Bayesian Model Comparison for Nonlinear State Space Models without Posterior Sampling

Joshua C.C. Chan
Purdue University

Wei Zhang*
Johns Hopkins University

November 2026

Abstract

Marginal likelihoods are central to Bayesian comparison of nonlinear state space models, but commonly used estimators require posterior simulation for each model, with computational costs that grow with the sample size T. We propose a deterministic estimator that requires no posterior simulation. A grid filter integrates out the latent states, and a Laplace approximation with a second-order correction integrates over the static parameters. Under derivative-growth and posterior-concentration conditions, the corrected Laplace approximation has error $O(T^{-2})$, and the size of the correction indicates the error that remains. In stochastic volatility experiments, the estimator is more accurate than Chib's method, the Gelfand–Dey estimator and importance sampling with a fitted proposal at equal computing time for $T \geqslant 2{,}000$. Applications to two exchange rates and to $T = 19{,}266$ five-minute stock returns provide the strongest support for heavy-tailed observations and a transitory volatility factor.

JEL classification: C11, C15, C52, C58

Keywords: marginal likelihood, Bayesian model comparison, Laplace approximation, nonlinear state space models, stochastic volatility, high-frequency data

*Wei Zhang gratefully acknowledges financial support through her postdoctoral fellowship at the Global Macro Lab, Johns Hopkins University.

1 Introduction

Nonlinear state space models are widely used to study time-varying volatility, counts, durations, and other latent dynamics in macroeconomics and finance. High-frequency financial data can supply tens of thousands of observations on a single asset. Models of volatility spillovers, heavy tails, jumps and multiple volatility factors require a few latent processes and non-Gaussian observations. Empirical work then has to compare many competing specifications on long series. These models have two sets of unknowns: a latent state path, whose dimension grows with the sample size T, and a small vector of static parameters $\boldsymbol{\theta}$. Bayesian model comparison is based on the marginal likelihood, which integrates over both. Integrating out the states gives the integrated likelihood $Z_T(\boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta})$, and integrating this against the prior density $p(\boldsymbol{\theta})$ gives the marginal likelihood, $m_T = \int Z_T(\boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$.

Computing this nested integral accurately is challenging. In nonlinear and non-Gaussian models, $Z_T(\boldsymbol{\theta})$ is generally unavailable analytically and must be evaluated by filtering or simulation over the full sample. Then computing the marginal likelihood adds an outer integration over the static parameters, which is costly to evaluate. Simulation-based methods require either repeated evaluations of $Z_T(\boldsymbol{\theta})$ across parameter values or posterior simulation over the parameters and latent states, with a sampler written for each model. Each sweep and each likelihood evaluation takes longer as T grows, so the computing time for a given accuracy grows with T.

We propose a deterministic estimator that requires no posterior sampler and computes the marginal likelihood in two steps. First, a numerical-integration filter on a grid integrates out the states, exploiting the Markov structure of the state space model. Although the number of latent states grows with the sample size T, the number of static parameters remains fixed. Integrating out the states leaves a fixed, low-dimensional integral over the static parameters, which the second step approximates by a Laplace expansion around the posterior mode. Under regularity conditions, its accuracy improves as the posterior of the

static parameters concentrates with increasing T. In our implementation, mode finding and derivative calculations require a few hundred to several thousand likelihood evaluations, many of which can run in parallel batches. Therefore, the estimator is particularly attractive for long samples, where simulation-based integration is most costly.

At moderate sample sizes, however, skewness and other departures from posterior normality can limit the accuracy of the ordinary Laplace approximation. We therefore add a second-order correction based on the third and fourth derivatives of the log integrated likelihood and the log prior at the posterior mode. This correction captures the leading effects of posterior nonnormality while retaining the computational advantages of a local approximation. It also acts as a diagnostic, because the error that remains after the correction is of the order of its square: we show in the Monte Carlo experiments that the absolute error of the corrected log marginal likelihood is between 0.5 and 7 times the squared correction in all 300 data sets. A small correction indicates an accurate estimate, and a correction of order one can serve as an indicator that the expansion is not reliable along some direction of the parameter space. In that case, the mode and curvature already computed provide the proposal for importance sampling over the static parameters, which replaces the expansion.

Our theoretical analysis characterizes the correction and its error in terms of the integrated likelihood. We express the third and fourth derivatives of its logarithm through expectations and joint cumulants under the smoothing distribution, which links the correction to the latent-state structure. Under explicit derivative-growth and posterior-concentration conditions, our main theorem shows that the correction reduces the approximation error in the log marginal likelihood from order T^{-1} to order T^{-2}. We separate this approximation error from the numerical errors in the mode, curvature and derivatives. The theory applies to a fixed, low-dimensional parameter vector with a posterior concentrated around a unique interior mode. The states are integrated out by a numerical-integration filter on a grid (Kitagawa, 1987; Fridman and Harris, 1998). For a scalar or bivariate state it gives the integrated likelihood to machine precision, so the Laplace expansion is the only source of error of the

estimator.

In Monte Carlo experiments on stochastic volatility and three related models, we measure the approximation error against quadrature benchmarks accurate to 10^{-4}. The correction removes 64% to 88% of the Laplace error at $T = 500$ and 94% to 99% at $T = 8{,}000$, leaving log errors below 0.08 and 0.003. At equal computing time, the proposed estimator is more accurate than Chib's method, the Gelfand–Dey estimator and importance sampling with the cross-entropy method from $T = 2{,}000$ at every persistence level. At $T = 8{,}000$ its error is 9 to 400 times smaller than the standard error of the cross-entropy method, the best of the three. The errors of Chib's method and the Gelfand–Dey estimator rise steeply with T, because each MCMC sweep takes longer. On data generated from three bivariate stochastic volatility models, the proposed estimator selects the true model in 149 of 150 series; the exception is a near-tie on data from the simplest model, which the other two include as a special case.

We then present two applications. The first compares three bivariate stochastic volatility specifications drawn from Yu and Meyer (2006) using $T = 4{,}000$ daily AUD/USD and NZD/USD returns. Starting from a constant-correlation baseline, the other two models separately introduce volatility spillovers and heavy-tailed observations. Yu and Meyer (2006) compare their specifications by the deviance information criterion on weekly exchange-rate data, and to our knowledge marginal likelihoods have not previously been reported for them. In our sample, the heavy-tailed model receives the strongest support, with a log Bayes factor of 110 relative to the baseline, against 98 for volatility spillovers. The second-order term is below 0.2 for every model, so the remaining approximation error is far smaller than these differences.

The second application compares seven stochastic volatility models for $T = 19{,}266$ five-minute returns on the S&P 500 ETF over 2025: the standard stochastic volatility model and its extensions with heavy tails, with jumps and with a second volatility factor. The two-factor model is the most favored by the data. The two extensions that change the dynamics

of the volatility—a transitory factor and a jump in the volatility—have log Bayes factors of 69 to 73 relative to the baseline, against 12 to 25 for heavy tails and return jumps. The second-order term is at most 0.31, so the remaining approximation error is far smaller than these differences.

A key advantage of the proposed approach is that comparing many models on many samples is feasible. We therefore fit all seven models at the end of each month, 77 fits in all, to see whether the ranking is stable over time. The ranking does change through the year: the model with jumps in volatility is favored most strongly after the April and October tariff episodes, and the two-factor model from June to August. The second-order term identifies three of the 77 fits at which the expansion is unreliable, each time for the variance of rare jumps. Because the mode and curvature at those fits are already available, the proposal for importance sampling over the static parameters can be constructed at no additional cost, and the resulting estimator is, to our knowledge, new to nonlinear state space model comparison.

The paper contributes to two strands of literature. The first is simulation-based marginal likelihood estimation, including Chib's method (Chib, 1995; Chib and Jeliazkov, 2001), bridge sampling (Meng and Wong, 1996), importance sampling (Geweke, 1989; Chan and Eisenstat, 2015; Chan, Yu, and Zhang, 2026) and sequential Monte Carlo (Creal, 2012; Chopin, Jacob, and Papaspiliopoulos, 2013). All of these methods simulate over the parameters, and most require posterior draws from MCMC. Our contribution is to approximate the integral over the static parameters directly by a Laplace approximation with a second-order correction.

The second is the literature on Laplace approximations and posterior expansions. The second-order correction is the next term of Laplace's method (Tierney and Kadane, 1986; Kass, Tierney, and Kadane, 1990; Shun and McCullagh, 1995); our contribution is its formulation and error analysis for nonlinear state space models after the states have been integrated out, with the derivatives expressed as smoothing cumulants. This differs from integrated nested Laplace approximation (Rue, Martino, and Chopin, 2009), which approxi-

mates the latent-field integral first, and from the higher-order conditional Laplace approach of Ruli, Sartori, and Ventura (2016). In score-driven filtering, Hansen and Tong (2026) show that the implicit update of Lange, van Os, and van Dijk (2026) is a local approximation to the Bayesian update of a time-varying parameter; our expansion concerns the integral over the static parameters.

The rest of the paper is organized as follows. Section 2 describes the proposed approach. Section 3 states the theoretical results. Section 4 measures the accuracy of the proposed estimator on generated data and how often it selects the true model. Sections 5 and 6 present the applications, and Section 7 concludes.

2 Marginal Likelihood Approximation

The proposed approach combines a grid filter for the integrated likelihood with a Laplace approximation to the integral over the static parameters. We first formulate this integral in unconstrained coordinates. Then we derive the Laplace approximation and its second-order correction in standardized coordinates, and describe the grid filter that evaluates the integrated likelihood at the mode. Finally, we explain how the mode and the derivatives are computed, how the approximation error enters the estimator, and when importance sampling over the static parameters takes the place of the expansion.

Let $\mathbf{y} = (\mathbf{y}_1', \dots, \mathbf{y}_T')'$ denote the observations and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1', \dots, \boldsymbol{\alpha}_T')'$ a latent state path. Conditional on a static parameter vector $\boldsymbol{\theta}$, the states form a Markov chain with transition density $p(\boldsymbol{\alpha}_t \,|\, \boldsymbol{\alpha}_{t-1}, \boldsymbol{\theta})$ and initial density $p(\boldsymbol{\alpha}_1 \,|\, \boldsymbol{\theta})$. The observations are conditionally independent given the states, with density $p(\mathbf{y}_t \,|\, \boldsymbol{\alpha}_t, \boldsymbol{\theta})$. The integrated likelihood and the marginal likelihood are, respectively,

$$Z_T(\boldsymbol{\theta}) = \int p(\mathbf{y} \,|\, \boldsymbol{\alpha}, \boldsymbol{\theta})\, p(\boldsymbol{\alpha} \,|\, \boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\alpha}, \qquad m_T = \int Z_T(\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta},$$

where $p(\boldsymbol{\theta})$ is a proper prior. Model comparison needs to use the integrated likelihood Z_T, because criteria computed from the likelihood conditional on the states are unreliable (see, e.g., Chan and Grant, 2015; Li, Yu, and Zeng, 2020).

Our running example is the stochastic volatility (SV) model

$$y_t = \mathrm{e}^{h_t/2}\varepsilon_t, \qquad h_t = \mu + \phi(h_{t-1} - \mu) + \sigma\eta_t, \qquad h_1 \sim \mathcal{N}\big(\mu, \sigma^2/(1-\phi^2)\big), \tag{1}$$

with ε_t and η_t independent standard normal. The state is the scalar log-volatility h_t, and the static parameters are $\boldsymbol{\theta} = (\mu, \phi, \sigma^2)'$, where ϕ and σ have the constrained supports $|\phi| < 1$ and $\sigma > 0$.

A Gaussian approximation, such as that underlying the Laplace method, does not respect these constrained supports. Therefore, we transform the static parameters to unconstrained coordinates to remove parameter boundaries and potentially reduce posterior asymmetry, thereby improving approximation accuracy. Specifically, let $\boldsymbol{\psi} = g(\boldsymbol{\theta}) \in \mathbb{R}^d$ be a smooth bijection, and let $p_{\boldsymbol{\psi}}(\boldsymbol{\psi})$ denote the transformed prior density. In these coordinates, define the log posterior kernel as

$$f(\boldsymbol{\psi}) = \log Z_T(\boldsymbol{\psi}) + \log p_{\boldsymbol{\psi}}(\boldsymbol{\psi}),$$

where $Z_T(\boldsymbol{\psi})$ denotes the integrated likelihood evaluated at $\boldsymbol{\theta} = g^{-1}(\boldsymbol{\psi})$. The marginal likelihood can then be written as $m_T = \int_{\mathbb{R}^d} \mathrm{e}^{f(\boldsymbol{\psi})}\, \mathrm{d}\boldsymbol{\psi}$. For the stochastic volatility model (1), we use $\boldsymbol{\psi} = (\mu, \operatorname{atanh}\phi, \log\sigma^2)'$, which maps the restricted supports of the persistence parameter $\phi \in (-1, 1)$ and the variance $\sigma^2 > 0$ to the real line. Online Appendix B records all transformations and their Jacobian terms used in this paper.

2.1 Laplace Approximation and the Second-Order Correction

Let $\widehat{\boldsymbol{\psi}}$ denote the posterior mode, the maximizer of the log posterior kernel f. Define $\boldsymbol{\Lambda} = -\nabla^2 f(\widehat{\boldsymbol{\psi}})$ as the negative Hessian matrix at the mode, assumed positive definite, and let $\mathbf{A} = \boldsymbol{\Lambda}^{-1/2}$ be its symmetric inverse square root. The ordinary Laplace approximation replaces the log posterior kernel f by its quadratic expansion around the mode and integrates the resulting Gaussian kernel. This gives the approximation $m_T \approx m_1$, where

$$\begin{aligned} m_1 &= \int_{\mathbb{R}^d} \exp\left\{ f(\widehat{\boldsymbol{\psi}}) - \frac{1}{2}(\boldsymbol{\psi} - \widehat{\boldsymbol{\psi}})' \boldsymbol{\Lambda} (\boldsymbol{\psi} - \widehat{\boldsymbol{\psi}}) \right\} \mathrm{d}\boldsymbol{\psi} \\ &= (2\pi)^{d/2} |\boldsymbol{\Lambda}|^{-1/2} \, \mathrm{e}^{f(\widehat{\boldsymbol{\psi}})} = (2\pi)^{d/2} |\boldsymbol{\Lambda}|^{-1/2} \, p_{\boldsymbol{\psi}}(\widehat{\boldsymbol{\psi}}) \, Z_T(\widehat{\boldsymbol{\psi}}). \end{aligned} \tag{2}$$

This approximation is exact when f is globally quadratic. Otherwise, its error reflects the terms omitted from this quadratic expansion. To isolate these terms, introduce the standardized coordinates $\mathbf{u} = \mathbf{A}^{-1}(\boldsymbol{\psi} - \widehat{\boldsymbol{\psi}})$, or equivalently $\boldsymbol{\psi} = \widehat{\boldsymbol{\psi}} + \mathbf{A}\mathbf{u}$. Define

$$R(\mathbf{u}) = \underbrace{f(\widehat{\boldsymbol{\psi}} + \mathbf{A}\mathbf{u})}_{\text{log posterior kernel}} - \underbrace{\left[f(\widehat{\boldsymbol{\psi}}) - \tfrac{1}{2}\mathbf{u}'\mathbf{u} \right]}_{\text{quadratic approximation}}$$

as the difference between the log posterior kernel and its quadratic approximation in standardized coordinates. It can be shown that the ratio of the marginal likelihood to its ordinary Laplace approximation is

$$\frac{m_T}{m_1} = \mathbb{E}\big[\mathrm{e}^{R(\mathbf{G})}\big], \tag{3}$$

where $\mathbf{G} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. The correction follows by expanding R and integrating the resulting terms against the standard Gaussian density (Barndorff-Nielsen and Cox, 1989, Chapter 6); Kass, Tierney, and Kadane (1990) use the same expansion, in coordinates rescaled by the square root of the sample size, to establish the validity of posterior expansions. Because the linear term vanishes at the mode and the quadratic term has been removed, the Taylor

expansion starts at order three:

$$R(\mathbf{u}) = C_3(\mathbf{u}) + C_4(\mathbf{u}) + \cdots, \qquad C_k(\mathbf{u}) = \frac{1}{k!}\, D^k f(\widehat{\boldsymbol{\psi}})[\mathbf{Au}, \ldots, \mathbf{Au}],$$

where $D^k f(\widehat{\boldsymbol{\psi}})[\cdot]$ is the kth derivative of f at $\widehat{\boldsymbol{\psi}}$ as a multilinear form. The coefficients of C_3 and C_4 are the standardized derivative tensors

$$a_{ijk} = D^3 f(\widehat{\boldsymbol{\psi}})[\mathbf{Ae}_i, \mathbf{Ae}_j, \mathbf{Ae}_k], \qquad b_{ijkl} = D^4 f(\widehat{\boldsymbol{\psi}})[\mathbf{Ae}_i, \mathbf{Ae}_j, \mathbf{Ae}_k, \mathbf{Ae}_l],$$

where $\mathbf{e}_i$ is the ith column of $\mathbf{I}_d$, so that $\mathbf{Ae}_i$ is the ith column of $\mathbf{A}$. Taking the expectation in (3) term by term, the odd-order terms vanish by the symmetry of the Gaussian, and the leading correction to $m_T/m_1 = 1$ is $\mathbb{E}[C_4(\mathbf{G})] + \frac{1}{2}\mathbb{E}[C_3(\mathbf{G})^2]$. In terms of the tensors, this second-order correction is

$$\Delta_2 = \frac{1}{8}\sum_{i,j} b_{iijj} + \frac{1}{12}\sum_{i,j,k} a_{ijk}^2 + \frac{1}{8}\sum_i \Big(\sum_j a_{ijj}\Big)^2, \tag{4}$$

where the first sum is $\mathbb{E}[C_4(\mathbf{G})]$ and the second and third sums together are $\frac{1}{2}\mathbb{E}[C_3(\mathbf{G})^2]$, the two contractions of the square of the cubic form under the Gaussian sixth moment. In the usual case in which $\boldsymbol{\Lambda}$ grows linearly in T, the standardized tensors satisfy $a_{ijk} = O(T^{-1/2})$ and $b_{ijkl} = O(T^{-1})$, so $\Delta_2 = O(T^{-1})$ and the terms left out are of order T^{-2}. Our main theorem in Section 3 gives conditions under which $\log m_T = \log m_1 + \Delta_2 + O(T^{-2})$. This correction is well known. The three contractions in (4) are the level-one terms of the expansion of Shun and McCullagh (1995) for integrals in d dimensions, enumerated by McCullagh (1987, pp. 254–256). For $d = 1$, (4) reduces to the correction $f^{(4)}(\widehat{\psi})/(8\Lambda^2) + 5f^{(3)}(\widehat{\psi})^2/(24\Lambda^3)$ of Tierney and Kadane (1986), with $\Lambda = -f''(\widehat{\psi})$.

The proposed estimator replaces $Z_T(\widehat{\boldsymbol{\psi}})$ in (2) by an evaluation $\widehat{Z}_T(\widehat{\boldsymbol{\psi}})$ and adds the correction:

$$\log \widehat{m}_2 = \tfrac{d}{2}\log(2\pi) - \tfrac{1}{2}\log|\boldsymbol{\Lambda}| + \log p_{\boldsymbol{\psi}}(\widehat{\boldsymbol{\psi}}) + \log \widehat{Z}_T(\widehat{\boldsymbol{\psi}}) + \Delta_2. \tag{5}$$

We write $\widehat{m}_1$ for the same expression without Δ_2. The subscripts distinguish the ordinary and corrected approximations, while the hats indicate use of an evaluated likelihood. Implementation combines an evaluator of the integrated likelihood with numerical optimization and derivative calculations.

2.2 Evaluating the Integrated Likelihood at the Mode

The proposed approach requires an evaluator that is accurate at one point and smooth enough in $\boldsymbol{\psi}$ for the derivatives in (4) to be computed by finite differences. Several established methods can compute the integrated likelihood. Numerical-integration filters propagate the filtering density on a grid (Kitagawa, 1987), and exact filters exist for special structures, such as the noncentral-gamma state dynamics of Creal (2017).[1] Importance samplers for the state path (Durbin and Koopman, 1997) and particle filters (see Creal, 2012, for a survey) give simulation estimates, but a simulation estimate is noisy and its noise enters the finite differences directly.[2] We therefore use a deterministic evaluator, the numerical-integration filter on a grid. The models in this paper have scalar or bivariate states; a trivariate state is also feasible, and the cost rises by a factor equal to the number of grid points in the third dimension.

For ease of exposition, consider a scalar state, and let $\alpha_i = \alpha_{\min} + i\,\delta$, $i = 1, \ldots, N$, be equally spaced grid points. Let $\mathbf{K}$ be the $N \times N$ matrix with entries $K_{ij} = p(\alpha_i \,|\, \alpha_j, \boldsymbol{\theta})\,\delta$, the transition density between grid points times the trapezoidal weight δ, and let $\mathbf{D}_t = \operatorname{diag}\big(p(\mathbf{y}_t \,|\, \alpha_1, \boldsymbol{\theta}), \ldots, p(\mathbf{y}_t \,|\, \alpha_N, \boldsymbol{\theta})\big)$ be the diagonal matrix of the observation densities at the grid points. Further define

$$\mathbf{a}_1 = \mathbf{D}_1 \left(p(\alpha_1 \,|\, \boldsymbol{\theta}), \ldots, p(\alpha_N \,|\, \boldsymbol{\theta})\right)', \qquad \mathbf{a}_t = \mathbf{D}_t \mathbf{K} \mathbf{a}_{t-1}, \quad t = 2, \ldots, T,$$

[1] Numerical-integration filters are also developed or used by Fridman and Harris (1998), Watanabe (1999), Langrock, MacDonald, and Zucchini (2012), Bégin and Boudreault (2021) and Farmer (2021).

[2] Importance samplers for the state path include those of Daníelsson and Richard (1993), Shephard and Pitt (1997), Richard and Zhang (2007), McCausland (2012) and Chan (2023); particle filters include those of Andrieu, Doucet, and Holenstein (2010) and Pitt, dos Santos Silva, Giordani, and Kohn (2012).

so that $\mathbf{a}_t$ approximates the joint density $p(\mathbf{y}_1, \ldots, \mathbf{y}_t, \alpha_t = \alpha_i)$ of the observations to date and the state at the grid points, and $Z_T(\boldsymbol{\theta}) \approx \delta\, \mathbf{1}'\mathbf{a}_T$. To avoid underflow, each $\mathbf{a}_t$ is divided by $c_t = \delta\, \mathbf{1}'\mathbf{a}_t$, the one-step-ahead predictive density $p(\mathbf{y}_t \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_{t-1}, \boldsymbol{\theta})$, after which $\mathbf{a}_t$ is the filtering density at the grid points and $\log Z_T(\boldsymbol{\theta}) \approx \sum_{t=1}^{T} \log c_t$. This is the numerical-integration filter of Kitagawa (1987), the grid-based filter of the nonlinear filtering literature; we call it the grid filter. To our knowledge, it has not been used for marginal likelihood evaluation.

The filter is accurate because the integrand at each step is analytic and decays in both tails. For such an integrand, the error of the trapezoidal rule decays exponentially as the spacing δ shrinks (Trefethen and Weideman, 2014), so a spacing of a fraction of the width of the transition and observation densities gives $\log Z_T$ to machine precision. We provide further details of the grid design and accuracy checks in Online Appendix B. Each step requires a multiplication by $\mathbf{K}$, so the cost is linear in T and in the number of grid points.

2.3 Mode Finding and the Approximation Error

We first describe how we compute the posterior mode $\widehat{\boldsymbol{\psi}}$, the curvature $\boldsymbol{\Lambda}$ and, for $\widehat{m}_2$, the third and fourth derivatives in (4). We then outline how the errors of these quantities and of the likelihood evaluation enter the proposed estimator. With the grid filter, $\log Z_T$ is deterministic and can be evaluated to an accuracy of 10^{-11}. We use Newton–Raphson iterations in the standardized coordinates $\mathbf{u}$, with gradients and Hessians computed by central differences. The negative Hessian at the mode is $\boldsymbol{\Lambda}$. The correction (4) needs only the third derivatives a_{ijk} and the fourth derivatives of the form b_{iijj}, which we compute by central differences on a sparse design in $\mathbf{u}$. The whole computation requires a few hundred to several thousand evaluations of $\log Z_T$. The evaluations within each Newton iteration and within the derivative calculation are independent, so they can be run in parallel batches. With enough cores, the wall-clock cost is a few dozen evaluation times. Online Appendix B

provides the implementation details and the evaluation counts.

Given $\widehat{\boldsymbol{\psi}}$, $\boldsymbol{\Lambda}$ and Δ_2, the error of the proposed estimator is exactly

$$\log \widehat{m}_2 - \log m_T = \underbrace{\log \widehat{Z}_T(\widehat{\boldsymbol{\psi}}) - \log Z_T(\widehat{\boldsymbol{\psi}})}_{\text{evaluation}} + \underbrace{\Delta_2 - \log \mathbb{E}\left[\mathrm{e}^{R(\mathbf{G})}\right]}_{\text{approximation}}. \tag{6}$$

The evaluation term is zero to machine precision with the grid filter. The error of the estimator is therefore the approximation term, the error of the second-order Laplace expansion, which is deterministic given the data. We show in the next section that under regularity conditions on f it is of order T^{-2}, while Δ_2 is of order T^{-1}. The remaining error is therefore of the order of Δ_2^2. The size of the second-order term thus indicates the accuracy of the estimate for the data at hand. We report it with every estimate. Because (6) holds for any $\widehat{\boldsymbol{\psi}}$ and positive definite $\boldsymbol{\Lambda}$, an inexact mode or curvature enters through the approximation term and adds no separate source of error.[3]

The second-order term also serves as a diagnostic for the reliability of the expansion. The remainder is of the order of Δ_2^2, so when $|\Delta_2|$ exceeds one the indicated remainder is of the same order as the correction itself, and the expansion is not reliable. In such cases we keep $\widehat{\boldsymbol{\psi}}$ and $\boldsymbol{\Lambda}$ and estimate the integral over the static parameters by importance sampling from a Student-t proposal centered at the mode in the standardized coordinates (Geweke, 1989); the sampler requires no fitting step and no sampling over the states. To our knowledge, this importance-sampling estimator of the marginal likelihood has not been used for nonlinear state space models.

[3]The finite-difference errors in the derivatives are of the order of the remainder and are reduced by extrapolation; see Online Appendix B.

3 Theoretical Results

We provide the theoretical results behind the approximation term in (6). Proposition 1 expresses the derivatives of the log integrated likelihood that enter the second-order term as cumulants under the smoothing distribution of the states. Theorem 1 states the conditions under which the corrected approximation has an error of order T^{-2}.

Let $\mathcal{C}(\boldsymbol{\psi}, \boldsymbol{\alpha}) = \log p(\mathbf{y} \,|\, \boldsymbol{\alpha}, \boldsymbol{\psi}) + \log p(\boldsymbol{\alpha} \,|\, \boldsymbol{\psi})$ denote the complete-data log density, so that $Z_T(\boldsymbol{\psi}) = \int \mathrm{e}^{\mathcal{C}(\boldsymbol{\psi}, \boldsymbol{\alpha})} \, \mathrm{d}\boldsymbol{\alpha}$. Let $p(\boldsymbol{\alpha} \,|\, \mathbf{y}, \boldsymbol{\psi})$ denote the smoothing distribution. For a fixed direction $\mathbf{w} \in \mathbb{R}^d$, write $c_j = D^j_{\mathbf{w}} \, \mathcal{C}(\boldsymbol{\psi}, \boldsymbol{\alpha})$ for the jth directional derivative in $\boldsymbol{\psi}$ at fixed $\boldsymbol{\alpha}$. Write $\mathbb{E}_{\boldsymbol{\psi}}$, $\mathrm{Var}_{\boldsymbol{\psi}}$, $\mathrm{Cov}_{\boldsymbol{\psi}}$ and $\mathrm{cum}_{\boldsymbol{\psi}}$ for moments and joint cumulants under $p(\boldsymbol{\alpha} \,|\, \mathbf{y}, \boldsymbol{\psi})$. The third-order joint cumulants are central moments, $\mathrm{cum}_{\boldsymbol{\psi}}(c_1, c_1, c_2) = \mathbb{E}_{\boldsymbol{\psi}}[(c_1 - \mathbb{E}_{\boldsymbol{\psi}} c_1)^2 (c_2 - \mathbb{E}_{\boldsymbol{\psi}} c_2)]$, and $\mathrm{cum}_{\boldsymbol{\psi}}(c_1, c_1, c_1, c_1) = \mathbb{E}_{\boldsymbol{\psi}}[(c_1 - \mathbb{E}_{\boldsymbol{\psi}} c_1)^4] - 3\mathrm{Var}_{\boldsymbol{\psi}}(c_1)^2$. Differentiating Z_T under the integral sign gives the score identity $D_{\mathbf{w}} \log Z_T = \mathbb{E}_{\boldsymbol{\psi}}[c_1]$ and the observed-information identity of Louis (1982), $D^2_{\mathbf{w}} \log Z_T = \mathbb{E}_{\boldsymbol{\psi}}[c_2] + \mathrm{Var}_{\boldsymbol{\psi}}(c_1)$. The next result extends them to the third and fourth derivatives.

Proposition 1 (Smoothing cumulants)**.** *For a fixed $\boldsymbol{\psi}$, suppose that $\mathcal{C}$ is four times continuously differentiable in its first argument, and that $\mathbb{E}_{\boldsymbol{\psi}'} \exp(\sum_{j=1}^4 t_j c_j)$, with the c_j evaluated at $\boldsymbol{\psi}'$, is finite for $(\mathbf{t}, \boldsymbol{\psi}')$ in a neighborhood of $(\mathbf{0}, \boldsymbol{\psi})$ and can be differentiated under the integral sign there. Then, at $\boldsymbol{\psi}$,*

$$
\begin{aligned}
D^3_{\mathbf{w}} \log Z_T &= \mathbb{E}_{\boldsymbol{\psi}}[c_3] + 3\,Cov_{\boldsymbol{\psi}}(c_1, c_2) + cum_{\boldsymbol{\psi}}(c_1, c_1, c_1), \\
D^4_{\mathbf{w}} \log Z_T &= \mathbb{E}_{\boldsymbol{\psi}}[c_4] + 4\,Cov_{\boldsymbol{\psi}}(c_1, c_3) + 3\,Var_{\boldsymbol{\psi}}(c_2) + 6\,cum_{\boldsymbol{\psi}}(c_1, c_1, c_2) \\
&\quad + cum_{\boldsymbol{\psi}}(c_1, c_1, c_1, c_1).
\end{aligned} \tag{7}
$$

The identities in (7) follow from the general rule for the derivative of a joint cumulant

(Skovgaard, 1986),

$$D_{\mathbf{w}}\mathrm{cum}_{\boldsymbol{\psi}}(g_1,\ldots,g_k)=\sum_{i=1}^{k}\mathrm{cum}_{\boldsymbol{\psi}}(g_1,\ldots,D_{\mathbf{w}}g_i,\ldots,g_k)+\mathrm{cum}_{\boldsymbol{\psi}}(g_1,\ldots,g_k,c_1), \tag{8}$$

which for $k=1$ is $D_{\mathbf{w}}\mathbb{E}_{\boldsymbol{\psi}}[g]=\mathbb{E}_{\boldsymbol{\psi}}[D_{\mathbf{w}}g]+\mathrm{Cov}_{\boldsymbol{\psi}}(g,c_1)$. We prove the proposition in Online Appendix A.

We assume a standard regularity condition of Laplace's method, the counterpart of a unique, interior maximum likelihood estimator with a nonsingular information matrix. In the unconstrained coordinates $\boldsymbol{\psi}$ a maximum is interior whenever it exists.

Assumption 1 (Mode)**.** *The function f is twice continuously differentiable, attains its maximum on $\mathbb{R}^d$ only at $\widehat{\boldsymbol{\psi}}$, and $\boldsymbol{\Lambda}=-\nabla^2 f(\widehat{\boldsymbol{\psi}})$ is positive definite.*

The next assumption is stated on the set $\mathcal{N}_T=\{\widehat{\boldsymbol{\psi}}+\mathbf{A}\mathbf{u}:\|\mathbf{u}\|\leqslant\varrho\sqrt{T}\}$ for a fixed $\varrho>0$.

Assumption 2 (Derivatives)**.** *The function f is six times continuously differentiable on an open set containing $\mathcal{N}_T$, and there are constants $K_3,\ldots,K_6$ not depending on T, with $\varrho K_3\leqslant 3/2$, such that, for $k=3,\ldots,6$,*

$$\sup_{\boldsymbol{\psi}\in\mathcal{N}_T}\sup_{\|\mathbf{w}\|=1}\left|D^k f(\boldsymbol{\psi})[\mathbf{A}\mathbf{w},\ldots,\mathbf{A}\mathbf{w}]\right|\leqslant K_k\,T^{1-k/2}.$$

Assumption 2 fixes the rates at which the derivatives of f grow with T, in the standardized coordinates in which the Laplace approximation is a standard Gaussian. Along any unit direction the kth standardized derivative is of order $T^{1-k/2}$: the third is $O(T^{-1/2})$, the fourth $O(T^{-1})$, and the fifth and sixth $O(T^{-3/2})$ and $O(T^{-2})$. These are the rates of a log-likelihood built from T weakly dependent observations, and they make the expansion of R in (3) an expansion in powers of $T^{-1/2}$, with $C_k=O(T^{1-k/2})$, $\Delta_2=O(T^{-1})$ and a remainder of order T^{-2}. The bounds are required uniformly on $\mathcal{N}_T$ because the proof uses them both on the ball in $\mathbf{u}$ where it expands f and on the shell between that ball and $\mathcal{N}_T$.

The rates hold when the information grows linearly in T. If the eigenvalues of $\mathbf{\Lambda}$ are of order T, then $\|\mathbf{A}\|$ is of order $T^{-1/2}$, $\mathcal{N}_T$ has a fixed size in the original coordinates, and the assumption follows when the unstandardized derivatives of f up to order six are of order T there. For the likelihood part of f these derivatives are the smoothing cumulants of Proposition 1 and their analogues up to order six, which are of order T when the derivatives c_j have moments bounded uniformly on $\mathcal{N}_T$ and the smoothing distribution forgets the past at a geometric rate, that is, when the dependence between the states at dates s and t decays like $\rho^{|t-s|}$ for some $\rho < 1$; the prior does not depend on T and its log density is smooth, so it contributes derivatives that are bounded on $\mathcal{N}_T$. The condition $\varrho K_3 \leqslant 3/2$ is a choice of ϱ, since K_3 cannot increase when ϱ decreases; the price is that the next assumption is then required on a smaller set.

Assumption 3 (Concentration)**.** *There is a constant C_0 not depending on T such that $\int_{\mathcal{N}_T^c} e^{f(\boldsymbol{\psi})}\, d\boldsymbol{\psi} \leqslant C_0\, m_1\, T^{-2}$ for all large T.*

Assumption 3 is a weak form of posterior concentration. Under the Laplace regularity conditions of Kass, Tierney, and Kadane (1990) it holds with an exponential rate in place of T^{-2}.

Theorem 1 (Second-order expansion)**.** *Under Assumptions 1–3,*

$$\log m_T = \log m_1 + \Delta_2 + O(T^{-2}),$$

with Δ_2 as in (4)*, and $\Delta_2 = O(T^{-1})$. The implied constant and the threshold on T depend only on d, ϱ, $K_3, \ldots, K_6$ and C_0.*

We prove the theorem in Online Appendix A. The proof expands R to sixth order on a ball of radius $T^{1/8}$ in $\mathbf{u}$, where the Gaussian expectation of the expansion gives Δ_2 and all other terms are of order T^{-2}. It then bounds the contributions of the shell between that ball and $\mathcal{N}_T$, and of $\mathcal{N}_T^c$. The argument follows Kass, Tierney, and Kadane (1990), with the

regularity conditions written in standardized coordinates.

4 Accuracy and Model Selection on Generated Data

In this section we measure the accuracy of the proposed approach against quadrature benchmarks and compare it with three commonly used simulation-based estimators at equal computing time. We then generate data from three bivariate stochastic volatility models, which we also compare in the exchange-rate application, and record how often the true model has the largest estimated marginal likelihood.

4.1 Design and Benchmark

We consider four commonly used nonlinear or non-Gaussian state space models. The first is the stochastic volatility (SV) model in (1). The other three vary only in their observation equations: the Student-t (SV-t) model has $\varepsilon_t \sim \mathcal{T}_\nu$ with unit scale, the stochastic volatility in mean (SV-M) model (Koopman and Hol Uspensky, 2002; Chan, 2017) has $y_t = \lambda \mathrm{e}^{h_t} + \mathrm{e}^{h_t/2}\varepsilon_t$, and the Poisson model has counts $y_t \sim \text{Poisson}(\mathrm{e}^{h_t})$.

Data are generated from the stochastic volatility model with $\mu = 0$ and (ϕ, σ) equal to $(0.90, 0.35)$, $(0.95, 0.25)$ or $(0.98, 0.15)$. The related models use $(\phi, \sigma) = (0.95, 0.25)$, with $\nu = 5$, $\lambda = 0.5$ and, for the Poisson model, $\mu = 2$ and $\sigma = 0.22$. We consider five sample sizes, $T \in \{500, 1{,}000, 2{,}000, 4{,}000, 8{,}000\}$, and take the first T observations of each generated path, so the samples are nested. We generate ten replications for each of the six model and parameter combinations, and the five sample sizes then give 300 data sets. The replications show the size of the approximation error and its dependence on T. The ranges reported below are ranges over replications, not confidence intervals.

The benchmark value of $\log m_T$ for each data set comes from a nested quadrature over the three or four coordinates of $\boldsymbol{\psi}$, a trapezoidal rule in $\operatorname{atanh}\phi$ with a Gauss–Hermite rule over

the remaining coordinates at each node. It is very accurate for two reasons. The integrand at each node is the integrated likelihood from the grid filter, which is accurate to 10^{-11}, so the only error is that of the outer quadrature. Since the outer quadrature is over a few smooth coordinates, refining its rules changes the value by less than 10^{-4} (see Online Appendix B). The proposed approach uses the same grid filter: the optimization routine gives $\widehat{\boldsymbol{\psi}}$, $\boldsymbol{\Lambda}$ and the derivatives in (4). The residual $\log m_T - \log m_1 - \Delta_2$ reported below is the negative of the approximation term in (6).

Finally, we consider the following priors: $\mu \sim \mathcal{N}(0, 10)$, $(1 + \phi)/2 \sim \mathcal{B}(20, 1.5)$ and $\sigma^2 \sim \mathcal{IG}(2.5, 0.1)$ in every model; $\lambda \sim \mathcal{N}(0, 1)$ in the stochastic volatility in mean model; and $\log(\nu - 2) \sim \mathcal{N}(2, 1)$ in the Student-t model. The coordinates are $\boldsymbol{\psi} = (\mu, \operatorname{atanh} \phi, \log \sigma^2)'$, extended by $\psi_4 = \lambda$ or $\psi_4 = \log(\nu - 2)$ in the two models with $d = 4$ parameters, and the Jacobian of the transform is in $p_{\boldsymbol{\psi}}$.

4.2 Accuracy of the Proposed Estimator

We first report the accuracy of the proposed estimator under the simulation design discussed above. Specifically, for each model, persistence level and sample size, Figure 1 plots the Laplace error $\log(m_T/m_1)$ and the residual $\log(m_T/m_1) - \Delta_2$ after the second-order term, as medians and ranges over the replications. The errors are in the logarithm, so they are approximately relative errors in m_T. For example, 0.01 corresponds to an error of 1% in the marginal likelihood and in any Bayes factor computed from it.

The left panel of Figure 1 shows that the Laplace approximation is reasonably accurate for every model and persistence level. Measured against the quadrature benchmark, it understates the log marginal likelihood by at most 0.33 at $T = 500$, 0.09 at $T = 4{,}000$ and 0.05 at $T = 8{,}000$. The error increases with the persistence, and the three related models behave like the stochastic volatility model at the same persistence.

The right panel shows that the second-order term removes between 64% and 99% of this

error across the data sets. The residual falls roughly at the rate T^{-2} of Theorem 1, at which each doubling of T divides it by four. The second-order term also indicates the size of the residual: in all 300 data sets the residual lies between 0.5 and 7 times Δ_2^2, with median 1.5, so we use about $2\Delta_2^2$ as the indication of the remaining error. The second-order term Δ_2 itself is at most 0.25 in these data sets, so the indicated error is at most about 0.12, and no data set comes near the threshold of one at which the expansion stops being reliable.

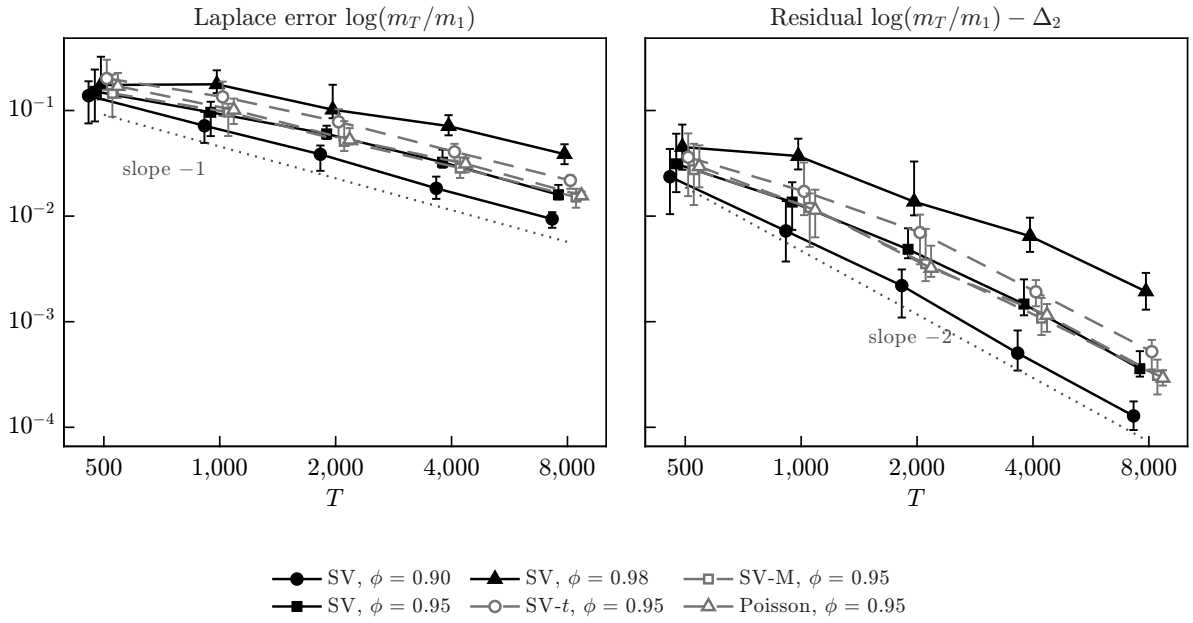


Figure 1: Laplace error $\log(m_T/m_1)$ (left) and residual after the second-order term $\log(m_T/m_1) - \Delta_2$ (right) against the sample size, on logarithmic scales. Markers are medians over the replications and bars their ranges, for the SV, SV-t, SV-M and Poisson models. The dotted lines have slopes -1 and -2.

4.3 Comparison with Simulation-Based Estimators

We now compare the proposed estimator with three commonly used marginal likelihood estimators on data sets simulated from the stochastic volatility model in Section 4.1: importance sampling with the cross-entropy method of Chan and Eisenstat (2015), Chib's method (Chib, 1995), and the modified harmonic mean of Gelfand and Dey (1994) with the weighting density of Geweke (1999). All three require posterior draws from MCMC.

The three estimators share one posterior sampler, the auxiliary mixture sampler of Kim, Shephard, and Chib (1998). We obtain 20,000 posterior draws after a burn-in period of 5,000. The cross-entropy method fits a Gaussian proposal in $\boldsymbol{\psi}$ to these draws. It then evaluates the integrated likelihood at 10,000 draws from this proposal. Chib's method computes the posterior ordinate at the posterior mean of the draws, from the full run and two reduced runs. The Gelfand–Dey estimator evaluates the integrated likelihood at each of the 20,000 posterior draws, as Chan and Grant (2015) recommend for latent variable models. Because the auxiliary mixture sampler targets the mixture model, these two estimators reweight the draws to the exact model as in Kim, Shephard, and Chib (1998); the importance weights of the cross-entropy method correct for the mixture approximation directly. All four estimators use the grid filter for the integrated likelihood, so the comparison isolates the integration over the parameters.

To compare the estimators at equal cost, we rescale the numerical standard errors of the three simulation-based estimators to a computing budget of one minute on one core. We measure each standard error with the fixed numbers of draws given above, together with the computing time of each estimator on one core.[4] Since the number of draws is proportional to the computing time, a numerical standard error is inversely proportional to the square root of the computing time. Figure 2 plots these rescaled standard errors next to the error of the proposed estimator, by sample size and persistence level.

At this budget the error of the proposed estimator falls as T grows. The errors of Chib's method and the Gelfand–Dey estimator rise steeply with T, and that of the cross-entropy method rises by a factor of about two to five. The proposed estimator is therefore the most accurate of the four from $T = 1{,}000$ at $\phi = 0.90$ and from $T = 2{,}000$ at $\phi \geqslant 0.95$. At $T = 4{,}000$ its residual is 2 to 47 times smaller than the standard error of the cross-entropy method, the most accurate of the three simulation-based estimators, and at $T = 8{,}000$ it is

[4]The computations of this section are done in MATLAB. The timings are single-thread times on a laptop computer with an Intel Core Ultra 7 255U processor.

9 to 400 times smaller.

The simulation-based estimators lose accuracy as T grows for two reasons. First, each MCMC sweep and each likelihood evaluation takes longer at larger T. A fixed budget therefore allows fewer draws as T grows, and from $T = 4{,}000$ the three runs of Chib's method alone exceed it. The proposed estimator requires a number of likelihood evaluations that does not grow with T, and it takes 2.5 to 63 seconds, within the budget at $T \leqslant 4{,}000$ and about the budget at $T = 8{,}000$. Second, the reweighting of Chib's method and the Gelfand–Dey estimator to the exact model keeps a smaller share of the draws as T grows: at $T = 4{,}000$ the effective sample size is 0.1 to 3 percent of the draws, and at $T = 8{,}000$ it is below 0.5 percent. Their standard errors therefore grow with T even at a fixed number of draws, and at $T \geqslant 4{,}000$ their realized errors exceed these standard errors by up to a factor of six (see Online Appendix B). The cross-entropy method requires no reweighting. Its standard error at a fixed number of draws falls with T as the posterior becomes closer to Gaussian, which offsets much of the rising cost of each draw.

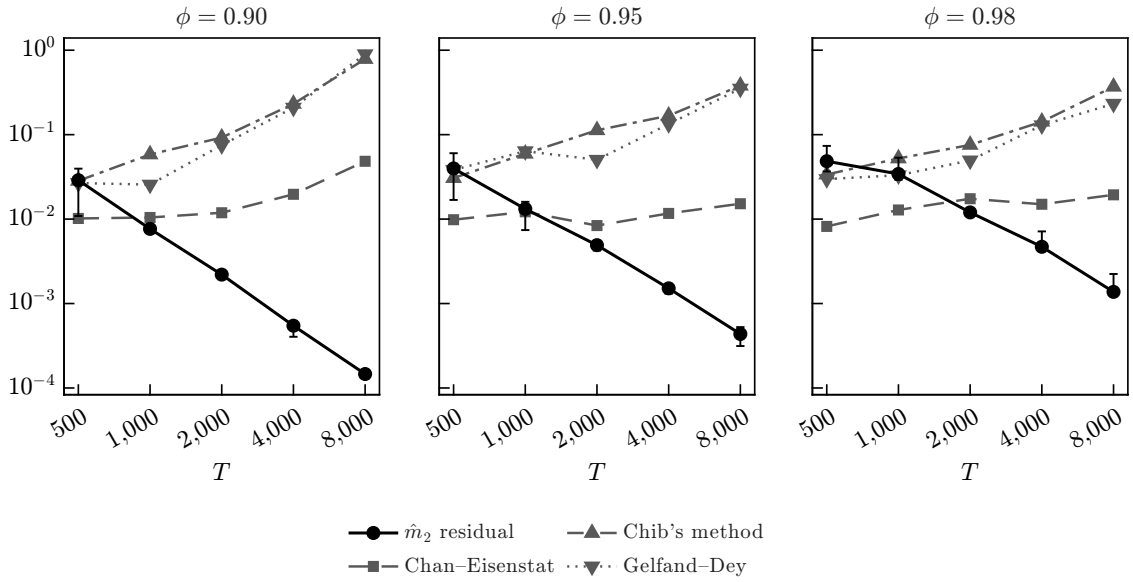


Figure 2: Error of the four marginal likelihood estimators (in log scale) at a computing budget of one minute on one core, against the sample size, for the stochastic volatility model at three persistence levels. For $\widehat{m}_2$ the error is the residual, as median and range over the replications; for the other estimators it is the numerical standard error scaled to the budget.

Overall, these results show that the proposed estimator is accurate for long time series at a small computing cost. Its error is roughly proportional to T^{-2}, and it takes about a minute or less on one core at every sample size. The simulation-based estimators require longer runs at larger T to reach a given accuracy. The advantage of the proposed estimator therefore grows with T.

4.4 Model Selection for Bivariate Stochastic Volatility

The Monte Carlo experiments above use scalar states, for which a quadrature benchmark exists. The models of the exchange-rate application have bivariate states and seven or eight static parameters, for which no benchmark is available. We assess how often the proposed estimator selects the true model by marginal likelihood. We compare three of the bivariate stochastic volatility models of Yu and Meyer (2006), members of the multivariate stochastic volatility family reviewed by Asai, McAleer, and Yu (2006). These models are costly to estimate by MCMC. No auxiliary mixture samplers in the tradition of Kim, Shephard, and Chib (1998) have been developed for them. Yu and Meyer (2006) estimate these models using single-move samplers in WinBUGS—which Asai, McAleer, and Yu (2006) note can be numerically inefficient—and compare them by the deviance information criterion.

We write a pair of returns as $\mathbf{y}_t = (y_{1t}, y_{2t})'$ with $y_{it} = \mathrm{e}^{h_{it}/2}\varepsilon_{it}$, $i = 1, 2$. In the constant-correlation model (CC), the two log-volatilities follow independent AR(1) processes as in (1), and $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \varepsilon_{2t})' \sim \mathcal{N}(\mathbf{0}, \boldsymbol{P})$, where $\boldsymbol{P}$ is a correlation matrix with off-diagonal element ρ. CC has a two-dimensional state and seven static parameters. Each of the other two models changes one feature of CC. The Granger-causality model (GC) adds the lagged first log-volatility to the second transition, $h_{2,t+1} = \mu_2 + \phi_2(h_{2t} - \mu_2) + \phi_{21}(h_{1t} - \mu_1) + \sigma_2\eta_{2t}$. The heavy-tailed model (CC-t) replaces the Gaussian $\boldsymbol{\varepsilon}_t$ by a bivariate Student-t with scale matrix $\boldsymbol{P}$ and $\nu > 2$ degrees of freedom, so that $\mathrm{Var}(\boldsymbol{\varepsilon}_t) = \nu\boldsymbol{P}/(\nu - 2)$. Each has eight static parameters.

We use the same priors as in Section 4.1 for $(\mu_i, \phi_i, \sigma_i^2)$. The other priors are $\rho \sim \mathcal{U}(-1, 1)$, $\phi_{21} \sim \mathcal{N}(0, 1)$ and $\log(\nu - 2) \sim \mathcal{N}(2, 1)$. The coordinates of CC are $\boldsymbol{\psi} = (\mu_1, \operatorname{atanh} \phi_1, \log \sigma_1^2, \mu_2, \operatorname{atanh} \phi_2, \log \sigma_2^2, \operatorname{atanh} \rho)'$. GC adds $\psi_8 = \phi_{21}$ and CC-t adds $\psi_8 = \log(\nu - 2)$. We generate data from each of the three models, with the parameters at their posterior modes for the exchange-rate data of Section 5. For each model we generate 50 series of $T = 1{,}000$ observations, a quarter of the exchange-rate sample, which makes the models harder to distinguish. We fit all three models to every series, with $\log \widehat{m}_2$ from the grid filter as the log marginal likelihood.

For each series, Figure 3 plots the log Bayes factors of the other two models against the true model. The true model is selected in 149 of the 150 series. The exception is a series generated from CC, on which GC is selected by 0.66—posterior odds of about two to one. Both other models include CC as a special or limiting case: GC when $\phi_{21} = 0$ and CC-t as $\nu \to \infty$. On data from CC, the more flexible models fit as well as CC. They have only one additional static parameter, so the penalty for their complexity is modest. The median log Bayes factors against CC are therefore only -4.1 for GC and -3.3 for CC-t. In contrast, when the true model has a feature that the other two lack, it is selected in every series. For data from GC and CC-t, the median log Bayes factors against the true model range from -16.0 to -20.5, and the 95th percentiles are -6.8 or below.

These results show that the proposed estimator reliably identifies the true model among the three bivariate specifications. The one exception reflects a property of the Bayes factor: when the true model is nested in a more flexible one, the evidence for the restricted model accumulates only slowly with the sample size.

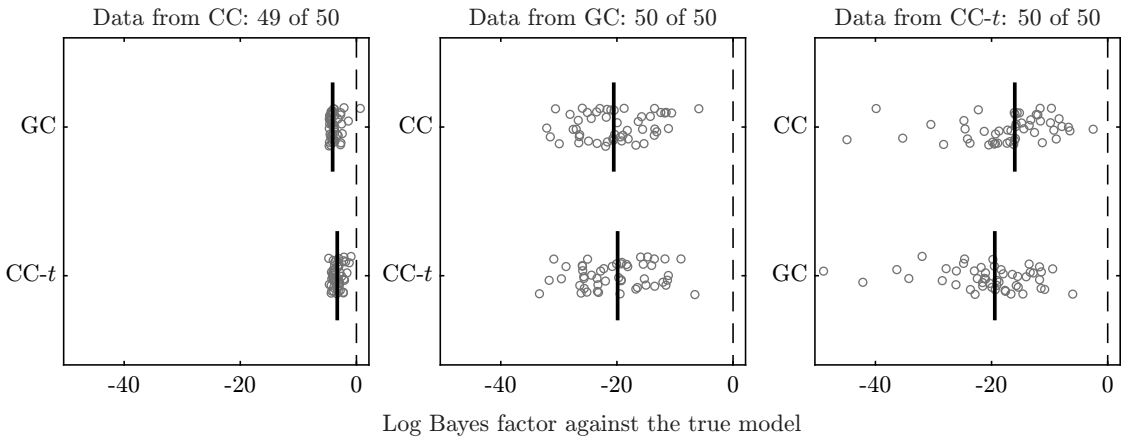


Figure 3: Log Bayes factors of the other two models against the true model on 50 data sets of $T = 1{,}000$ observations generated from CC (left), GC (middle) and CC-t (right). Each dot is one series, jittered vertically, and the bar is the median; values below zero favor the true model.

5 Application: Bivariate SV for Exchange Rates

This section applies the proposed approach to three bivariate stochastic volatility models of Yu and Meyer (2006) for two exchange rates. The models have bivariate states and seven or eight static parameters. To our knowledge, the marginal likelihood has never been reported for these models. Posterior sampling for them is itself difficult—no auxiliary mixture sampler exists, and the existing samplers are single-move and numerically inefficient (Asai, McAleer, and Yu, 2006). We describe the data and the models, then present the Bayes factors and the computing cost.

5.1 Data and Models

We use daily log returns on the Australian and New Zealand dollars against the US dollar, constructed from the European Central Bank's daily euro reference rates. The sample spans 4,000 trading days from February 16, 2011, to October 2, 2026. Returns are expressed in percent and demeaned over the full ECB series, from January 1999 to October 2026. The two series have standard deviations of approximately 0.65 and 0.68 percent and a sample

correlation of 0.83.

We compare the three bivariate stochastic volatility models introduced in Section 4.4. They share the observation equation $y_{it} = \mathrm{e}^{h_{it}/2}\varepsilon_{it}$ and first-order autoregressive log-volatility dynamics. The baseline constant-correlation model (CC) assumes Gaussian return shocks with correlation ρ. Each alternative extends the baseline in one direction. The Granger-causality model (GC) allows lagged AUD log-volatility to enter the NZD log-volatility equation. The heavy-tailed model (CC-t) replaces the Gaussian return shocks with a bivariate Student-t distribution with ν degrees of freedom.

The integrated likelihood of each model is evaluated on a product grid in (h_{1t}, h_{2t}), and the derivatives in (4) come from the sparse design outlined in Section 2.3.[5] Online Appendix B provides the details of the optimization settings.

5.2 Bayes Factors and Computing Cost

Both extensions receive strong support relative to the constant-correlation baseline, and heavy-tailed observations provide the larger improvement. Table 1 shows log Bayes factors against CC of 110.3 for CC-t and 97.7 for GC. The heavy-tailed model is the best supported model of the three, whereas Yu and Meyer (2006), who use weekly returns, rank the heavy-tailed model below CC by the deviance information criterion (DIC). This is not surprising as daily returns typically have heavier tails than weekly returns. The parameter estimates help explain these gains. CC-t accommodates heavy tails with a degrees-of-freedom parameter of $\nu = 7.9$, while leaving the volatility parameters close to their baseline estimates. In addition, the data show a strong volatility spillover from the Australian dollar to the New Zealand dollar ($\phi_{21} = 0.83$), in line with their results. Once this spillover is taken into account, the coefficient ϕ_2 on the own lag of NZD log-volatility falls to 0.003.

[5]The transition kernel of the grid filter is the product of two one-dimensional banded kernels, or a block-banded kernel for GC. One evaluation takes between 4 and 14 seconds on one core at this T, and the evaluations of each optimization step run in parallel on 16 workers, in MATLAB on a desktop computer with an Intel Core i7-14700 processor. The sparse design has 770 points in seven dimensions and 1,168 in eight.

Table 1: Bivariate stochastic volatility models for the AUD/USD and NZD/USD returns, $T = 4{,}000$. The parameter values are those at the posterior mode. The column Δ_2 is the second-order term, and the column $\log \mathrm{BF}$ gives the log Bayes factor against CC, whose $\log \widehat{m}_2$ is $-5{,}474.86$. Evaluations are the number of likelihood evaluations, and minutes are the wall-clock time on 16 workers.

Model	ϕ_1	ϕ_2	ρ	ϕ_{21}	ν	Δ_2	$\log \mathrm{BF}$	Evaluations	Minutes
CC	0.982	0.984	0.836	–	–	0.15	0	1,976	12
GC	0.977	0.003	0.845	0.829	–	0.08	97.7	4,330	129
CC-t	0.985	0.987	0.837	–	7.9	0.18	110.3	3,266	32

The second-order term Δ_2 is between 0.08 and 0.18 for the three models, well below the threshold of one; the remaining approximation error indicated by $2\Delta_2^2$ is below 0.07 for each log marginal likelihood. The log Bayes factors in Table 1 are therefore accurate to about 0.1, and the ranking is unaffected by the approximation error.

The computational cost is modest even for a large sample of $T = 4{,}000$. The three models require between 2,000 and 4,300 likelihood evaluations each, taking from 12 minutes to about two hours on 16 workers. MCMC-based alternatives would require posterior simulation for each specification, with additional reduced runs for Chib's method. For a Gelfand–Dey estimator based on the integrated likelihood, evaluating the likelihood at tens of thousands of posterior draws would itself be costly. At 4 to 14 seconds per evaluation, 20,000 evaluations would take about one to three days on a single core, compared with 2 to 17 hours of single-core time for the optimization underlying our approximation.

6 Application: SV for Intraday Returns

Our second application examines whether extreme intraday returns are better explained by heavy-tailed observations, jumps, or short-lived movements in volatility. We compare seven stochastic volatility models, four with a scalar state and three with two volatility factors, on $T = 19{,}266$ five-minute returns on the SPDR S&P 500 ETF (SPY) over 2025. The models have seven to eleven static parameters. We first describe the data and the models, report

the Bayes factors and the computing cost, and then examine how the ranking of the models develops through the year.

6.1 Data and Models

The data are five-minute log returns in percent on the SPY in the regular session, 9:30 to 16:00, on the 247 full trading days of 2025. The price of each interval is its last trade and the price at 9:30 is the first trade of the day, so there are 78 returns per day and no overnight return. The returns are demeaned over the sample.[6]

The kurtosis of the returns is 153. Without the four most volatile days, April 7 to 10, the autocorrelation of the squared returns is 0.23 at lag 1, 0.12 at lag 20 and still 0.11 at lag 78, one trading day later. It halves within 20 lags and then barely declines, which suggests a fast-moving volatility component in addition to a persistent one. Volatility also varies with the time of day: the root mean square return is 0.13 percent in the first five minutes of the session, 0.08 percent between 12:00 and 13:00 and 0.10 percent in the last five minutes, which suggests a recurring time-of-day component could be beneficial. The extreme returns also motivate specifications with heavy tails and with jumps.

We compare seven specifications that accommodate these features in different ways. The baseline model is the stochastic volatility model in (1), denoted SV. The Student-t specification, SV-t, replaces its Gaussian return shock with $\varepsilon_t \sim \mathcal{T}_\nu$ with unit scale. The remaining models introduce return jumps, volatility jumps, or a second volatility factor. These extensions build on specifications considered by Chernov, Gallant, Ghysels, and Tauchen (2003) for daily stock index returns.

The return-jump model, SVJ, adds an occasional shock to the observation equation, as in

[6]The trades are from the TAQ millisecond consolidated trades file on Wharton Research Data Services (WRDS), 201.6 million in the year. We drop trades whose condition codes mark them as out of sequence, canceled, corrected or outside regular hours, and we exclude the three early-close days (July 3, November 28 and December 24).

Chib, Nardari, and Shephard (2002):

$$y_t = \mathrm{e}^{h_t/2}\varepsilon_t + b_t\zeta_t,$$

where the jump indicator b_t equals one with probability ω and zero otherwise, the jump size is $\zeta_t \sim \mathcal{N}(0, \sigma_\zeta^2)$, and both are independent over time and of the other shocks. Given h_t, the return is a mixture of two normal distributions with mean zero, variances e^{h_t} and $\mathrm{e}^{h_t} + \sigma_\zeta^2$, and weights $1-\omega$ and ω. The state remains the scalar h_t.

In the fourth model (SVJV) the same jump also enters the log-volatility, as in the model of Eraker, Johannes, and Polson (2003) with jumps in returns and volatility:

$$y_t = \mathrm{e}^{h_t/2}\varepsilon_t + b_t\zeta_t, \qquad h_t = \mu + \phi(h_{t-1} - \mu) + \sigma\eta_t + b_t\xi_t, \qquad \xi_t \sim \mathcal{N}(\mu_\xi, \sigma_\xi^2),$$

so that a jump in the return arrives together with a step in the log-volatility, which then decays at the rate ϕ. Given h_{t-1}, the pair (y_t, h_t) is a mixture of two components with weights $1-\omega$ and ω: the no-jump transition with the observation density $\mathcal{N}(0, \mathrm{e}^{h_t})$, and the transition shifted by ξ_t with the observation density $\mathcal{N}(0, \mathrm{e}^{h_t} + \sigma_\zeta^2)$. The state is again the scalar h_t, with h_1 from the stationary distribution of the recursion without jumps. Through the shared jump the observation density depends on (h_{t-1}, h_t), and the grid filter sums over the two components at each date.

The last three models split the log-volatility into a persistent and a transitory factor:

$$y_t = \mathrm{e}^{(h_{1t}+h_{2t})/2}\varepsilon_t, \qquad h_{1t} = \mu + \phi_1(h_{1,t-1} - \mu) + \sigma_1\eta_{1t}, \qquad h_{2t} = \phi_2 h_{2,t-1} + \sigma_2\eta_{2t},$$

where η_{1t} and η_{2t} are independent standard normal and both factors start from their stationary distributions. The restriction $\phi_2 < \phi_1$ makes h_{1t} the persistent factor. The two-factor stochastic volatility model (TF) has a Gaussian ε_t, the two-factor Student-t model (TF-t) has $\varepsilon_t \sim \mathcal{T}_\nu$ with unit scale, and the two-factor jump model (TFJ) adds the jump $b_t\zeta_t$ of the

third model to the return. Their state is two-dimensional, and the grid filter evaluates the integrated likelihood on a product grid.

All seven models account for the pronounced time-of-day pattern in intraday volatility (Andersen and Bollerslev, 1997). We add to the log-volatility in the observation equation a deterministic component $\kappa(\tau_t)$ with coefficients $\kappa_1, \ldots, \kappa_4$. The component accounts for the recurring daily pattern, so that the latent factors describe only the fluctuations around it. The coefficients are estimated jointly with the dynamics, as in Stroud and Johannes (2014), so the Bayes factors do not depend on a pattern estimated beforehand.[7]

We use the same priors on μ, ϕ, σ^2 and ν as in Section 4.1, with (ϕ_1, σ_1^2) in the role of (ϕ, σ^2) in the two-factor models. The jump probability has the prior $\omega \sim \mathcal{B}(2, 100)$, with mean 0.02, and the jump variance has the prior $\sigma_\zeta^2 \sim \mathcal{IG}(2.5, 0.1)$. The volatility jump has the priors $\mu_\xi \sim \mathcal{N}(0, 4)$ and $\sigma_\xi^2 \sim \mathcal{IG}(2.5, 0.1)$. In the two-factor models, $\phi_2 \,|\, \phi_1 \sim \mathcal{U}(-1, \phi_1)$ and $\sigma_2^2 \sim \mathcal{IG}(2.5, 0.1)$. The coefficients of the time-of-day term have the priors $\kappa_1, \ldots, \kappa_4 \sim \mathcal{N}(0, 1)$. We use the same coordinates as in Section 4.1, extended by $(\operatorname{logit}\omega, \log\sigma_\zeta^2)$ in the jump models, by $(\mu_\xi, \log\sigma_\xi^2)$ in the volatility-jump model and by $(w, \log\sigma_2^2)$ in the two-factor models, where $w = \log(\operatorname{atanh}\phi_1 - \operatorname{atanh}\phi_2)$ with $\phi_2 < \phi_1$. The coefficients $\kappa_1, \ldots, \kappa_4$ are coordinates in every model, without transformation.

6.2 Bayes Factors and Computing Cost

For all seven models, the marginal likelihood estimator is $\log\widehat{m}_2$ from the grid filter, with the derivatives in (4) from the sparse design of Section 2.3. Table 2 reports, for each model, the volatility and jump parameters at the posterior mode, the second-order term, the log Bayes factor against SV and the computing cost.[8] Heavy tails and return jumps improve

[7] The component has the first two harmonics of the trading day, $\kappa(\tau_t) = \kappa_1\cos(2\pi\tau_t) + \kappa_2\sin(2\pi\tau_t) + \kappa_3\cos(4\pi\tau_t) + \kappa_4\sin(4\pi\tau_t)$, where $\tau_t = j_t/78$ and $j_t \in \{1, \ldots, 78\}$ indexes the five-minute intervals of the trading day. It multiplies the variance of the non-jump return component by $\mathrm{e}^{\kappa(\tau_t)}$ and enters the observation density only, so the filters are unchanged.

[8] The coefficients of the time-of-day term differ by less than 0.02 across the seven models and are omitted from the table; in SV they are $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (0.30, 0.36, 0.16, 0.23)$. The fitted pattern rises from 2.9

the log marginal likelihood over SV by 25.1 and 12.3, respectively. Larger gains come from more flexible volatility dynamics: SVJV and TF improve on SV by 68.9 and 73.2. The parameter estimates help explain why these extensions are useful. SV has a single persistence parameter of 0.987 (an implied half-life of 4.4 hours). TF separates a persistent component with $\phi_1 = 0.996$ (a half-life of about two trading days) from a transitory component with $\phi_2 = 0.834$ (a half-life of nineteen minutes). The innovation variance of the transitory component is ten times that of the persistent component, so large volatility movements can occur and dissipate quickly.

Table 2: Stochastic volatility models for the five-minute SPY returns, $T = 19{,}266$. The parameter values are those at the posterior mode; in the two-factor models, ϕ and σ^2 belong to the persistent factor and ϕ_2 and σ_2^2 to the transitory factor. The column Δ_2 is the second-order term, and the column $\log \mathrm{BF}$ gives the log Bayes factor against SV, whose $\log \widehat{m}_2$ is 25,352.52. Evaluations are the number of likelihood evaluations, and minutes are the wall-clock time on eight workers.

Model	ϕ	σ^2	ϕ_2	σ_2^2	ν	ω	σ_ζ^2	μ_ξ	σ_ξ^2	Δ_2	$\log \mathrm{BF}$	Evaluations	Minutes
SV	0.987	0.040	–	–	–	–	–	–	–	0.02	0	2,794	3
SV-t	0.990	0.028	–	–	15.9	–	–	–	–	0.04	25.1	3,538	4
SVJ	0.988	0.034	–	–	–	0.0035	0.047	–	–	0.07	12.3	4,668	8
SVJV	0.989	0.019	–	–	–	0.0068	0.032	1.33	0.054	0.31	68.9	9,044	19
TF	0.996	0.010	0.834	0.098	–	–	–	–	–	0.07	73.2	5,342	57
TF-t	0.996	0.009	0.854	0.083	55.3	–	–	–	–	0.09	69.4	7,230	138
TFJ	0.996	0.010	0.847	0.087	–	0.0010	0.077	–	–	0.15	71.8	8,550	208

SVJV also accommodates abrupt volatility increases, with an estimated jump probability implying approximately 131 expected jump events over the year and a mean log-volatility jump of 1.33. Their effects, however, decay with a half-life of about five hours. The log Bayes factor of 4.3 in favor of TF over SVJV suggests that the shorter-lived transitory component provides a better account of these movements.

Once the transitory volatility factor is included, adding heavy tails or return jumps reduces the log marginal likelihood by about 3.7 and 1.4, respectively. The parameter estimates also suggest less need for either feature: the degrees-of-freedom parameter increases from 15.9 in

times its minimum at the open to 3.5 times at about 10:00, falls to the minimum at 14:30 and rises again to 2.7 times at the close. The computations of this section are done in MATLAB on a desktop computer with an Intel Core i9-12900KS processor. The evaluations of each optimization step run on eight workers.

SV-t to 55.3 in TF-t, while the jump probability falls from 0.0035 in SVJ to 0.0010 in TFJ, corresponding to approximately 67 and 20 expected jumps over the year. The second-order term Δ_2 ranges from 0.02 to 0.31 across the seven models, well below the threshold of one. The remaining approximation error indicated by $2\Delta_2^2$ is therefore below 0.2 for every log marginal likelihood, and it does not affect any of the comparisons above.

6.3 Model Comparison Through the Year

The four models TF, TFJ, TF-t and SVJV have log Bayes factors against SV within 4.3 of one another. Over 19,266 returns, such differences may reflect small gains accumulated steadily through the sample or large gains concentrated in a few episodes such as the tariff shock of April, when jumps are known to matter (Eraker, Johannes, and Polson, 2003; Maheu and McCurdy, 2004). We therefore examine whether TF is favored consistently through the year or whether the best model changes over time. For $s < t$, $\log m_t - \log m_s$ is the predictive log likelihood of the observations between s and t, and its difference between two models is their predictive log Bayes factor over that period (Geweke and Amisano, 2010). We fit each model on the sample up to the end of each month from February through December. The sample to the end of February serves as the training sample, and the differences between consecutive fits are the predictive log Bayes factors of the ten months from March to December.[9]

The proposed approach makes this exercise computationally feasible. It requires 77 marginal likelihoods, seven models at eleven month ends. Each is a mode search with a few thousand likelihood evaluations, started from the mode of the previous month end, and no posterior sampler is needed. A simulation-based estimator would require a posterior simulation for every model at every month end.

Figure 4 shows the log Bayes factors of SVJV, TF-t and TFJ against TF, the best model in the full sample, on the sample up to each month end; a value above zero means that the

[9]The returns are demeaned over the full sample, so the exercise decomposes the full-sample Bayes factors retrospectively and is not a real-time comparison.

model is favored over TF on the sample available at that date. The paths start at the log Bayes factors on the training sample, -1.0, -0.3 and -1.9, and end at the full-sample values of Table 2, about -4.3, -3.7 and -1.4. We also mark the events on the days when a five-minute return on the SPY exceeds 0.5 percent in absolute value: the tariffs on Canadian steel and aluminum (March 11), the tariff episode of April 4 to 16, the FOMC announcements of May 7 and October 29, the Jackson Hole speech (August 22) and the China tariff threat (October 10).

The gains of SVJV over TF mostly coincide with the marked events, except the Jackson Hole speech. The strongest evidence in its favor comes from two months: its log Bayes factor against TF rises by 6.8 in April, the month of the tariff episode, and by 13.6 in October, which contains the China tariff threat and an FOMC announcement. It also rises by 2.0 in March and by 1.9 in May, the months of the tariffs on Canadian steel and aluminum and of another FOMC announcement. For all three models the net change over the ten months is small relative to these monthly movements, and TF remains favored in the full sample, as it was at the end of the training sample.

The 77 fits also illustrate how the second-order term serves as a diagnostic. It is at most 0.45 at 74 of the fits and exceeds one at three, each time because the variance of the jump sizes is informed by only 25 to 110 jumps. At those three fits we use the mode and curvature already computed to construct the importance-sampling proposal, as described in Section 2.3, and the resulting estimate has a numerical standard error of 0.04. Online Appendix C reports the monthly log Bayes factors and the details of the three fits.

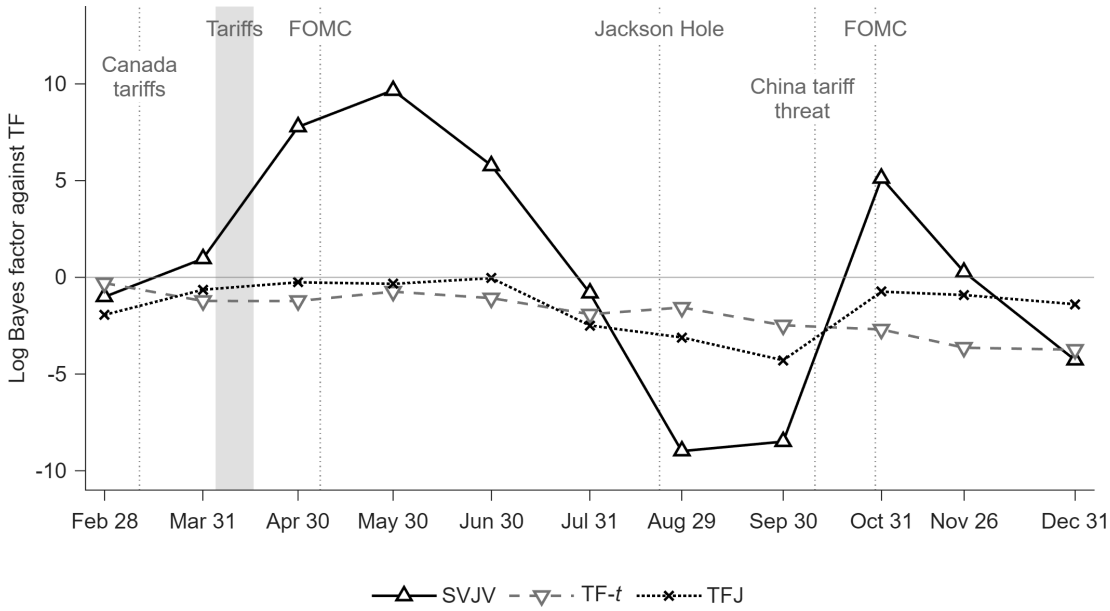


Figure 4: Log Bayes factors of SVJV, TF-t and TFJ against TF on the sample up to each month end of 2025, from the end of the training sample on February 28; above zero the model is favored over TF on the sample so far. The marked events are the days with a five-minute return on the SPY above 0.5 percent in absolute value during the 9:30 – 16:00 trading window: the tariffs on Canadian steel and aluminum (March 11), the tariff episode of April 4 to 16, the FOMC announcements of May 7 and October 29, the Jackson Hole speech (August 22) and the China tariff threat (October 10).

7 Concluding Remarks

We propose a deterministic estimator of the marginal likelihood of nonlinear state space models. A grid filter integrates out the states, and a Laplace approximation with its second-order term integrates over the static parameters. No posterior sampler is required, and the second-order term indicates the remaining error. For stochastic volatility models the estimator is accurate to within 0.01 in the log marginal likelihood at $T = 4{,}000$, and at equal computing time it is more accurate than three commonly used simulation-based estimators from $T = 2{,}000$. In the applications, heavy tails are the best supported extension for the exchange rates, and a transitory volatility factor is decisively supported for the intraday returns.

There are two interesting directions for future research. The first concerns models with

states of four or more dimensions. The grid filter is then too costly, and a particle filter would have to evaluate the integrated likelihood, so the mode and the curvature would be found from noisy values. The second is real-time use. Updating the mode and curvature at every observation would take only a few likelihood evaluations, so the predictive likelihood m_{T+1}/m_T and the posterior model weights could be updated each period. This would give real-time Bayesian model averaging and density forecast comparison for nonlinear state space models without a posterior sampler for any model.

Declaration of Generative AI

We acknowledge the use of Claude Code and Codex for editing, brainstorming and checking the code. A previous version of the paper was reviewed by Refine.ink, and its suggestions are incorporated in the current draft. All AI-assisted suggestions were carefully reviewed, and we take full responsibility for the final content and accuracy of this paper.

Data Availability Statement

The exchange rate data are the European Central Bank's daily euro reference rates, which are publicly available from the ECB Data Portal at `https://data.ecb.europa.eu/data/datasets/EXR`. The intraday data are from the NYSE Trade and Quote (TAQ) database on Wharton Research Data Services. They are proprietary and cannot be redistributed. The code and the exchange rate data will be provided as online supplementary material.

References

Andersen, T. G., and T. Bollerslev (1997): "Intraday Periodicity and Volatility Persistence in Financial Markets," *Journal of Empirical Finance*, 4, 115–158.

Andrieu, C., A. Doucet, and R. Holenstein (2010): "Particle Markov Chain Monte Carlo Methods," *Journal of the Royal Statistical Society Series B*, 72(3), 269–342.

Asai, M., M. McAleer, and J. Yu (2006): "Multivariate Stochastic Volatility: A Review," *Econometric Reviews*, 25(2–3), 145–175.

Barndorff-Nielsen, O. E., and D. R. Cox (1989): *Asymptotic Techniques for Use in Statistics*. Chapman and Hall, London.

Bégin, J.-F., and M. Boudreault (2021): "Likelihood Evaluation of Jump-Diffusion Models Using Deterministic Nonlinear Filters," *Journal of Computational and Graphical Statistics*, 30(2), 452–466.

Burden, R. L., and J. D. Faires (2011): *Numerical Analysis*. Brooks/Cole, Boston, 9th edn.

Chan, J. C. C. (2017): "The Stochastic Volatility in Mean Model with Time-Varying Parameters: An Application to Inflation Modeling," *Journal of Business and Economic Statistics*, 35(1), 17–28.

——— (2023): "Comparing Stochastic Volatility Specifications for Large Bayesian VARs," *Journal of Econometrics*, 235(2), 1419–1446.

——— (2026): "Exact Rejection Sampling for Non-Gaussian State Space Models," arXiv:2608.21619.

Chan, J. C. C., and E. Eisenstat (2015): "Marginal Likelihood Estimation with the Cross-Entropy Method," *Econometric Reviews*, 34(3), 256–285.

Chan, J. C. C., and A. L. Grant (2015): "Pitfalls of Estimating the Marginal Likelihood Using the Modified Harmonic Mean," *Economics Letters*, 131, 29–33.

Chan, J. C. C., and I. Jeliazkov (2009): "Efficient Simulation and Integrated Likelihood Estimation in State Space Models," *International Journal of Mathematical Modelling and Numerical Optimisation*, 1(1–2), 101–120.

Chan, J. C. C., X. Yu, and W. Zhang (2026): "Bayesian Model Comparison for Large Bayesian VARs after the COVID-19 Pandemic," *Journal of Econometrics*, 256, 106072.

Chernov, M., A. R. Gallant, E. Ghysels, and G. Tauchen (2003): "Alternative Models for Stock Price Dynamics," *Journal of Econometrics*, 116, 225–257.

Chib, S. (1995): "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90(432), 1313–1321.

Chib, S., and I. Jeliazkov (2001): "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96(453), 270–281.

Chib, S., F. Nardari, and N. Shephard (2002): "Markov Chain Monte Carlo Methods for Stochastic Volatility Models," *Journal of Econometrics*, 108(2), 281–316.

Chopin, N., P. E. Jacob, and O. Papaspiliopoulos (2013): "SMC^2: An Efficient Algorithm for Sequential Analysis of State Space Models," *Journal of the Royal Statistical Society Series B*, 75(3), 397–426.

Creal, D. (2012): "A Survey of Sequential Monte Carlo Methods for Economics and Finance," *Econometric Reviews*, 31(3), 245–296.

——— (2017): "A Class of Non-Gaussian State Space Models with Exact Likelihood Inference," *Journal of Business and Economic Statistics*, 35(4), 585–597.

Daníelsson, J., and J.-F. Richard (1993): "Accelerated Gaussian Importance Sampler with Application to Dynamic Latent Variable Models," *Journal of Applied Econometrics*, 8(S1), S153–S173.

Durbin, J., and S. J. Koopman (1997): "Monte Carlo Maximum Likelihood Estimation for Non-Gaussian State Space Models," *Biometrika*, 84(3), 669–684.

Eraker, B., M. Johannes, and N. Polson (2003): "The Impact of Jumps in Volatility and Returns," *Journal of Finance*, 58(3), 1269–1300.

Farmer, L. E. (2021): "The Discretization Filter: A Simple Way to Estimate Nonlinear State Space Models," *Quantitative Economics*, 12(1), 41–76.

Fridman, M., and L. Harris (1998): "A Maximum Likelihood Approach for Non-Gaussian Stochastic Volatility Models," *Journal of Business and Economic Statistics*, 16(3), 284–291.

Gelfand, A. E., and D. K. Dey (1994): "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society Series B*, 56(3), 501–514.

Geweke, J. (1989): "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57(6), 1317–1339.

——— (1999): "Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication," *Econometric Reviews*, 18(1), 1–73.

Geweke, J., and G. Amisano (2010): "Comparing and evaluating Bayesian predictive distributions of asset returns," *International Journal of Forecasting*, 26(2), 216–230.

Hansen, P. R., and C. Tong (2026): "Tweedie's Formula and Score-Driven Updating," arXiv:2605.15902.

Kass, R. E., L. Tierney, and J. B. Kadane (1990): "The Validity of Posterior Expansions Based on

Laplace's Method," in *Bayesian and Likelihood Methods in Statistics and Econometrics*, ed. by S. Geisser, J. S. Hodges, S. J. Press, and A. Zellner, pp. 473–488. North-Holland, Amsterdam.

Kim, S., N. Shephard, and S. Chib (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), 1032–1041.

Koopman, S. J., and E. Hol Uspensky (2002): "The Stochastic Volatility in Mean Model: Empirical Evidence from International Stock Markets," *Journal of Applied Econometrics*, 17(6), 667–689.

Lange, R.-J., B. van Os, and D. van Dijk (2026): "Implicit Score-Driven Filters for Time-Varying Parameter Models," *Journal of Econometrics*, 255, 106251.

Langrock, R., I. L. MacDonald, and W. Zucchini (2012): "Some Nonstandard Stochastic Volatility Models and Their Estimation Using Structured Hidden Markov Models," *Journal of Empirical Finance*, 19(1), 147–161.

Li, Y., J. Yu, and T. Zeng (2020): "Deviance Information Criterion for Latent Variable Models and Misspecified Models," *Journal of Econometrics*, 216(2), 450–493.

Louis, T. A. (1982): "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society Series B*, 44(2), 226–233.

Maheu, J. M., and T. H. McCurdy (2004): "News Arrival, Jump Dynamics, and Volatility Components for Individual Stock Returns," *The Journal of Finance*, 59(2), 755–793.

McCausland, W. J. (2012): "The HESSIAN Method: Highly Efficient Simulation Smoothing, in a Nutshell," *Journal of Econometrics*, 168(2), 189–206.

McCullagh, P. (1987): *Tensor Methods in Statistics*. Chapman and Hall, London.

Meng, X.-L., and W. H. Wong (1996): "Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6(4), 831–860.

Pitt, M. K., R. dos Santos Silva, P. Giordani, and R. Kohn (2012): "On Some Properties of Markov Chain Monte Carlo Simulation Methods Based on the Particle Filter," *Journal of Econometrics*, 171(2), 134–151.

Richard, J.-F., and W. Zhang (2007): "Efficient High-Dimensional Importance Sampling," *Journal of Econometrics*, 141(2), 1385–1411.

Rue, H., S. Martino, and N. Chopin (2009): "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations," *Journal of the Royal Statistical Society Series B*, 71(2), 319–392.

Ruli, E., N. Sartori, and L. Ventura (2016): "Improved Laplace Approximation for Marginal Likelihoods," *Electronic Journal of Statistics*, 10(2), 3986–4009.

Shephard, N., and M. K. Pitt (1997): "Likelihood Analysis of Non-Gaussian Measurement Time Series," *Biometrika*, 84(3), 653–667.

Shun, Z., and P. McCullagh (1995): "Laplace Approximation of High Dimensional Integrals," *Journal of the Royal Statistical Society Series B*, 57(4), 749–760.

Skovgaard, I. M. (1986): "A Note on the Differentiation of Cumulants of Log Likelihood Derivatives," *International Statistical Review*, 54(1), 29–32.

Stroud, J. R., and M. S. Johannes (2014): "Bayesian Modeling and Forecasting of 24-Hour High-Frequency Volatility," *Journal of the American Statistical Association*, 109(508), 1368–1384.

Tierney, L., and J. B. Kadane (1986): "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81(393), 82–86.

Trefethen, L. N., and J. A. C. Weideman (2014): "The Exponentially Convergent Trapezoidal Rule," *SIAM Review*, 56, 385–458.

Watanabe, T. (1999): "A Non-Linear Filtering Approach to Stochastic Volatility Models with an Application to Daily Stock Returns," *Journal of Applied Econometrics*, 14(2), 101–121.

Yu, J., and R. Meyer (2006): "Multivariate Stochastic Volatility Models: Bayesian Estimation and Model Comparison," *Econometric Reviews*, 25(2–3), 361–384.

Online Appendix A Proofs

Proof of Proposition 1. Write $Z = Z_T(\boldsymbol{\psi})$ and $\pi(\boldsymbol{\alpha}) = \mathrm{e}^{\mathcal{C}(\boldsymbol{\psi},\boldsymbol{\alpha})}/Z$. For smooth $g_1, \dots, g_k$ let $\mathcal{K}(\mathbf{t}) = \log \mathbb{E}_{\boldsymbol{\psi}} \exp(\sum_i t_i g_i)$ be their joint cumulant generating function, so that $\mathrm{cum}_{\boldsymbol{\psi}}(g_1, \dots, g_k) = \partial_{t_1} \cdots \partial_{t_k} \mathcal{K}(\mathbf{0})$. Differentiating under the integral sign,

$$D_{\mathbf{w}} \mathcal{K}(\mathbf{t}) = \widetilde{\mathbb{E}}_{\mathbf{t}} \Big[\sum_i t_i D_{\mathbf{w}} g_i + c_1 \Big] - \mathbb{E}_{\boldsymbol{\psi}}[c_1],$$

where $\widetilde{\mathbb{E}}_{\mathbf{t}}$ is the expectation under the density proportional to $\exp(\sum_i t_i g_i)\pi(\boldsymbol{\alpha})$, and $\widetilde{\mathbb{E}}_{\mathbf{t}}[q] = \partial_s \log \mathbb{E}_{\boldsymbol{\psi}} \exp(\sum_i t_i g_i + sq)|_{s=0}$ for any q. Applying $\partial_{t_1} \cdots \partial_{t_k}$ at $\mathbf{t} = \mathbf{0}$ gives (8). The term $t_i D_{\mathbf{w}} g_i$ contributes $\mathrm{cum}_{\boldsymbol{\psi}}(g_1, \dots, D_{\mathbf{w}} g_i, \dots, g_k)$, and the term c_1 contributes $\mathrm{cum}_{\boldsymbol{\psi}}(g_1, \dots, g_k, c_1)$. Differentiating Z under the integral sign gives $D_{\mathbf{w}} \log Z = \mathbb{E}_{\boldsymbol{\psi}}[c_1]$ directly. Since $D_{\mathbf{w}} c_j = c_{j+1}$, repeated application of (8) gives

$$\begin{aligned}
D_{\mathbf{w}} \mathbb{E}_{\boldsymbol{\psi}}[c_1] &= \mathbb{E}_{\boldsymbol{\psi}}[c_2] + \mathrm{Var}_{\boldsymbol{\psi}}(c_1), \\
D_{\mathbf{w}} \mathrm{Var}_{\boldsymbol{\psi}}(c_1) &= 2\mathrm{Cov}_{\boldsymbol{\psi}}(c_1, c_2) + \mathrm{cum}_{\boldsymbol{\psi}}(c_1, c_1, c_1), \\
D_{\mathbf{w}} \mathrm{Cov}_{\boldsymbol{\psi}}(c_1, c_2) &= \mathrm{Var}_{\boldsymbol{\psi}}(c_2) + \mathrm{Cov}_{\boldsymbol{\psi}}(c_1, c_3) + \mathrm{cum}_{\boldsymbol{\psi}}(c_1, c_1, c_2), \\
D_{\mathbf{w}} \mathrm{cum}_{\boldsymbol{\psi}}(c_1, c_1, c_1) &= 3\mathrm{cum}_{\boldsymbol{\psi}}(c_1, c_1, c_2) + \mathrm{cum}_{\boldsymbol{\psi}}(c_1, c_1, c_1, c_1),
\end{aligned}$$

and (7) follows by summing. □

Proof of Theorem 1. Let φ denote the $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ density, let $\rho_T = T^{1/8}$, and write $\beta_k = K_k T^{1-k/2}$. Take T large enough that $\rho_T \leqslant \varrho\sqrt{T}$ and $\sum_{k=3}^{6} \beta_k \rho_T^k / k! \leqslant \log 2$. By (3), $m_T/m_1 = I_1 + I_2 + I_3$, where I_1 is the integral of $\mathrm{e}^R \varphi$ over the ball $\|\mathbf{u}\| \leqslant \rho_T$, I_2 is the integral over the shell $\rho_T < \|\mathbf{u}\| \leqslant \varrho\sqrt{T}$, and $I_3 = m_1^{-1} \int_{\mathcal{N}_T^c} \mathrm{e}^f$. Assumption 3 gives $I_3 \leqslant C_0 T^{-2}$.

For $\|\mathbf{u}\| \leqslant \varrho\sqrt{T}$, the segment from $\mathbf{0}$ to $\mathbf{u}$ lies in $\mathcal{N}_T$ because $\mathcal{N}_T$ is convex. Taylor's theorem along this segment gives $R(\mathbf{u}) = D^3 f(\boldsymbol{\xi})[\mathbf{Au}, \mathbf{Au}, \mathbf{Au}]/6$ for some $\boldsymbol{\xi} \in \mathcal{N}_T$, so $|R(\mathbf{u})| \leqslant \beta_3 \|\mathbf{u}\|^3/6 \leqslant (\varrho K_3/6)\|\mathbf{u}\|^2 \leqslant \frac{1}{4}\|\mathbf{u}\|^2$ by Assumption 2. Hence $\mathrm{e}^{R(\mathbf{u})} \varphi(\mathbf{u}) \leqslant (2\pi)^{-d/2} \mathrm{e}^{-\|\mathbf{u}\|^2/4}$

on $\mathcal{N}_T$, and

$$I_2 \leqslant (2\pi)^{-d/2} \int_{\|\mathbf{u}\|>\rho_T} \mathrm{e}^{-\|\mathbf{u}\|^2/4}\,\mathrm{d}\mathbf{u} = \frac{2^{d/2}\,\Gamma(d/2, \rho_T^2/4)}{\Gamma(d/2)} \leqslant M_d\,\rho_T^{d-2}\,\mathrm{e}^{-\rho_T^2/4}$$

for a constant M_d, where $\Gamma(\cdot,\cdot)$ is the upper incomplete gamma function. This bound is $o(T^{-2})$.

On the ball, Taylor's theorem to sixth order gives $R = C_3 + C_4 + C_5 + R_6$. Here $|C_k(\mathbf{u})| \leqslant \beta_k\|\mathbf{u}\|^k/k!$ for every $\mathbf{u}$ by Assumption 2 at $\boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}$, and $|R_6(\mathbf{u})| \leqslant \beta_6\|\mathbf{u}\|^6/720$ on the ball by the same assumption on $\mathcal{N}_T$. On the ball $|R(\mathbf{u})| \leqslant \sum_{k=3}^{6} \beta_k\rho_T^k/k! = O(T^{-1/8})$, and $\mathrm{e}^R = \sum_{j=0}^{4} R^j/j! + E_5$ with $|E_5| \leqslant |R|^5\mathrm{e}^{|R|}/120 \leqslant |R|^5/60$. Each monomial in the expansion of $\sum_{j\leqslant 4} R^j/j!$, and the bound on E_5, is dominated by a constant times a product of factors $\beta_{k_i}\|\mathbf{u}\|^{k_i}$. The integral of such a product against φ over the ball is at most the corresponding Gaussian moment over $\mathbb{R}^d$, which is finite, times $\prod_i \beta_{k_i} = O(T^{\sum_i(1-k_i/2)})$. The terms of order $T^{-1/2}$ and $T^{-3/2}$ are C_3, C_5, C_3C_4 and $C_3^3/6$. They are odd functions of $\mathbf{u}$ and integrate to zero over the ball. The terms of order T^{-1} are C_4 and $\frac{1}{2}C_3^2$. Every other term, including E_5, is of order T^{-2} or smaller. Therefore

$$I_1 = \int_{\|\mathbf{u}\|\leqslant\rho_T} \bigl\{1 + C_4(\mathbf{u}) + \tfrac{1}{2}C_3(\mathbf{u})^2\bigr\}\varphi(\mathbf{u})\,\mathrm{d}\mathbf{u} + O(T^{-2}),$$

and extending the integral to $\mathbb{R}^d$ changes it by at most a constant times $(1 + \beta_4\rho_T^4 + \beta_3^2\rho_T^6)\rho_T^{d-2}\mathrm{e}^{-\rho_T^2/2}$, which is $o(T^{-2})$. The Gaussian moments give $\mathbb{E}[C_4(\mathbf{G})] = \frac{1}{8}\sum_{i,j} b_{iijj}$, because $\mathbb{E}[G_iG_jG_kG_l]$ is the sum over the three pairings of the indices. They also give $\frac{1}{2}\mathbb{E}[C_3(\mathbf{G})^2] = \frac{1}{12}\sum_{i,j,k} a_{ijk}^2 + \frac{1}{8}\sum_i(\sum_j a_{ijj})^2$, because $\mathbb{E}[G_iG_jG_kG_lG_mG_n]$ is the sum over the fifteen pairings, of which six pair each index of one cubic factor with an index of the other and nine pair one index within each factor. Hence $m_T/m_1 = 1 + \Delta_2 + O(T^{-2})$. By polarization, i.e., the identity that expresses a symmetric multilinear form through its diagonal values, Assumption 2 at $\boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}$ bounds every entry of the standardized tensors, so $\Delta_2 = O(T^{-1})$, and $\log(1 + \Delta_2 + r) = \Delta_2 + O(T^{-2})$ for $r = O(T^{-2})$. All constants depend

only on d, ϱ, $K_3, \ldots, K_6$ and C_0. $\square$

Online Appendix B Implementation Details

This appendix gives the implementation details behind the results in Sections 4 to 6. We first describe the grid filter for the scalar-state models and the quadrature benchmark, then the mode search and the sparse design for the derivatives in Δ_2. The last two subsections cover the MCMC-based marginal likelihood estimators of Section 4.3 and the parameter transformations with their Jacobians.

B.1 Grid Filter and Benchmark

We first describe the grid filter for the models in Section 4, whose state is scalar with transition $h_t = \mu + \phi(h_{t-1} - \mu) + \sigma\eta_t$. At a given $\boldsymbol{\theta}$, Newton's method with the banded precision matrix gives the smoothing mode $\widehat{\mathbf{h}}$ and the marginal standard deviations $\widehat{s}_t$ of the Gaussian approximation to $p(\mathbf{h} \,|\, \mathbf{y}, \boldsymbol{\theta})$. The grid has spacing $\delta = \min(\sigma/2, 0.05)$ and runs from $\min_t \widehat{h}_t - 12 \max_t \widehat{s}_t - 6\sigma$ to $\max_t \widehat{h}_t + 12 \max_t \widehat{s}_t + 6\sigma$. With a spacing of at most half the transition standard deviation, the aliasing error of the trapezoidal rule for the Gaussian kernel is at most of order $\mathrm{e}^{-8\pi^2}$. The product grids of the bivariate models in Section 5 have spacing $\min(\sigma_i/2, 0.1)$ in each state, and those of the two-factor models in Section 6 have spacing $0.6\sigma_i$.

One evaluation of $\log Z_T$ takes between 0.01 seconds at $T = 500$ and 0.19 seconds at $T =$ 8,000 on one core. For a linear Gaussian model, whose integrated likelihood is available in closed form, the filter agrees with the exact value to 3×10^{-12}. On the 300 data sets in Section 4, it also agrees with an independent simulation estimate of $\log Z_T(\widehat{\boldsymbol{\psi}})$, from an importance sampler for the state path with the bounded-weight proposal of Chan (2026); the largest difference is 0.001.

The benchmark value of $\log m_T$ in Section 4 is a nested quadrature in $\boldsymbol{\psi}$. The outer rule is a trapezoidal rule in $z = \operatorname{atanh}\phi$ with step 0.5 posterior standard deviations, extended until the integrand falls below 10^{-9} of its maximum. The integrand is analytic in z with Gaussian-like tails, so the trapezoidal rule converges exponentially and the error of the outer rule is far smaller than the truncation at 10^{-9}. The inner rule is a Gauss–Hermite rule over the remaining coordinates with 8 points per dimension (6 when the inner integral is three-dimensional), centered at the conditional mode and scaled by the conditional curvature. Refining both rules changes $\log m_T$ by at most 4.2×10^{-5} on the 240 data sets with $T \leqslant 4{,}000$, and by at most 3×10^{-9} on the nine stochastic volatility data sets of the comparison at $T = 8{,}000$. For the Student-t model at $T \leqslant 2{,}000$, where the conditional posterior of ψ_4 has a long right tail, the inner rule in ψ_4 is a trapezoidal rule.

B.2 Mode Search and Sparse Design

The mode search uses Newton steps in the standardized coordinates $\mathbf{u} = \mathbf{A}^{-1}(\boldsymbol{\psi} - \widehat{\boldsymbol{\psi}})$, with the gradient and Hessian from central differences.[10] Each step is evaluated at nine scales in one parallel batch and the best is taken, because for some models the full Newton step is too short along a curved ridge of the posterior, as for GC. Where the Hessian is not negative definite, the step follows the gradient. In Section 6 it takes the modified Newton direction described below instead. The iteration stops when the standardized step is below 10^{-7} in the experiments and 10^{-6} in the applications. For the bivariate models in Section 5, the search starts from a point constructed from univariate fits. Each volatility process starts at the posterior mode of a univariate stochastic volatility model fitted to its own series, and the correlation at the sample correlation.

The derivatives in (4) come from central differences of $R(\mathbf{u})$ at $\mathbf{u} = \mathbf{0}$ on a sparse design with step $\ell = 0.25$: the points $\pm\ell\mathbf{e}_i$ and $\pm2\ell\mathbf{e}_i$ on each axis give a_{iii} and b_{iiii}, the four

[10] Fourth-order central differences of step 0.1, with Richardson extrapolation for the mixed second derivatives. The nine scales are $0.25, 0.5, 1, 1.5, 2, 3, 4, 6, 8$.

points $\pm\ell\mathbf{e}_i \pm \ell\mathbf{e}_j$ in each coordinate plane give a_{iij}, a_{ijj} and b_{iijj}, and the eight points $\pm\ell\mathbf{e}_i \pm \ell\mathbf{e}_j \pm \ell\mathbf{e}_k$ in each coordinate cube give a_{ijk}. The design is sparse because (4) needs only the third derivatives and the fourth derivatives with paired indices, so its points lie on the coordinate axes, planes and cubes, $6d + 8\binom{d}{2} + 16\binom{d}{3}$ in all, against 5^d for a full central-difference grid. These differences have an error of order ℓ^2, which under Assumption 2 changes Δ_2 by $O(T^{-2})$, the order of the remainder in Theorem 1; repeating them at step 2ℓ and taking $(4D_\ell - D_{2\ell})/3$ for each derivative D removes this leading term (Burden and Faires, 2011, Section 4.2). In the experiments the optimization requires between 258 and 760 evaluations per data set, of which the design takes 58 for $d = 3$ and 136 for $d = 4$.

In the two-factor stochastic volatility models in Section 6 the Hessian is indefinite over most of the path from the start, along a ridge in $(\operatorname{atanh}\phi_1, w)$, where gradient steps advance slowly. The search therefore takes a modified Newton step there: the eigenvalues of the standardized Hessian are replaced by the negatives of their absolute values, clipped to $[-4, -0.25]$, and the step is capped at two standardized units. The three two-factor searches converge in 10, 11 and 10 iterations.

B.3 MCMC-Based Marginal Likelihood Estimators

Chib's method, the Gelfand–Dey estimator and the cross-entropy method require posterior draws. For the stochastic volatility model in Section 4.3, we obtain them with the auxiliary mixture sampler of Kim, Shephard, and Chib (1998) in conjunction with the precision-based draw of the volatility path (Chan and Jeliazkov, 2009). The sampler uses the seven-component mixture, Gibbs steps for μ and σ^2, and the independence Metropolis–Hastings step of Kim, Shephard, and Chib (1998) for ϕ. The full run starts from $\phi = 0.95$, $\sigma^2 = 0.1$ and μ equal to the logarithm of the sample variance.

Chib's method evaluates the posterior ordinate at the posterior mean of the full run: the ordinate of μ comes from the full run, that of σ^2 from a reduced run with μ fixed, and

that of ϕ from a second reduced run, normalized by quadrature. The Gelfand–Dey estimator uses the truncated normal of Geweke (1999) at the 95% ellipsoid in $\boldsymbol{\psi}$, and the cross-entropy method a Gaussian in $\boldsymbol{\psi}$ with the mean and covariance of the draws of the full run. Because the auxiliary mixture sampler targets the mixture approximation, Chib's method and the Gelfand–Dey estimator reweight each draw to the exact model. The weight is $\prod_{t=1}^{T} p(\varepsilon_t^*)/q(\varepsilon_t^*)$, where $\varepsilon_t^* = \log y_t^2 - h_t$, p is its exact $\log \chi_1^2$ density and q the seven-component mixture. The cross-entropy method uses the integrated likelihood of the exact model and needs no reweighting.

Numerical standard errors are computed by batch means with 20 batches. The computing times in Section 4.3 are single-thread times on an idle core, extrapolated from the time of one sweep of each run and of one likelihood evaluation at each T and ϕ. At $T = 4{,}000$, where the reweighting keeps 0.1 to 3 percent of the draws, the largest errors are 0.76 for Chib's method and 0.13 for the Gelfand–Dey estimator, 6.2 and 4.5 times their standard errors. At $T = 8{,}000$ the reweighting keeps at most 0.5 percent of the draws, and the largest errors are 1.45 and 0.37, with standard errors of up to 1.7 and 0.11. The errors of the cross-entropy method are at most 0.020 at every T.

B.4 Parameter Transformations and Jacobians

This subsection records the transformations that map the parameters of each model to the unconstrained coordinates $\boldsymbol{\psi}$, and the terms that they add to $\log p_{\boldsymbol{\psi}}$. For a scalar parameter θ with prior density $p(\theta)$ and coordinate $\psi = g(\theta)$, the prior density of ψ is $p(g^{-1}(\psi))\,|\mathrm{d}g^{-1}(\psi)/\mathrm{d}\psi|$. The coordinates of each model are listed in Sections 4.1, 4.4 and 6.1, and $\log p_{\boldsymbol{\psi}}$ is the sum over the coordinates of the terms below. Six cases cover all the models.

A parameter with a normal prior enters as itself and adds no Jacobian: the levels μ, μ_1 and μ_2, the coefficients λ and ϕ_{21}, the time-of-day coefficients $\kappa_1, \ldots, \kappa_4$, the mean μ_ξ of

the volatility jump, and the degrees of freedom through $\psi = \log(\nu - 2)$, on which the prior $\mathcal{N}(2, 1)$ is placed directly. Under this prior $\nu - 2$ is log-normal, with density

$$p(\nu) = \frac{1}{(\nu - 2)\sqrt{2\pi}} \exp\left\{-\frac{(\log(\nu - 2) - 2)^2}{2}\right\}, \qquad \nu > 2,$$

whose median is $2 + \mathrm{e}^2 \approx 9.4$ and whose mean is $2 + \mathrm{e}^{5/2} \approx 14.2$.

A persistence parameter with prior $(1 + \phi)/2 \sim \mathcal{B}(a, b)$ enters as $z = \operatorname{atanh} \phi$. Since $1 + \tanh z = \mathrm{e}^z / \cosh z$, $1 - \tanh z = \mathrm{e}^{-z} / \cosh z$ and $\mathrm{d}\phi/\mathrm{d}z = \cosh^{-2} z$,

$$\log p(z) = (a - b)z - (a + b) \log \cosh z - (a + b - 1) \log 2 - \log B(a, b),$$

where $B(a, b)$ is the beta function. This case covers ϕ, ϕ_1 and ϕ_2 in the models with independent persistence parameters.

A variance with prior $\sigma^2 \sim \mathcal{IG}(\nu_0, S_0)$, whose density is $S_0^{\nu_0} \Gamma(\nu_0)^{-1} (\sigma^2)^{-\nu_0 - 1} \mathrm{e}^{-S_0/\sigma^2}$, enters as $s = \log \sigma^2$. Since $\mathrm{d}\sigma^2/\mathrm{d}s = \mathrm{e}^s$,

$$\log p(s) = \nu_0 \log S_0 - \log \Gamma(\nu_0) - \nu_0 s - S_0 \mathrm{e}^{-s}.$$

This case covers σ^2, σ_1^2, σ_2^2, σ_ζ^2 and σ_ξ^2.

The correlation with prior $\rho \sim \mathcal{U}(-1, 1)$ enters as $r = \operatorname{atanh} \rho$, with $\mathrm{d}\rho/\mathrm{d}r = \cosh^{-2} r$, so $\log p(r) = -\log 2 - 2 \log \cosh r$.

The jump probability with prior $\omega \sim \mathcal{B}(a_\omega, b_\omega)$ enters as $k = \operatorname{logit} \omega$. Since $\mathrm{d}\omega/\mathrm{d}k = \omega(1 - \omega)$, $\log \omega = k - \log(1 + \mathrm{e}^k)$ and $\log(1 - \omega) = -\log(1 + \mathrm{e}^k)$,

$$\log p(k) = a_\omega k - (a_\omega + b_\omega) \log(1 + \mathrm{e}^k) - \log B(a_\omega, b_\omega).$$

The second persistence parameter of the two-factor models, with prior $\phi_2 \,|\, \phi_1 \sim \mathcal{U}(-1, \phi_1)$ and density $1/(1 + \phi_1)$, enters as $w = \log(z_1 - z_2)$ with $z_i = \operatorname{atanh} \phi_i$, so that $\phi_2 = \tanh(z_1 - \mathrm{e}^w)$

and $\partial\phi_2/\partial w = -\mathrm{e}^w \cosh^{-2}(z_1 - \mathrm{e}^w)$. The conditional prior density of w given z_1 is

$$\log p(w \mid z_1) = w - 2\log\cosh(z_1 - \mathrm{e}^w) - \log(1 + \tanh z_1),$$

and z_1 itself carries the term of the persistence case above.

Online Appendix C Monthly Bayes Factors

This appendix reports the monthly estimates behind Section 6.3 and the importance-sampling estimates at the three month-end fits at which $|\Delta_2|$ exceeds one. Both models with such fits have $d = 11$ parameters.

C.1 Monthly Predictive Comparisons

Table 3 gives the log Bayes factors against SV by period. On the training sample, every model but SVJ is already favored over SV by 10 to 12, and the ranking over the ten months from March to December is close to the full-sample one. The last column shows that $|\Delta_2|$ does not exceed 0.45 at any month-end fit but three: SVJV at the end of June, with $\Delta_2 = 3.41$ at $t = 9{,}516$, and TFJ at the end of August and of September, with 1.49 and -3.27 at $t = 12{,}792$ and 14,430.

At these three fits the table uses the importance-sampling estimate of Section C.2 in place of $\log \widehat{m}_2$. The replacement lowers the June entry of SVJV from 7.8 to 4.3 and raises its July entry from -0.1 to 3.3; for TFJ it moves about 1.4 from August to September and about 3.0 from October to September. A month-end value enters the increment of its own month with a positive sign and that of the following month with a negative sign, so the cumulative row does not depend on the three values.

Table 3: Log Bayes factors against SV by period, five-minute SPY returns of 2025. The first row is the training sample, each later row is the predictive log Bayes factor of the month, and the last two rows are the sum over March to December and the full-sample value. The last column is the largest $|\Delta_2|$ among the seven fits at the month end. Entries marked ^{a} use the importance-sampling estimate of Section C.2.

Period	SV-t	SVJ	SVJV	TF	TF-t	TFJ	$\max\|\Delta_2\|$
Training, to Feb 28	10.39	5.44	11.23	12.23	11.93	10.29	0.32
March	−2.30	1.70	0.53	−1.43	−2.36	−0.13	0.37
April	0.84	0.05	8.71	1.90	1.89	2.28	0.35
May	7.17	4.95	11.51	9.62	10.10	9.54	0.45
June	4.20	3.38	4.31^a	8.11	7.78	8.40	3.41
July	1.53	−4.03	3.34^a	10.02	9.19	7.58	0.41
August	8.90	2.55	5.93	14.11	14.45	13.51^a	1.49
September	−4.63	−5.41	3.29	2.80	1.88	1.59^a	3.27
October	1.14	5.35	20.80	7.17	6.95	10.73^a	0.23
November, to Nov 26	−2.06	−0.40	−1.19	3.67	2.72	3.50	0.26
December, from Nov 27	−0.09	−1.33	0.41	4.97	4.86	4.49	0.31
March to December	14.70	6.82	57.65	60.93	57.49	61.48	
Full sample	25.08	12.25	68.88	73.16	69.42	71.77	

C.2 The Three Fits with a Large Second-Order Term

The large second-order terms at the three fits are not numerical artifacts: a second implementation of the grid filter reproduces Δ_2 to 10^{-4} at the two TFJ fits. Recomputing (4) without one coordinate at a time attributes almost all of Δ_2 to the log variance of the jump sizes, $\log \sigma_\xi^2$ in SVJV and $\log \sigma_\zeta^2$ in TFJ (see Table 4).

At the three fits we estimate $\mathbb{E}[\mathrm{e}^{R(\mathbf{G})}]$ in (3) by importance sampling over the static parameters, as described in Section 2.3. The proposal is a multivariate Student-t distribution with

five degrees of freedom in the standardized coordinates, and we use 2,000 draws.[11] With draws $\mathbf{u}^{(m)}$ from the proposal density q, the average of $\mathrm{e}^{R(\mathbf{u}^{(m)})}\varphi(\mathbf{u}^{(m)})/q(\mathbf{u}^{(m)})$ estimates m_T/m_1, where φ is the standard normal density.

We highlight two results. First, the importance sampler is accurate: the numerical standard error is 0.04 at each of the three fits, and at four adjacent month-end fits of the same models, where $|\Delta_2|$ is below one, the importance-sampling estimate is within 0.15 of $\log \widehat{m}_2$. Second, at the three fits the estimate differs from $\log \widehat{m}_2$ by about $-\Delta_2$ (see Table 4), so the second-order term correctly identifies the fits at which the expansion should not be used.

Table 4: The three month-end fits at which $|\Delta_2|$ exceeds one. Omission is the change in Δ_2 when the log jump variance is left out of (4), Std. dev. is its local Gaussian standard deviation, and IS is the importance-sampling estimate of $\log m_T$ (Student-t proposal with five degrees of freedom, 2,000 draws, numerical standard error 0.04).

Model	Date	t	Δ_2	Omission	Std. dev.	IS $-\log \widehat{m}_1$	IS $-\log \widehat{m}_2$
SVJV	June 30	9,516	3.41	3.28	1.23	−0.06	−3.47
TFJ	August 29	12,792	1.49	1.36	0.79	0.05	−1.43
TFJ	September 30	14,430	−3.27	−3.38	1.12	−0.31	2.97

[11] A draw at which a parameter falls outside the range of the grid receives zero weight; this happens at 27 of the 2,000 draws at the SVJV fit of June and at most 10 draws elsewhere.