Y often cannot be derived in closed form from the transition functions and error distributions. In the appendix, we explain how to handle this within the existing RLI framework and how the same approach carries over to the algorithms proposed in this paper.

[1] A prominent example for MAR data is mixed-frequency data, where time series are observed periodically but at different frequencies. This is common in macroeconomic applications with monthly, quarterly, or annual series, or with series aggregated over several periods and then reported at a lower frequency. For instance, the high-frequency series $y_{t\,t=0}^T$ might be monthly, while $x_{t\,t=0}^T$ is recorded annually. Then the observation set is $\bar{\mathcal{T}} = 0, 12, 24, \ldots, T$, entirely independent of the values taken by $x_{t\,t=0}^T$ because the mechanism depends only on t and is therefore exogenous.

[2] An example of NMAR data involves prices: in scanner data, prices are often observed only when the corresponding product is purchased Erdem, Keane and Sun (1999). Because purchase decisions depend directly on price, price observation is endogenous. A similar setting with wholesale steel prices is studied in Hall and Rust (2021) and motivates our analysis in Section 4.

2.3 Parameter Estimation using Recursive Likelihood Integration

To use state-space models, e.g., for policy experiments or to evaluate counterfactuals, one typically needs to estimate the parameters θ of the model first. A popular approach is maximum likelihood estimation; however, for models other than the linear-Gaussian, the likelihood forms an integral over the latent sequence that is not available in closed form. Moreover, because estimation requires evaluating the likelihood for many candidate values of θ, a fast and accurate approximation of this integral is essential.

The recursive likelihood integration (RLI) algorithm provides such an approximation by applying deterministic numerical integration and interpolation methods. Furthermore, being fully deterministic, it avoids issues related to noisy objective functions and their optimization, which are inherent in some alternative approximation and estimation methods. Since the methods we develop in Section 3.1 build on similar ideas than RLI, we consider it worthwhile to briefly outline them here.

Given a model (1)-(2) that admits probability densities for the data, we define the likelihood of the parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$ as the density of the data seen as function of the parameter and use it for estimating θ. In the full-data case $\bar{\mathcal{T}} = 0, \ldots, T$, the likelihood is the standard product of the period-wise densities,

$$L\left(\theta | \{x_t, y_t\}_{t=0}^T\right) = P_\theta^{x_{0:T}, y_{0:T}}\left(\{x_t, y_t\}_{t=0}^T\right) = \prod_{t=1}^T P_\theta^{X,Y}\left(y_t, x_t | y_{t-1}, x_{t-1}\right). \tag{5}$$

However, with x_t only observed occasionally, the likelihood forms an integral over the missing observations:[3]

$$\begin{aligned} L\left(\theta | y_{0:T}\right) &= P_\theta^{y_{0:T}}\left(y_{0:T}\right) = \mathbb{E}_{x_{0:T}}\left[P_\theta^{y_{0:T}|x_{0:T}}\left(y_{0:T}|x_{0:T}\right)\right] \\ &= \int_{\mathcal{S}_x^{T+1}} P_\theta^{X,Y}\left(y_{0:T}|x_{0:T}\right) dP_\theta^{x_{0:T}}\left(x_{0:T}\right) \\ &= \int_{\mathcal{S}_x^{T+1}} P_\theta^{X,Y}\left(\{x_t, y_t\}_{t=0}^T\right) dx_{0:T} \\ &= \int_{\mathcal{S}_x^{T+1}} \prod_{t=1}^T P_\theta\left(y_t, x_t | y_{t-1}, x_{t-1}\right) dx_{0:T}. \end{aligned} \tag{6}$$

The likelihood integral is typically unavailable in closed form and must be approximated. In particular for non-linear models, the integral must be computed numerically or by simulation. Although the integrand is a product of terms, the integral does not factor, making approximation of the integral challenging: each $P_\theta\left(y_t, x_t | y_{t-1}, x_{t-1}\right)$ in-

[3] Recall that we assumed $\bar{\mathcal{T}} = \emptyset$ for the rest of this paper, hence, no occasional observations of x_t are part of the dataset.

volves two integration variables x_t, x_{t-1}; due to this pairwise entanglement the integral has dimension $(T+1)$.

As a consequence, many practitioners use simulation-based integration methods. However Monte Carlo approximations have probabilistic error $O_p(N_{eval}^{-1/2})$ and thus demand substantial computational effort for high accuracy; furthermore, to our knowledge they are not designed to incorporate occasional observations, making estimation using these methods also statistically inefficient. In this paper, we use particle smoothing as benchmark for simulation methods; we provide more details on it in the Appendix.

Estimation requires repeatedly evaluating the data density at many candidate parameters, so speed is critical. This holds for likelihood-based estimation—where we search for a maximizer by iterating over the parameter space—and for Bayesian estimation—where we form the posterior by evaluating the likelihood at draws of θ from a prior or proposal distribution. When optimization is challenging (e.g., due to a large or poorly conditioned parameter space) or when the prior/proposal is difficult to sample from, these computational demands are magnified.

Recursive likelihood integration (RLI) by Reich (2018) is a recursive algorithm that approximates the likelihood efficiently and, as shown by Gilch et al. (2025), seamlessly incorporates occasional observations. Its recursive structure decomposes the likelihood integral into T lower-dimensional—but nested—integrals. RLI approximates this sequence of nested integrals by alternating numerical integration and interpolation. Importantly, the approximation error can be controlled by utilizing fast-converging numerical integration and interpolation methods, yielding better convergence rates than related simulation methods. [4] Additionally, the approximated function is deterministic and therefore doesn't suffer from simulation noise.

RLI is the basis for similar algorithms we develop in Section 3. We provide details on its derivation and implementation in the Appendix.

2.4 Two Frameworks for Inference for the Missing Sequence

We base inference for the missing sequence on the state-space model-induced law $P_\theta^{x_{0:T}|y_{0:T}}$. In Section 2.4.1, we treat the parameter as known, so plug-in inference accounts only for model-induced randomness. However, θ is typically estimated from the same data $y_{0:T}$ so plug-in estimates are subject to parameter uncertainty. In Section 2.4.2, we formalize predictive inference as a frequentist framework for estimating the missing sequence under parameter uncertainty.

[4] To be concrete, the approximation error of the deterministic integration and interpolation methods accumulates linearly with the time-series length T; however we consider T to be fixed in this paper.(Gilch, Reich and Wilms, 2025) show how to adjust RLI for $T \to \infty$ to also obtain asymptotically small approximation errors.

2.4.1 Plug-in Inference

When treating θ as known (or estimated with negligible uncertainty), inference on the missing sequence $x_{0:T}$ is based on the conditional (smoothing) law

$$P_\theta^{x_{0:T}|y_{0:T}}(x_{0:T} \mid y_{0:T}) = \frac{P_\theta^{X,Y}(\{x_t, y_t\}_{t=0}^T)}{P_\theta(y_{0:T})}, \tag{7}$$

i.e., the joint distribution of the latent states given the observed data $y_{0:T} \equiv \{y_t\}_{t=0}^T$ under the model. The law $P_\theta^{x_{0:T}|y_{0:T}}(\cdot \mid y_{0:T})$ quantifies how compatible each entire path $x_{0:T}$ is with the realized finite sample $y_{0:T}$; different data would yield a different posterior over paths. Because state and measurement equations contain random shocks, there is no one-to-one mapping from $y_{0:T}$ to $x_{0:T}$: even at a single time t, multiple state values are a priori compatible with the same observation and with neighboring states. Conditioning on $y_{0:T}$ assigns to each candidate value $x_t = x$ a posterior probability (mass/density), and these assignments across all t and all paths make up $P_\theta^{x_{0:T}|y_{0:T}}(\cdot \mid y_{0:T})$. Since $P_\theta^{x_{0:T}|y_{0:T}}$ is high-dimensional,[5] we summarize it through timewise functionals (e.g., posterior mean, mode, and credible sets for x_t) and through simultaneous credible (plug-in prediction) sets that constrain the entire latent path.

Asymptotically, the posterior $P_\theta^{x_{0:T}|y_{0:T}}$ over the missing sequence does not concentrate on the realized sequence; instead, at any fixed time t the marginal smoothing law converges to a limit distribution conditional on the infinite sample $y_{0:\infty}$. Recall that we are in a time-series setting, so the asymptotic framework means letting the sample length grow ($T \to \infty$). Even with infinite data, the model-induced randomness at a given period t does not vanish. Exactly as in finite samples, multiple state values remain compatible with the observations and the neighboring states. This limit is still data-dependent: it is a random measure determined by the realized $y_{0:\infty}$. Correspondingly, features of the finite-sample smoothing distribution (mean, mode, quantiles, and plug-in prediction sets) converge to the same features under $P_\theta^{x_{0:T}|y_{0:\infty}}$; in particular, the widths of plug-in prediction sets converge rather than vanish.

Furthermore, future observations y_s with $s \gg t$ have a negligible effect on the marginal law of x_t: under standard signal stability (geometric ergodicity of $\{x_t\}_{t=0}^T$) and informative observation channel assumptions on $\{x_t, y_t\}_{t=0}^T$, their influence on the smoothing distribution decays exponentially in the distance $s-t$ (heuristically like ρ^{s-t} for some $0 < \rho < 1$). Therefore, adding very distant observations refines inference at time t only exponentially little; the uncertainty at t does not collapse as $T \to \infty$ but stabilizes to the limit law $P_\theta^{x_{0:T}|y_{0:\infty}}$. Note that the limit depends on the full realized data $y_{0:\infty}$, yet,

[5] Note that while formally $(T+1)$-dimensional, the Markov property of the underlying time series implies typically fast decaying covariances between far away periods, hence the covariance matrix of $P_\theta^{x_{0:T}|y_{0:T}}$ has near-zero entries far away from the diagonal and most of the mass of the distribution, while technically high-dimensional, concentrates around a lower-dimensional subspace of $\mathbb{R}^{(T+1)}$

by exponential forgetting, it depends chiefly on the adjacent observations $Y_{t-\ell:t+\ell}$, with the effect of more distant data decaying like ρ^ℓ.

It turns out that in the Kalman case (linear transition and measurement functions Ψ_θ, Φ_θ with Gaussian noises $P_\theta^\eta, P_\theta^\varepsilon$), these assumptions on the time series $\{x_t, y_t\}_{t=0}^T$ are benign. Indeed, we have $X_t \mid y_{0:T} \sim \mathcal{N}(\mu_{t|T}, \Sigma_{t|T})$, with $\mu_{t|T}$ a linear functional of $y_{0:T}$ whose weights decay geometrically with temporal distance (under a stable transition[6]). The covariances are data-independent and admit a steady state: under standard detectability and stabilizability and nondegeneracy conditions,[7] the filtering covariance converges to the unique stabilizing solution of the discrete algebraic Riccati equation, and the Rauch-Tung-Striebel smoother covariance converges to a positive-definite limit Σ_∞. Thus, uncertainty about X_t does not vanish but stabilizes.

Because the full posterior distribution over the possible paths (7) is high-dimensional, researchers typically report (i) a smoothed point path together with (ii) a simultaneous plug-in prediction (credible) set to express model-induced uncertainty over the missing sequence. For the former, we use either the posterior mean $\bar{x}_t := \mathbb{E}_\theta[X_t \mid Y_{0:T}]$ for all $t \in \bar{\mathcal{T}}$ or the posterior mode (MAP). The mean is the MMSE estimator under squared loss; the MAP corresponds to a frequentist "mode" summary and is often easier to compute via optimization.

For the latter, we seek a random set $\hat{X}_\alpha(\theta, y_{0:T}) \subset \mathcal{S}_x^{(T+1)}$ such that it contains the entire latent path with posterior probability at least $1-\alpha$:

$$P_\theta^{x_{0:T}|y_{0:T}}\Big(\, x_{0:T} \in \hat{X}_\alpha(\theta, y_{0:T}) \;\Big|\; y_{0:T}\Big) \;\geq\; 1-\alpha. \tag{8}$$

This condition enforces simultaneous (pathwise) coverage and is stronger than having $(1-\alpha)$ pointwise intervals at each t. As (8) does not pin down a unique set, a construction rule is required.

In Section 3.1.2 we define the one-scale class of prediction bands, which includes many common prediction band constructions, and show how to compute the unique exact plug-in prediction band in this class.

2.4.2 Predictive Inference

To include uncertainty from parameter estimation, we move to a frequentist framework and define prediction sets using predictive inference, naturally augmenting the previous plug-in prediction set.

The plug-in prediction set from the previous section is overconfident whenever the parameter θ is estimated with error (typically from the same sample $y_{0:T}$). In some applic-

[6] In the univariate case, this would mean the lag-1-autocorrelation of x_t is smaller than one in absolute terms, $|\operatorname{Corr}(x_t, x_{t-1})| < 1$.

[7] Again, in the univariate case, these are satisfied if the lag-1-autocorrelation of x_t is smaller than one in absolute terms.

ations this uncertainty is negligible—e.g., when θ is estimated from very large and informative samples—so researchers sometimes forgo any correction (**¡add-large-sample-cite¿**). In many time-series settings, however, samples are short (macro data), or model dimensionality is large relative to the available information, so estimation uncertainty in $\hat{\theta}$ remains material even for substantial T.

Two prominent remedies are Bayesian inference and parametric bootstrap methods. The Bayesian route introduces a prior on θ and propagates posterior uncertainty about θ into the law of the latent states; while principled, it requires prior choices that some practitioners prefer to avoid. Some frequentist alternatives therefore rely on simulation-based calibration: the parametric bootstrap re-generates the series $\{x_t, y_t\}_{t=0}^{T}$ under the fitted model, re-estimates θ, and uses the resulting empirical distribution of $\hat{\theta}$ to quantify parameter uncertainty and fold it into smoothing or prediction sets (e.g. Pfeffermann and Tiller, 2005; Rodriguez and Ruiz, 2009; Rodríguez and Ruiz, 2012).

We also adopt a frequentist predictive framework (cf. Geisser (1993)) to incorporate parameter uncertainty but avoid any simulation. For each level $\alpha \in (0, 1)$, a prediction set is a random set $S_\alpha(y_{0:T})$ satisfying uniformly for all possible parameters θ

$$\forall\, \theta \in \Theta : \qquad \liminf_{T\to\infty} P_\theta^{X,Y}\left(x_{0:T} \in S_\alpha(y_{0:T})\right) \;\geq\; 1-\alpha. \tag{9}$$

This probability is frequentist: imagine repeatedly generating full series $y_{0:T}$ and the associated latent path $x_{0:T}$ from the model and recomputing the set $S_\alpha(y_{0:T})$ each time. Thus, coverage averages over the sampling of $y_{0:T}$ and the induced randomness in $\hat{\theta}(y_{0:T})$.[8]

Unlike the plug-in case—where one reports a conditional distribution given $y_{0:T}$ under a known θ—with unknown parameters an exact conditional target such as

$$P_\theta\big(x_{0:T} \in S_\alpha(y_{0:T}) \;\big|\; y_{0:T}\big) = 1-\alpha \quad \text{a.s. in } y_{0:T} \tag{10}$$

is generally unattainable. Intuitively, conditioning on the realized $y_{0:T}$ removes sampling variability and forces the set based on that $y_{0:T}$ to achieve $1-\alpha$ uniformly in the unknown θ; to "hedge" against unlikely (θ, Y) combinations, one must inflate S_α substantially—often to the point of being uninformative.[9] We therefore target unconditional predictive coverage—i.e., averaging over repeated sampling of $(x_{0:T}, y_{0:T})$ under P_θ—which naturally accounts for the randomness of $\hat{\theta}(y_{0:T})$ and assigns negligible weight to low-probability (θ, Y) corners. In principle one would derive the joint finite-sample law of $(\hat{\theta}_T,\, x_{0:T} \mid y_{0:T})$ to calibrate S_α, but in state-space models this is typically intractable. Therefore,

[8] In applications where per-time reporting is preferred, we later also consider simultaneous prediction bands $\{B_{t,\alpha}(Y)\}_{t=0}^{T}$ as a common subclass of prediction sets.

[9] In distribution-free settings this tension is formalized by impossibility results for exact conditional coverage (e.g., in conformal prediction, citetVovk). Parametric structure relaxes but does not eliminate the issue: demanding exact conditional validity typically yields very large sets unless the model admits special pivots or ancillary reductions.

practically, $S_\alpha(y_{0:T})$ will be constructed from the estimator $\hat{\theta}_T = \hat{\theta}(y_{0:T})$ together with an explicit adjustment for its sampling uncertainty, e.g., using a confidence set for $\hat{\theta}$.

Taking asymptotics in T has two consequences for interpretation: First, as already emphasized in the plug-in setting, smoothing uncertainty reflects model-induced randomness. Therefore, even with abundant data, the predictive set does not shrink to a singleton (the set only containing the true sequence). Second, with unknown θ, the coverage guarantee we target is approximate: in practice, $S_\alpha(y_{0:T})$ attains its nominal level only as T grows. This is the usual tension between the theoretical properties of large-sample frequentist inference and the finite-sample demands of the researcher's reality.

Even in the linear-Gaussian (Kalman) case, finite-sample exact predictive inference with unknown θ is generally out of reach: closed forms for the sampling law of $\hat{\theta}_T$ are unavailable and there is no useful low-dimensional sufficient statistic to condition on. Exceptions exist in special structures (e.g., pure scale parameters leading to χ^2/t pivots via innovations). What is convenient here is asymptotics: scores/information and sensitivity recursions are explicit, so projected-information LR sets and θ-inflated simultaneous bands are straightforward, and parametric bootstrap calibration is easy.

Finally, a natural goal is that the prediction set under parameter certainty converges, in large samples, to the plug-in set we defined in the previous subsection; yet, such convergence is not automatic. As $T \to \infty$, parameter uncertainty vanishes under standard regularity, so the true θ is effectively known and we can compute the plug-in set evaluated at the true parameter $S^\star_\alpha(y_{0:T}; \theta)$. That target is stronger than the unconditional predictive target because it delivers conditional coverage given $y_{0:T}$. The question, then, is whether the prediction set we construct for the finite-sample, unknown-θ case converges to this plug-in (true-θ) solution as T grows. This would be desirable: even though exact conditional coverage is unattainable in finite samples, it would be recovered in the limit. However, the unconditional predictive target admits multiple asymptotic solutions; by definition these need only satisfy unconditional coverage in the limit. The plug-in (true-θ) set is one such solution—since it is also unconditionally valid—but it is not the only one. Therefore, the predictive inference construction must explicitly encode a limit preference, selecting the plug-in (true-θ) limit among the admissible unconditional limits. Without such a tie-breaker, a predictive procedure may converge to a different valid limit that is typically more conservative than the plug-in set.

In Section 3.2 we develop and implement a constrained optimization algorithm, whose solution is a prediction set achieving predictive inference as defined in (9) under standard regularity. Moreover, our construction is designed to deliver the desired asymptotic, i.e., our prediction set converges asymptotically to the exact plug-in prediction set defined in the previous subsection.

3 Contribution

We contribute along two fronts for state-space models with missing data. First, we develop an RLI-type algorithm for fast, accurate approximation of plug-in point estimators and prediction bands providing a computationally cheaper deterministic alternative to simulation-based methods. Second, we construct a prediction band union—which explicitly incorporates parameter uncertainty—by solving a sequence of constrained optimization problems. These pieces are complementary: plug-in bands can be overconfident in small samples, which the uncertainty-aware bands correct, while the optimization routine repeatedly calls the plug-in bands and is therefore only computationally feasible because the latter can be approximated efficiently

3.1 Improving plug-in estimation using numerical approximation

Existing approaches for computing point estimators and plug-in prediction bands deliver only Monte-Carlo accuracy $O(N_{eval}^{-1/2})$ and struggle to incorporate occasional observations. Wald-type bands with Bonferroni corrections are an ad hoc alternative but are typically over-conservative. We develop RLI-based deterministic algorithms that compute plug-in point estimators and prediction bands with polynomial error convergence rates, enabling tight, well-calibrated bands; a stochastic-volatility study on simulated data illustrates these gains.

3.1.1 Computing plug-in point estimators is computationally challenging

Two standard point estimators are the mean and the mode (MAP) of the conditional distribution of the missing sequence. Simulation-based methods converge slowly—typically at rate $O(N_{eval}^{-1/2})$—so accurate computation is expensive, and they also fail to exploit occasional observations of x_t, leading to statistically inefficient estimation.

The two most common candidates for point estimators of the missing sequence are the mode and the mean of the conditional distribution, as defined in Equation (7). The mode,

$$\hat{x}_{0:T} = \text{argmax}_{x_{0:T} \in \mathcal{S}_x^{T+1}} P_\theta^{X,Y}(x_{0:T}|y_{0:T}) \tag{11}$$

$$= \text{argmax}_{x_{0:T} \in \mathcal{S}_x^{T+1}} \frac{P_\theta^{X,Y}(\{x_t, y_t\}_{t=0}^T)}{P_\theta(y_{0:T})} \tag{12}$$

$$= \text{argmax}_{x_{0:T} \in \mathcal{S}_x^{T+1}} \prod_{t \in \bar{\mathcal{T}}} P_\theta\left(x_t, y_t | x_{t-1}, y_{t-1}\right), \tag{13}$$

is the most likely sequence given the data: Analogous to the maximum-likelihood estimator, it has the highest density among all possible sequences, and is therefore standard in

the literature. Computationally, the mode is the solution to a single $(T+1)$-dimensional maximization problem, where the objective function requires only computing the joint density of all variables, because the denominator is a scaling factor; hence, no integration is needed.[10] [11]

Alternatively, the mean of the conditional density,

$$\bar{x}_{0:T}(\theta, y_{0:T}) \equiv \mathbb{E}_\theta[x_{0:T} \mid y_{0:T}], \tag{14}$$

is often considered, particularly in Bayesian estimation. For each $t \in \bar{\mathcal{T}}$, we have a separate integral

$$\bar{x}_t(\theta, y_{0:T}) \equiv \int_{\mathcal{S}_x^{T+1}} \tilde{x}_t \, P_\theta(\tilde{x}_{0:T} \mid y_{0:T}) \, d\tilde{x}_{0:T} \tag{15}$$

which integrates x_t against the smoothing density over the full latent path. As for the likelihood, the integral is usually not available in closed form and the integrand cannot be decomposed easily to break up the integral: hence, numerical approximation of the integral is both necessary and computationally challenging. Additionally, computing the mean sequence requires approximating $(T+1)$ separate integrals, one for each time period.

For the actual computation, particle methods can be used but they have two problems: First, their asymptotic error rate is the usual Monte Carlo error rate, $O_p(N_{eval}^{-1/2})$, so a faster algorithm is desirable.[12] Moreover, just as for approximating the likelihood given occasional observations of x_t, we are not aware of particle methods that natively incorporate occasional observations into their sampling scheme (see Gilch et al. (2025) for a review of this literature). Hence, estimation of the mean is also statistically inefficient.

[10] The numerator of the smoothing density is the joint distribution of both series $\{x_t\}_{t=0}^T$ and $\{y_t\}_{t=0}^T$, i.e., the product of the transition densities $P_\theta(y_t, x_t \mid y_{t-1}, x_{t-1})$. The denominator is the density of the data, i.e., the likelihood of θ given the data. Therefore, if we want to evaluate the value of the conditional density of the missing sequence (and not just find its mode), we need to compute the likelihood, e.g., using RLI. However, this must be done only once for given θ and data; once it is available, evaluating the density at any candidate sequence is cheap.

[11] In this paper, we do not make a formal statement about the complexity of the former, but we argue that numerical solvers typically find the maximum fairly quickly. In particular when the relevant densities are differentiable, which is usually the case because economic models are often defined using smooth functions.

[12] The particles produced by particle filters cannot be used to compute the mean integrals (15) for $t \in \bar{\mathcal{T}}$ because they are generated by conditioning only on the previous observations $y_{0:t}$, thereby ignoring the information in $y_{t+1:T}$ when estimating x_t. Therefore, particle smoothing reweights the filtered particles to approximate the distribution conditional on all observations $y_{0:T}$. The integral (15) can then be approximated by the (weighted) average of the smoothed particles, simultaneously for each $t \in \bar{\mathcal{T}}$; i.e., a single filtering-smoothing run yields all the required expectations. This reweighting incurs additional function evaluations per particle and thus increases total computational effort; the asymptotic probabilistic error rate is still $O_p(N_{eval}^{-1/2})$ though.

3.1.2 Computing plug-in prediction bands is computationally challenging

Conditional prediction bands given the data and a fixed parameter are not unique, so we focus on the one-scale class, where a single scale c determines the band and the exact plug-in set is uniquely identified by a monotone coverage equation. Computing that coverage entails a high-dimensional integral—akin to those for the likelihood and the mean—so simulation yields only $O(N_{eval}^{-1/2})$ accuracy and handles occasional observations poorly, while ad hoc Wald/Bonferroni bands avoid integration at the cost of misspecification and conservatism. This motivates a fast, high-accuracy algorithm for evaluating coverage and recovering the exact plug-in band.

In Section 2.4.1, we defined the plug-in prediction set as a random set based on a fixed parameter θ and the data $y_{0:T}$ such that it has coverage $1-\alpha$ w.r.t. the conditional distribution of the missing sequence. However, this definition is too general to deliver a unique prediction set for a given α. Therefore, we restrict the space of admissible sets to well-defined classes in which the exact plug-in set is unique (under mild monotonicity and continuity assumptions on the models transition densities). The literature has focused on two such classes, each admitting a unique set at level α: the one-scale class of prediction bands (Montiel Olea and Plagborg-Møller, 2018) and highest density regions (HDRs).Both can be implemented by imposing additional constraints on the plug-in set $\hat{X}$. In this paper, we focus on the former as it delivers easier-to-visualize prediction bands.

In general, prediction bands B_θ are hyperrectangles in the space of missing sequences $\mathcal{S}_x^{(T+1)}$, meaning they can be written as a tensor product of bounded intervals

$$B_\theta\left(\theta, y_{0:T}\right) \equiv \bigtimes_{t \notin \bar{\mathcal{T}}} \left[\underline{\tilde{x}}_t\left(\theta, y_{0:T}\right), \bar{\tilde{x}}_t\left(\theta, y_{0:T}\right)\right]. \tag{16}$$

This allows for a one-to-one projection of such sets into the standard visualization as a one-dimensional band along the time axis. Hence, prediction bands don't exhibit a projection error, which is usually arising when projecting high-dimensional sets into such a visualization.[13] The one-scale class of prediction bands consists of prediction bands, hence exploiting the lack of a projection error, and additionally constrains the possible

[13]The projection error is the difference in probability coverage between an T-dimensional set $A \subset \mathbb{R}^T$ and its projection into a one-dimensional representation $\bar{A}$. By the latter we mean a series of (one-dimensional) intervals $[\underline{a}_t, \bar{a}_t]$ s.t. $\underline{a}_t = \min_{a \in A} a_t$ and $\bar{a}_t = \max_{a \in A} a_t$. In particular in the time series context, formally high-dimensional sets such as A are usually represented through such projections, i.e., through a "band" with upper bound $\bar{a}_t$ and lower bound $\underline{a}_t$ for each t. However, this representation, as T-dimensional set, is larger than A and therefore incures the projection error $e(A, P) = P(\bar{A}) - P(A) \geq 0$ for any probability measure on $\mathbb{R}^T$. By considering only prediction bands as admissible prediction sets, this projection error becomes 0 for all candidate prediction sets and, hence, the visualization of the prediction set coincides with true set and actually achieves correct coverage.

intervals to scale jointly, making it is easy to find the unique exact set in this class:

$$B_\theta(c) = B_\theta(c, \theta, y_{0:T}) \equiv \bigtimes_{t \notin \bar{\mathcal{T}}} [\mathring{x}_t - c\,\sigma_t,\ \mathring{x}_t + c\,\sigma_t]\,, \tag{17}$$

where $\mathring{x}$ is one of the point estimators defined earlier, $\sigma_t = \sqrt{\mathrm{Var}_\theta[x_t \mid y_{0:T}]}$ is the conditional standard deviation of x_t given the data, and c is the common scale applied simultaneously to all $t \notin \bar{\mathcal{T}}$. In the following, we drop the dependence on θ and $y_{0:T}$ to simplify notation. Montiel Olea and Plagborg-Møller (2018) define this class for conducting simultaneous inference on multiple parameters and impulse response functions of VARs—a computationally less demanding, though less general, setting—and argue that this class in fact includes many commonly used prediction bands in empirical work.

Importantly, under suitable assumptions on $P_\theta^{X,Y}$,[14] the coverage of bands in the one-scale class of prediction bands is strictly increasing and continuous in c. Hence there is a unique $c^*(\alpha)$ such that $B_\theta(c^*(\alpha))$ attains exact coverage $1-\alpha$ with respect to the conditional distribution of the missing sequence.[15]

In particular, the prediction band with exact coverage $1-\alpha$ is the unique solution to the root-finding problem

$$R(c) \equiv P_\theta^{x_{0:T}|y_{0:T}}(x_{0:T} \in B_\theta(c) \mid y_{0:T}) - (1-\alpha) = 0, \tag{19}$$

with solution $c^*(\alpha)$. As the coverage function R is strictly increasing and continuous,[16] any numerical solver can find $c^*(\alpha)$ fairly quickly if we can evaluate the coverage of $B_\theta(c)$. In practice, we bracket c on $[0, c_{\max}]$ so that coverage at $c_{\max}$ definitely exceeds $1-\alpha$, then use bisection or Newton.

However, similar to the likelihood and the mean, the coverage poses a computational

[14] This includes: no plateaus, no point masses.

[15] Alternatively, the highest density region (HDR) minimizes the size of the prediction set in the missing-sequence space but can suffer from a large projection error. It is defined by a lower bound on the smoothing density,

$$hdr(\tau) \ = \ \left\{ x_{0:T} \in \mathcal{S}_x^{(T+1)} \ : \ P_\theta^{x_{0:T}|y_{0:T}}(x_{0:T} \mid y_{0:T}) \geq \tau \right\}, \tag{18}$$

and therefore includes only those sequences with the highest probability given the observed data. As with the one-scale class, under suitable assumptions on P_θ, the coverage of $hdr(\tau)$ varies monotonically and continuously with τ (specifically, it is strictly decreasing in τ). Hence there exists a unique $\tau^*(\alpha)$ such that $hdr\big(\tau^*(\alpha)\big)$ attains exact coverage $1-\alpha$ with respect to the conditional distribution of the missing sequence. However, even though the HDR is minimal in the sequence space $\mathcal{S}_x^{(T+1)}$ for a given coverage level $1-\alpha$, its projection onto a one-dimensional prediction band no longer has exact coverage and is typically conservative. Because the projected band is what is usually reported in practice, we focus on the one-scale class of prediction bands in the main text and relegate use cases and computation of HDRs to the appendix.

[16] Under suitable assumptions on P_θ; e.g., no plateaus on $\partial B_\theta(c)$ and $\sigma_t > 0$ for all $t \notin \bar{\mathcal{T}}$. See also Footnote 14.

challenge, because it is a $(T+1)$-dimensional integral over the hyperrectangle,

$$P_\theta^{x_{0:T}|y_{0:T}}(x_{0:T} \in B_\theta(c) \mid y_{0:T}) = \int_{B_\theta(c)} P_\theta^{x_{0:T}|y_{0:T}}(\tilde{x}_{0:T} \mid y_{0:T})\, d\tilde{x}_{0:T}, \tag{20}$$

which generally does not decompose into one-dimensional integrals for the same reason as the likelihood in Section 2.3.

Existing approaches either compute the integral using smoothed particles or avoid it altogether, in which case they are not valid for simultaneous coverage. With particle smoothing, a plug-in prediction band can be implemented, but the approximation error decays at the same (slow) rate, $O_p\left(N_{eval}^{-1/2}\right)$, as for the plug-in mean, and the method is not designed to incorporate occasional observations, making the resulting bands looser than necessary. Other approaches do not compute the coverage at all but rely on Wald-type bands: these fall within our one-scale class, yet the scale c is chosen as the critical value of a Normal or t-distribution. Concretely, they set $\mathring{x}$ to the (posterior) mean and σ_t to the (approximate) marginal standard deviation, then take c as the univariate critical value—thereby ignoring the joint dependence that simultaneous coverage must account for. This works if the model is genuinely close to linear and Gaussian, but when the model is far from linear-Gaussian such ad hoc choices can be very inaccurate. Even when a linear-Gaussian approximation seems reasonable, this approach is overconfident because it yields pointwise prediction bands. Adjusting these to be simultaneous, e.g., via Bonferroni, tends to be overly conservative (Montiel Olea and Plagborg-Møller, 2018).

In the following section, we show how to evaluate (20) deterministically via a RLI-type algorithm, based on the concepts of Section 2.3, that also handles occasional observations.

3.1.3 Our recursive numerical integration algorithm enables fast and exact plug-in inference

We develop RLI-type algorithms—using the same alternating sequence of integration and interpolation—to compute point estimators and exact plug-in prediction bands. Our method is fast, substantially reducing the number of model evaluations needed for high accuracy, and it naturally incorporates occasional observations. In particular, it yields tight prediction bands in contrast to heuristic alternatives.

Recall that we use the mean of the conditional density as the point estimator for $x_{0:T}$ (Equation (14)) in this paper, and that a particle smoother can compute it, but only at the usual Monte Carlo rate, $O_p(N_{eval}^{-1/2})$. The same is true for the coverage of a prediction band. We propose an RLI-type algorithm to approximate the integrals in (15) and (20) faster and more accurately.

For a given period $t' \notin \bar{\mathcal{T}}$, the integral for $\mathbb{E}_\theta[x_{t'} \mid y_{0:T}]$ can be written recursively,

analogous to the likelihood integral, as

$$f_t^\theta(x_{t-1}) = \int_{\mathcal{S}_x} g_{t'}^\theta(\tilde{x}_t, x_{t-1})\, P_\theta(y_t, \tilde{x}_t \mid y_{t-1}, x_{t-1})\, f_{t+1}^\theta(\tilde{x}_t)\, d\tilde{x}_t, \tag{21}$$

for $t = 0, \ldots, T-1$. When $t \in \bar{\mathcal{T}}$, this integral collapses to evaluation at the observed state x_t. Starting this recursion at T with

$$f_T^\theta(x_{t-1}) = \int_{\mathcal{S}_x} g_{t'}^\theta(\tilde{x}_T, x_{T-1})\, P_\theta(y_T, \tilde{x}_T \mid y_{T-1}, x_{T-1})\, d\tilde{x}_T, \tag{22}$$

and iterating backwards, we obtain

$$\mathbb{E}_\theta[x_{t'} \mid y_{0:T}] = \begin{cases} f_1^\theta(x_0), & \text{if } 0 \in \bar{\mathcal{T}}, \\ f_0^\theta, & \text{else.} \end{cases} \tag{23}$$

The additional integrand relative to the likelihood recursion is

$$g_{t'}^\theta(\tilde{x}_t, x_{t-1}) = \begin{cases} \tilde{x}_t, & \text{if } t = t', \\ 1, & \text{else.} \end{cases} \tag{24}$$

Thus, using the same quadrature scheme as for RLI, but inserting the corresponding $g_{t'}^\theta$ for each $t' \in \bar{\mathcal{T}}$, yields an approximation of the mean path. In particular, smoothness is preserved and we retain a polynomial convergence rate:[17]

$$\max_{t=0,\ldots,T} \left| \bar{x}_t - \bar{x}_t^{N_{eval}} \right| \;=\; O\big(N_{eval}^{-r}\big)\,. \tag{25}$$

Note that this recursion also delivers other moments of the smoothing distribution, e.g., via $g_{t'}^\theta(\tilde{x}_t, x_{t-1}) = \tilde{x}_t^2$ if $t = t'$ and 1 otherwise for the second moment (and thus the variance), and $g_{t'}^\theta(\tilde{x}_t, x_{t-1}) = \tilde{x}_t\, x_{t-1}$ if $t = t'$ and 1 otherwise for the autocovariance of the

[17] Recall three aspects of how we report asymptotic error rates. First, we use N_{eval}—the number of function evaluations per recursion step—as the accuracy knob that we increase for higher accuracy. For our RLI-type algorithm, N_{eval} includes the evaluations needed for numerical integration (N_Q) and interpolation ($N_{\mathcal{I}}$). In particle filtering/smoothing, it includes the evaluations per forward-filtering/backward-smoothing step; therefore, it is analogous to the total number of particles, and the error rate is also stated in terms of this number.

In contrast, the total computational effort, N_{total}, is the number of function evaluations required to compute the full object (e.g., the likelihood, the mean sequence, or the coverage of a prediction band). It scales linearly with N_{eval}, although the multiplicative factor may itself depend nonlinearly on T. Finally, we treat T as fixed in this paper, reflecting the empirical setting in which a researcher has a given dataset of length T and must perform inference for that particular sample. Accordingly, we omit all T-dependence in the reported convergence rates. In particular, for rates it does not matter whether we write N_{eval} or N_{total}, since the constant factor relating them is absorbed by Landau-O notation.

For reference, one can choose a rule for increasing $N_{eval} = N_{eval}(T)$ with T such that the approximation rate also holds asymptotically, i.e., for $T \to \infty$. Such rates are discussed in Gilch et al. (2025) and are usually less than linear speed for RLI-type algorithms.

two consecutive states $x_{t'-1}$ and $x_{t'}$. This becomes relevant when constructing confidence bands based on the variance of $x_{t'}$ for each $t' \in \bar{\mathcal{T}}$. However, while any function depending only on two consecutive states can be handled in this way, further adjustments are needed for more general moments and functionals, which we leave for future research.

To compute the coverage integral (20) for plug-in prediction bands, we also use an RLI-type recursive representation of the integral, this time adjusting the integration bounds (rather than the integrand) to

$$f_t^\theta(x_{t-1}) = \int_{\underline{a}_t}^{\bar{a}_t} P_\theta(y_t, \tilde{x}_t \mid y_{t-1}, x_{t-1})\, f_{t+1}^\theta(\tilde{x}_t)\, d\tilde{x}_t. \tag{26}$$

Recall that we defined the bounds of the one-scale class in Section 3.1.2 as $\underline{a}_t = \mathring{x}_t - c\,\sigma_t$ and $\bar{a}_t = \mathring{x}_t + c\,\sigma_t$.

As before, the integral is obtained at the final recursion step,

$$\hat{P}_\theta^{x_{0:T}|y_{0:T},N_{eval}}\big(x_{0:T} \in B_\theta(c) \mid y_{0:T}\big) = \begin{cases} f_1^\theta(x_0), & \text{if } x_0 \in \bar{\mathcal{T}}, \\ f_0^\theta, & \text{else.} \end{cases} \tag{27}$$

Based on this recursion, we again use the same combination of numerical integration and interpolation to approximate the full integral. This approximation retains the same convergence rates as the RLI approximation of the likelihood, i.e.,

$$\left| P_\theta^{x_{0:T}|y_{0:T}}\big(x_{0:T} \in B_\theta(c) \mid y_{0:T}\big) - \hat{P}_\theta^{x_{0:T}|y_{0:T},N_{eval}}\big(x_{0:T} \in B_\theta(c) \mid y_{0:T}\big) \right| = O\big(N_{eval}^{-r}\big), \tag{28}$$

for regularity level r. Note that the prediction-band integral is easier to compute than the likelihood integral because no change of variables is needed to obtain good approximations even for a small number of integration nodes. [18]

Given this algorithm to compute the coverage of plug-in prediction bands, determining the correct band in the one-scale class for a given prediction level α reduces to a single root-finding problem over the scale c of B_θ, where the solution c^* satisfies

$$P_\theta^{x_{0:T}|y_{0:T}}\big(x_{0:T} \in B_\theta(c^*(\alpha)) \mid y_{0:T}\big) - (1-\alpha) \;=\; 0. \tag{29}$$

Solving this is straightforward because the coverage is monotonically increasing and continuous in c (provided $P_\theta^{x_{0:T}|y_{0:T}}$ has no point masses), so there is a unique root that standard methods (e.g., bisection or Newton) can find quickly.

[18] Recall that, for the likelihood approximation, a suitable change of variables helps achieve accuracy with few quadrature nodes by correctly centering the integrand. In contrast, for the coverage of the plug-in prediction band, each recursion step f_t integrates over a bounded interval regardless of the model's probability distributions. Thus we can simply apply a linear variable transformation to the domain $[-1, 1]$ and then use a Gauss-Legendre rule on the transformed integral. Unless the joint distribution is highly skewed, Gauss-Legendre already provides a good approximation for few integration nodes.

Moreover, if $P_\theta^{x_{0:T}|y_{0:T}}$ is differentiable, replacing the coverage integral by its RLI approximation is unproblematic: the root $\hat{c}^*$ of the approximate equation converges to the true root c^* at the same rate as the coverage approximation in (28). In particular,

$$\hat{c}^* \;\to\; c^* \quad \text{at rate} \quad O\big(N_{eval}^{-r}\big)\,, \tag{30}$$

so the combination of a fast-converging coverage approximation and a simple root search yields plug-in bands with exact coverage $1-\alpha$ faster than particle smoothing methods.

To summarize, under sufficient regularity of the joint conditional distribution $P_\theta^{x_{0:T}|y_{0:T}}$, we can provide fast-converging approximation algorithms for both the mean and the exact plug-in prediction band, whose convergence rates significantly improve over existing methods based on simulation.

3.1.4 We demonstrate our algorithm for a stochastic volatility model

We demonstrate the numerical performance of our RLI-type approximation for both the mean and the plug-in prediction band for the volatility series x_t in the stochastic-volatility model. We show that the relative approximation error of the mean integral decays much faster under the RLI-type algorithm than under the FFBS particle smoother. The same holds true for the plug-in prediction band, despite its computation involving a root-finding problem.

Stochastic volatility models are a central tool in financial econometrics and serve as a natural test case for assessing estimation methods. We benchmark the numerical performance of our RLI-based approach against standard simulation-based methods and confirm the expected computational speed-ups predicted by our theoretical results.

We consider the simplest version of a stochastic volatility model with one observed variable (e.g., an asset return series) and one unobserved variable, the time-varying and serially correlated volatility x_t. Its state-measurement equations are

$$y_t = \mu + e^{x_t/2}\eta_t, \tag{31}$$

$$x_t = \rho\, x_{t-1} + \varepsilon_t, \tag{32}$$

with $\eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_y^2)$ and $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\sigma_x^2)$. With parameter vector $\theta = (\mu,\rho,\sigma_y,\sigma_x)$, this stochastic volatility model is a straightforward instance of the state-space setup in Equations (1)-(2), with fully tractable $P_\theta^\eta, P_\theta^\varepsilon, \Psi_\theta$, and Φ_θ. The transition densities are available in closed form. A key feature of the model is time-varying volatility, which captures periods of high vs. low uncertainty in the data. It does so continuously, with ρ governing the persistence of high- and low-volatility regimes, thereby avoiding the more restrictive regime-switching approach with a discrete number of clearly separated regimes.

Obviously, this is a bare-bones specification, but countless richer versions exist, in-

cluding additional explanatory variables, (seasonal) trends, and so forth. However, the main complication is already present here: the latent state enters the measurement equation highly nonlinearly, so linearization-based approximations can quickly deviate from the true dynamics. Consequently, there is a large literature on estimating stochastic volatility models, with a strong focus on particle methods.

Recall that the primary objective of our algorithm is to reduce the number of function evaluations needed to compute the mean or the prediction band. In elaborate structural economic models, the functions Ψ_θ and Φ_θ can both be expensive (e.g., if they require solving an optimization problem or evaluating an expectation). Since, in macro models, the transition equation for the latent state x_t, i.e., Φ_θ, usually contains these complex components, we take the number of evaluations of Φ_θ as our measure of computational effort, N_{eval}.[19] This is consistent with our convergence-rate derivations for the RLI-type algorithm and the particle smoother: in the former, Ψ_θ and Φ_θ (or the corresponding densities) are always needed synchronously; in the latter, Φ_θ is required in both filtering and smoothing, implying the rate $O_p(N_{eval}^{-1/2})$ but with different constants.[20] In this section, we use a Stochastic Volatility model with inexpensive densities, and report cost in function evaluations so the metric carries over to more complex settings.

We use the relative approximation error to display how well each algorithm approximates the true solution for a given total number of function evaluations, N_{total}. For a functional $I(f)$, e.g., the mean or the coverage integral or the root c^* scaling the exact plug-in prediction band, and an approximation $\hat{I}_{N_{total}}(f)$ using N_{total} evaluations of f, define

$$e(I(f), N_{total}) \equiv \left| \frac{I(f) - \hat{I}_{N_{total}}(f)}{I(f)} \right|. \tag{33}$$

The relative error can be read as "digits of accuracy where it matters," rather than simply counting decimals. In our setting, the true value $I(f)$ is usually not available in closed form.Therefore, in our implementation we use a high-N_{total} run of the RLI-type algorithm to obtain a near-exact benchmark for $I(f)$ and substitute this benchmark into the formula above.

We report the approximation error against N_{total} to visualize the actual computational effort needed to achieve a target accuracy. This is necessary because the N_{total} is proportional to N_{eval} for both the RLI-type algorithm and Particle smoothing, but with different proportionality constants. Hence, the actual N_{total} needed to reach the target accuracy does not only depend on N_{eval} but also on this linear constant. However, we

[19] We treat additional overhead from setting up the algorithm as a fixed cost that becomes negligible both for large models and with efficient implementations. If Ψ_θ is the dominant cost in a given application, we define N_{eval} analogously as the number of evaluations of Ψ_θ.

[20] If Ψ_θ is the dominant cost in a given application, then smoothing is less costly because no new evaluations of Ψ_θ are needed in the backwards smoothing step.

can still compare our empirical rates with the theoretical ones, because these different constants effectively only imply an earlier or later start of the asymptotic error whereas the slope is the same.

Figure 1 reports relative approximation errors for the plug-in mean path computed by our RLI-type algorithm and by a basic FFBS particle smoother. We focus on plug-in means (true θ, no parameter uncertainty) and consider two horizons, $T \in \{10, 100\}$. The former is representative of settings with frequent occasional observations (e.g., short gaps), while the latter matches macro/financial series at quarterly or annual frequency. We simulate the full data—no real data in this subsection—, initializing it with draws from the stationary distribution, and use $S = 10$ Monte Carlo replications to smooth the reported rates (deterministic for RLI-type, but helpful for the probabilistic FFBS errors).

RLI-type uses Gaussian-Hermite quadrature for the integration steps and cubic splines for interpolation. In this smooth model, Gaussian-Hermite can achieve near-exponential convergence, while cubic splines achieve their nominal rate $r = 4$. Therefore, we keep N_Q practically fixed and increase $N_{\mathcal{I}}$ to balance the two errors and approach the optimal full rate (theoretically close to $r = 4$).[21] As a baseline, we implement a standard FFBS particle smoother following Chopin and Papaspiliopoulos (2020).

We compute approximations of each integral with increasing N_{eval} (and, thus, N_{total}), i.e., we use an increasing number of quadrature and interpolation nodes for the RLI-type method and an increasing number of particles for FFBS. To obtain the full mean sequence, RLI-type must evaluate all T integrals in (15) separately; hence we sum function evaluations over all computations. In contrast, the FFBS algorithm requires only one run to generate particles that can be used to compute the entire sequence.

We report convergence in two ways. First, for each approximation run with N_{total} function evaluations, we report the mean error, averaging over t and s:

$$\bar{e}(I(f), N_{total}) \;=\; \frac{1}{ST} \sum_{t=0}^{T} \sum_{s=1}^{S} e(I_{st}(f), N_{total}) \,, \tag{34}$$

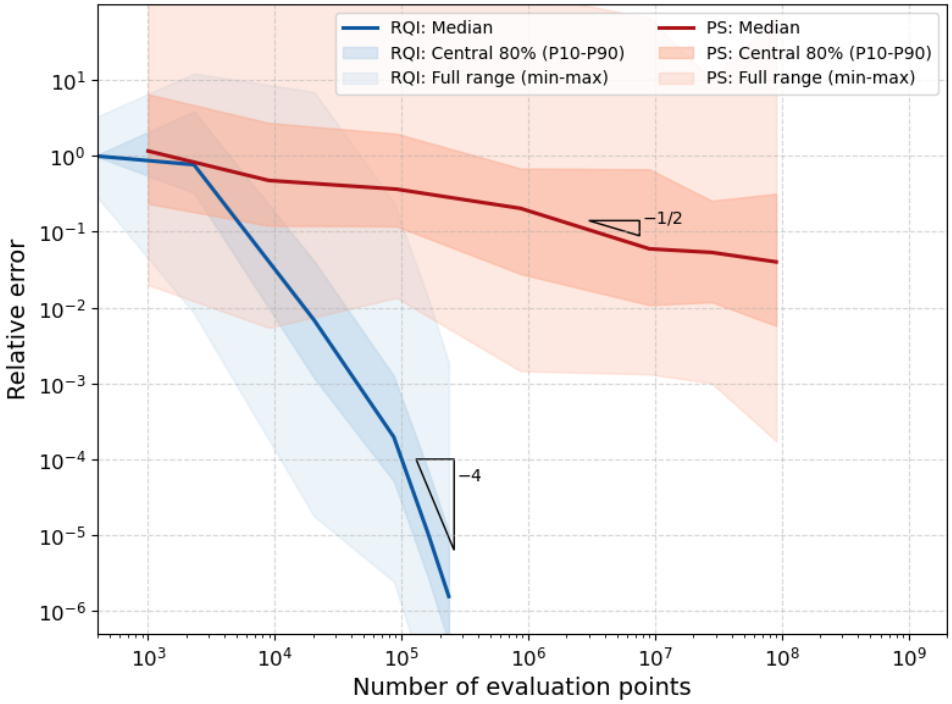
where $I_{st}(f)$ is the mean integral for the s-th simulated data set and period t, and f is the function whose evaluations we count—in this case, Φ_θ.[22] Second, we report the variation in approximation errors by displaying the full range (min-max across t, s) of errors at each N_{total}.

Errors are plotted on log-log axes, as is standard to display convergence: increasing N_{total} by one order of magnitude should add r digits of accuracy. From the theory above, we expect slopes near $r \approx 4$ for RLI-type (in this smooth Stochastic Volatility setting) and $r = 1/2$ for FFBS. For readability, we overlay reference lines with the approximate

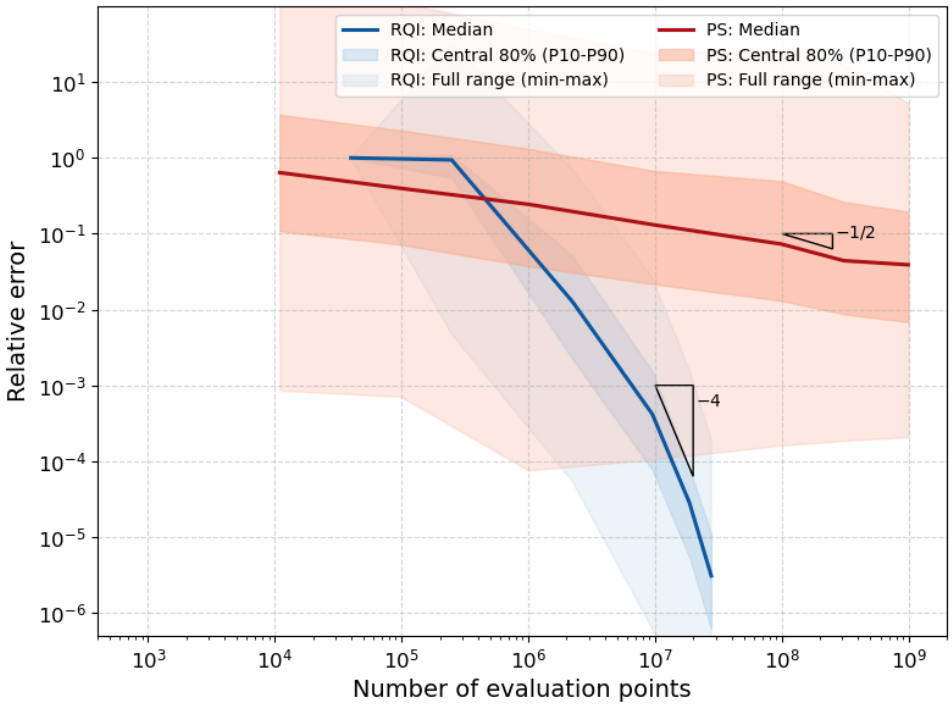
[21] See Reich (2018) for more details on balancing integration and interpolation.

[22] Recall that it is not the integral over Φ_θ, but that Φ_θ is needed to compute the mean, and only its evaluations matter for computational effort.

Figure 1: Numerical performance of RLI-based mean approximation vs. particle smoother



(a) Horizon $T = 10$



(b) Horizon $T = 100$

Notes: We compute the mean sequence for the time-varying volatility $\{x_t\}_{t=0}^{T}$ for the Stochastic Volatility model (Eqs. (31)-(32)) and $S = 10$ simulated data sets of length $T = 10$ and $T = 100$ respectively. We use our RLI-type algorithm as well as a standard FFBS bootstrap Particle smoother, each with increasing total numbers of function evaluations N_{total}. We report the mean as well as the range of the relative approximation error as defined in Eq. (33), taken over all simulated data sets and all periods of the mean sequence. For reference, we provide triangles with slope $r = -4$ and $r = -1/2$ indicating the theoretical convergence rate for RQI (with Gaussian quadrature and cubic interpolation) and particle smoothing respectively.

empirical slopes.

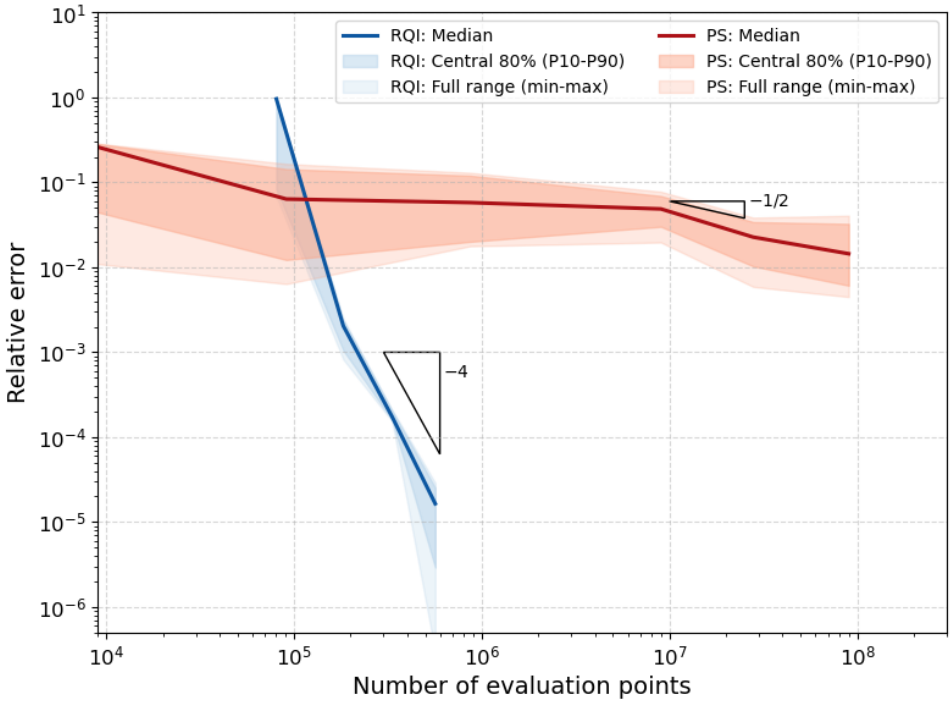
As expected, RLI-type converges much faster to the true mean, with an empirical rate close to $r \approx 4$. It achieves the same approximation errors as particles with far fewer function evaluations; given more evaluations, it reaches accuracy levels that are not realistically attainable with particle methods. There is a short "burn-in" region reflecting pre-asymptotic error typical of deterministic quadrature (**numerics-literature**), but beyond that the log-log slope stabilizes near the theoretical benchmark. RLI-type is also reliable in the worst case: the maximal error (over simulations s and periods t) drops quickly, whereas particle smoothing can exhibit large errors for some periods even with many particles.

To conclude, for uniformly good approximation (e.g., at least two digits of relative accuracy), RLI-type should be preferred: it is faster for a given accuracy and attains higher accuracy for a given computational budget. If only a very rough approximation is needed, the FFBS particle smoother may be the cheaper alternative. Practically, this difference stems from work reuse: particle smoothing can leverage a single run for many summaries, whereas RLI-type evaluates integrals separately (e.g., T expectations for the mean and multiple coverage integrals across candidate c for bands). This structural overhead makes RLI-type less attractive in the low-accuracy regime. That said, our current implementation is not fully optimized: shared recursion segments (e.g., the first $T - t$ steps when computing means for all t) can be cached and reused, which would cut evaluations substantially. We view such reuse as a straightforward avenue for further speedups.

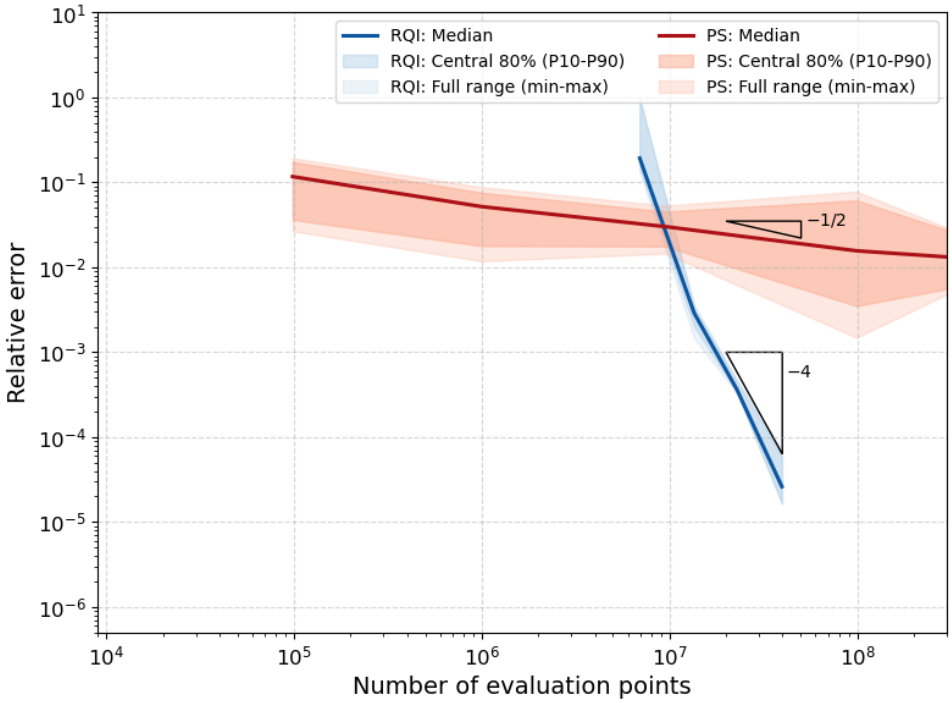
Next we show the numerical performance of our RLI-type algorithm for plug-in prediction bands: We compute plug-in prediction bands for simulated data at coverage level $1 - \alpha = 0.95$. As before, we consider plug-in inference (true θ, no parameter uncertainty), two horizons $T \in \{10, 100\}$, and S Monte Carlo replications with independently simulated datasets.

We again use Gauss-Hermite quadrature and cubic splines for the RLI-type method, and a basic FFBS particle smoother to obtain smoothed particles. For plug-in prediction bands, recall that they are obtained as the solution of the root-finding problem (19). This means that, for each simulated dataset, we first compute the mean sequence and the per-period variances (the fixed ingredients of the one-scale class). Second, a bisection algorithm iterates over the scale c, computing the coverage of the candidate band $B_\theta(c)$—i.e., approximating a coverage integral for each candidate c via the recursion in (26)—until it finds the root c^* at which coverage equals $1 - \alpha$. Since our RLI-type method computes each of these integrals separately, the computational effort for the exact plug-in band on a given draw equals the sum of function evaluations across all these approximations. By contrast, for FFBS we reuse the smoothed particles from a single run to approximate both the mean/variance ingredients and the coverage for different c (by computing the

Figure 2: Numerical performance of RLI-based prediction band approximation vs. Particle Smoother: Scales



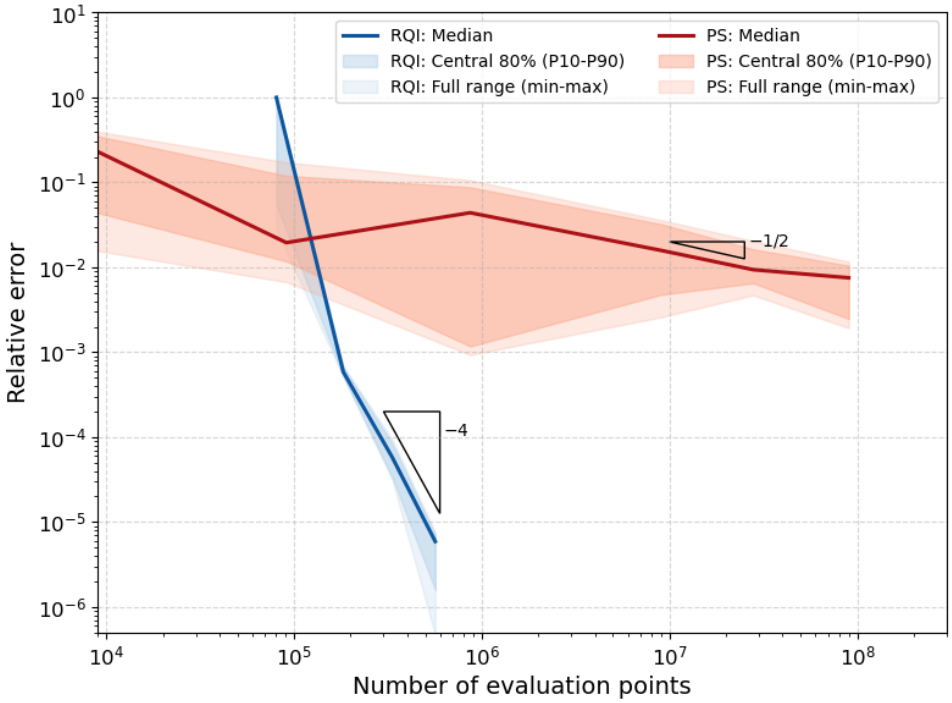
(a) Horizon $T = 10$



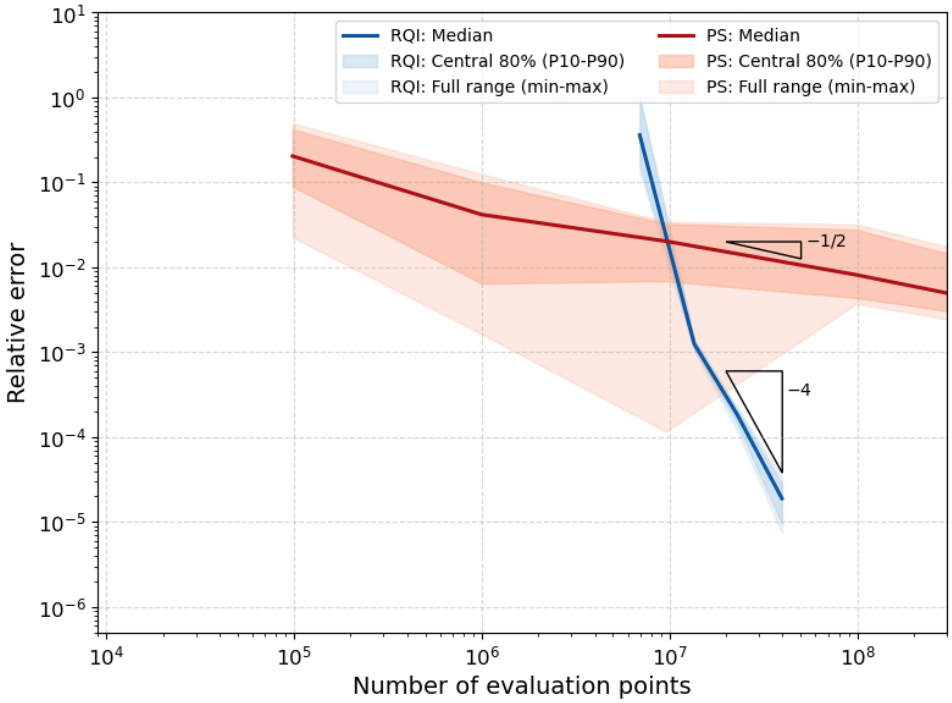
(b) Horizon $T = 100$

Notes: We compute the scale for the exact plug-in prediction band in the one-scale class of prediction bands as defined in Eq. (29) for the time-varying volatility $\{x_t\}_{t=0}^{T}$ for the Stochastic Volatility model (Eqs. (31)-(32)) and $S = 10$ simulated data sets of length $T = 10$ and $T = 100$ respectively. We use our RLI-type algorithm as well as a standard FFBS bootstrap Particle smoother, each with increasing total numbers of function evaluations N_{total}. We report the mean relative approximation error as defined in Eq. (33), taken over all simulated data sets. For reference, we provide triangles with slope $r = -4$ and $r = -1/2$ indicating the theoretical convergence rate for RQI (with Gaussian quadrature and cubic interpolation) and particle smoothing respectively.

Figure 3: Numerical performance of RLI-based prediction band approximation vs. Particle Smoother: Coverage



(a) Horizon $T = 10$



(b) Horizon $T = 100$

Notes: We compute the coverage of tthe exact plug-in prediction band in the one-scale class of prediction bands as defined in Eq. (29) for the time-varying volatility $\{x_t\}_{t=0}^{T}$ for the Stochastic Volatility model (Eqs. (31)-(32)) and $S = 10$ simulated data sets of length $T = 10$ and $T = 100$ respectively. We use our RLI-type algorithm as well as a standard FFBS bootstrap Particle smoother, each with increasing total numbers of function evaluations N_{total}. We report the mean relative approximation error as defined in Eq. (33), taken over all simulated data sets. For reference, we provide triangles with slope $r = -4$ and $r = -1/2$ indicating the theoretical convergence rate for RQI (with Gaussian quadrature and cubic interpolation) and particle smoothing respectively.27

weighted share of smoothed particles inside the band for a candidate c).

We report relative approximation errors for two targets—scale error and coverage error. For the exact one-scale band, we identify the solution by its scale $c^*(y_{0:T})$. For each dataset $y^s_{0:T}$ we compute a high-accuracy RLI reference c^*_s, and then report the mean relative error

$$\bar{e}\big(c^*, N_{eval}\big) \;\equiv\; \frac{1}{S}\sum_{s=1}^{S} e(c^*_s, N_{eval})\,, \tag{35}$$

suppressing the dependence on α in the notation. Additionally, we report the relative error in the achieved coverage of the bands delivered by each method. Specifically, using the RLI-type method at high accuracy, we evaluate the coverages $P_\theta^{x_{0:T}|y_{0:T}}\big(B(c^*_s) \mid y^s_{0:T}\big)$ of the approximated plug-in bands for both methods and compare them to the target $1-\alpha = 0.95$. We then average over datasets:

$$\bar{e}\big(P_\theta^{x_{0:T}|y_{0:T}}(B(c^*) \mid y^{1:S}_{0:T}), N_{eval}\big) \;\equiv\; \frac{1}{S}\sum_{s=1}^{S} e\Big(P_\theta^{x_{0:T}|y_{0:T}}\big(B(c^*_s) \mid y^s_{0:T}\big), N_{eval}\Big)\,. \tag{36}$$

This second measure tests whether errors from approximating the mean, variances, and c^* might cancel—especially for particle smoothing—so that the resulting band differs from the true exact one-scale band yet still attains the correct coverage.

As with the point estimator, we find that our RLI-based method is much faster, even though it requires recomputation at each iteration step of the optimizer, whereas the particle smoother reuses a single set of particles. Under suitable regularity conditions on P_θ, this advantage is largely model-agnostic: model complexity is handled in the integration, while the optimization concerns only the strictly increasing, continuously differentiable coverage function. Consistent with this regularity, the errors for both metrics—convergence in c^* and in achieved coverage converge at virtually the same rate.

3.2 Implementing parameter uncertainty with a constrained optimization problem

We address predictive inference under parameter uncertainty by taking the prediction set to be the union of plug-in bands across a confidence set for θ, and we compute this union via a series of constrained optimization problems. In simulation, the plug-in band is overconfident, whereas the prediction-band union attains the target coverage, albeit somewhat conservatively.

3.2.1 The prediction band union

We define the prediction band union and establish a coverage lower bound, which we use to calibrate the union to a target coverage level under parameter uncertainty. We then derive a series of small, smooth constrained optimization problems that standard solvers can handle, and use them to compute the projected prediction band union.

Predictive inference for the full latent path $x_{0:T}$, i.e., acknowledging parameter uncertainty, as discussed in Section 2.4, is difficult for two reasons: First, for nonlinear state-space models, the joint law of $(x_{0:T}, \theta)$ is usually not available in closed form, so in a frequentist setting without assumed prior over θ, one cannot integrate out θ to construct the exact prediction set under parameter uncertainty or to evaluate its coverage analytically. Second, the property (9) imposes only one scalar coverage constraint on a high-dimensional set, so the target prediction set (as in the plug-in case) is not unique.

It is therefore natural to start from the plug-in construction and then augment it to account for parameter uncertainty. In this paper, we construct a prediction set under parameter uncertainty by taking the union of plug-in prediction bands across a confidence set for θ. For a target level α, our building blocks are plug-in bands with conditional coverage $1-\tilde{\alpha}$[23]. Define

$$\hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T}) \equiv \bigcup_{\theta\in\hat{\Theta}_\gamma(y_{0:T})} \hat{X}_{\tilde{\alpha}}(\theta, y_{0:T}), \tag{37}$$

where $\hat{\Theta}_\gamma(y_{0:T})$ is a $(1-\gamma)$-level confidence set for θ.

The prediction-band union admits a simple lower bound in the predictive inference problem: if the plug-in bands attain conditional level $1-\tilde{\alpha}$ for every fixed θ and the confidence set $\hat{\Theta}_\gamma(y_{0:T})$ has (finite-sample) level $1-\gamma$, then for any $\theta \in \Theta$,

$$\begin{aligned}
&P_\theta^{X,Y}\left(x_{0:T} \in \hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T})\right) && (38)\\
&\quad = P_\theta^{X,Y}\left(x_{0:T} \in \hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T}), \theta \in \hat{\Theta}_\gamma(y_{0:T})\right) && (39)\\
&\qquad + P_\theta^{X,Y}\left(x_{0:T} \in \hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T}), \theta \notin \hat{\Theta}_\gamma(y_{0:T})\right) && (40)\\
&\quad \geq P_\theta^{X,Y}\left(x_{0:T} \in \hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T}), \theta \in \hat{\Theta}_\gamma(y_{0:T})\right) && (41)\\
&\quad \geq P_\theta^{X,Y}\left(x_{0:T} \in \hat{X}_{\tilde{\alpha}}(\theta, y_{0:T}), \theta \in \hat{\Theta}_\gamma(y_{0:T})\right) && (42)\\
&\quad = \mathbb{E}_\theta^{y_{0:T}}\left[\mathbf{1}_{\{\theta\in\hat{\Theta}_\gamma(y_{0:T})\}} P_\theta^{x_{0:T}|y_{0:T}}\left(x_{0:T} \in \hat{X}_{\tilde{\alpha}}(\theta, y_{0:T}) \,|y_{0:T}\right)\right] && (43)\\
&\quad = (1-\tilde{\alpha}) P_\theta^{y_{0:T}}\left(\theta \in \hat{\Theta}_\gamma(y_{0:T})\right) && (44)\\
&\quad = (1-\tilde{\alpha})(1-\gamma) && (45)
\end{aligned}$$

[23] Since the target of the final prediction set is α, we write $\tilde{\alpha}$ for the coverage level used inside the plug-in bands to keep the two coverages distinct.

The key step (42) uses that, on the event $\{\theta \in \hat{\Theta}_\gamma(y_{0:T})\}$, the plug-in band $\hat{X}_{\tilde{\alpha}}(\theta, y_{0:T})$ is a subset of the union $\hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T})$. Hence, choosing $\tilde{\alpha}$ and γ such that

$$(1-\tilde{\alpha})(1-\gamma) = 1-\alpha \tag{46}$$

ensures that $\hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T})$ has predictive coverage at least $1-\alpha$. The bound holds uniformly over $\theta \in \Theta$ provided both ingredients (plug-in conditional validity and $\hat{\Theta}_\gamma$ validity) hold uniformly; in practice, $\hat{\Theta}_\gamma$ is typically only asymptotically valid, so (45) holds with an $o(1)$ remainder as $T \to \infty$. In this paper, we take $\hat{\Theta}_\gamma$ from likelihood-ratio test (LR) inversion, which enjoys classic frequentist validity asymptotically.[24]

This calibration tends to be conservative: Empirically, values of γ larger than those solving $(1-\tilde{\alpha})(1-\gamma) = 1-\alpha$ often still yield near-nominal predictive coverage. The conservatism stems from our proof technique: we (i) discard the branch $\{\theta \notin \hat{\Theta}_\gamma(y_{0:T})\}$ and (ii) replace the union by the single plug-in set at θ, each step shrinking the probability event (see the two inequalities Eqs. (41)-(42)). Tightening the prediction band union would require a sharper lower bound—one that retains more of the discarded mass or exploits dependence between events—or an alternative set construction that achieves the predictive guarantee with less slack.

Even with the conservative lower bound (45), we still must choose $(\tilde{\alpha}, \gamma)$, and this choice directly affects the size of the PU band. In fact, $\tilde{\alpha}$ and γ are underdetermined because any pair satisfying (46) attains the same lower bound, yet different pairs yield different prediction-band unions with different sizes. For a given sample, a principled choice is to search for the tuple $(\tilde{\alpha}, \gamma)$, which minimizes a size functional of the prediction band union (e.g., total width) subject to the condition (46):

$$\min_{0\le\tilde{\alpha},\gamma\le 1} \operatorname{size}\Big(\hat{X}^{\cup}_{\gamma,\,\tilde{\alpha}}(y_{0:T})\Big), \tag{49}$$

$$\text{s.t.} \quad (1-\tilde{\alpha})(1-\gamma) = 1-\alpha \tag{50}$$

Practically, rearranging the constraint to solve for $\gamma(\tilde{\alpha}) = 1 - \frac{1-\alpha}{1-\tilde{\alpha}}$ reduces the calibration to a one-dimensional search in $\tilde{\alpha}$, which may be solved by Bisection or Newton methods (recall that we assumed sufficient regularity before). Monotonicity is clear in the extremes:

[24] The confidence set $\hat{\Theta}_\gamma$ can be derived by inverting the likelihood-ratio test (LR):

$$\hat{\Theta}_\gamma(y_{0:T}) \;=\; \Big\{\theta :\; \log L(\theta \mid y_{0:T}) \;\ge\; \log L(\hat{\theta} \mid y_{0:T}) - \tfrac{1}{2}\,\chi^2_{p,\,1-\gamma}\Big\}, \tag{47}$$

with $L(\theta \mid y_{0:T})$ as in (6), $\hat{\theta}$ the MLE, $p = \dim(\theta)$, and $\chi^2_{p,\,1-\gamma}$ the $(1-\gamma)$-quantile of χ^2_p. This set enjoys classic frequentist validity asymptotically

$$\forall\, \theta \in \Theta : \quad \lim_{T\to\infty} P_\theta^{X,Y}\Big\{\theta \in \hat{\Theta}_\gamma(Y_{0:T})\Big\} = 1-\gamma, \tag{48}$$

where the probability is with respect to repeated sampling of the full time series $\{x_t, y_t\}_{t=0}^T$ at fixed T. The LR confidence set leverages the full likelihood rather than a purely quadratic (Wald) approximation.

decreasing $\tilde{\alpha}$ (more conservative plug-in bands) or decreasing γ (larger confidence set) both enlarge $\hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T})$, but the optimal trade-off is data dependent.

If evaluating the full frontier $(\tilde{\alpha}, \gamma(\tilde{\alpha}))$ is computationally expensive, we consider two types of heuristics: First, a simple and effective heuristic is the square-root split,

$$1 - \tilde{\alpha} \;=\; 1 - \gamma \;=\; \sqrt{1-\alpha}\,, \tag{51}$$

which balances the two uncertainty sources so that neither dominates. When prior or empirical intuition about the relative tail behavior of $\hat{\theta}$ versus the plug-in bands for $x_{0:T}$ is available, one may "tilt" the split accordingly by assigning a larger share of the product $1-\alpha$ to the relatively better-behaved component (e.g., heavier latent tails $\Rightarrow$ increase $1-\gamma$, lighten $1-\tilde{\alpha}$, and vice versa).

Our second heuristic exploits the fact that, in large sample, parameter uncertainty shrinks while the model-induced uncertainty in $x_{0:T}$ does not vanish to minimize the width of the prediction band union. For fixed γ, the LR confidence set $\hat{\Theta}_\gamma(y_{0:T})$ shrinks as $T \to \infty$. Hence, the asymptotically optimal target is to let $\gamma \to 0$ and $\tilde{\alpha} \to \alpha$: this minimizes the contribution of parameter uncertainty while keeping the plug-in coverage at the desired level. In finite samples, $\gamma = 0$ would produce an uninformative LR set, but asymptotically $\hat{\theta} \to \theta$ and $\hat{\Theta}_\gamma(y_{0:T})$ collapses, so the set union becomes minimal.

In practice, we therefore seek a schedule $\gamma(T) \downarrow 0$ such that the confidence sets still contract. Along the calibration frontier $(1-\tilde{\alpha}(T))(1-\gamma(T)) = 1-\alpha$, set

$$1 - \tilde{\alpha}(T) = \frac{1-\alpha}{1-\gamma(T)} \quad \Longrightarrow \quad \tilde{\alpha}(T) \uparrow \alpha \text{ as } \gamma(T) \downarrow 0. \tag{52}$$

A sufficient condition for contraction is that the LR "radius" obeys $\chi^2_{p,\,1-\gamma(T)} = o(T)$, because standard quadratic expansion of the log-likelihood yields a diameter for $\hat{\Theta}_{\gamma(T)}(y_{0:T})$ of order $O_p\Big(\sqrt{\chi^2_{p,\,1-\gamma(T)}/T}\Big)$. Using the tail behavior $\chi^2_{p,\,1-\gamma} \sim 2\log(1/\gamma)$ as $\gamma \downarrow 0$, any choice $\gamma(T) = T^{-k}$ with $k > 0$ gives

$$\sqrt{\frac{\chi^2_{p,\,1-\gamma(T)}}{T}} \;\sim\; \sqrt{\frac{2k\log T}{T}} \;\longrightarrow\; 0, \tag{53}$$

so the LR confidence sets shrink despite the decreasing confidence level.[25]

In particular, under such a schedule, $\hat{\Theta}_{\gamma(T)}(y_{0:T}) \Rightarrow \{\theta\}$ and $\tilde{\alpha}(T) \to \alpha$. Con-

[25] Pathwise nesting $\hat{\Theta}_{\gamma(T)} \subset \hat{\Theta}_{\gamma(T')}$ for all $T \geq T'$ is not guaranteed, because increasing T tightens curvature (shrinking sets) while decreasing $\gamma(T)$ loosens the threshold (expanding sets). The schedule above ensures asymptotic contraction (diameter $o_p(1)$), which suffices for our limit statements.

sequently, the prediction-band union

$$\hat{X}^{\cup}_{\gamma(T),\,\tilde{\alpha}(T)}(y_{0:T}) \;=\; \bigcup_{\theta\in\hat{\Theta}_{\gamma(T)}(y_{0:T})} \hat{X}_{\tilde{\alpha}(T)}(\theta, y_{0:T}) \tag{54}$$

converges to the plug-in band at the true parameter and target level, i.e., to $\hat{X}_{\alpha}(\theta, y_{0:T})$.

With the theoretical setup in place, we now turn to a fully deterministic implementation of the prediction-band union. We show that the smallest hyperrectangle enclosing the union can be obtained via a series of constrained optimization problems. These problems are straightforward to solve—even for complex, nonlinear state-space models—because they can leverage our RLI-type algorithm from the previous section.

As an $(T+1)$-dimensional set, the prediction-band union $\hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T})$ generally has an irregular shape: each $\hat{X}_{\tilde{\alpha}}(\theta, y_{0:T})$ is a band, i.e., a hyperrectangle in the space of possible missing sequences, $\mathbb{R}^{(T+1)}$, but their union need not be. We therefore report its minimal axis-aligned (hyperrectangular) envelope

$$\hat{X}^{\square}_{\gamma,\tilde{\alpha}}(y_{0:T}) \;\equiv\; \{x :\; \underline{x}_t \le x_t \le \bar{x}_t \text{ for all } t\}\,, \tag{55}$$

where

$$\bar{x}_t \equiv \max_{x\in\hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T})} x_t, \tag{56}$$

$$\underline{x}_t \equiv \min_{x\in\hat{X}^{\cup}_{\gamma,\tilde{\alpha}}(y_{0:T})} x_t. \tag{57}$$

This introduces a projection error (see Footnote **??**) but yields a computable object that preserves simultaneous, pathwise interpretation. Operationally, computing this projection reduces to $2T$ scalar extremizations, which we cast as smooth constrained programs next.

In particular, we compute each envelope boundary $\bar{x}_t, \underline{x}_t$ as the solution of a smooth

constrained program.[26] For the upper boundary, this is

$$\bar{x}_t = \max_{x\in\mathbb{R}^{T+1},\, c>0,\, \theta\in\Theta} x_t \tag{58}$$

$$\text{s.t.} \quad x \in B_\theta(c), \tag{59}$$

$$P_\theta^{x_{0:T}|y_{0:T}}\big(B_\theta(c) \mid y_{0:T}\big) = 1-\tilde{\alpha}, \tag{60}$$

$$\theta \in \hat{\Theta}_\gamma(y_{0:T}). \tag{61}$$

The lower boundary $\underline{x}_t$ is obtained by replacing "max" with "min". Note that, we are not maximizing x_t over a precomputed union; rather, we allow any $x \in \mathbb{R}^{T+1}$ but enforce that x belongs to some plug-in band of level $1-\tilde{\alpha}$ and that the corresponding parameter lies in $\hat{\Theta}_\gamma(y_{0:T})$. This makes the feasible set of (58)-(61) exactly the (implicit) union, so the optimizer attains the true envelope boundary. Furthermore, constraint (63) hides two tasks: (a) evaluating $\hat{\sigma}_t^2$ (an integral of the form (15), handled via our RLI-type approximation) and (b) obtaining the point path $\mathring{x}_t$. If the point path is the mean sequence, both pieces use the same RLI machinery; if it is the MAP, i.e., the most likely sequence, the problem becomes bilevel (outer (x_t, c, θ), inner MAP), which is generally hard but practical with software that handles implicit functions, e.g., CasADi (Andersson et al., 2019).

Otherwise, the resulting program is a small, smooth nonlinear optimization that off-the-shelf solvers handle well, e.g., with automatic differentiation (AD); in practice, supplying analytic/AD gradients and using warm starts across t yields stable, fast solves. In particular, even for complex state-space models, our RLI-type algorithm delivers all required components (point estimator, variance, and plug-in coverage) quickly and with high accuracy.

We can refine the constrained optimization problem further to

$$\bar{x}_t = \max_{x_t\in\mathbb{R}, c\in\mathbb{R}_{>0}, \theta\in\Theta} x_t \tag{62}$$

$$\text{s.t.} \qquad -\hat{\sigma}_t \le \frac{x_t - \mathring{x}_t}{c} \le \hat{\sigma}_t, \tag{63}$$

$$P_\theta^{x_{0:T}|y_{0:T}}\left(B_\theta(c)|y_{0:T}\right) \le 1-\alpha, \tag{64}$$

$$L\left(\theta|y_{0:T}\right) \ge L(\hat{\theta}|y_{0:T}) - \frac{1}{2}\chi^2_{p,1-\gamma} \tag{65}$$

By refining the constraints in (62)-(65), we reduce the decision dimension of the max-

[26] Forming $\hat{X}^{\square}_{\gamma,\tilde{\alpha}}(y_{0:T})$ by iterating over plug-in bands on a grid of $\hat{\Theta}_\gamma(y_{0:T}) \subset \mathbb{R}^p$ is ineffective. First, $\hat{\Theta}_\gamma(y_{0:T})$ contains infinitely many θ, so any discretization risks missing the true extremizers. Second, the width and location of $\hat{X}_{\tilde{\alpha}}(\theta, y_{0:T})$ can be non-monotone in θ; extremal coordinates $\max x_t$ or $\min x_t$ may occur at interior points, not just on the boundary of $\hat{\Theta}_\gamma$. A boundary-only net is therefore unsound, while a full interior net must be exponentially fine in p to be safe—prohibitively costly and still vulnerable to gaps.

imization from $T + p + 2$[27] to $2 + p$, i.e., we no longer optimize over the full sequence x, only its t-th component x_t, along with c and θ. Likewise, the number of inequality constraints drops from $2T + 4$ to just 4, since we no longer need to enforce band bounds for all off-target periods of the candidate sequence.

We justify the simplifications in (63)-(65) one by one. For (63), we have replaced the abstract condition for x to be included in the plug-in prediction band $B_\theta(c)$ by its concrete bounds. Note that due to the rectangular shape of $B_\theta(c)$, the maximization in dimension (i.e. period) t is independent of all other periods. Hence, we can discard all other periods x_s, $s \neq t$, of the candidate sequence x and it suffices to constrain it in the t-th dimension.[28] Second, because the objective is to increase x_t, the optimizer will "push" the band outward by enlarging c.[29] Without a coverage cap, c (and thus $\bar{x}_t$) would "blow up". It therefore suffices to impose an upper bound on the plug-in coverage, $P_\theta^{x_{0:T}|y_{0:T}}\big(B_\theta(c) \mid y_{0:T}\big) \leq 1-\tilde{\alpha}$ (analogously with $1 - \alpha$ if one calibrates directly to the product frontier). The lower bound is immaterial in this maximization since the optimal c is the largest feasible one. Importantly, this also holds when minimizing, i.e., for computing the lower bound $\underline{x}_t$ since the plug-in prediction bands are designed to be symmetric. Finally, the LR-based confidence set is defined by the inequality $\log L(\theta \mid y_{0:T}) \geq \log L(\hat{\theta} \mid y_{0:T}) - \frac{1}{2}\chi^2_{p,1-\gamma}$. We can enforce this directly for each candidate (x_t, c, θ) during optimization, avoiding the separate computation (or discretization) of $\hat{\Theta}_\gamma(y_{0:T})$ while retaining exactly the same feasible θ.

Leaving aside any issues with the time series setting and occasional observations, a possible alternative to our approach could be bootstrap methods, which may be used to resample the data and construct a prediction band union based on the recomputed parameters. However, this approach requires truncating the empirical parameter distribution somewhere and it is a priori unclear how one would compute the resulting quantiles over the set of plug-in prediction bands (for the recomputed parameters).

3.2.2 Simulation Experiments

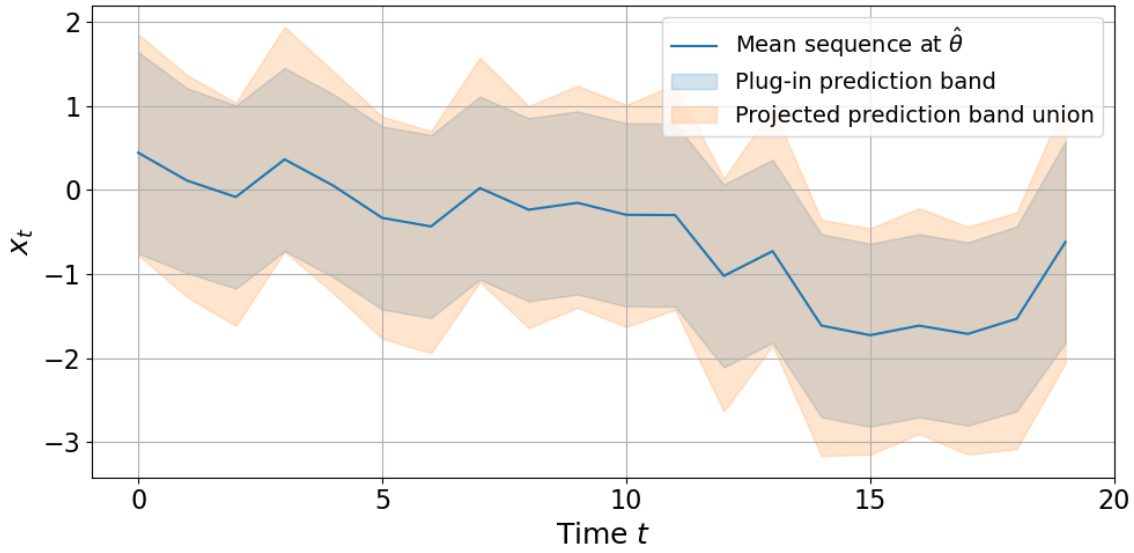
Finally, we confirm the validity and necessity of our approach in a simulation experiment for a linear Gaussian model by demonstrating that the plug-in prediction band is indeed overconfident, i.e., inadequate for inference under parameter uncertainty, and then showing that the projected prediction band union does satisfy our coverage target.

[27] $T + 1$ coordinates for the candidate path x, p for the parameter vector θ, and 1 for the scale c.

[28] This coordinate-wise reduction fails for non-rectangular plug-in sets (e.g., highest-density regions; see Footnote 15).

[29] Reich and Judd (2020) follow a similar approach to obtain confidence bands for functions of counterfactual parameters by maximizing an implicitly defined likelihood using MPEC.

Figure 4: Mean path and the projected prediction band union for the latent state $\{x_t\}_{t=0}^{19}$ in one simulated dataset



Notes: We plot the mean sequence $\{\bar{x}_t\}_{t=0}^{19}$ (solid blue line) as well as the plug-in prediction band (shaded blue) and the projected band union (shaded orange) for the linear-Gaussian model defined in Eqs. (66) and (67). The mean sequence and the plug-in prediction band union are computed using the maximum likelihood estimator $\hat{\theta}$. The plug-in prediction band satisfies 95% plug-in coverage as defined in Eq. (8), the prediction band union satisfies predictive coverage as defined in Eq. (45) with $\tilde{\alpha} = \gamma \approx 0.0253$.

We study a linear-Gaussian state-space model

$$y_t = \beta\, x_t + \eta_t, \tag{66}$$

$$x_t = \rho\, x_{t-1} + \varepsilon_t, \tag{67}$$

with $\eta_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_y^2)$ and $\varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_x^2)$. We take $x_0 \sim \mathcal{N}\big(0, \sigma_x^2/(1-\rho^2)\big)$ (stationary initialization).

In this setting, Kalman filter/smoother (KF/RTS) yields the exact point path (we use the mean) and pointwise standard deviations, and provides the exact log-likelihood $L(\theta \mid y_{0:T})$ for $\theta = (\rho, \sigma_x)$. This removes approximation error so we can isolate the effect of parameter uncertainty. In more complex models, the methods from Section 3.1 replace KF/RTS. One quantity still requires an outer search: the plug-in band scale $c^\star$ cannot be read off from KF marginals, so we compute it by monotone root-finding of (19) using our RLI-type algorithm.

We vary two uncertain parameters, ρ and σ_x, fixing $\beta = 1$ and $\sigma_y = 0.5$. The parameter space is $\Theta = [0.6, 0.99] \times [0.1, 2]$, with truth $(\rho, \sigma_x) = (0.9, 0.5)$. For $T = 20$ (short but informative; $T = 10$ is too small and $T = 100$ is computationally heavy) we draw $S = 100$ datasets.

For each dataset $y_{0:T}^{(s)}$ we perform three steps: (i) estimate $\hat{\theta}^{(s)}$ by maximizing the KF

likelihood over Θ; (ii) compute the plug-in prediction band at level $1-\tilde{\alpha}$ by solving (19) (posterior mean path + KF variances); (iii) compute the projected PU band by solving (62) $2T$ times (upper/lower for each t). Unless stated otherwise we use the square-root split to target $1-\alpha = 0.95$,

$$1-\tilde{\alpha} = 1-\gamma = \sqrt{1-\alpha} \approx 0.9747 \quad (\tilde{\alpha} = \gamma \approx 0.0253), \tag{68}$$

and we also report a conservative variant with $\tilde{\alpha} = \gamma = 0.05$.

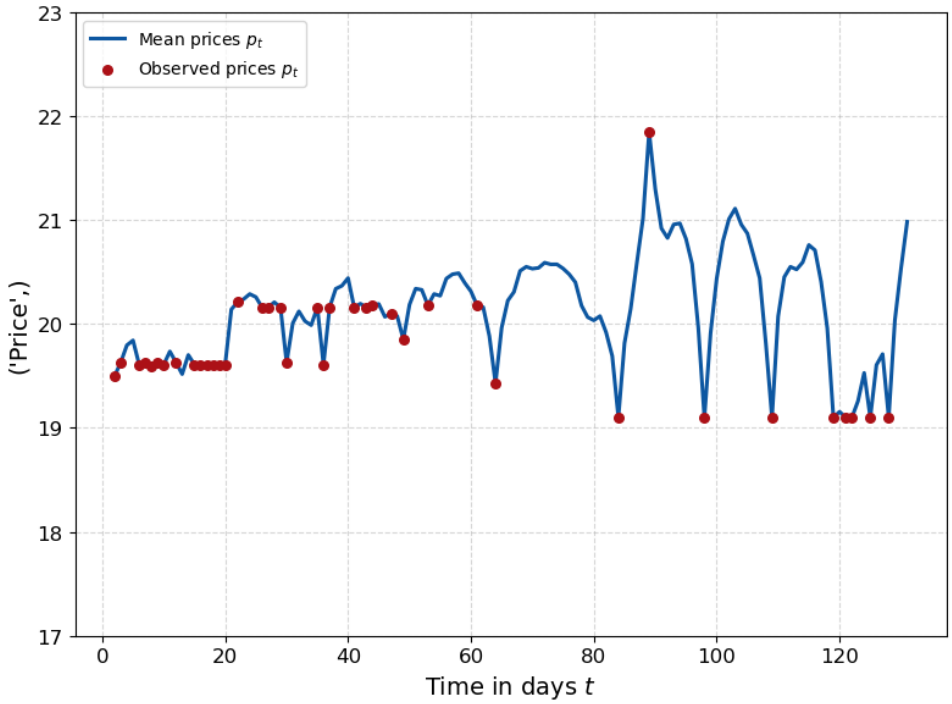
We find that the plug-in band is overconfident, whereas the projected prediction-band union attains the coverage goal but is somewhat conservative. Concretely, the plug-in band covers the true sequence $x_{0:T}^{true}$ in 85 out of $S = 100$ simulations. By contrast, the projected prediction-band union covers $x_{0:T}^{true}$ in 99 simulations. These results show two things: first, relying solely on the plug-in band is inappropriate because it underestimates the uncertainty about the missing sequence inherent in the data; second, the projected prediction-band union corrects this undercoverage but overshoots, yielding conservative sets. Finally in Figure 4, to visualize full predictive inference in this example, we plot the mean estimator and the plug-in prediction band, both evaluated at the maximum likelihood estimator $\hat{\theta}$, as well as the projected prediction band union for one simulated data set.

4 Estimating prices in a steel-trade model

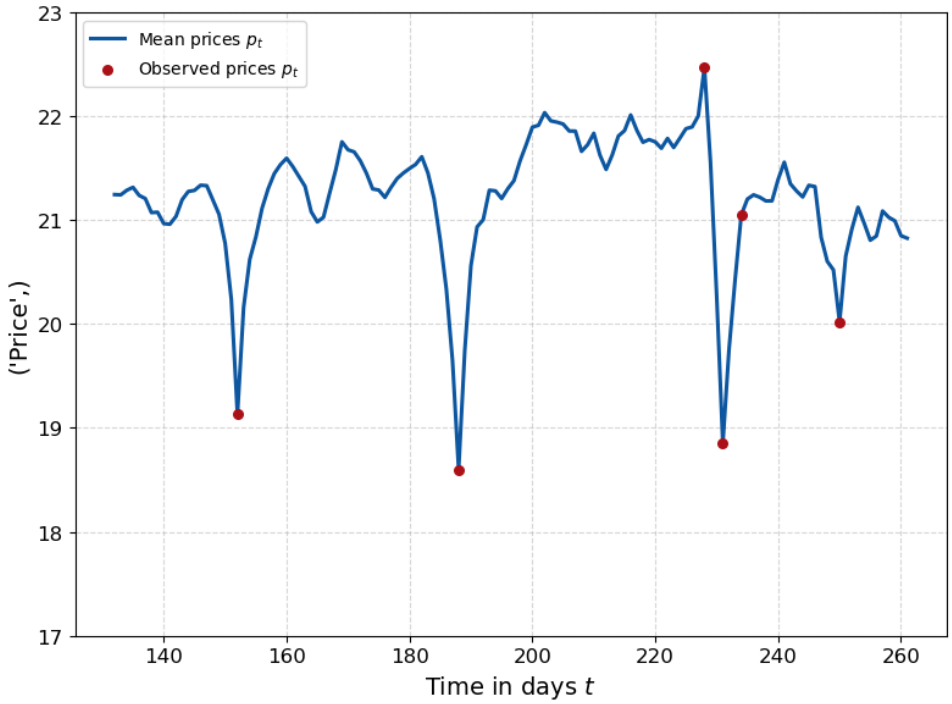
To demonstrate the applicability of our methods in real examples, we use it to infer a sequence of occasionally observed prices in the steel trading model of Hall and Rust (2021). There, a steel-trading company buys items on the wholesale market, and re-sells them to retail customers. The inventory management of the company is assumed to implement the solution to a dynamic profit optimization problem, which trades-off stockpile available for sale against cost minimization (attempting to "buy low").

Importantly, the data set is limited to the steel trading company's transactions, and is not augmented by data from the wholesale market. Therefore, it contains only occasional observations of the wholesale price p_t, namely for periods where the firm made an actual purchase; additionally, this observability pattern is endogenous: The company is less likely to stock-up their inventory when the price is high (but might be forced to do so if running low), and vice versa. The model admits a state-space representation as in (1)-(2), allowing us to apply the methods we developed in the previous section to estimate the band of "reasonable" wholesale prices for the non-observed periods, given the data: we report the mean path and the projected prediction band union, thereby delivering full predictive inference for p_t under both model-induced randomness and parameter uncertainty.

Figure 5: Mean path for the wholesale price sequence $\{p_t\}_{t=0}^{260}$
TBD: Projected prediction band union analogous to Figure 4.



(a) Horizon $t = 0, ..., 130$



(b) Horizon $t = 130, ..., 260$

Notes: We estimate the series of wholesale prices p_t for 3/4-inch steel plate in our steel-trade data set and using the reduced-form model presented in Hall and Rust (2021). Red dots indicate observations of p_t, the solid blue line represents the mean path for p_t. The upper panel plots the evolution for periods $t = 0, ..., 130$, the lower panel for $T = 130, ..., 260$.
TBD: Boundaries of the projected prediction band union at the coverage level $\alpha = 0.95$.

5 Conclusion

TBD.

References

Andersson, J. A. E., Gillis, J., Horn, G., Rawlings, J. B. and Diehl, M. (2019). CasADi – A software framework for nonlinear optimization and optimal control. *Mathematical Programming Computation* **11** (1), 1–36.

Aruoba, S. B., Bocola, L. and Schorfheide, F. (2017). Assessing DSGE model nonlinearities. *Journal of Economic Dynamics and Control* **83**, 34–54.

Aruoba, S. B., Cuba-Borda, P., Higa-Flores, K., Schorfheide, F. and Villalvazo, S. (2021). Piecewise-linear approximations and filtering for DSGE models with occasionally-binding constraints. *Review of Economic Dynamics* **41**, 96–120.

Barndorff-Nielsen, O. and Cox, D. R. (1996). Prediction and asymptotics. *Bernoulli* **2** (4), 319–340.

Blevins, J. R. (2015). Sequential Monte Carlo Methods for Estimating Dynamic Microeconomic Models. *Journal of Applied Econometrics* **31** (5), 773–804.

Chopin, N. and Papaspiliopoulos, O. (2020). *An Introduction to Sequential Monte Carlo*. Springer International Publishing.

Cox, D. R. (1975). Prediction Intervals and Empirical Bayes Confidence Intervals. *Journal of Applied Probability* **12** (S1), 47–55.

Dubé, J.-P., Hitsch, G. J. and Manchanda, P. (2005). An Empirical Model of Advertising Dynamics. *Quantitative Marketing and Economics* **3** (2), 107–144.

Durbin, J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika* **89** (3), 603–616.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods: Second Edition*. Oxford University Press.

Engle, R. and Patton, A. (2001). What good is a volatility model? *Quantitative Finance* **1** (2), 237–245.

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.

Erdem, T. and Keane, M. P. (1996). Decision-Making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets. *Marketing Science* **15** (1), 1–20.

Fernandez-Villaverde, J., Rubio-Ramirez, J. F. and Santos, M. S. (2006). Convergence Properties of the Likelihood of Computed Dynamic Models. *Econometrica* **74** (1), 93–119.

Fernández-Villaverde, J. and Rubio-Ramírez, J. F. (2007). Estimating Macroeconomic Models: A Likelihood Approach. *Review of Economic Studies* **74** (4), 1059–1087.

Geisser, S. (1993). *Predictive inference*. Chapman and Hall.

Gilch, A., Lanz, A., Müller, P., Reich, G. and Wilms, O. (2025). "Small Data": Inference with Occasionally Observed States. *Management Science*.

Gilch, A., Reich, G. and Wilms, O. (2025). Asymptotic Properties of the Maximum Likelihood Estimator under Occasionally Observed States. Working paper.

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.

Hamilton, J. D. (1986). A standard error for the estimated state vector of a state-space model. *Journal of Econometrics* **33** (3), 387–397.

Herbst, E. and Schorfheide, F. (2014). SEQUENTIAL MONTE CARLO SAMPLING FOR DSGE MODELS. *Journal of Applied Econometrics* **29** (7), 1073–1098.

Keane, M. P. (1994). A Computationally Practical Simulation Estimator for Panel Data. *Econometrica* **62** (1), 95.

Keane, M. P. and Wolpin, K. I. (1994). The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence. *The Review of Economics and Statistics* **76** (4), 648.

— (1997). The Career Decisions of Young Men. *Journal of Political Economy* **105** (3), 473–522.

Kim, S., Shepherd, N. and Chib, S. (1998). Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies* **65** (3), 361–393.

Kitagawa, G. (1987). Non-Gaussian State-Space Modeling of Nonstationary Time Series. *Journal of the American Statistical Association* **82** (400), 1032.

Montiel Olea, J. L. and Plagborg-Møller, M. (2018). Simultaneous confidence bands: Theory, implementation, and an application to SVARs. *Journal of Applied Econometrics* **34** (1), 1–17.

Norets, A. (2009). Inference in Dynamic Discrete Choice Models With Serially orrelated Unobserved State Variables. *Econometrica* **77** (5), 1665–1682.

Pakes, A. (1986). Patents as Options: Some Estimates of the Value of Holding European Patent Stocks. *Econometrica* **54** (4), 755.

Pfeffermann, D. and Tiller, R. (2005). Bootstrap Approximation to Prediction MSE for State–Space Models with Estimated Parameters. *Journal of Time Series Analysis* **26** (6), 893–916.

Quenneville, B. and Singh, A. C. (2000). Bayesian Prediction Mean Squared Error for State Space Models with Estimated Parameters. *Journal of Time Series Analysis* **21** (2), 219–236.

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

Reich, G. and Judd, K. L. (2020). Efficient Likelihood Ratio Confidence Intervals using Constrained Optimization. *Available at SSRN*.

Rodriguez, A. and Ruiz, E. (2009). Bootstrap prediction intervals in state–space models. *Journal of Time Series Analysis* **30** (2), 167–178.

Rodríguez, A. and Ruiz, E. (2012). Bootstrap prediction mean squared errors of unobserved states based on the Kalman filter with estimated parameters. *Computational Statistics and Data Analysis* **56** (1), 62–74.

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

— (1988). Maximum Likelihood Estimation of Discrete Control Processes. *SIAM Journal on Control and Optimization* **26** (5), 1006–1024.

Shephard, N. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika* **84** (3), 653–667.

Young, G. A. and Smith, R. L. (2005). *Essentials of statistical inference*. Vol. 16. Cambridge University Press.

ONLINE APPENDIX

More details on particle filtering and RLI

Simulation based methods

A popular way to approximate high-dimensional integrals such as that in the likelihood (6) is Monte Carlo simulation. These methods are not directly affected by the curse of dimensionality, so their asymptotic accuracy is essentially independent of the integral's dimension $(T+1)$. The basic idea is to draw N_{eval} samples $x_{0:T}$ from $P_\theta^{x_{0:T}}$ and average:

$$L\left(\theta|y_{0:T}\right) = \int_{\mathcal{S}_x^{T+1}} P_\theta^{y_{0:T}|x_{0:T}}\left(y_{0:T}|x_{0:T}\right) dP_\theta^{x_{0:T}}\left(x_{0:T}\right) \approx \frac{1}{N_{eval}} \sum_{i=1}^{N_{eval}} P_\theta^{y_{0:T}|x_{0:T}}\left(y_{0:T}|x_i\right) \quad (69)$$

Regardless of the dimension of $x_{0:T}$, and under mild regularity conditions, the simulation error satisfies

$$\left| L\left(\theta|y_{0:T}\right) - \frac{1}{N_{eval}} \sum_{i=1}^{N_{eval}} P_\theta^{y_{0:T}|x_{0:T}}\left(y_{0:T}|x_i\right) \right| = \sqrt{\mathrm{Var}_\theta^{x_{0:T}}\left[P_\theta^{y_{0:T}|x_{0:T}}\left(y_{0:T}|x_i\right)\right]} O_p\left(N_{eval}^{-1/2}\right) \quad (70)$$

as $N_{eval} \to \infty$.

However, sampling directly from $P_\theta^{x_{0:T}}$ is generally nontrivial. Consequently, methods such as particle filtering or Gibbs sampling do not draw from the full joint distribution at once, but instead iteratively sample from conditional distributions. In particle filtering, one uses the following recursion for $t = 0, \ldots, T$ to obtain samples $x_t^i \sim P_\theta\left(x_t|y_{0:t}\right)$:

$$P_\theta\left(x_t|y_{0:t-1}\right) = \int_{\mathcal{S}_x} P_\theta\left(x_t|x_{t-1}\right) P_\theta\left(x_{t-1}|y_{0:t-1}\right) dx_{t-1} \quad (71)$$

$$P_\theta\left(y_t|y_{0:t-1}\right) = \int_{\mathcal{S}_x} P_\theta\left(y_t|x_t\right) P_\theta\left(x_t|y_{0:t-1}\right) dx_t \quad (72)$$

$$P_\theta\left(x_t|y_{0:t}\right) = \frac{P_\theta\left(y_t|x_t\right) P_\theta\left(x_t|y_{0:t-1}\right)}{P_\theta\left(y_t|y_{0:t-1}\right)}. \quad (73)$$

In particular, the filtering algorithm simultaneously computes the conditional densities for the data, $P_\theta\left(y_t|y_{0:t-1}\right)$, yielding the likelihood as

$$L\left(\theta|y_{0:T}\right) = \prod_{t=0}^{T} P_\theta\left(y_t|y_{0:t}\right). \quad (74)$$

Regardless of implementation-specific optimizations, simulation attains only the Monte Carlo rate (70). In practice, this means that in order to gain one additional digit of accur-

acy, one typically requires about 100 times more evaluations of the integrand. Moreover, because the error bound is probabilistic, the achieved accuracy is not even guaranteed (even though probability bounds on the maximum error can be stated); similarly, the estimate of the integral itself is subject to simulation noise. Therefore, simulation-based approaches are mainly justified if either function evaluations are extremely cheap, or if the researcher is satisfied with a very coarse (and noisy) approximation of the object of interested. However, this is not the case for many relevant applications in economics and finance. For example, many models feature optimizing agents, where the transition functions Ψ_θ, Φ_θ from Equations (1)-(2)—and hence the transition density $P_\theta^{X,Y}$—are not available in closed form; then, obtaining a sample x_t requires solving the agents' optimization problems numerically, making function evaluations expensive, and thus motivating methods with faster convergence. Also, many solvers that are regularly employed to optimize the likelihood strongly benefit from noise-free objectives; in the context of estimation, this issue is distinctive if either data is scarce or if the likelihood cannot strongly discriminate multiple sets of parameter values due to the way the model is formulated (sometimes referred to as "poor identification"), and thus the maximum of the likelihood is ambiguous and thus hard to pinpoint. Both issues give deterministic approximation methods a potential edge over simulation in practical applications.

Recursive Likelihood Integration (RLI)

On the other hand, the RLI algorithm addresses both challenges: the slow convergence of simulation and the incorporation of occasional observations of x_t. It does so via two components: (i) a recursive formulation of the integral and (ii) a per-step combination of numerical integration and interpolation. The recursion is

$$f_t^\theta(x_{t-1}) = \int_{\mathcal{S}_x} P_\theta\left(y_t, \tilde{x}_t \mid y_{t-1}, x_{t-1}\right) f_{t+1}^\theta(\tilde{x}_t), d\tilde{x}_t \tag{75}$$

for $t = 0, \ldots, T$,[30] so that the final step $t = 1$ yields the likelihood:

$$L\left(\theta | y_{0:T}\right) = \begin{cases} f_1^\theta(x_0) & \text{if } x_0 \in \bar{\mathcal{T}}, \\ f_0^\theta & \text{else.} \end{cases} \tag{76}$$

To handle occasional observations, Gilch et al. (2025) modify (75) by replacing the integral over x_t with evaluation at the observed x_t. For multivariate states with partial observations, the integral is restricted to the unobserved components. See Gilch et al. (2025) for details.

Given the recursive formulation (75), RLI approximates the conditional integral at

[30] This mirrors the classic filtering recursion in (71) but avoids computing $P_\theta\left(x_t \mid y_{0:t-1}\right)$ as a ratio with an additional integral in the denominator.

each step using a numerical integration rule. Unlike simulation-based methods, these rules use a fixed set of nodes x_i and weights w_i tailored to weighted one-dimensional integrals,

$$\int_A f(x)\omega(x)dx \approx \sum_{i=1}^{N_Q} w_i f(x_i), \tag{77}$$

so there is a one-to-one correspondence between the choice of rule (x_i, w_i) and the pair (A, ω) specifying the domain and weight function.

Reich (2018) propose a change of variables for the integration variable $\tilde{x}_t$ which, intuitively, shifts the integration nodes toward the mass of the integrand.[31] Concretely, we seek a map ξ such that $\tilde{x}_t = \xi(z_t, x_{t-1})$ and

$$P_\theta\left(y_t, \tilde{x}_t | y_{t-1}, x_{t-1}\right) = g_t\big(\xi(z_t, x_{t-1}), x_{t-1}\big)\omega(z_t). \tag{79}$$

This reparametrizes the integral in terms of z_t and factorizes the conditional density of $\tilde{x}_t$ into a function g_t times the quadrature weight ω. We then approximate $f_t^\theta(x_{t-1})$ by

$$f_t^\theta(x_{t-1}) = \int_{\mathcal{S}_x} g_t\big(\xi(z_t, x_{t-1})x_{t-1}\big)\omega(z_t), f_{t+1}^\theta\big(\xi(z_t, x_{t-1})\big)\xi(z_t)dz_t \tag{80}$$

$$\approx \tilde{f}_t^\theta(x_{t-1}) \equiv \sum_{i=1}^{N_Q} g_t\big(\xi(z_{t,i}, x_{t-1}), x_{t-1}\big)\tilde{f}_{t+1}^\theta\big(\xi(z_{t,i}, x_{t-1})\big), \xi(z_{t,i}). \tag{81}$$

At this point, a direct application of the change-of-variables scheme can cause an exponential growth in function evaluations. Because ξ depends on x_{t-1}, the argument x_{t-1} of $\tilde{f}_t^\theta$ also enters $\tilde{f}_{t+1}^\theta$. One step earlier, $\tilde{f}_t^\theta$ is evaluated at $x_{t-1} = \xi(z_{t-1,j}, x_{t-2})$, so the dependence of $\tilde{f}_{t+1}^\theta$ on x_{t-1} becomes a dependence on x_{t-2}. Consequently, at time $t+1$ we must evaluate $\tilde{f}_{t+1}^\theta$ for all combinations $(z_{t,i}, z_{t-1,j})$, $i, j = 1, \ldots, N_Q$. Iterating this argument shows that, under the naive scheme, the number of evaluations of $\tilde{f}_t^\theta$ grows as N_Q^t for $t = 0, \ldots, T$, i.e., computational cost increases exponentially in T.

RLI solves this issue by using interpolation: we interpolate the function $\tilde{f}_t^\theta$ using a

[31] It is suboptimal to approximate f_t naively by dividing and multiplying by a weighting function and then applying a rule designed for $A = \mathcal{S}_x$:

$$f_t^\theta(x_{t-1}) = \int_{\mathcal{S}_x} \frac{P_\theta\left(y_t, \tilde{x}_t | y_{t-1}, x_{t-1}\right) f_{t+1}^\theta(\tilde{x}_t)}{\omega(\tilde{x}_t)}\omega(\tilde{x}_t)d\tilde{x}_t \quad \approx \sum_{i=1}^{N_Q} w_i \frac{P_\theta\left(y_t, \tilde{x}_{t,i} | y_{t-1}, x_{t-1}\right) f_{t+1}^\theta(\tilde{x}_{t,i})}{\omega(\tilde{x}_{t,i})}. \tag{78}$$

The mismatch between the integrand and the weight ω induces large errors because nodes $\tilde{x}_{t,i}$ cluster where ω is large. With small N_Q, the integrand is evaluated mainly where it is small. Although this error vanishes asymptotically, it can be substantial unless N_Q is large. (This mirrors the need for importance sampling when the support of a function and the sampling density have little overlap.)

fixed number, $N_{\mathcal{I}}$, of interpolation nodes

$$\hat{f}_t^\theta(x_{t-1}) \approx \mathcal{I}^{N_{\mathcal{I}}}(\tilde{f}_t^\theta)(x_{t-1}) \tag{82}$$

$$= \mathcal{I}^{N_{\mathcal{I}}}\left(\sum_{i=1}^{N_Q} g_t\big(\xi(z_{t,i},\cdot), x_{t-1}\big)\hat{f}_{t+1}^\theta\big(\xi(z_{t,i},\cdot)\big), \xi(z_{t,i})\right)(x_{t-1}) \tag{83}$$

The function $\hat{f}_t^\theta$ can be evaluated for any value $x_{t-1} = x$ without triggering new evaluations of $\tilde{f}_{t+1}^\theta$ and therefore avoids the exponential "blow-up" of function evaluations. Constructing $\hat{f}_t^\theta$ requires $N_{eval} = N_Q N_{\mathcal{I}}$ evaluations of the integrand: for each of the $N_{\mathcal{I}}$ interpolation nodes, the integrand has to be evaluated N_Q times. Finally, replacing f_t^θ by $\hat{f}_t^\theta$ in the recursion (75) yields the likelihood approximation

$$\hat{L}^{N_{eval}}(\theta|y_{0:T}) = \begin{cases} \hat{f}_1^\theta(x_0) & \text{if } x_0 \in \bar{\mathcal{T}}, \\ \hat{f}_0^\theta & \text{else.} \end{cases} \tag{84}$$

Numerical integration and interpolation has the advantage that they can achieve polynomial converging approximation errors, $O(N_Q^{-r_Q})$ and $O(N_{\mathcal{I}}^{-r_I})$ respectively. Reich (2018) shows that the RLI recursion (83), i.e., alternating interpolation and integration to approximate the sequence of nested integrals $\{f_t^\theta\}_{t=0}^T$, preserves these polynomial convergence rates. For fixed T, it yields following approximation error rate:

$$\left|L(\theta|y_{0:T}) - \hat{L}^{N_{eval}}(\theta|y_{0:T})\right| = O\left(N_{eval}^{-r}\right) \tag{85}$$

where $r = \frac{r_Q r_I}{r_Q + r_I}$.