Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Fall 2023

**Nicolas N. Koch**

# On-Line Comparison of Volatility Models via Sequential Monte Carlo

Submission Date: 2 April 2024

Co-Advisors:
Advisor: Prof. Dr. Peter L. Bühlmann

...and of course shortly after kindergarten,
you learned how to solve such an equation.

– *Josef Teichmann*

Equations are often easy to write down,
but rarely easy to compute.

– *Josef Teichmann*

# Preface

First words and acknowledgements.

# Abstract

Short summary of your thesis.

# Contents

## List of Figures

## List of Tables

# Notation

Explain your symbols and abbreviations.

# Chapter 1

# Introduction

Volatility is key quantity for pricing options, creating hedges, quantitative trading strategies, ...

Non-Bayesian model selection may not generalize well to new datasets

Online estimation estimation allows incorporating new data after fitting model without re-fitting, constant memory/computation cost. Valuable in finance, where data arrives continuously. Offline estimation results are conditional on length of time series; not clear how best model dependent on sample size

# Chapter 2

# Volatility Models

A volatility model is a model for the conditional standard deviation of the next value of a path given the past. Formally, we consider discrete $\mathbb{R}$-valued price processes with non-negative time indices, $S : \mathbb{N}_0 \to \mathbb{R}$, adapted to the filtration $\mathcal{F}_t = \sigma(\{X_{t'}, t' \leq t\})$. Based on the price process, define $(X_t)$ such that $X_t := \Delta \log S_{t-1} = \log \frac{S_t}{S_{t-1}}$, which are the *log-returns*. Conventionally, volatility models are constructed to model the latter.

$$\begin{aligned} X_t &= V_t Z_t \\ V_t &= V(t, X_{1:t-1}, W_{1:t-1}) \end{aligned}$$

inherently mean-zero, make no attempt at modeling expectation

**Lemma 2.0.0.1** (Ito's Lemma)**.** *Consider a stochastic process $(S_t)$ that satisfies the stochastic differential equation $dS_t = \mu_t \, dt + \sigma_t \, dW_t$. Let $f(t, S_t)$ be a twice differentiable function mapping to $\mathbb{R}$. Then,*

$$df = \left(\partial_t f + \mu_t \partial_S f + \frac{1}{2}\sigma_t^2 \partial_S^2 f\right) dt + \sigma_t \partial_S f \, dW_t$$

$$\mathrm{d} \log S_t = \frac{\sigma_t}{S_t} \mathrm{d}\widetilde{W}_t$$

## 2.1 Deterministic Volatility Models

GBM

$$\begin{aligned} \mathrm{d}S_t &= \mu S_t \, \mathrm{d}t + \sigma S_t \, \mathrm{d}W_t \\ X_t &= \sigma Z_t \end{aligned}$$

Ornstein-Uhlenbeck

$$\begin{aligned} \mathrm{d}S_t &= \kappa(\mu - S_t) \, \mathrm{d}t + \sigma \, \mathrm{d}W_t \\ X_t &= \frac{\sigma}{S_{t-1}} Z_t \end{aligned}$$

GARCH

$$\begin{aligned} X_t &= \sigma_t Z_t \\ \sigma_t^2 &= \omega + \alpha X_{t-1}^2 + \beta \sigma_{t-1}^2 \end{aligned}$$

$$\sigma_t^2 \;=\; \omega \sum_{j=1}^{t} \beta^{j-1} + \alpha \sum_{j=1}^{t} \beta^{j-1} X_{t-j}^2 + \beta^{t-1} \sigma_1^2$$

E-GARCH

$$\begin{aligned} X_t &= \sigma_t Z_t \\ \log \sigma_t^2 &= \omega + \alpha\Big(|Z_{t-1}| - \;\mathbb{E}[|Z_{t-1}|]\Big) + \gamma \widetilde{Z}_{t-1} + \beta \log \sigma_t^2 \end{aligned}$$

GJR-GARCH

$$\begin{aligned} X_t &= \sigma_t Z_t \\ \sigma_t^2 &= \omega + \Big(\alpha + \gamma 1(X_{t-1} > 0)\Big) X_{t-1}^2 + \beta \sigma_{t-1}^2 \end{aligned}$$

## 2.2 Stochastic Volatility Models

Stochastic volatility

$$\begin{aligned} X_t &= V_t Z_t \\ \log V_t^2 &= \omega(1-\alpha) + \alpha \log V_{t-1}^2 + \xi \widetilde{Z}_t \end{aligned}$$

$Z_t, \widetilde{Z}_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $\mathbb{C}\text{ov}(Z_t, \widetilde{Z}_t) = \rho$

Heston model

$$\begin{aligned} \mathrm{d}S_t &= \mu S_t \,\mathrm{d}t + V_t S_t \,\mathrm{d}W_t \\ \mathrm{d}V_t^2 &= \kappa(\nu - V_t^2)\,\mathrm{d}t + \xi V_t \,\mathrm{d}\widetilde{W}_t \end{aligned}$$

$$\begin{aligned} X_t &= V_t Z_t \\ \log V_t &= \log V_{t-1}^2 + \kappa \left( \frac{\nu}{V_{t-1}^2} - 1 \right) - \frac{1}{2} \frac{\xi^2}{V_{t-1}^2} + \frac{\xi}{V_{t-1}} \widetilde{Z}_t \end{aligned}$$

$\mathbb{C}\text{ov}(Z_t, \widetilde{Z}_t) = \rho$

SABR model

$$\begin{aligned} \mathrm{d}S_t &= V_t S_t^{\beta} \,\mathrm{d}W_t \\ \mathrm{d}V_t^2 &= \alpha V_t^2 \,\mathrm{d}\widetilde{W}_t \end{aligned}$$

$$\begin{aligned} X_t &= V_t S_{t-1}^{\beta-1} Z_t \\ \log V_t^2 &= \alpha \widetilde{Z}_t \end{aligned}$$

Quintic Ornstein-Uhlenbeck

$$\begin{aligned} \mathrm{d}S_t &= V_t S_t \,\mathrm{d}W_t \\ \mathrm{d}V_t^2 &= \sqrt{\xi}\frac{p(X_t)}{\sqrt{\mathbb{E}[p(X_t)^2]}}, \quad p(x) = \alpha_0 + \alpha_1 x + \alpha_3 x^3 + \alpha_5 x^5 \\ \mathrm{d}X_t &= \varepsilon^{-1}(H - 1/2)X_t \,\mathrm{d}t + \varepsilon^{H-1/2}\,\mathrm{d}W_t \end{aligned}$$

## 2.3 Neural Volatility Models

Neural GARCH

$$\begin{aligned} X_t &= \sigma_t Z_t \\ \sigma_t^2 &= f_\phi(X_{t-1}, \sigma_{t-1}^2) \end{aligned}$$

Neural stochastic volatility

$$\begin{aligned} X_t &= V_t Z_t \\ V_t^2 &= f_\phi(V_{t-1}, \widetilde{Z}_t) \end{aligned}$$

## 2.4 Path-Dependent Volatility

The models introduced thus far all make one strong assumption: that the volatility today is uniquely determined by metrics of yesterday. In general, volatility may depend on the entire past path of observed as well as latent processes.

Consider first volatility models in which path-dependence enters only through the price process itself, so that $\sigma_t^2 = f(X_{1:t-1})$ for some function $f : \mathcal{X}^t \to \mathbb{R}_+$.

One approach is to handcraft a specific form in which path-dependence enters the model. For instance, Guyon and Lekeufack (2023) propose a model wherein volatility is a simple affine function of a measure of the recent trend $\mathcal{T}$ and recent volatility $\Sigma$,

$$\sigma_t = \beta_0 + \beta_1 \mathcal{T}_t + \beta_2 \Sigma_t$$

where they define

$$\begin{aligned} \mathcal{T}_t &:= \sum_{t' \leq t} k_1(t - t') X_{t'} \\ \Sigma_t^2 &:= \sum_{t' \leq t} k_2(t - t') X_{t'}^2 \end{aligned}$$

for some kernel functions $k_1, k_2 : \mathbb{R}_+ \to \mathbb{R}_+$.

As in the previous section, neural networks allow to take a more data-driven approach also for path-dependent settings. *Recurrent neural networks* (RNNs) maintain a hidden state $\boldsymbol{h}_t$ which is recursively updated over time and hence are theoretically able to capture all relevant information of a path of observations.

A simple application of RNNs to deterministic volatility models might look as follows:

$$\begin{aligned} X_t &= \sigma_t Z_t \\ \sigma_t^2 &= f_\phi(\boldsymbol{h}_t) \\ \boldsymbol{h}_t &= g_\phi(\boldsymbol{h}_{t-1}, X_{t-1}) \end{aligned}$$

Similar applications have been proposed for stochastic volatility, e.g. by Luo, Zhang, Xu, and Wang (2018).

Although RNNs are theoretically able to learn functions on path spaces, they are limited by numerical issues in practice. Since the hidden state update $g_\phi(\cdot)$ is applied recursively, gradient information tends to be lost over time (”vanishing gradient problem”), and hence long-range dependencies are only very inefficiently learned. The more refined GRU or LSTM architectures are thus often preferred in practice. However, the idea of path-dependence is the same, and as is more illustrated in detail in Sect. **?**

# Chapter 3

# Bayesian Model Assessment, Selection, & Comparison

## 3.1 Basics of Bayesian Statistics

$$\pi(\theta \,|\, x) \;=\; \frac{\pi(\theta)\, p(x \,|\, \theta)}{p(x)}$$

## 3.2 Model Assessment

Goals: (i) identify ”correct” model, (ii) identify model with superior predictive accuracy

in-sample predictive error: downward biased, even for Bayesian models; cross-validation popular in practice, but computationally expensive; information criteria adjust for bias

Bayesian cross-validation: same as standard cross-validation, but assess mean predictive density on test set $p(X_{test}) = \int p(X_{test} \,|\, \theta)\, \mathrm{d}\pi(\theta \,|\, X_{train})$

asymptotically (for $N \to \infty$) xyIC equivalent to cross-validation (also for time series??)

**Definition 3.2.0.1** (Evidence)**.** *Consider a probabilistic, parametric model for the data-generating process of a dataset $X_{1:t}$ parametrized by $\theta \in \Theta$ with prior law $\pi_0(\theta)$. The evidence $Z_t$ of the model given data $X_{1:t}$ is*

$$Z_t \;:=\; \int_\Theta p(X_{1:t} \,|\, \theta)\; d\pi_0(\theta)$$

Equivalently, the evidence of a model can also be seen as the marginal likelihood $p(X_{1:t})$ of the data or as the expected likelihood of the dataset, $\mathbb{E}_{\pi_0}\Big[p(X_{1:t} \,|\, \theta)\Big]$.

### 3.2.1 Information Criteria

**Definition 3.2.1.1** (Bayesian Information Criterion)**.**

**Definition 3.2.1.2** (Deviance Information Criterion).

$$\begin{aligned} DIC &:= -2\log p(X \mid \hat{\theta}) + 2k_{\mathit{eff}} \\ k_{\mathit{eff}} &= 2\left(\log p(X \mid \hat{\theta}) - \mathbb{E}_{\pi_t}\left[\log p(X \mid \theta)\right]\right) \end{aligned}$$

**Definition 3.2.1.3** (Watanabe–Akaike Information Criterion).

plug-in predictive distribution

$$p(X_{t+1} \mid X_{1:t}) \;=\; \int_{\Theta} p(X_{t+1} \mid X_{1:t}, \theta)\, \mathrm{d}\pi_t(\theta \mid X_{1:t})$$

## 3.3 Model Comparison

### 3.3.1 Bayes Factors

# Chapter 4

# Sequential Monte Carlo Methods

If $\mathcal{V}$ was finite, the posterior is similarly tractable and can be computed following a procedure known as the Baum-Welch algorithm.

## 4.1 Basics of Monte Carlo Estimation

Consider a random variable $\theta \in \Theta$ distributed according to (cumulative) distribution function $\pi$. On the one hand, we are interested in the expectation of a test function $f : \Theta \to \mathbb{R}^d$,

$$\mathbb{E}_\pi[f(\theta)] \;=\; \int_\Theta f(\theta)\,\mathrm{d}\pi(\theta)$$

Depending on $\Theta$, $f$, and $\pi$, this integral quickly becomes impossible to solve. The central dogma of Monte Carlo estimation is to exploit a approximation of $\pi$ based on individual samples, or "particles" therefrom, giving rise to its particle approximation, or empirical distribution function

$$\hat{\pi}(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} 1(\theta \le \theta_i)$$

Since $\hat{\pi}$ is piecewise constant and always jumps by $\frac{1}{n}$, such a particle approximation of $\pi$ reduces the problem to a much more tractable summation task, namely

$$\mathbb{E}_{\hat{\pi}}[f(\theta)] \;=\; \int_\Theta f(\theta)\,\mathrm{d}\hat{\pi}(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} f(\theta_i)$$

This basic idea can be applied in any setting where distributions can be approximated by particle approximations and test functions can be evaluated pointwise. For instance, under a Bayesian paradigm, we are interested in the *evidence* of a model given some data $x$, which is defined as

$$p(x) \;=\; \mathbb{E}_\pi[p(x \mid \theta)] \;=\; \int_\Theta p(x \mid \theta)\,\mathrm{d}\pi(\theta)$$

$$\widehat{p}(x) \;=\; \frac{1}{n}\sum_{i=1}^{n} p(x \mid \theta_i)$$

i.e. the average likelihood of the data over the sampled parameters.

## 4.2 Importance Sampling

Consider the following trivial identities:

$$\pi(\theta) \;=\; \int_{\theta' \leq \theta} \mathrm{d}\pi(\theta') = \int_{\theta' \leq \theta} \frac{\mathrm{d}\pi(\theta')}{\mathrm{d}\nu(\theta')}\,\mathrm{d}\nu(\theta') \;=\; \int_{\theta' \leq \theta} w(\theta')\,\mathrm{d}\nu(\theta')$$

and hence for $x_i \overset{\text{iid}}{\sim} q$, $i = 1, ..., n$,

$$\hat{\pi}(\theta) \;=\; \frac{1}{N}\sum_{i=1}^{N} w(\theta_i) 1(\theta \leq \theta_i)$$

where $\nu$ is another distribution function called the "proposal" and $w(\theta) := \frac{\mathrm{d}\pi(\theta)}{\mathrm{d}\nu(\theta)}$ is the Radon-Nikodym derivative of $\pi$ w.r.t. $\nu$. This suggests that if it is not possible to sample from $\pi$, but from $\nu$ and to evaluate $w(\theta)$ pointwise, we can still use Monte Carlo techniques to estimate quantities of $\pi$.

The weight $w(\theta)$ corrects for the fact that $\theta$ was drawn from the wrong distribution. Intuitively, if a specific particle $\theta_i$ is common under the target $\pi$ but rare under the proposal $\nu$, it needs to be given high weight, and vice versa.

Formally, the only requirement is that $w$ is always well-defined, i.e. that $\sup_\theta \frac{\mathrm{d}\pi(\theta)}{\mathrm{d}\nu(\theta)} \leq M$ for some $M < \infty$. This in turn requires that $\mathrm{d}\nu(\theta) = 0$ only when $\mathrm{d}\pi(\theta) = 0$, implying that the support of $\nu$ must contain the support of $\pi$. Additionally, $\nu$ must have fatter tails than $\pi$ (if the latter's support is unbounded). Otherwise, if e.g. the density of $\nu$ decayed more rapidly towards zero than that of $\pi$, $w$ would of course grow indefinitely.

In many practical applications, $\pi$ is only known up to a multiplicative factor. That is, we know $\gamma(\theta)$ such that

$$\frac{1}{Z}\gamma(\theta) \;=\; \pi(\theta)$$

where $Z := \int_\Theta \mathrm{d}\gamma(\theta)$ is an unknown normalizing constant ensuring that $\pi(\theta) \overset{\theta\to\infty}{\longrightarrow} 1$. In this case, one has to work with the unnormalized weights $W(\theta) := \frac{\mathrm{d}\gamma(\theta)}{\mathrm{d}\nu(\theta)}$. It turns out that these are unbiased estimates of the unknown normalizing constant $Z$:

$$\mathbb{E}_\nu\left[W(\theta)\right] \;=\; \int_\Theta W(\theta)\,\mathrm{d}\hat{\pi}(\theta) \;=\; \int_\Theta \mathrm{d}\gamma(\theta) \;=:\; Z$$

suggesting use of the estimator

$$\widehat{Z} \;:=\; \frac{1}{N}\sum_{i=1}^{N} W(\theta_i)$$

In the context of model selection, $Z$ is itself of prime interest, and hence this is a relevant result.

$$\mathbb{V}_\nu(\widehat{Z}) \;=\; \frac{1}{N}\mathbb{V}_\nu\left(W(\theta)\right)$$

Since, if $\nu = \pi$, the unnormalized weights are always 1, and hence $\widehat{Z}$ has zero variance, it is clear that the optimal proposal should be as close to the target as possible.

$$\nu^* \;=\; \pi$$

$$\hat{\pi}(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{N}\frac{W(\theta_i)}{\sum_{i=1}^{n} W(\theta_i)}1(\theta \leq \theta_i)$$

$$\text{ESS} \;:=\; \frac{N}{1+\mathbb{V}_{\nu}(W(\theta))}$$

## 4.3 Sequential Importance Sampling

Suppose our unknown quantity $\theta$ is a parameter in the distribution $p(x)$ of $x$. Observing data $x_{1:t} = \{x_1, ..., x_t\}$ carrying information about this quantity gives rise to the target distribution $\pi_t(\theta)$

If all the observations $y_{1:t}$ arrive simultaneously, batch inference is a popular approach. Therein, one attempts to estimate the distribution by computing resp. approximating $\pi_t$ directly. Some of the most popular batch inference methods include Markov chain Monte Carlo (MCMC) and Hamiltonian Monte Carlo (HMC).

When a new observation $x_{t+1}$ arrives, however, batch inference methods have no direct way to use $\pi_t$. Consequently, when data arrives sequentially, one is left with no other choice than running the entire analysis again every time. This is a severe limitation in fields including finance, where often data arrives and hence models have to be updated daily if not more frequently.

In sequential Bayesian inference, the idea is to construct intermediate distributions and sequentially update them to incorporate new observations. That is, they are inherently on-line by construction. This is of primary interest when a model needs to be estimated today but new data will inevitably arrive tomorrow. It has also been demonstrated to have attractive properties in batch inference.

$$X_t \,|\, (X_{t-1} = x_{t-1}, \theta) \;\sim\; P_{\theta}(\cdot \,|\, x_{t-1})$$

where $P_{\theta}(\cdot) = P(\cdot \,|\, \theta)$.

$$\pi_t(\theta) \;:=\; \pi(\theta \,|\, x_{1:t})$$

$$\begin{aligned}
\pi_t(\theta) &= \frac{1}{Z_t}\gamma_t(\theta); \quad Z_t \;=\; \int_{\Theta} \mathrm{d}\gamma_t(\theta) \\
\mathrm{d}\gamma_t(\theta) &= P_{\theta}(x_{1:t})\,\mathrm{d}\pi_0(\theta) \;=\; \mathrm{d}\pi_0(\theta)\cdot\prod_{k=1}^{t} P_{\theta}(x_k \,|\, x_{k-1})
\end{aligned}$$

Through another trivial identity, we obtain a recursive expression for the weights, namely

$$\begin{aligned}
W_t(\theta) \;:=\; \frac{\mathrm{d}\gamma_t(\theta)}{\mathrm{d}\nu_t(\theta)} &= \frac{\mathrm{d}\gamma_{t-1}(\theta)}{\mathrm{d}\nu_{t-1}(\theta)}\cdot\frac{\mathrm{d}\gamma_t(\theta)\,/\,\mathrm{d}\gamma_{t-1}(\theta)}{\mathrm{d}\nu_t(\theta)\,/\,\mathrm{d}\nu_{t-1}(\theta)} \\
&= W_{t-1}(\theta)\cdot\alpha_t(\theta)
\end{aligned}$$

where $\alpha_t(\theta) := \frac{\mathrm{d}\gamma_t(\theta)\,/\,\mathrm{d}\gamma_{t-1}(\theta)}{\mathrm{d}\nu_t(\theta)\,/\,\mathrm{d}\nu_{t-1}(\theta)}$ is the weight update from $t-1$ to $t$. If use a constant proposal distribution at each iteration, ignoring new information from the observations $x_{1:t}$, then $\alpha_t$ corresponds to the conditional likelihood of the $t^{\text{th}}$ observation

$$\frac{\mathrm{d}\gamma_t(\theta)}{\mathrm{d}\gamma_{t-1}(\theta)} = P_\theta(x_t \,|\, x_{t-1})$$

An estimate for the normalizing constant is obtained akin to standard importance sampling

$$\widehat{Z}_t = \frac{1}{n}\sum_{i=1}^{N} W_t(\theta_i)$$

## 4.4 Resampling

**Algorithm 1** Iterated Batch Importance Sampling (IBIS)

```
    Input $\pi_0$, $P$, $K_t$, resample
$\theta_0^{(i)} \sim \pi_0(\theta)$                                              ▷ generate particles from prior
$W_0^{(i)} \leftarrow 1$                                                       ▷ initialize uniform weights
for $t = 1$ to $T$ do
    if ESS < $\mathrm{ESS}_{\min}$ then
        $\tilde\theta_{t-1}^{(i)} \leftarrow$ resample($\theta_{t-1}^{(1:N)}; w_{t-1}^{(1:N)}$)   ▷ resample particles
        $\widetilde{W}_{t-1}^{(i)} \leftarrow 1$                                     ▷ reset weights
    else
        $\tilde\theta_{t-1}^{(i)} \leftarrow \theta_{t-1}^{(i)}$                     ▷ maintain particles
        $\widetilde{W}_{t-1}^{(i)} \leftarrow W_{t-1}^{(i)}$                         ▷ maintain weights
    end if
    $\theta_t^{(i)} \sim K_t(\cdot \,|\, \tilde\theta_{t-1}^{(i)})$                  ▷ propagate particles
    $\alpha_t^{(i)} \leftarrow P(x_t \,|\, x_{t-1}; \theta_t^{(i)})$                 ▷ compute weight update
    $W_t^{(i)} \leftarrow \widetilde{W}_{t-1}^{(i)} \cdot \alpha_t^{(i)}$            ▷ update weights
    $w_t^{(i)} \leftarrow W_t^{(i)} / \sum_{j=1}^N W_t^{(j)}$                        ▷ re-normalize weights
end for
```

In the original paper, Chopin (2002) proposes to use an adaptive MCMC scheme, wherein the target is approximated through a Gaussian distribution with the empirical mean $\hat\mu$ and covariance $\widehat\Sigma$ of the particles. New particles are then proposed via a Gaussian random walk, so that $K_t(\cdot \,|\, \theta) = \mathcal{N}(\theta, \widehat\Sigma)$

## 4.5 Particle Filtering

In stochastic models, the evolution of $(X_t)$ depends that of another, latent stochastic process $(V_t)$. The models proposed in Sect. 2 are all of the form

$$\begin{aligned} X_t \,|\, (X_{t-1} = x_{t-1}, V_t = v_t) &\sim P(\cdot \,|\, x_{t-1}, v_t) \\ V_t \,|\, V_{t-1} = v_{t-1} &\sim K(\cdot \,|\, v_{t-1}) \end{aligned}$$

where $P$ is a likelihood function for the observations and $K$ is a transition kernel. In stochastic volatility and other models, $V_t$ can be interpreted as the current "state" of the system, hence such models are referred to as *state-space models.* For the moment, suppose that $P$ and $K$ do not depend on any (unknown) parameters.

Our target distribution is that of the latent path $v_{1:t}$,

$$\pi_t(v_{1:t}) \;=\; \frac{1}{Z_t}\prod_{j=1}^{t} K_\theta(v_j \,|\, v_{j-1}) P_\theta(x_j \,|\, x_{j-1}, v_j) \;=\; \frac{1}{Z_t}\gamma_t(v_{1:t})$$

where we define prior distributions $P(x_1 \,|\, x_0, v_1) := \eta_X(x_1 \,|\, v_1)$ and $K(v_1 \,|\, v_0) := \eta_V(v_1)$. From this distribution on the path space, we can deduce all distributions of interest through integration. These include the *filtering distribution* $\pi_t(v_t \,|\, v_{1:t-1})$, evidence $p(x_{1:t}) = \int \pi_t(v_t \int)$

We can again readily extend the sequential importance sampling framework to approximate the new target distribution. Borrowing the expression from the static setting, we have the recursion for the weights

$$W_t(v_{1:t}) \;=\; W_{t-1}(v_{1:t}) \cdot \frac{\mathrm{d}\gamma_t(v_t \,|\, v_{1:t-1})}{\mathrm{d}\nu_t(v_t \,|\, v_{1:t-1})}$$

where the numerator of the weight update is again just a conditional likelihood, this time of the new latent quantity $v_t$ given the previous ones, $v_{1:t-1}$, and the observations $x_{1:t}$. It is given by

$$\mathrm{d}\gamma_t(v_t \,|\, v_{1:t-1}) \;=\; P(x_t \,|\, v_t)K(v_{1-t} \,|\, v_t)$$

This gives rise to the so called *guided particle filter* outlined in Alg. 4.5.

**Algorithm 2** (Guided) Particle Filter

```
    Input $\eta_V$, $\eta_X$, $P$, $K$, $\mathrm{ESS}_{\min}$, resample
  $v_1^{(i)} \sim \eta_V(v_1)$                                      ▷ generate particles from prior
  $W_1^{(i)} \leftarrow 1$                                          ▷ initialize weights uniformly
  $w_1^{(i)} \leftarrow N^{-1}$
  for $t = 2$ to $T$ do
      $\mathrm{ESS} = 1/\sum_{j=1}^N (W_t^{(i)})^2$                     ▷ compute ESS
      if $\mathrm{ESS} < \mathrm{ESS}_{\min}$ then
          $\widetilde{v}_{t-1}^{(i)} \sim$ resample$(v_{t-1}^{(1:N)}; w_{t-1}^{(1:N)})$    ▷ resample particles
          $\widetilde{W}_{t-1}^{(i)} \leftarrow 1$                            ▷ reset weights
      else
          $\widetilde{v}_{t-1}^{(i)} \leftarrow v_{t-1}^{(i)}$                        ▷ maintain particles
          $\widetilde{W}_{t-1}^{(i)} \leftarrow W_{t-1}^{(i)}$                        ▷ maintain weights
      end if
      $v_t^{(i)} \sim K(\cdot \,|\, \widetilde{v}_{t-1}^{(i)})$                         ▷ propagate particles
      $\alpha_t^{(i)} \leftarrow K(v_t^{(i)} \,|\, v_{t-1}^{(i)})P(x_t \,|\, x_{t-1}, v_t^{(i)}) \,/\, \mathrm{d}\nu_t(v_t \,|\, v_{t-1})$   ▷ compute weight update
      $W_t^{(i)} \leftarrow \widetilde{W}_{t-1}^{(i)} \cdot \alpha_t^{(i)}$                   ▷ update weights
      $w_t^{(i)} \leftarrow W_t^{(i)} / \sum_{j=1}^N W_t^{(j)}$                 ▷ re-normalize weights
  end for
```

In particle filtering, unlike in IBIS, particles do not necessarily have to be moved after resampling to introduce variation, since they are propagated either way.

A problem that remains is that of the choice of proposal distribution. Akin to standard importance sampling, the optimal proposal is again the true law,

$$\mathrm{d}\nu_t^*(v_t \,|\, v_{t-1}) \;=\; \pi_t(v_t \,|\, v_{t-1}) \;=\; \frac{P(x_t \,|\, v_t)K(v_t \,|\, v_{t-1})}{\int P(x_t \,|\, v_t)K(v_t \,|\, v_{t-1})\,\mathrm{d}x_t}$$

However, this is often intractable in practice. A popular alternative is to ignore the information about $v_t$ brought by the $x_t$ and simply use $\nu_t(v_t \,|\, v_{t-1}) = K(v_t \,|\, v_{t-1})$. This is referred to as the *bootstrap proposal.* Naturally, its inefficiency is smaller whenever $x_t$ not very informative about $v_t$.

## 4.6 SMC Samplers

$$\begin{aligned}
\pi_t(v_{1:t}, \theta) &= \frac{1}{Z_t}\gamma_t(v_{1:t}, \theta) \\
\mathrm{d}\gamma_t(v_{1:t}, \theta) &= L(x_{1:t} \,|\, v_{1:t}, \theta)\,\mathrm{d}\pi_0(v_{1:t}, \theta) \\
&= L(x_1 \,|\, v_1, \theta)\,\mathrm{d}\pi_0(v_1, \theta)\prod_{k=2}^{t} L(x_k \,|\, v_k, \theta)K(v_k \,|\, v_{k-1}, \theta) \\
Z_t &= \int L(x_{1:t} \,|\, v_{1:t}, \theta)\,\mathrm{d}\pi_0(v_{1:t}, \theta)
\end{aligned}$$

$$W_t(v_{1:t}) \;=\; \frac{\mathrm{d}\gamma_t(v_{1:t})}{\mathrm{d}\nu_t(v_{1:t})}$$

$$\begin{aligned}
W_t(v_{1:t}, \theta) \;:=\; \frac{\mathrm{d}\gamma_t(v_{1:t}, \theta)}{\mathrm{d}\nu_t(v_{1:t}, \theta)} &= \frac{\mathrm{d}\gamma_{t-1}(v_{1:t-1}, \theta)}{\mathrm{d}\nu_{t-1}(v_{1:t-1}, \theta)} \cdot \frac{\mathrm{d}\gamma_t(v_t \,|\, v_{1:t-1}, \theta)}{\mathrm{d}\nu_t(v_t \,|\, v_{1:t-1}, \theta)} \\
&= W_{t-1} \cdot \frac{\mathrm{d}\gamma_t(v_t \,|\, v_{t-1})\,\mathrm{d}\gamma_t(x_t \,|\, x_{t-1}, v_t)}{\mathrm{d}\nu_t(v_t \,|\, v_{t-1})\,\mathrm{d}\nu_t(x_t \,|\, x_{t-1}, v_t)}
\end{aligned}$$

### 4.6.1 Rao-Blackwell Particle Filtering

**Theorem 4.6.1.1** (Rao-Blackwell)**.**

$$\mathbb{E}\left[\left(\mathbb{E}\left[\hat{\theta}(X_{1:t}, \tau) \,|\, \tau(X_{1:t})\right] - \theta\right)^2\right] \;\leq\; \mathbb{E}\left[\left(\hat{\theta}(X_{1:t}) - \theta\right)^2\right]$$

**Lemma 4.6.1.2** (Student-t distribution)**.** *If* $\tau \sim$ *Gamma*$(\nu/2, \nu/2)$ *and* $X \,|\, \tau \;\sim\; \mathcal{N}(0, \tau^{-1})$, *then* $X \sim t_\nu$.

# Chapter 5

# Reservoir Computing

## 5.1 Random Neural Networks

*random feature models*,

*Extreme Learning Machines* (ELMs). An ELM is a shallow feedforward neural network with randomly drawn inner weights. In the one-dimensional case, we have

$$\begin{aligned} \mathrm{ELM}(\boldsymbol{x}) &:= \boldsymbol{\beta}^\top \phi(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}) + \beta_0 \\ \boldsymbol{A}_{ij}, \boldsymbol{b}_i &\overset{\text{iid}}{\sim} \end{aligned}$$

*echo state network*

## 5.2 Signature Methods

**Definition 5.2.0.1** (Signature)**.** *Let $\boldsymbol{u} \in BV([0,T];\mathbb{R}^p)$. The signature component associated with the word $w \in \{0, ..., d\}^k$ of length $k \geq 0$ is*

$$\mathcal{S}^{(w)}(\boldsymbol{u}) := \int_{0 \leq t_1 \leq \cdots \leq t_k \leq t} du_{t_1}^{(i_1)} \cdots du_{t_k}^{(i_k)}$$

$\mathcal{S}((t, \boldsymbol{u}))$ is point-separating. Hence signature effective feature extraction method for time series and well suited to capture path-dependency.

$$\begin{aligned} X_t &= \sigma_t Z_t \\ \sigma_t &= \beta_0 + \boldsymbol{\beta}^\top \mathcal{S}(X_{1:t}) \end{aligned}$$

**Theorem 5.2.0.2** ((Cuchiero, Primavera, and Svaluto-Ferro, 2022))**.**

$$\left| f(\boldsymbol{u}) - \ell(\mathcal{S}(\boldsymbol{u})) \right| \leq \varepsilon$$

Continuous path functionals can be uniformly approximated on compact sets by linear functionals of the time-extended signature evaluated at the final time

**Theorem 5.2.0.3** (Cuchiero et al., 2021)**.** $\mathcal{W}^k = \{0, ..., d\}^k$

$$f(X_t) \;=\; \sum_{k=1}^{\infty} \sum_{(i_1,\ldots,i_k)\,\in\,\mathcal{W}^k} V_{i_1}\cdots V_{i_k} f(X_0)\cdot \mathcal{S}_t^{(i_1,\ldots,i_k)}(\boldsymbol{u})$$

Signature path

$$\mathbb{S}(\boldsymbol{u}) \;:=\; \Big(\mathcal{S}_{t'}(\boldsymbol{u})\Big)_{0\le t'\le t}$$

Truncated signature

$$\mathcal{S}^M(\boldsymbol{u}) \;:=\; \int_{0\le t_1\le\cdots\le t_k\le t} \mathrm{d}u_{t_1}^{i_1}\cdots \mathrm{d}u_{t_k}^{i_k}$$

Let $f(x) = x^2$, then $\mathbb{E}[f(X_t)] = V_t$.

$$\mathbb{E}\left[X_t^2\right] \;=\; \sigma_t^2 \;=\; \sum_{k=1}^{M} \sum_{(i_1,\ldots,i_k)\,\in\,\mathcal{W}^k} \frac{\sigma_{i_1}}{S_{i_1}}\cdots\frac{\sigma_{i_k}}{S_{i_k}} X_0^2\cdot \mathbb{E}\left[\mathcal{S}_t^{(i_1,\ldots,i_k)}(W)\;\right]$$

Randomized signature

**Theorem 5.2.0.4** ((Cuchiero, Gonon, Grigoryeva, Ortega, and Teichmann, 2021))**.** *Consider random matrices $\boldsymbol{A}_i \in \mathbb{R}^{q\times q}$, random vectors $\boldsymbol{b}_i \in \mathbb{R}^q$, and activation function $\phi : \mathbb{R} \to \mathbb{R}$ (applied component-wise). For a random vector $\widetilde{\mathcal{S}} \in \mathbb{R}^q$ which is the solution to the controlled ordinary differential equation*

$$d\widetilde{\mathcal{S}}_t \;=\; \sum_{i=1}^{d} \phi\Big(\boldsymbol{A}_i\widetilde{\mathcal{S}}_t + \boldsymbol{b}_i\Big)\, du_t^{(i)}$$

*it holds that*

$$(1-\varepsilon)\left\|\mathcal{S}(\boldsymbol{u}_t) - \mathcal{S}(\boldsymbol{u}_t)\right\|_2 \;\le\; \left\|\widetilde{\mathcal{S}}(\boldsymbol{u}) - \widetilde{\mathcal{S}}(\boldsymbol{u}')\right\|_2 \;\le\; (1+\varepsilon)\left\|\mathcal{S}(\boldsymbol{u}) - \mathcal{S}(\boldsymbol{u})\right\|_2$$

Expected randomized signature

$$\mathbb{E}\left[\widetilde{\mathcal{S}}(W)\right] \;=\; \mathbb{E}\left[\int \phi\Big(\boldsymbol{A}\widetilde{\mathcal{S}} + \boldsymbol{b}\Big)\,\mathrm{d}W\right]$$

# Chapter 6

# Summary

Summarize the presented work. Why is it useful to the research field or institute?

## 6.1 Future Work

Possible ways to extend the work.

# Bibliography

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika 89*(3), 539–552.

Cuchiero, C., L. Gonon, L. Grigoryeva, J.-P. Ortega, and J. Teichmann (2021). Expressive power of randomized signature. *35th Conference on Neural Information Processing Systems (NeurIPS 2021).*

Cuchiero, C., F. Primavera, and S. Svaluto-Ferro (2022). Universal approximation theorems for continuous functions of cadlag paths and levy-type signature models. *arXiv preprint*.

Guyon, J. and J. Lekeufack (2023). Volatility is (mostly) path-dependent. *Quantitative Finance*.

Luo, R., W. Zhang, X. Xu, and J. Wang (2018). A neural stochastic volatility model. *Proceedings of the AAAI Conference on Artificial Intelligence 32*(1).

# Appendix A

# Complementary information

# Appendix B

# Yet another appendix....