Nonparametric Bayesian estimation of a Hölder continuous diffusion coefficient

Shota Gugushvili[a], Frank van der Meulen[b], Moritz Schauer[c] and Peter Spreij[d]

[a]*Wageningen University & Research*
[b]*Delft University of Technology*
[c]*Leiden University*
[d]*University of Amsterdam and Radboud University Nijmegen*

Abstract. We consider a nonparametric Bayesian approach to estimate the diffusion coefficient of a stochastic differential equation given discrete time observations over a fixed time interval. As a prior on the diffusion coefficient, we employ a histogram-type prior with piecewise constant realisations on bins forming a partition of the time interval. Specifically, these constants are realizations of independent inverse Gamma distributed randoma variables. We justify our approach by deriving the rate at which the corresponding posterior distribution asymptotically concentrates around the data-generating diffusion coefficient. This posterior contraction rate turns out to be optimal for estimation of a Hölder-continuous diffusion coefficient with smoothness parameter $0 < \lambda \leq 1$. Our approach is straightforward to implement, as the posterior distributions turn out to be inverse Gamma again, and leads to good practical results in a wide range of simulation examples. Finally, we apply our method on exchange rate data sets.

1 Introduction

1.1 Problem description

Stochastic differential equations (SDEs) have been widely used as models in numerous applications ranging from physics (see, for example, Allen (2007)) to engineering (see Wong and Hajek (1985)) and to finance (see Musiela and Rutkowski (2005)). We assume observations from an SDE of the form

$$dX_t = b_0(t, X_t) \, dt + s_0(t) \, dW_t, \quad X_0 = x, t \in [0, T], \tag{1.1}$$

with a drift coefficient b_0, (deterministic) dispersion coefficient s_0, and a deterministic initial condition x. Here X is real valued and W is a Brownian motion. We assume observations

$$\mathcal{X}_n = \{X_{t_{0,n}}, \ldots, X_{t_{n,n}}\}$$

from the solution X to (1.1) are available, where $t_{i,n} = \frac{i}{n}T$, $i = 0, \ldots, n$. Our aim is to estimate s_0 nonparametrically within the Bayesian setup.

Model (1.1) covers the case of linear SDEs, such as the popular Ornstein-Uhlenbeck process; see, for example, Section 5.6 in Karatzas and Shreve (1988). Among references that study (1.1) as models for log returns of asset prices, we mention Lutz (2010) and Mishura (2015), but the model has applications far beyond this context as well. A seemingly more general SDE

$$dX_t = \widetilde{b}_0(t, X_t) \, dt + s_0(t) f_0(X_t) \, dW_t, \quad X_0 = x, t \in [0, T], \tag{1.2}$$

can be reduced to the form (1.1) through a simple transformation of X_t, namely

$$Y_t = x + \int_0^{X_t} \frac{1}{f_0(u)} \, du,$$

provided f_0 is known and sufficiently regular; see p. 186 in Soulier (1998). Financial practitioners often are content with a model of the type (1.2) as a simple and useful generalisation of the Black–Scholes model; see, for example, pp. 7–14 in Gatheral (2006) for Dupire, Derman and Kani's pioneering work on local volatility. In particular, a discretely observed geometric Brownian motion with time-varying coefficients is also a special case of (1.2), once one passes to the corresponding log returns (see Taleb (1997) for additional information and applications). As the drift in (1.1) is allowed to be non-linear, the distribution of X_t is in general not Gaussian and may well exhibit heavy tails, which is attractive from the point of view of financial applications. Finally, in some practical applications it is genuinely important to employ a time-dependent diffusion coefficient; a real data example is given in Section 7.

1.2 Related literature

Statistical inference for SDEs is a well studied and very active field of research, that is far from saturation. Relevant literature can be divided into two categories: works dealing with parametric and works dealing with nonparametric methods. Parametric approaches specify parametric forms for the drift and diffusion coefficients of SDEs. When these specifications use correct functional forms, such methods attain a higher statistical efficiency over the nonparametric ones. On the other hand, nonparametric approaches, where one only assumes qualitative features of the drift and diffusion coefficients, guard one against model misspecification, which may have dramatically negative consequences for valid inference; also, nonparametric techniques may suggest plausible parametric models in those situations where these models cannot be derived from the first principles; see, for example, Silverman (1986). For parametric approaches to inference in SDE models, see, for example, Chapter 2 in Kutoyants (2004), Chapter 3 in Iacus (2008), and references therein. Nonparametric statistical inference for SDEs of the type studied in the present work has been considered in Genon-Catalot, Laredo and Picard (1992), Hoffmann (1997) and Soulier (1998) within the frequentist setup, while Gugushvili and Spreij (2014b) and Gugushvili and Spreij (2016) have explored the problem from the Bayesian perspective. Although the nonparametric methods these papers study are implementable in principle, these works are primarily of theoretical nature and practical performance of the corresponding approaches is not clear. Furthermore, except Gugushvili and Spreij (2014b) and Gugushvili and Spreij (2016), there is hardly any other work available on estimation of the dispersion coefficient (or diffusion coefficient) from the nonparametric Bayesian point of view, which constitutes the central topic of our paper. In this context, we can mention only a theoretical contribution Nickl and Söhl (2017) and a practically oriented paper Batz, Ruttor and Opper (2018), but the models considered there, as well as the sampling scheme, are different from ours, and the theory developed in Nickl and Söhl (2017) does not cover the approach in Batz, Ruttor and Opper (2018). On a general level, apart from a philosophical appeal for Bayesians, advantages of a Bayesian approach include automatic quantification of uncertainty in parameter estimates through Bayesian credible sets, and the fact that it is a fundamentally likelihood-based method (see Berger and Wolpert (1988)). Furthermore, recent practical advances made in nonparametric Bayesian estimation of the drift coefficient, see, for example, van der Meulen, Schauer and van Zanten (2014), van der Meulen and Schauer (2017) and Papaspiliopoulos et al. (2012), would suggest that comparable results can be obtained for estimation of the diffusion coefficient too. We note, however, that from an implementational point of view nonparametric Bayesian estimation of a dispersion coefficient is very different from drift coefficient estimation: the latter

fundamentally relies on the equivalence of laws of continuously observed diffusion processes that have the same diffusion coefficient, which is not applicable when the diffusion coefficient itself is unknown and is a parameter to be estimated.

1.3 Approach and results

The main practical challenges for Bayesian inference in SDE models from discrete observations are an *intractable likelihood* and absence of a closed form expression for the posterior distribution, which complicates considerably the inference; see, for example, Roberts and Stramer (2001), Elerian, Chib and Shephard (2001), Fuchs (2013) and van der Meulen and Schauer (2017). We circumvent these difficulties by intentionally misspecifying the drift coefficient, and employing a (conjugate) histogram-type prior on the diffusion coefficient, that has piecewise constant realisations on bins forming a partition of $[0, T]$ (this is different from Gugushvili and Spreij (2014b) and Gugushvili and Spreij (2016), where the drift b_0 is in fact zero, and other priors are used). Due to this, our nonparametric Bayesian method to estimate the dispersion coefficient s_0 in (1.1) is easily implemented, fast and requires little fine-tuning from the user. We demonstrate its good practical performance on a wide range of simulated data examples and we apply it on real data from finance, yielding interesting conclusions.

On the theoretical side, we investigate the asymptotic performance of our Bayesian procedure from the frequentist point of view. Theoretical analysis of Bayesian procedures for inference in SDE models from discrete observations is in general challenging; see, for example, the contributions van der Meulen and van Zanten (2013), Gugushvili and Spreij (2014a) and Nickl and Söhl (2017) for an impression, albeit in settings different from ours. We consider the 'infill' asymptotics with the time T horizon staying fixed and the number of observations n in the interval $[0, T]$ increasing; this asymptotic regime is standard in the literature and can be thought of as reasonably satisfied in many financial applications. Complicating factors for a theoretical analysis in our setting are due to influence of the unknown drift coefficient b_0, which we have intentionally misspecified. We address this through an argument based on Girsanov's theorem. The main theoretical result we obtain tells us that the drift misspecification is asymptotically harmless and our procedure for estimating the diffusion coefficient is consistent at rate $n^{-\beta}$ in the L_2-norm, with the precise value of β depending on the smoothness of the true dispersion coefficient. The corresponding posterior contraction rate is optimal for estimation of Hölder smooth dispersion coefficients of order $0 < \lambda \leq 1$.

1.4 Organisation of this paper

Section 2 contains the model specification and a detailed description of our nonparametric Bayesian approach. Section 3 contains the theoretical results. Our Bayesian method depends on a hyperparameter, the number of bins forming the partition of the interval $[0, T]$, and Sections 4 and 5 discuss possible practical methods of its choice. In Section 6 we investigate practical performance of our method via simulations and provide illustrations of our theory from Section 3. Examples with real data are studied in Section 7. Proofs of the results from Section 3 can be found in Section 8. In the Appendix, we state and prove an additional theoretical result.

1.5 Frequently used notation

We denote by $\| \cdot \|_2$ the L_2-norm with respect to the Lebesgue measure on the Borel sets of $[0, 1]$. We use the following notation to compare two sequences $\{a_n\}$ and $\{b_n\}$ of positive real numbers: $a_n \lesssim b_n$ (or $b_n \gtrsim a_n$) means that there exists a constant $C > 0$ that is independent of n and is such that $a_n \leq C b_n$. As a combination of the two we write $a_n \asymp b_n$ if both $a_n \lesssim b_n$

and $a_n \gtrsim b_n$. We will also write $a_n \gg b_n$ to indicate that $a_n/b_n \to \infty$ as $n \to \infty$. By $a \vee b$ we denote the maximum of two numbers a and b.

We let $\mathbb{P}_{b_0,s_0}$ denote the law of the path $(X_t : t \in [0, T])$ from (1.1) under the true parameter values (b_0, s_0). In particular, the notation $\mathbb{P}_{0,s_0}$ is used for such a law when the drift coefficient is equal to zero. We denote the prior distribution on the dispersion coefficient by Π_n (with n the number of observations) and write the posterior as $\Pi_n(\cdot \mid \mathcal{X}_n)$. We denote the posterior expectation and variance by $\mathbb{E}_{\Pi_n}(\cdot \mid \mathcal{X}_n)$ and $\mathbb{V}\mathrm{ar}_{\Pi_n}(\cdot \mid \mathcal{X}_n)$, respectively.

2 Assumptions and Bayesian setup

We summarise the assumptions on our statistical model.

Assumption 1. Assume that

(a) the model (1.1) is given with $x = 0$ and $T = 1$;
(b) the drift coefficient satisfies a linear growth condition and is Lipschitz in its second argument: for some $K > 0$ it holds that

$$|b_0(t, x)|^2 \leq K(1 + |x|^2), \quad \forall x \in \mathbb{R},$$

$$|b_0(t, x) - b_0(t, y)| \leq K|x - y|, \quad \forall t \in [0, 1], \forall x, y \in \mathbb{R};$$

(c) the dispersion coefficient s_0 is Hölder continuous on $[0, 1]$ with Hölder constant L and Hölder exponent $\lambda \in (0, 1]$, $|s_0(u) - s_0(v)| \leq L|u - v|^\lambda$ for all $u, v \in [0, 1]$, and is bounded away from zero and (by continuity also from infinity);
(d) a discrete time sample

$$\mathcal{X}_n = \{X_{t_{0,n}}, \ldots, X_{t_{n,n}}\}$$

from the solution X to (1.1) is available, where $t_{i,n} = i/n$, $i = 0, \ldots, n$. For future reference, we also define $Y_{i,n} = X_{t_{i,n}} - X_{t_{i-1,n}}$.

Under Assumption 1, equation (1.1) admits a unique strong solution, see, for example, Theorems 2.5 and 2.9 of Section 5.2 in Karatzas and Shreve (1988). The minimal regularity conditions on the dispersion coefficient s_0 in Assumption 1(c) are needed for our asymptotic statistical theory to work; see Section 3. The sampling scheme in Assumption 1(d) is referred to as the high-frequency data setting and is a popular asymptotic setup for inference in SDE models, see, for example, Dette, Podolskij and Vetter (2006), Florens-Zmirou (1993), Genon-Catalot, Laredo and Picard (1992), Hoffmann (1997), Hoffmann (1999b), Jacod (2000) and Soulier (1998). Financial data are often of a similar type, see Sabel, Schmidt-Hieber and Munk (2015). As remarked in Ignatieva and Platen (2012), p. 1334, "while the estimation of a drift coefficient function is theoretically of major interest in finance, for instance, for portfolio optimization, it can practically rarely be achieved. It is a fact that accurate estimation of drift coefficient functions requires considerably longer time series than typically available". This example, then, provides an instance of a setting when the assumption $T \to \infty$ cannot be thought to be reasonably satisfied. Other sampling schemes have been also considered in the literature on nonparametric inference for SDE models, for example, when the time horizon $T \to \infty$, but the distance Δ between observation times stays fixed (the so-called low frequency data setting; see, for example, Gobet, Hoffmann and Reiß (2004) and Nickl and Söhl (2017)). Alternatively, one can assume observations are available at times $\Delta, 2\Delta, \ldots$, and consider the case $\Delta \to 0$, $T = n\Delta \to \infty$, see, for example, Hoffmann (1999a) (this is again referred to as the high-frequency data setting). The question which of the asymptotic

regimes reasonably applies to a given dataset in practice can be decided only on the case by case basis.

Our first observation is that under the sampling scheme as in Assumption 1(d), consistent estimation of the drift coefficient b_0 is impossible; see, for example, Mai (2014), p. 919. Furthermore, in many contexts, for example, when pricing financial derivatives, knowledge of the drift coefficient is in fact of no interest, whereas the dispersion coefficient is of paramount importance; see Musiela and Rutkowski (2005). This motivates us to completely ignore the drift coefficient in our estimation procedure by intentionally misspecifying the model and acting as if the drift were equal to zero. We will justify this in Section 3. Then the pseudo-likelihood associated with our observations is Gaussian and is given by

$$L_n(s) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi \int_{t_{i-1,n}}^{t_{i,n}} s^2(u)\, du}} \psi \left(\frac{Y_{i,n}}{\sqrt{\int_{t_{i-1,n}}^{t_{i,n}} s^2(u)\, du}} \right) \right\}, \tag{2.1}$$

where $\psi(u) = \exp(-u^2/2)$. Gaussian pseudo-likelihood is a widely used object in statistics, see, for example, Section 5.2 in Brockwell and Davis (2002), Dimitriou-Fakalou (2014) and Hualde and Robinson (2011) for some examples. Setting the drift to zero in our setting has a practical advantage of obtaining a simple and tractable expression for the pseudo-likelihood.

With Π_n denoting a prior on the dispersion coefficient, provided all the involved quantities are suitably measurable, Bayes' theorem gives that the posterior probability of any measurable subset $S \subset \mathcal{S}$ of dispersion coefficients is given by

$$\Pi_n(S \mid X_{t_{0,n}} \ldots, X_{t_{n,n}}) = \frac{\int_S L_n(s) \Pi_n(\mathrm{d}s)}{\int_{\mathcal{S}} L_n(s) \Pi_n(\mathrm{d}s)},$$

where $\mathcal{S}$ denotes a space on which the prior Π_n is defined. From the above display various point estimates of s_0 can be obtained, such as the posterior mean.

It follows from (2.1) that the likelihood depends on the parameter of interest only through the integrals $\int_{t_{i-1,n}}^{t_{i,n}} s^2(u)\, du$ and is otherwise 'blind' to precise values the diffusion coefficient takes inside the intervals $[t_{i-1,n}, t_{i,n}]$. Consequently, it appears natural to a priori model the diffusion coefficient as piecewise constant on intervals $[t_{i-1,n}, t_{i,n}]$. However, some smoothing should also be performed, and this can be achieved by aggregated several neighbouring intervals $[t_{i-1,n}, t_{i,n}]$. Thus, to construct the prior, we proceed as follows: Let m be an integer smaller than n. Then we can uniquely write $n = mN + r$ with $0 \le r < m$, and in fact $N = \lfloor \frac{n}{m} \rfloor$. Both m and N will depend on n (and we also write m_n and N_n to emphasize this when appropriate). With this assumption, we have bins $B_k = [t_{m(k-1),n}, t_{mk,n})$, $k = 1, \ldots, N-1$ and $B_N = [t_{m(N-1),n}, 1]$. Note that the length of B_k is then equal to m/n for $k \le N-1$, whereas B_N has length $1 - t_{m(N-1),n} = \frac{r+m}{n} < \frac{2m}{n}$. For notational convenience later on, we also write

$$B_k = [a_{k-1}, a_k), \quad k = 1, \ldots, N-1,$$

$$B_N = [a_{N-1}, a_N].$$

Let $s = \sum_{k=1}^{N_n} \xi_k \mathbf{1}_{B_k}$. The prior Π_n on the dispersion coefficient s is defined by putting a prior on the coefficients ξ_k's. Since

$$s^2 = \sum_{k=1}^{N_n} \xi_k^2 \mathbf{1}_{B_k} = \sum_{k=1}^{N_n} \theta_k \mathbf{1}_{B_k}, \tag{2.2}$$

where we have put $\theta_k = \xi_k^2$, equivalently one can place the prior on the coefficients θ_k's of the diffusion coefficient s^2.

We call the prior Π_n a histogram-type prior. Conceptually somewhat similar priors have already been employed in the nonparametric Bayesian density estimation context, as well as the Poisson intensity estimation context, see, for example, Scricciolo (2003, 2004, 2007), Arjas and Heikkinen (1997), Heikkinen and Arjas (1998), Castillo and Nickl (2014), Castillo and Rousseau (2015) and Giné and Nickl (2011), but our problem is rather different from density or Poisson intensity estimation and requires the use of many different ideas. In practice, for example, in financial applications, one can hardly hope to estimate the diffusion coefficient to a very fine degree of detail, and in that respect using a prior with piecewise constant realisations does not appear to be unnatural. We note that a practical frequentist approach to nonparametric volatility estimation in Sabel, Schmidt-Hieber and Munk (2015) likewise produces piecewise constant estimates. There is yet another practical advantage in using priors with piecewise constant realisations: financial time series often exhibit jumps, accurate detection of which is a delicate task. Due to their localised nature, our Bayesian estimates of the dispersion coefficient are likely to quickly recover from negative effects of a moderate number of undetected jumps.

If the prior coefficients $\theta_1, \ldots, \theta_N$ are independent and have an inverse gamma $\mathrm{IG}(\alpha, \beta)$ distribution with parameters $\alpha, \beta > 0$, which will henceforth be our assumption, then the posterior is conjugate, as stated in the next lemma.

Lemma 1. *Assume $\theta_1, \ldots, \theta_N$ are independent with the inverse gamma $\mathrm{IG}(\alpha, \beta)$ distribution. Then $\theta_1, \ldots, \theta_N$ are a posteriori independent and, for $k = 1, \ldots, N-1$,*

$$\theta_k \mid \mathcal{X}_n \sim \mathrm{IG}(\alpha + m/2, \beta + nZ_k/2).$$

with

$$Z_k = \sum_{i=(k-1)m+1}^{km} Y_{i,n}^2,$$

whereas

$$\theta_N \mid \mathcal{X}_n \sim \mathrm{IG}\big(\alpha + (m+r)/2, \beta + nZ_N/2\big),$$

with

$$Z_N = \sum_{i=(N-1)m+1}^{n} Y_{i,n}^2.$$

Proof. Write $\theta = (\theta_1, \ldots, \theta_N)$. The likelihood, considered as a function of θ, is given by

$$L_n(\theta) = \prod_{i=1}^{n} \phi\left(0; Y_{i,n}, \int_{t_{i-1}}^{t_i} s^2(u)\,\mathrm{d}u\right)$$

$$\propto \theta_N^{-(m+r)/2} \exp\left(-\frac{nZ_N}{2\theta_N}\right) \prod_{k=1}^{N-1} \theta_k^{-m/2} \exp\left(-\frac{nZ_k}{2\theta_k}\right).$$

Here $\phi(x; \mu, \sigma^2)$ denotes the density of the normal distribution with mean μ and variance σ^2 evaluated at point x. As the prior satisfies $p(\theta_1, \ldots, \theta_N) \propto \prod_{k=1}^{N} \theta_k^{-(\alpha+1)} e^{-\beta/\theta_k}$, the result easily follows. $\qquad\square$

Posterior computations thus turn out to be elementary with our approach. For instance, the posterior mean of s^2 can be obtained from the posterior means of θ_k's. Recall that the $\mathrm{IG}(\alpha, \beta)$

distribution has mean $\beta/(\alpha - 1)$ and variance $\beta^2/(\alpha - 1)^2(\alpha - 2)$ (if finite, to which end one must have $\alpha > 2$). Then, e.g., for $k < N$ and $m \geq 2$, the posterior mean of θ_k is equal to

$$\mathbb{E}_{\Pi_n}(\theta_k \mid \mathcal{X}_n) = \frac{\beta + nZ_k/2}{\alpha + m/2 - 1}. \tag{2.3}$$

Conceptually the posterior mean of s^2 in this context is similar to a regressogram; see, for example, Examples 4.5 and 5.24 in Wasserman (2006). However, the Bayesian approach deals with the *entire posterior distribution* and does not reduce to a point estimate, such as the posterior mean.

Remark 1. Instead of a histogram-type representation in (2.2), one could have tried to base the prior on some other series representation for s^2. At first sight, for example, splines are a sensible choice to that end. However, enforcing spline basis functions to have disjoint supports with endpoints at $t_{n,i}$'s does not appear to be a natural procedure, whereas other choices would not have lead to simple posterior computations as in Lemma 1.

3 Asymptotic theory

3.1 Generalities

A highly desirable property of a Bayesian procedure, in particular from the frequentist point of view, is that the posterior asymptotically concentrates around the true parameter value. In fact, studying the rate at which the posterior contracts around the true parameter is similar to studying convergence rates of frequentist estimators. Some of by now classical references, where general conditions for derivation of posterior contraction rates are given, include Ghosal, Ghosh and van der Vaart (2000), Ghosal and van der Vaart (2007) and Shen and Wasserman (2001). However, we will follow a rather different and more direct path of the proof.

For $\varepsilon > 0$, we denote by

$$U_{s_0, \varepsilon} = \{s \in \mathcal{S}_n \colon \|s - s_0\|_2 < \varepsilon\}$$

the L_2-neighbourhood of s_0 of radius ε.

In the next proposition, we show that without loss of generality one may assume $b_0 = 0$ in the proofs. This proposition also explains why ignoring the drift in our estimation procedure by intentionally setting it to zero still leads to consistent Bayesian estimation of s_0. The corresponding theoretical argument relies on an application of Girsanov's theorem (see, e.g., Section 3.5 in Karatzas and Shreve (1988)). We would like to stress the fact that given the simplicity of our prior and intentional misspecification of the likelihood, the possibility of consistent estimation of s_0 with our approach is not obvious and requires a thorough investigation.

Proposition 1. *Let Assumption 1 hold and assume that for $\varepsilon_n \to 0$*

$$\mathbb{E}_{0,s_0}\big[\Pi_n\big(U_{s_0,\varepsilon_n}^c \mid \mathcal{X}_n\big)\big] \to 0$$

as $n \to \infty$. Then also

$$\mathbb{E}_{b,s_0}\big[\Pi_n\big(U_{s_0,\varepsilon_n}^c \mid \mathcal{X}_n\big)\big] \to 0.$$

3.2 Posterior contraction rates

As we consider the asymptotics $n \to \infty$, we take the number of bins N to depend on the sample size n, and indicate this in our notation by writing N_n. Then also m depends on n, and we write m_n to emphasise this dependence.

Assume that Assumption 1 holds for the remainder of this subsection. The following theorem shows that posterior contracts at rate $n^{-\lambda/(2\lambda+1)}$ in L_2.

Theorem 1. *Assume $N_n \asymp n^{1/(2\lambda+1)}$. If we let $\varepsilon_n \asymp n^{-\lambda/(2\lambda+1)}$, then for any sequence h_n tending to infinity (as $n \to \infty$) we have*

$$\mathbb{E}_{b_0,s_0}\big[\Pi_n\big(\|s^2 - s_0^2\|_2 \geq h_n \varepsilon_n \mid \mathcal{X}_n\big)\big] \to 0$$

as $n \to \infty$.

In the next theorem, we give the posterior contraction rate for the sup-norm.

Theorem 2. *Assume $N_n \asymp n^{1/(2\lambda+1)}$. If we let $\tilde{\varepsilon}_n \asymp n^{-\lambda/(2\lambda+2)}$, then for any sequence h_n tending to infinity (as $n \to \infty$) we have*

$$\mathbb{E}_{b_0,s_0}\Big[\Pi_n\Big(\sup_{x\in[0,1]} |s^2(x) - s_0^2(x)| \geq h_n \tilde{\varepsilon}_n \mid \mathcal{X}_n\Big)\Big] \to 0$$

as $n \to \infty$.

3.3 Discussion

Now we provide some discussion on the obtained theoretical results.

Remark 2. Establishing posterior contraction in the L_2-metric is rather natural, as $\int_0^1 s_0^2(t)\, \mathrm{d}t$ is the quadratic variation of the process $(X_t : t \in [0,1])$ over the interval $[0,1]$. It is also the variance of X_1 when the drift coefficient b_0 is a deterministic function depending only on time.

Remark 3. The inequality

$$\|s^2 - s_0^2\|_2 \geq \kappa \|s - s_0\|_2,$$

valid for s_0 satisfying Assumption 1 (here $\kappa > 0$ is a lower bound of s_0), implies that the rate of Theorem 1 is also valid with $\|s - s_0\|_2$. A similar remark applies to the posterior contraction in Theorem 2.

Remark 4. A comparison with the frequentist minimax convergence rate in Hoffmann (1997) shows that the posterior for the diffusion coefficient contracts at the optimal rate in the L_2-metric (strictly speaking, the results in the latter paper are given for $B_{p,q}^s$-Besov smooth diffusion coefficients with $s > 1$, but general arguments for derivation of lower bounds in our statistical setup are classical and work also in the Hölder setting). With histogram-type priors considered in this work no further improvement in the posterior contraction rate, when $\lambda = 1$, is possible beyond $n^{-1/3}$, even if the function s_0 is smoother than a Lipschitz function. An intuitive reason for this is that realisations of our histogram-type priors are too rough for this; cf. p. 629 in Scricciolo (2007).

Remark 5. There exists an excellent and deep reference on Bayesian inference in misspecified infinite-dimensional models, namely Kleijn and van der Vaart (2006). That paper provides some additional intuition why our approach is still successful despite the model misspecification. Summarised somewhat simplistically, the results from Kleijn and van der Vaart

(2006) say that the posterior in misspecified statistical models asymptotically concentrates around that value from the parameter space that is closest to the 'true' parameter value in the sense of the minimal Kullback-Leibler distance between respective probability distributions. Asymptotically the observation scheme as in Assumption 1(d) is almost as good as observing the process X continuously over the interval $[0, 1]$. On the other hand, the laws corresponding to paths $(X_t : t \in [0, 1])$ with two different diffusion coefficients are mutually singular, see Theorem 3.24 in §III.3d, Jacod and Shiryaev (2003), with a consequence that the corresponding Kullback-Leibler divergence is infinite. Hence, irrespective of the prior assumptions on the drift coefficient, the posterior for the dispersion coefficient (equivalently, diffusion coefficient) should concentrate around the 'true' dispersion coefficient s_0 (equivalently, the true diffusion coefficient s_0^2), for it is precisely this parameter value that yields finite Kullback-Leibler divergence between the laws of the 'true' and various misspecified models.

Remark 6. The result in Theorem 2 has to be compared to similar results on estimation of the volatility (not necessarily a deterministic function of time, as in our paper) in the L_∞-norm. Such results have been obtained in Aït-Sahalia and Jacod (2014), Kanaya and Kristensen (2016), Kristensen (2010) and Malliavin and Mancino (2009). As observed in Aït-Sahalia and Jacod (2014), p. 273, 'near to nothing is known on this topic'. Malliavin and Mancino (2009) establish consistency of their Fourier-based estimator of volatility, without deriving its convergence rate. The asymptotics considered in Kanaya and Kristensen (2016) are different from the one in the present work, which makes a direct comparison difficult. Finally, Theorem 3.5 in Kristensen (2010) gives a convergence rate of a kernel estimator of volatility; when translated to our setting, the corresponding optimal convergence rate, up to a log factor, is $n^{-\lambda/(2\lambda+1)}$. This is a faster rate than the rate obtained in our Theorem 2. However, the suboptimal rate in Theorem 2 is likely to be an artefact of our proof, specifically a somewhat rough bound employed in the second inequality of (8.5). Unlike our method, the kernel estimator in Kristensen (2010) suffers from the boundary bias problem, which necessitates studying its asymptotic properties on a time interval strictly contained in $[0, T]$.

Remark 7. In Soulier (1998), the following frequentist estimator of s_0^2 is introduced,

$$\hat{s}^2(t) = \sum_{i=1}^{n} K_h(t, t_i) Y_{i,n}^2,$$

where K is a kernel function, a constant $h > 0$ is a bandwidth, and $K_h(s, t) = \frac{1}{h} K(\frac{t-s}{h})$ is a rescaled kernel. Suppose now K is a boxcar kernel, $K(u) = (1/2)1_{|u| \leq 1}$, see p. 55 in Wasserman (2006). Then

$$\hat{s}^2(t) = \frac{1}{2h} \sum_{i \,:\, |t - t_{i,n}| \leq h} Y_{i,n}^2.$$

A brief reflection shows that for $k < N$, n large, and $h = m/(2n)$ (half the bin length), $\hat{s}^2$ is quite similar to (2.3), the difference being that in that formula averaging occurs over individual bins, while here one averages locally over observations in a neighbourhood of each time point t. We note, however, that the asymptotic theory in Soulier (1998) does not cover the asymptotics of the posterior mean as in (2.3), and also that the regularity conditions of that paper are different from ours. On the other hand, practical computation of the kernel estimator $\hat{s}^2$ would typically require from a user some form of data binning, cf. Appendix D.2 in Wand and Jones (1995), so that from this point of view the posterior mean and the estimator $\hat{s}^2$ are closely related.

Remark 8. We have already pointed out in the introduction that our setup covers more general SDE models than those with deterministic dispersion coefficients, see equation (1.2); this follows from an application of Itô's formula. Under regularity conditions, a further generalisation of our results is possible to the case when the dispersion coefficient s_0 is in fact a stochastic process independent of the driving Wiener process in (1.1) (we cite an interesting example in this context: a widely known stochastic volatility model arising as a diffusion limit of a GARCH(1, 1) process, see Nelson (1990)). Namely, in this case one can simply follow the Bayesian methodology we described in Section 2 with no further changes, acting as if the dispersion coefficient were a deterministic function. Despite (yet another) purposeful misspecification, the resulting Bayesian procedure is consistent. This can be established by combining arguments in the present paper with the ones similar to those in Kristensen (2010), that deal with a kernel volatility estimator. Indeed, careful examination shows that one of the main steps in our proofs is what can be termed a Bayesian bias-variance decomposition, see the proof of Theorem 1. Both the 'bias' and 'variance' terms there are analysed using techniques similar to those employed in for example, kernel regression or density estimation problems, whence a possibility for further generalisations. Space considerations preclude us from studying this interesting question in detail in the present work.

4 Bin number selection via DIC

In this section, we describe a method of choosing N that is based on the Deviance Information Criterion (DIC) of Spiegelhalter et al. (2002); see also Spiegelhalter et al. (2014) and Gelman, Hwang and Vehtari (2014). Further discussion is given in Section 6.

By M we denote the posterior mean of s^2; $\log L_n(M)$ is our notation for the log-likelihood evaluated at the posterior mean. Introduce the DIC measure of predictive accuracy,

$$\widehat{\text{elpd}}_{\text{DIC}} = \log L_n(M) - \nu_{\text{DIC}},$$

where "elpd" is an abbreviation for "expected log predictive density" and

$$\nu_{\text{DIC}} = 2\{\log L_n(M) - \mathbb{E}_{\Pi_n}(\log L_n(s)|\mathcal{X}_n)\}$$

is the effective number of parameters. Straightforward but tedious calculations employing Lemma 1 and properties of the (inverse) gamma distribution give that

$$\log L_n(M) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{T}{n}\right)$$
$$- \frac{1}{2}\sum_{k=1}^{N} m_k \log\left(\frac{\beta + nZ_k/2}{\alpha + m_k/2 - 1}\right) - \frac{n}{2T}\sum_{k=1}^{N} Z_k \frac{\alpha + m_k/2 - 1}{\beta + nZ_k/2},$$

where m_k denotes the number of observations in the bin B_k (with a harmless abuse of notation) and Z_k is as in Lemma 1. By similar calculations,

$$\nu_{\text{DIC}} = \frac{n}{T}\sum_{k=1}^{N} \frac{Z_k}{\beta + nZ_k/2} - \sum_{k=1}^{N} m_k\left\{\Psi\left(\alpha + \frac{m_k}{2}\right) - \log\left(\alpha + \frac{m_k}{2} - 1\right)\right\},$$

where Ψ is the digamma function. The formulae simplify even further, when $m_k = m$, $k = 1, \ldots, N$.

Now the idea consists in evaluating $\widehat{\text{elpd}}_{\text{DIC}}$ for a range of values of N, and choosing the one that maximises $\widehat{\text{elpd}}_{\text{DIC}}$. This aims at optimising the predictive behaviour of the model. We note that using predictive performance criteria for Bayesian model selection is a well-established practice in the case of finite-dimensional, parametric models (see, e.g., Gelman,

Hwang and Vehtari (2014)), but seems to be a new idea in the non-parametric Bayesian setting. DIC in some sense constitutes a Bayesian analogue of Akaike's AIC, and conceptually our proposal is similar to employing information criteria for smoothing parameter selection in the frequentist literature; see, for example, the widely cited work Hurvich, Simonoff and Tsai (1998). On the computational side, our DIC-based method is very simple to implement and does not require heavy computations. We test its practical performance in Section 6.

5 Bin number selection via marginal likelihood

In this section, we describe an alternative method of the bin number selection to the one we discussed in Section 4. This is based on maximising the marginal likelihood $\int_{\mathcal{S}} L_n(s) \Pi_n(\mathrm{d}s)$ as a function of N, which can be viewed as model evidence given the data. Model selection based on the marginal likelihood (or Bayes factors) is well-established in Bayesian statistics. See, for example, Wang (2012) for an application to smoothing parameter selection in the context of smoothing spline regression, which is conceptually related to choosing the bin number N in our problem. On a more general level, this is nothing else but an instance of a well-known empirical Bayes method (cf. Gelman et al. (2013), Section 5.1).

Since we identify a piecewise constant diffusion coefficient s^2 with its coefficients $\theta_1, \ldots, \theta_N$, using a priori independence of θ_k's and Fubini's theorem, the marginal likelihood in our setting can be evaluated as (here m_k was defined in Section 4 and is the number of observations in bin B_k)

$$
\begin{aligned}
\mathrm{ML}_N(\mathcal{X}_n) &= \prod_{k=1}^{N} \left\{ \int_{[0,\infty)} \left(\frac{2\pi}{n} \right)^{-m_k/2} \frac{\beta^\alpha}{\Gamma(\alpha)} \theta_k^{-(\alpha+m_k/2)-1} \right. \\
&\qquad \left. \times \exp\left(-\frac{1}{\theta_k}\left(\beta + \frac{nZ_k}{2} \right) \right) \mathrm{d}\theta_k \right\} \\
&\propto \frac{\beta^{\alpha N}}{\Gamma(\alpha)^N} \prod_{k=1}^{N} \frac{\Gamma(\alpha + m_k/2)}{(\beta + nZ_k/2)^{\alpha+m_k/2}}.
\end{aligned}
$$

From a numerical point of view, rather than using analytic tools for optimisation, we recommend plotting the values of ML_N versus its argument N, and performing graphical maximisation. This results in a computationally simple model selection procedure and we apply it in practice in Section 6.

6 Simulated data examples

In this section, we use simulations of diffusion processes with known drift and diffusion coefficient to gain insight into the numerical performance of our method. We are particularly interested in both the practical consequence of using a pseudo-likelihood ignoring the drift and the empirical rate of posterior contraction attainable in examples.

In the first subsection, we simulate realisations from the model for different dispersion and drift coefficients. Given subsamples of those realisations sampled at different rates we compute the posterior distribution of the dispersion function using the piecewise constant prior (histogram-type prior) with varying number of bins. As an illustration, plots of marginal posterior bands are compared with the true dispersion function. The marginal posterior bands are obtained by computing $1 - \alpha$ central posterior intervals (see Gelman et al. (2013), p. 33) separately for the coefficients θ_k's using Lemma 1.

By Proposition 1, assuming that there is no drift still leads to consistent Bayesian estimation of the dispersion coefficient, even if the data are from a diffusion process with nonzero drift. This is illustrated by our simulation results.

In the second subsection, we use Monte Carlo methods to determine the distribution of the distance between samples from the posterior distributions and the true dispersion coefficient. In Theorem 1, we showed that the posterior contraction rate in the L_2-norm is optimal for estimation of Hölder smooth dispersion coefficients of order $0 < \lambda \leq 1$ for the prior based on the inverse gamma distribution. The results of the Monte Carlo simulation agree with this. Furthermore, we numerically determine the rate of posterior convergence in the supremum norm for two examples. The simulation results in this case are less conclusive, but suggest that the posterior contraction rate in the L_∞-norm in Theorem 2 is possibly suboptimal.

Our analyses were done employing the programming language Julia, see Bezanson et al. (2017).

6.1 Influence of the drift

For this numerical experiment we simulated sample paths of the diffusion $(X_t : t \in [0, 1])$ where the true dispersion coefficient is given by one of

$$s_1(t) = 3/2 + \sin\big(2(4t - 2)\big) + 2\exp\big(-16(4t - 2)^2\big),$$

$$s_2(t) = W_t(\omega_0) + 1,$$

and the drift is given *by* one of

$$b_0(x) = 0,$$

$$b_1(x) = -10x + 20.$$

The function s_1 is a benchmark function used in Fan and Gijbels (1996) in the context of nonparametric regression, up to a vertical shift to ensure positivity. To define $s_2(t)$, we took a fixed realisation of a Wiener path starting in 1, with $W_t(\omega_0) > -1$ for $t \in [0, 1]$ (sampled on an equidistant grid with 800,001 points in $[0, 1]$). The function s_1 is Lipschitz continuous, while s_2 is Hölder continuous with coefficient essentially $\frac{1}{2}$.

Specifically, we used the Euler scheme on a grid with 800,001 equidistant points in the interval $[0, 1]$ to obtain a single diffusion path for each combination of drift and dispersion coefficients given above, which then was subsampled to obtain $n = 4000 \cdot 2^j + 1$, $j \in \{1, 2, 3\}$ observations each. As the prior on the coefficients on the individual bins we used independent $IG(0.1, 0.1)$ distributions.

Figures 1 and 3 show marginal 98% posterior bands for different combinations of bin number and observation regime for both dispersion coefficients when the drift is zero. Figure 2 shows the marginal posterior bands for s_1 which are obtained if an affine drift term $b_1(x) = -10x + 20$ is present, but neglected in the estimation procedure. Comparison with Figure 1 shows that presence of a strong nonzero drift hardly affects the obtained credible bands. Note that credible bands successfully recover the overall shape of the functions s_i, although the recovery is not too refined; however, it would be misleading to visually compare the results obtained in the SDE setting to for example, those obtainable in nonparametric regression, as the latter constitutes a much easier inferential problem. The functions s_i do not always pass through all the credible intervals, which is not surprising given the fact that these intervals are marginal. In general, construction of uniform confidence bands in nonparametric statistics is a long-studied and difficult problem, see Section 5.7 in Wasserman (2006); for a general perspective on nonparametric uniform confidence bands see Faraway (2016). Less is known about frequentist performance of nonparametric Bayesian confidence sets, although some interesting results have already been obtained in recent years, see, for example, Nickl
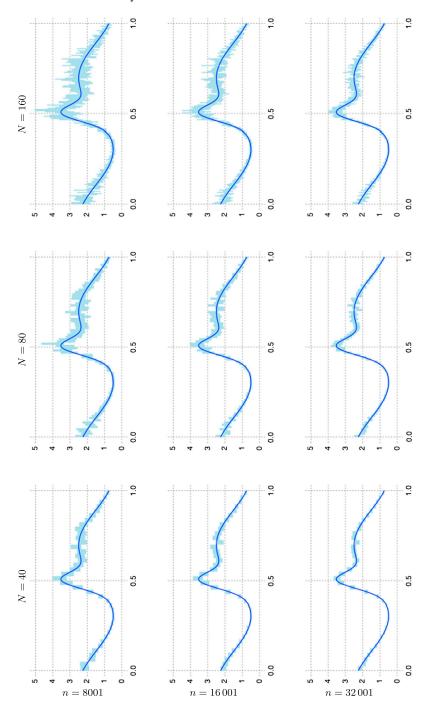
Figure 1 *Estimation results for s_1 with varying number of observations and bins, no drift. The solid curve is the true dispersion coefficient while the light blue areas are 98% marginal posterior bands. n is the number of observations, N is the number of bins.*

and Szabó (2016), Szabó, van der Vaart and van Zanten (2015b) and Szabó, van der Vaart and van Zanten (2015a). We do not address this issue in detail in this paper, but note that posterior contraction at an optimal rate does not automatically imply 'good' frequentist coverage properties of Bayesian credible sets. Following pp. 130–131 in Wasserman (2006) in a similar study in the case of histograms employed as nonparametric probability density estimators, in our case it is arguably more natural to consider performance of Bayesian credible bands at
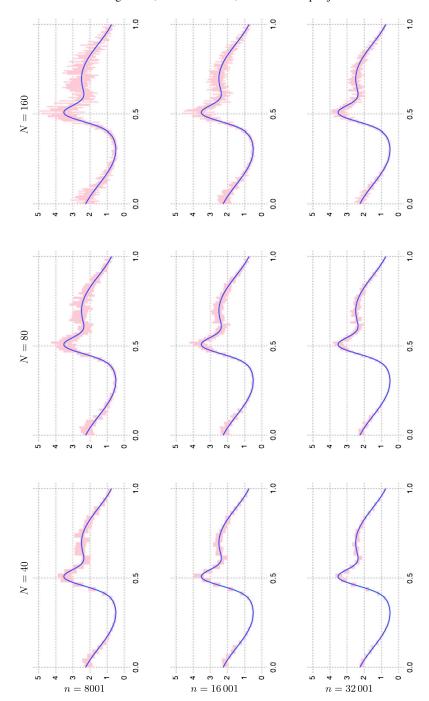
Figure 2 *Estimation for s_1 with varying number of observations and bins and drift $b_1(x) = -10x + 20$. The solid curve is the true dispersion coefficient while the light red areas are 98% marginal posterior bands. n is the number of observations, N is the number of bins.*

the resolution of the histogram-type priors, that is, for a histogramised version of a dispersion coefficient s, obtained as

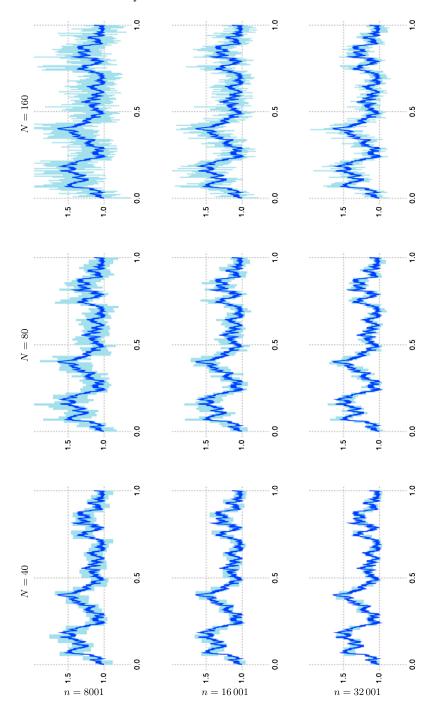$$\bar{s} = \sum_{k=1}^{N_n} \bar{s}_k \mathbf{1}_{B_k}$$

Figure 3 *Estimation results for s_2 with varying number of observations and bins, no drift. The solid curve is the true dispersion coefficient while the light blue areas are 98% marginal posterior bands. n is the number of observations, N is the number of bins.*

for

$$\bar{s}_k = \frac{1}{\Delta_k} \int_{B_k} s(t)\,\mathrm{d}t, \quad k = 1, \ldots, N_n,$$

where Δ_k denotes the length of the bin B_k. In a frequentist approach to nonparametric inference this constitutes an alternative to attempting to get rid of the bias of a nonparametric esti-
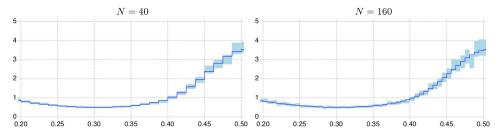
Figure 4 *Detail of Figure* 1, *panels for* $N = 40$ *and* $N = 160$, $n = 16,001$, *zoomed-in to show* $t \in [0.2, 0.5]$, *posterior credible bands only, with a histogramised* s_1 *superimposed.*

mator for confidence band construction purposes via an artificial device like undersmoothing. Figure 4 gives a detail of Figure 1 (panels for $N = 40$ and $N = 160$, $n = 16,001$, zoomed-in to show $t \in [0.2, 0.5]$, posterior credible bands only) with a histogramised s_1 superimposed. In comparison to Figure 1, the results appear to be visually even more pleasing, with the Bayesian credible band covering the curve s_1 in its entirety in the left panel. This suggests to give the number of bins N_n an additional interpretation of a resolution at which one is interested in learning properties of the function s. Obviously, this resolution cannot be made arbitrarily fine, as this would distort the frequentist consistency property of our nonparametric Bayesian procedure as expressed in our theoretical results from Section 3. We close this brief discussion on confidence bands by mentioning the fact that a number of authors have argued in favour of the so-called average coverage as a more natural concept of coverage of confidence bands than the uniform confidence bands, see Section 5.8 in Wasserman (2006) and references therein.

Note that the recovery is somewhat less accurate for function s_2, see Figure 3, than for function s_1, see Figure 1. This is in perfect agreement with our theoretical results from Section 3, that give a slower posterior contraction rate for less smooth functions.

Our posterior contraction theorems only specify that the optimal number of bins N_n is proportional to $n^{-\beta}$, where the exponent β depends on the smoothness λ of a function to be estimated. This does not give a directly applicable recipe on how to choose the proportionality constant. In practice, we recommend to use our theoretical results as guidance and to try out several choices of the number of bins, cf. Figures 1, 2 and 3. This is not unlike the scale-space smoothing approach in the frequentist literature, see, for example, Section 5.11 in Wasserman (2006). Furthermore, a useful point of reference in our setup is the number of non-overlapping neighbouring marginal posterior intervals. Figure 4 shows that if N is too small to capture adequately the curvature of a dispersion coefficient, neighbouring marginal posterior credible intervals tend to be disjoint. On the other hand, choosing too many bins leads to undersmoothing and erratic appearance of marginal credible intervals.

Numerical experiments similar to the above were also performed for other benchmark functions given in Fan and Gijbels (1996). As the results were similar, they are not reported here.

We also tested the performance of the procedure for choosing the number of bins as discussed in Section 4. We considered the case with dispersion coefficient s_1, drift b_1 and 8001 observations. This corresponds to the leftmost column of Figure 2. We computed $\widehat{\mathrm{elpd}}_{\mathrm{DIC}}$ for $N \in \{5, 10, 20, 40, 80, 160, 320\}$. The results are in Figure 5, from which it is seen that the criterion is maximised for $N = 40$. This corresponds to the bottom left panel of Figure 2. In Figure 6, we plot the results obtained with the procedure for choosing the number of bins as discussed in Section 5. Also this procedure suggests $N = 40$ as an optimal number of bins.
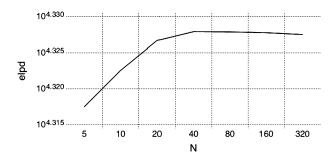
Figure 5 *A log plot of the* $\widehat{\text{elpd}}_{DIC}$*-values for estimating the dispersion coefficient* s_1 *versus different bin numbers* N*, n fixed at 8001, with drift* b_1.
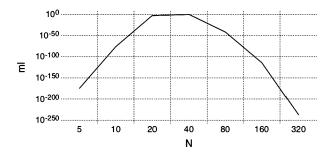

Figure 6 *A log plot of suitably scaled and shifted* ML*-values for estimating the dispersion coefficient* s_1 *versus different bin numbers* N*, n fixed at 8001, with drift* b_1.

6.2 Empirical contraction rates

Of particular interest is the empirical size of the L_2- and L_∞-balls containing most of the posterior mass. To assess this, we approximate the distribution of the L_2- or L_∞-distance between posterior samples and the truth by sampling from the posterior. We do this for four different realisations of the model denoted by $X(\omega_1), \ldots, X(\omega_4)$. Note that q being the 90%-quantile of this distribution entails that the L_2-ball respective L_∞-ball of size q contains 90% of the posterior mass. To be specific, we employ the following steps four times for each s_i, $i = 1, 2$ and both norms.

1. Simulate the diffusion $(X_t(\omega))_{t \in [0,1]}$ with dispersion coefficient s_i (without drift) on a grid with 800,001 points in the interval $[0, 1]$.
2. Subsample to obtain $n = 2500 \cdot 2^j + 1$, $j \in \{1, 2, 3, 4, 5\}$ observations each.
3. Draw $k = 1, \ldots, 2000$ samples $S_k^{i,n}$ from the posterior using $\log(N_n) = \log 5 + \frac{1}{2\lambda_i+1} \times \log(n)$ bins and determine the distance $\|S^{i,n} - s_i\|_2$ (respective $\log(N_n) = \log 5 + \frac{1}{2\lambda_i+1} \log(n/\log(n))$ bins for $\|S^{i,n} - s_i\|_\infty$).
4. Determine the 90% quantile $q^{i,n}(\omega)$ of the distance samples and plot as function of n.

Figure 7 shows the 90% quantile $q^i(n)$ on a log-log scale for the L_2-norm for $i = 1, 2$. The empirical findings for function s_1 agree very well with the exponent $\frac{1}{3} = \frac{\lambda}{2\lambda+1}$ for $\lambda = 1$ obtained in Theorem 1. The function s_2 is λ-Hölder smooth for any $\lambda < \frac{1}{2}$. The empirically determined exponent is in excellent agreement with the exponent $\frac{1}{4} = \frac{\lambda}{2\lambda+1}$ for $\lambda = \frac{1}{2}$ obtained in Theorem 1.

Figure 8 shows the 90% quantile $q^i(n)$ for the L_∞-norm. The number of bins N_n was chosen in analogy to nonparametric kernel regression, see Theorem 1.8 in Tsybakov (2009), as $\log(N_n) = \log 5 + \frac{1}{2\lambda_i+1} \log(n/\log(n))$. Here the results are less conclusive than in the
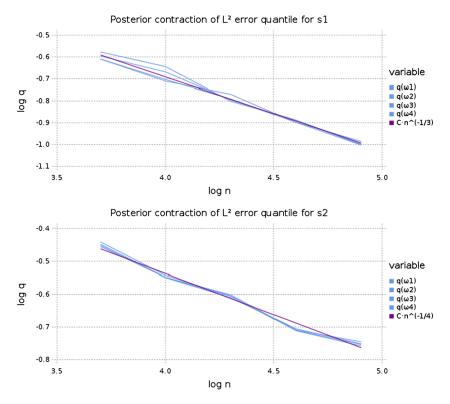
Figure 7 *Estimate of the contraction rate of the size of L_2-balls around s covering 90% of posterior probability mass by sampling (blue); compared with rate (purple). Top: Lipschitz continuous case, example s_1. Bottom: Hölder coefficient essentially $\frac{1}{2}$, example s_2.*

case of the L_2-norm. The empirical findings suggest the rate $(n/\log n)^{\frac{1}{3}}$ or similar for s_1, and the rate $(n/\log n)^{\frac{1}{4}}$ or similar for s_2.

7 Real data examples

In this section, we apply our Bayesian method on two real data examples and study its implications. The daily exchange rates (noon buying rates in New York City for cable transfers payable in foreign currencies) JPY/USD and USD/GBP from January 1, 1999, to March 20, 2010 are available as data sets DEXJPUS and DEXUSUK from Board of Governors of the Federal Reserve System (2016). We visualise the data in Figure 9 (time in this and subsequent figures corresponds to the physical time, with unit being a year).

These exchange rate time series were considered for example, in Hamrick et al. (2011). Based on discrete time observations assumed to have arisen from the solution of an SDE

$$dX_t = b(X_t)\,dt + \sigma(X_t)\,dW_t, \quad X_0 = x, t \in [0, T], \tag{7.1}$$

with space-dependent dispersion coefficient σ, Hamrick et al. (2011) proposed a maximum penalised quasi-likelihood method to estimate nonparametrically the diffusion coefficient σ^2.

Plots of both series appear to indicate that the data are nonstationary, and this is confirmed also by the outcomes of the augmented Dickey–Fuller test we performed using urca package in **R**, see R Core Team (2017) and Pfaff (2008). The test constitutes a standard unit root test in time series analysis, see, for example, Chapter 17 in Hamilton (1994) and Section 5.3 in Aragon (2011) for additional information. A similar conclusion on nonstationarity is reached in Hamrick and Taqqu (2009). As stationarity is no prerequisite for application of
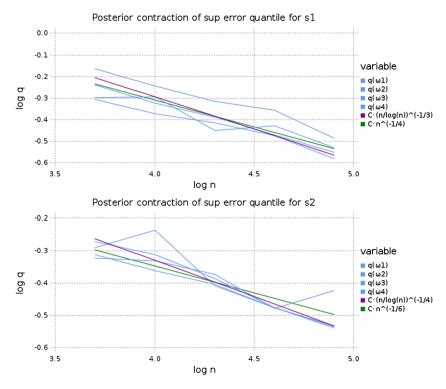
Figure 8 *Estimate of the contraction rate of the size of L_∞-balls around s covering 90% of posterior probability mass by sampling (blue); compared with likely rate (purple) and rate of Theorem 2 (green). Top: Lipschitz continuous case, example s_1. Bottom: Hölder coefficient essentially $\frac{1}{2}$, example s_2.*

our nonparametric Bayesian method, we do not pursue this question any further in the paper. We retrieved the estimates $\hat{\sigma}$ given in Hamrick et al. (2011) from the figures published in the digital version of the publication using WebPlotDigitizer, see Rohatgi (2015), and next used them to calculate the induced estimates $t \mapsto \hat{s}(t) = \hat{\sigma}(X_t)$ of the historical volatility at time t. In Figures 10 and 11, we contrast those induced estimates $\hat{s}$ with 90% marginal posterior bands for the deterministic dispersion coefficient s_0 that were obtained through our Bayesian procedure. Since nominal exchange rates vary widely with the denomination used, we employed a non-informative IG(0.001, 0.001) prior for coefficients θ_k's from Lemma 1. Our estimation results in Figures 10 and 11 show that the volatility was high in the final years of the decade 2000–2010, coinciding with the sub-prime mortgage crisis and the following recession. On the other hand, the model from Hamrick et al. (2011) does not appear to capture this fact. It is reassuring to see that our method recovers this relevant event from the data. Based on this observation, we believe that including time-dependence of the volatility into the model is more appropriate in this example.

A further confirmation of our findings comes from the change-point analysis of the data. For both the DEXJPUS and DEXUSUK series, a simple estimator for detection of a change-point in diffusivity of an SDE, see De Gregorio and Iacus (2008) and Section 4.3.1 in Iacus (2008), that is implemented in **R** in the sde package (see Iacus (2016)), yields a change-point in the volatility level before and after 2007, with volatility prior to 2007 being lower. This is in excellent agreement with findings using our Bayesian method. See Figure 9 below, where we plot the exchange rate data together with change point estimates, and compare to Figures 10 and 11. We note, however, that our Bayesian approach yields more, in that the method from De Gregorio and Iacus (2008) assumes that the volatility is constant before and after the change-point, which is not what our Bayesian marginal credible sets suggest.
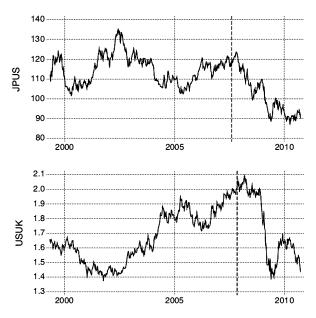
Figure 9 *Daily exchange rates JPY/USD (*top*) and USD/GBP (*bottom*) from January* 1, 1999 *to March* 20, 2010. *Change-point estimates are indicated with dashed vertical lines.*
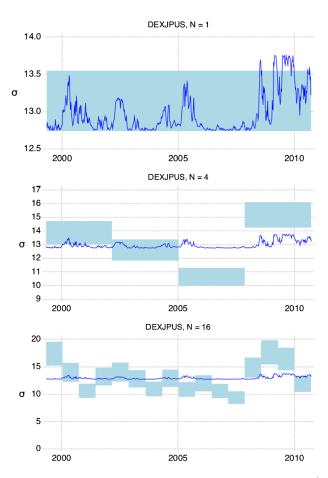

Figure 10 90% *marginal posterior band for the volatility of DEXJPUS. Plot of* $t \mapsto \hat{\sigma}(X_t)$ *as induced by the estimates in* Hamrick et al. (2011).
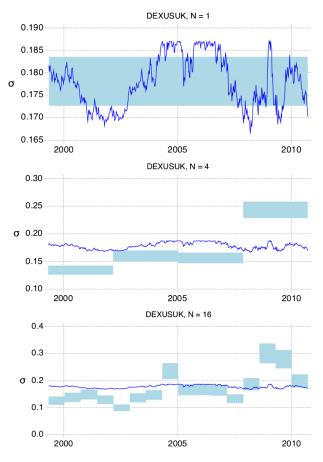
Figure 11 90% *marginal posterior band for volatility of DEXUSUK. Plot of* $t \mapsto \hat{\sigma}(X_t)$.

8 Proofs

Proof of Proposition 1. By Theorem 6.10 in Höpfner (2014) and our Assumption 1(b)–(c), the laws $\mathbb{P}_{b_0,s_0}$ and $\mathbb{P}_{0,s_0}$ of the path $(X_t : t \in [0, 1])$ are equivalent, the result that ultimately relies upon Girsanov's theorem. Let $Z = \frac{\mathrm{d}\mathbb{P}_{b_0,s_0}}{\mathrm{d}\mathbb{P}_{0,s_0}}$, and let $Z^n = \frac{\mathrm{d}\mathbb{P}^n_{b_0,s_0}}{\mathrm{d}\mathbb{P}^n_{0,s_0}}$ be the density of the respective laws of $\mathcal{X}_n$. By Theorem 2 on p. 245 in Skorohod (1964) and Corollary 2 on p. 246 there, $Z^n = \mathbb{E}_{0,s_0}(Z \mid \mathcal{X}_n)$. Then convergence of $\Pi_n(U^c_{s_0,\varepsilon_n} \mid \mathcal{X}_n)$ in $\mathbb{P}_{0,s_0}$-probability to zero implies that it also converges to zero in $\mathbb{P}_{b_0,s_0}$-probability. Indeed, fix $\eta > 0$, let $A^n = \{\Pi_n(U^c_{s_0,\varepsilon_n} \mid \mathcal{X}_n) > \eta\}$ and let $\varepsilon > 0$. Then

$$\mathbb{P}_{b_0,s_0}(A_n) = \mathbb{E}_{0,s_0}\big[Z^n \mathbf{1}_{A^n}\big] = \mathbb{E}_{0,s_0}\big[\mathbb{E}_{0,s_0}(Z\mathbf{1}_{A^n} \mid \mathcal{X}_n)\big] = \mathbb{E}_{0,s_0}[Z\mathbf{1}_{A^n}].$$

Choose $\delta > 0$ such that $\mathbb{P}_{0,s_0}(A) < \delta$ for any event A implies $\mathbb{P}_{b_0,s_0}(A) = \mathbb{E}_{0,s_0}[Z\mathbf{1}_A] < \varepsilon$, possible in view of Lemma 13.1 in Williams (1991). As eventually $\mathbb{P}_{0,s_0}(A^n) < \delta$, we have $\mathbb{P}_{b_0,s_0}(A_n) < \varepsilon$. $\square$

Lemma 2. *Define*

$$M(x) := \mathbb{E}_{\Pi_n}\big(s^2(x) \mid \mathcal{X}_n\big),$$

the posterior mean of $s^2(x)$. *Assume* $m_n \asymp n^{2\lambda/(2\lambda+1)}$ *(equivalently,* $N_n \asymp n^{1/(2\lambda+1)}$*). If we let* $\varepsilon_n \asymp m_n^{-1/2}$, *i.e.* $\varepsilon_n \asymp n^{-\lambda/(2\lambda+1)}$, *then for any sequence* h_n *tending to infinity we have for*

any fixed $x \in [0, 1]$

$$\mathbb{P}_{0,s_0}\big(|M(x) - s_0^2(x)| \geq \varepsilon_n h_n\big) \to 0$$

as $n \to \infty$.

Proof. Assume $x \in B_k$ for some $k < N$ (note that then $k = \lfloor \frac{nx}{m} \rfloor + 1$). The case $x \in B_N$ follows later on. We compute

$$M(x) = \mathbb{E}_{\Pi_n}(\theta_k \mid \mathcal{X}_n) = \frac{\beta + \frac{n}{2}\sum_{i=(k-1)m+1}^{km} Y_{i,n}^2}{\alpha + m/2 - 1}$$

$$= \frac{2\beta}{2(\alpha - 1) + m} + \frac{n\sum_{i=(k-1)m+1}^{km} Y_{i,n}^2}{2(\alpha - 1) + m}.$$

Note that

$$\mathbb{E}_{0,s_0}\left[\sum_{i=(k-1)m+1}^{km} Y_{i,n}^2\right] = \int_{t_{m(k-1)}}^{t_{mk}} s_0^2(u)\,\mathrm{d}u.$$

Hence,

$$b(x) := \mathbb{E}_{0,s_0}[M(x)] - s_0^2(x) = \frac{2\beta}{2(\alpha - 1) + m} + \frac{n\int_{t_{m(k-1)}}^{t_{mk}} s_0^2(u)\,\mathrm{d}u}{2(\alpha - 1) + m} - s_0^2(x).$$

We consider

$$\frac{n\int_{t_{m(k-1)}}^{t_{mk}} s_0^2(u)\,\mathrm{d}u}{2(\alpha - 1) + m} - s_0^2(x) = \frac{n\int_{t_{m(k-1)}}^{t_{mk}} s_0^2(u)\,\mathrm{d}u}{2(\alpha - 1) + m} - \frac{n}{m}\int_{t_{m(k-1)}}^{t_{mk}} s_0^2(x)\,\mathrm{d}u$$

$$= \frac{n}{m + 2(\alpha - 1)}\int_{t_{m(k-1)}}^{t_{mk}} (s_0^2(u) - s_0^2(x))\,\mathrm{d}u - \frac{2(\alpha - 1)s_0^2(x)}{m + 2(\alpha - 1)}.$$

As s_0 is bounded from above by some constant $\mathcal{K} > 0$, the last term is of order $\frac{1}{m}$. We continue with the integral expression. By Hölder continuity of s_0, it follows that $|s_0^2(u) - s_0^2(v)| \leq 2\mathcal{K}L|u - v|^\lambda$. Using the sharp bound (attained at $x = t_{m(k-1)}$ and $x = t_{mk}$)

$$\int_{t_{m(k-1)}}^{t_{mk}} |u - x|^\lambda\,\mathrm{d}u \leq \frac{1}{\lambda + 1}\left(\frac{m}{n}\right)^{\lambda+1},$$

we get the order bound (uniformly in $x \in B_k$)

$$|b(x)| \leq O\left(\frac{1}{m}\right) + O\left(\frac{n}{m}\left(\frac{m}{n}\right)^{\lambda+1}\right) = O\left(\frac{1}{m}\right) + O\left(\frac{m}{n}\right)^\lambda.$$

For the variance, we obtain (using that $Y_{i,n}$ and $Y_{j,n}$ are independent for $i \neq j$)

$$\mathbb{V}\mathrm{ar}_{0,s_0}[M(x)] = \left(\frac{n}{2(\alpha - 1) + m}\right)^2 \sum_{i=(k-1)m+1}^{km} \mathbb{V}\mathrm{ar}_{0,s_0}[Y_{i,n}^2]$$

$$= \left(\frac{n}{2(\alpha - 1) + m}\right)^2 2\sum_{i=(k-1)m+1}^{km}\left(\int_{t_{i-1}}^{t_i} s_0^2(u)\,du\right)^2$$

$$\leq C_\alpha \mathcal{K}^4\frac{1}{m} = O\left(\frac{1}{m}\right),$$

for a positive constant C_α, uniformly for $x \in B_k$. Balancing bias and standard deviation yields the order equality

$$O\left(\frac{1}{m}\right) + O\left(\frac{m}{n}\right)^\lambda = O\left(\frac{1}{\sqrt{m}}\right),$$

or

$$O(1) + O\left(\frac{m^{\lambda+1}}{n^\lambda}\right) = O(\sqrt{m}),$$

which gives $m^{\lambda+1/2} \asymp n^\lambda$, so that

$$m \asymp n^{\frac{2\lambda}{2\lambda+1}}. \tag{8.1}$$

A similar analysis holds for $x \in B_N$. We highlight the main steps for this case. For instance, we now get

$$M(x) = \mathbb{E}_{\Pi_n}(\theta_N|\mathcal{X}_n) = \frac{2\beta}{2(\alpha-1)+m+r} + \frac{nZ_N}{2(\alpha-1)+m+r},$$

and

$$\mathbb{E}_{0,s_0}[Z_N] = \int_{t_{m(N-1)}}^1 s_0^2(u)\,\mathrm{d}u.$$

This results in the bias

$$b(x) = O\left(\frac{1}{m}\right) + O\left(\frac{m+r}{n}\right)^\lambda = O\left(\frac{1}{m}\right) + O\left(\frac{m}{n}\right)^\lambda,$$

as $0 \le r < m$. Similarly, we get

$$\mathbb{V}\mathrm{ar}_{0,s_0}[M(x)] = O\left(\frac{1}{m+r}\right),$$

which is, using again $0 \le r < m$, of order $O(\frac{1}{m})$. Hence, balancing gives the same order relation $m \asymp n^{\frac{2\lambda}{2\lambda+1}}$ as for the case $k < N$ in (8.1).

With the above choice for m we obtain, for any $x \in [0,1]$, for the mean squared error that

$$\mathbb{E}_{0,s_0}\left[(M(x) - s_0^2(x))^2\right] \asymp n^{-\frac{2\lambda}{2\lambda+1}}.$$

Hence, upon taking $\varepsilon_n \asymp n^{-\frac{\lambda}{2\lambda+1}}$, the result follows from Chebyshev's inequality:

$$\mathbb{P}_{0,s_0}\left(|M(x) - s_0^2(x))| \ge \varepsilon_n h_n\right) \le \frac{\mathbb{E}_{0,s_0}[(M(x) - s_0^2(x))^2]}{\varepsilon_n^2 h_n^2} \lesssim \frac{1}{h_n^2} \to 0$$

as $n \to \infty$. $\qquad\square$

Corollary 1. *Under the assumptions of Lemma 2, we have*

$$\mathbb{E}_{0,s_0}\left[\|M - s_0^2\|_2^2\right] \asymp \varepsilon_n^2,$$

where $\|\cdot\|_2$ is the L_2-norm on $[0,1]$.

Proof. Simply note that

$$\mathbb{E}_{0,s_0}\left[\|M - s_0\|_2^2\right] = \int_0^1 \mathbb{E}_{0,s_0}\left[|M(x) - s_0(x)|^2\right]\mathrm{d}x$$

$$= \sum_{k=1}^{N_n} \int_{B_k} \mathbb{E}_{0,s_0}\left[|M(x) - s_0(x)|^2\right]\mathrm{d}x,$$

and apply the bounds derived in the proof of Lemma 2 on each bin B_k separately, $1 \leq k \leq N_n$. $\square$

Proof of Theorem 1. First assume $b_0 \equiv 0$. By Chebyshev's inequality, one has

$$\mathbb{E}_{0,s_0}\big[\Pi_n\big(\|s^2 - s_0^2\|_2 \geq h_n\varepsilon_n \mid \mathcal{X}_n\big)\big]$$

$$\leq (h_n\varepsilon_n)^{-2}\mathbb{E}_{0,s_0}\big[\mathbb{E}_{\Pi_n}\big(\|s^2 - s_0^2\|_2^2 \mid \mathcal{X}_n\big)\big]. \tag{8.2}$$

The bias-variance decomposition gives

$$\mathbb{E}_{0,s_0}\big[\mathbb{E}_{\Pi_n}\big(\|s^2 - s_0^2\|_2^2|\mathcal{X}_n\big)\big]$$

$$= \mathbb{E}_{0,s_0}\big[\|M - s_0^2\|_2^2\big] + \sum_{k=1}^{N_n}\int_{B_k}\mathbb{E}_{0,s_0}\big[\mathbb{V}\mathrm{ar}_{\Pi_n}\big(s^2(x) \mid \mathcal{X}_n\big)\big]\,\mathrm{d}x.$$

The behaviour of the first term on the right-hand side was derived in the proof of Lemma 2. For the second term, we consider $x \in B_k$ for $k < N$. As in the proof of Lemma 2, the case $x \in B_N$ is similar. We obtain

$$\mathbb{V}\mathrm{ar}_{\Pi_n}\big(s^2(x)|\mathcal{X}_n\big) = \mathbb{V}\mathrm{ar}_{\Pi_n}\big(\theta_k|\mathcal{X}_n\big) = \frac{(\beta + nZ_k/2)^2}{(\alpha + m/2 - 1)^2(\alpha + m/2 - 2)}$$

$$= O\Big(\frac{1}{m^3}\Big) + O\Big(\frac{n^2}{m^3}\Big)Z_k^2. \tag{8.3}$$

By Assumption 1(c), s_0 is bounded by a positive constant $\mathcal{K}$. We have

$$\mathbb{E}_{0,s_0}[Z_k^2] = \mathbb{V}\mathrm{ar}_{0,s_0}[Z_k] + \big(\mathbb{E}_{0,s_0}[Z_k]\big)^2$$

$$= \sum_{i=(k-1)m+1}^{km}\mathbb{V}\mathrm{ar}_{0,s_0}[Y_{i,n}^2] + \Big(\int_{B_k}s_0^2(u)\,\mathrm{d}u\Big)^2$$

$$\leq 2\sum_{i=(k-1)m+1}^{km}\Big(\int_{t_{i-1}}^{t_i}s^2(u)\,\mathrm{d}u\Big)^2 + \Big(\frac{m}{n}\Big)^2\mathcal{K}^4$$

$$\leq 2\sum_{i=(k-1)m+1}^{km}\mathcal{K}^4n^{-2} + \Big(\frac{m}{n}\Big)^2\mathcal{K}^4$$

$$= O\Big(\frac{m}{n^2}\Big) + O\Big(\frac{m}{n}\Big)^2 = O\Big(\frac{m}{n}\Big)^2.$$

This implies

$$\mathbb{E}_{0,s_0}\big[\mathbb{V}\mathrm{ar}_{\Pi_n}\big(s^2(x)|\mathcal{X}_n\big)\big] = O\Big(\frac{1}{m^3}\Big) + O\Big(\frac{n^2}{m^3}\frac{m^2}{n^2}\Big) = O\Big(\frac{1}{m}\Big).$$

Combining the above order bounds then yields

$$\mathbb{E}_{0,s_0}\big[\mathbb{E}_{\Pi_n}\big(\|s^2 - s_0^2\|^2|\mathcal{X}_n\big)\big] = \Big[O\Big(\frac{1}{m}\Big) + O\Big(\frac{m}{n}\Big)^\lambda\Big]^2 + O\Big(\frac{1}{m}\Big).$$

Balancing these terms was already done in the proof of Lemma 2, and gives the value for m depending on n as in display (8.1). This implies the stated result for $b_0 \equiv 0$. Proposition 1 implies the result is not only true for $b_0 \equiv 0$, but in general. $\square$

Proof of Theorem 2. We first prove

$$\mathbb{E}_{0,s_0}\Big[\sup_{x\in[0,1]}|M(x)-s_0^2(x)|\Big] \asymp \tilde{\varepsilon}_n. \tag{8.4}$$

For that, we first consider a selected bin B_k for $k < N$ (the analysis on B_N would yield the same order estimates) and $\sup_{x\in B_k}|M(x)-s_0^2(x)|$. Note that for $x \in B_k$ one has that $M(x)$ does not explicitly depend on x, being equal to $\mathbb{E}_{0,s_0}(\theta_k|\mathcal{X}_n)$. We thus occasionally write $M(x) = M_k$, whenever this is convenient. For any $x \in B_k$ one has, recalling $b(x) = \mathbb{E}_{0,s_0}[M(x)] - s_0^2(x)$,

$$|M(x)-s_0^2(x)|^2 \le 2|M(x)-\mathbb{E}_{0,s_0}[M(x)]|^2 + 2|\mathbb{E}_{0,s_0}[M(x)]-s_0^2(x)|^2$$
$$= 2|M_k - \mathbb{E}_{0,s_0}[M_k]|^2 + 2|b(x)|^2,$$

and hence

$$\sup_{x\in B_k}|M(x)-s_0^2(x)|^2 \le 2|M_k-\mathbb{E}_{0,s_0}[M_k]|^2 + 2\sup_{x\in B_k}|b(x)|^2.$$

Turning to $\sup_{x\in[0,1]}|M(x)-s_0^2(x)|$, we obtain from the above that

$$\sup_{x\in[0,1]}|M(x)-s_0^2(x)|^2 = \sup_k \sup_{x\in B_k}|M(x)-s_0^2(x)|^2$$
$$\le 2\sup_k|M_k-\mathbb{E}_{0,s_0}[M_k]|^2 + 2\sup_k\sup_{x\in B_k}|b(x)|^2$$
$$\le 2\sum_k|M_k-\mathbb{E}_{0,s_0}[M_k]|^2 + 2\sup_k\sup_{x\in B_k}|b(x)|^2, \tag{8.5}$$

and therefore

$$\mathbb{E}_{0,s_0}\Big[\sup_{x\in[0,1]}|M(x)-s_0^2(x)|^2\Big] \le 2\sum_k\mathbb{V}\mathrm{ar}_{0,s_0}[M_k] + 2\sup_k\sup_{x\in B_k}|b(x)|^2.$$

Note that in the proof of Lemma 2, we obtained the uniform order bounds, not depending on x and k, $b(x) = O(\frac{1}{m}) + O(\frac{m}{n})^\lambda$ and $\mathbb{V}\mathrm{ar}_{0,s_0}[M(x)] = O(\frac{1}{m})$. It follows that

$$\mathbb{E}_{0,s_0}\Big[\sup_{x\in[0,1]}|M(x)-s_0^2(x)|^2\Big] \le N_n O\Big(\frac{1}{m}\Big) + \Big(O\Big(\frac{1}{m}\Big) + O\Big(\frac{m}{n}\Big)^\lambda\Big)^2$$
$$\le O\Big(\frac{n}{m^2}\Big) + \Big(O\Big(\frac{1}{m}\Big) + O\Big(\frac{m}{n}\Big)^\lambda\Big)^2.$$

Balancing of the two summands, using $m = O(n^\alpha)$, is obtained for $\alpha = \frac{2\lambda+1}{2\lambda+2}$. This results in

$$\mathbb{E}_{0,s_0}\Big[\sup_{x\in[0,1]}|M(x)-s_0^2(x)|^2\Big] = O(n^{-\frac{\lambda}{\lambda+1}}),$$

so that

$$\mathbb{E}_{0,s_0}\Big[\sup_{x\in[0,1]}|M(x)-s_0^2(x)|\Big] = O(n^{-\frac{\lambda}{2\lambda+2}}),$$

which completes the proof of (8.4). The proof of the second statement of the theorem follows upon combining elements from the proofs of Theorems 1 and of this result.

Proposition 1 implies the result is not only true for $b_0 \equiv 0$, but in general. $\square$

Appendix

In this appendix, we present an alternative asymptotic frequentist analysis of our Bayesian procedure. A principal difference with the one in the main text is that our prior on the coefficients θ_k's is not bound to be inverse gamma. On the downside, the L_2-posterior contraction rate we obtain is slower than that in Theorem 1. This is due to the fact that in our arguments we cannot rely on the conjugacy of the inverse gamma prior anymore.

Definition 1. Let $\mathcal{S}_n$ denote a set of dispersion coefficients $s : [0, 1] \to [\kappa, \mathcal{K}]$, that are piecewise constant on the bins B_k:

$$s = \sum_{k=1}^{N_n} \xi_k \mathbf{1}_{B_k}. \tag{A.1}$$

The prior Π_n on the dispersion coefficient s is defined by putting a prior on the coefficients ξ_k's.

$$s^2 = \sum_{k=1}^{N_n} \xi_k^2 \mathbf{1}_{B_k} = \sum_{k=1}^{N_n} \theta_k \mathbf{1}_{B_k},$$

where we have set $\theta_k = \xi_k^2$.

A.1 General contraction rate

Theorem 3. *Let Assumption 1 hold with bounds $0 < \kappa \leq s_0(t) \leq \mathcal{K} < \infty$ for all $t \in [0, 1]$. Assume the prior Π_n is defined as the law of a random function s from (A.1), where the random variables $\kappa \leq \xi_k \leq \mathcal{K}$, $k = 1, \ldots, N_n$, are independent and identically distributed with a density that is bounded away from zero on the interval $[\kappa, \mathcal{K}]$. Then for any sequence $m_n \asymp n^{1-\alpha}$, equivalently $N_n \asymp n^\alpha$, with $\alpha \in [\frac{1}{2}, 1 - \frac{\lambda}{2}]$, there exists a constant $\widetilde{M} > 0$, such that for $\varepsilon_n = \widetilde{M} n^{-\beta} \log^\gamma n$ with $\beta = \frac{\lambda}{4}$ and arbitrary $\gamma > 1$,*

$$\mathbb{E}_{b_0, s_0}\big[\Pi_n\big(U_{s_0, \varepsilon_n}^c \,|\, \mathcal{X}_n\big)\big] \to 0 \tag{A.2}$$

as $n \to \infty$.

The essential term determining the posterior contraction rate is $n^{-\beta}$. Here β depends on λ in an increasing way: a smoother function s_0 allows for a faster contraction rate. The optimality of the choice of α, the exponent β and the condition $\gamma > 1$, at least within the context of our proof, is elaborated in Remark 11 on page 575. The best possible rate obtainable from Theorem 3 is achieved for $\lambda = 1$ and is (essentially) $n^{-1/4}$. The possibility that the posterior contraction rate for arbitrary priors is in fact faster than the one given in Theorem 3 cannot be discarded: although our proofs for this general result use intricate technical arguments, they still might be not sharp enough.

Remark 9. The statement of Theorem 3 is in some respect similar to that of Theorem 1 in Gugushvili and Spreij (2016). Significant differences are that Gugushvili and Spreij (2016) choose a different prior and assume that the function s_0 is differentiable; cf. our remarks in the introduction to the paper. Not only does this amount to difference in statements, but also proofs in the present paper are more involved than the ones in Gugushvili and Spreij (2016).

Remark 10. The inequality

$$\|s - s_0\|_2 \geq \frac{1}{2\mathcal{K}} \|s^2 - s_0^2\|_2,$$

valid for $s \in \mathcal{S}_n$ and s_0 satisfying Assumption 1, together with Theorem 3 implies that also the posterior for the diffusion coefficient s_0^2 contracts around the truth with the rate ε_n given in that theorem.

A.2 Proof of Theorem 3

Some of by now classical references, where general conditions for derivation of posterior contraction rates are given, include Ghosal, Ghosh and van der Vaart (2000), Ghosal and van der Vaart (2007) and Shen and Wasserman (2001). The proof of Theorem 3 follows the same roadmap as these papers, however without appealing directly to their results, as our statistical setup is somewhat different from the ones covered by those papers. In particular, note that the distribution of the observations $X_{t_{i,n}}$'s depends on the index n, which is not covered by the results in the above-mentioned references.

We first introduce some notation and definitions: $p_{i,n,s}$ and $p_{i,n,0}$ will be the densities of increments $Y_{i,n} = X_{t_{i,n}} - X_{t_{i-1,n}}$ under the parameter values s and s_0, with drift equal to $b_0 = 0$ in both cases. The notation $Z_{i,n,s}(Y_{i,n}) = \log(p_{i,n,s}(Y_{i,n})/p_{i,n,0}(Y_{i,n}))$ will stand for the log-likelihood ratio corresponding to a single observation $Y_{i,n}$. $R_n(s) = L_n(s)/L_n(s_0)$ will denote the likelihood ratio. Furthermore, in line with the notation in van de Geer (2000), we let

$$z_i = t_{i-1,n}, \qquad \mathcal{W}_i = 1 - \frac{Y_{i,n}^2}{\int_{t_{i-1,n}}^{t_{i,n}} s_0^2(u)\, \mathrm{d}u},$$

$$f_s(z) = \frac{\int_z^{z+1/n} [s_0^2(u) - s^2(u)]\, \mathrm{d}u}{\int_z^{z+1/n} s^2(u)\, \mathrm{d}u}.$$

Under the parameter pair $(0, s_0)$, the random variables $\mathcal{W}_i$'s are i.i.d. with zero mean and variance equal to two. Since their distributions do not depend on n, we take the liberty to omit an extra index n in our notation. For notational simplicity, we also omit an extra index n in z_i's and f_s, though, strictly speaking, it is still required there.

Proof of Theorem 3. As in the proofs of results from the main body of the paper, we may assume $b_0 = 0$: the statement for a general b_0 follows from this particular case, see Proposition 1. The general structure of the proof is similar to the one of Theorem 1 in Gugushvili and Spreij (2016) and ultimately Ghosal, Ghosh and van der Vaart (2000) and Shen and Wasserman (2001), but many details differ, as evidenced in particular by the proofs of the lemmas in Appendix A.3.

Write

$$\Pi_n\big(U_{s_0,\varepsilon_n}^c \mid X_{t_0,n}, \ldots, X_{t_n,n}\big) = \frac{\int_{U_{s_0,\varepsilon_n}^c} R_n(s)\Pi(\mathrm{d}s)}{\int_{\mathcal{S}} R_n(s)\Pi_n(\mathrm{d}s)} = \frac{\mathcal{N}_n}{\mathcal{D}_n}.$$

Let $\varepsilon > 0$. For any events E_n and F_n we have

$$\mathbb{P}_{0,s_0}\Big(\frac{\mathcal{N}_n}{\mathcal{D}_n} > \varepsilon\Big) \leq \mathbb{P}_{0,s_0}\Big(\Big\{\frac{\mathcal{N}_n}{\mathcal{D}_n} > \varepsilon\Big\} \cap E_n \cap F_n\Big)$$

$$+ \mathbb{P}_{0,s_0}\big(E_n^c\big) + \mathbb{P}_{0,s_0}\big(F_n^c\big), \tag{A.3}$$

which we shall use for suitably chosen E_n and F_n having the property $\mathbb{P}_{0,s_0}(E_n^c) \to 0$ and $\mathbb{P}_{0,s_0}(F_n^c) \to 0$.

Denote $S_n(s) = n^{-1} \log R_n(s)$ and note that $\mathcal{D}_n = \int_{\mathcal{S}_n} \exp(nS_n(s))\Pi_n(ds)$. As in Gugushvili and Spreij (2016), we write

$$S_n(s) = \frac{1}{2}\frac{1}{n}\sum_{i=1}^n \mathcal{W}_i f_s(z_i) + \frac{1}{2}\frac{1}{n}\sum_{i=1}^n [\log(1+f_s(z_i)) - f_s(z_i)].$$

Below and in Appendix A.3 we need, for arbitrary $\varepsilon > 0$, the neighbourhoods $V_{s_0,\varepsilon} = \{s \in \mathcal{S} : \|s - s_0\|_\infty < \varepsilon\}$, where $\|\cdot\|_\infty$ is the usual L_∞-norm. We will use these for $\varepsilon = \widetilde{\varepsilon}_n \asymp n^{-\beta} \log^\gamma n$. We have the following lower bound on $\mathcal{D}_n$,

$$\mathcal{D}_n \geq \int_{V_{s_0,\widetilde{\varepsilon}_n}} R_n(s)\Pi_n(ds) \geq \inf_{s \in V_{s_0,\widetilde{\varepsilon}_n}} R_n(s) \times \Pi_n(V_{s_0,\widetilde{\varepsilon}_n}).$$

Let

$$E_n = \left\{ \sup_{s \in V_{s_0,\widetilde{\varepsilon}_n}} \left| \frac{1}{n}\sum_{i=1}^n \mathcal{W}_i f_s(z_i) \right| \leq \delta_n \right\},$$

with $\delta_n = \widetilde{\varepsilon}_n^2$. Lemmas 4 and 6 from Appendix A.3 yield that on E_n one has

$$\inf_{s \in V_{s_0,\widetilde{\varepsilon}_n}} R_n(s) \geq \exp\left(-\frac{2\mathcal{K}^2}{\kappa^4} n\widetilde{\varepsilon}_n^2 - \frac{n}{2} \sup_{s \in V_{s_0,\widetilde{\varepsilon}_n}} \left| \frac{1}{n}\sum_{i=1}^n \mathcal{W}_i f_s(z_i) \right| \right)$$

$$\geq \exp\left(-\frac{2\mathcal{K}^2}{\kappa^4} n\widetilde{\varepsilon}_n^2 - \frac{n\widetilde{\varepsilon}_n^2}{2} \right),$$

whereas $\mathbb{P}_{0,s_0}(E_n^c) \to 0$. Hence on E_n we have, using the prior mass result of Lemma 3,

$$\mathcal{D}_n \geq \exp\left(-\left(\frac{2\mathcal{K}^2}{\kappa^4} + \frac{1}{2} + \overline{C} \right) n\widetilde{\varepsilon}_n^2 \right).$$

Next, we consider $\mathcal{N}_n$, for which we have the trivial upper bound

$$\mathcal{N}_n \leq \sup_{U_{s_0,\varepsilon_n}^c} R_n(s).$$

Let

$$F_n = \left\{ \sup_{s \in U_{s_0,\varepsilon_n}^c} R_n(s) \leq \exp(-c_1 n\varepsilon_n^2) \right\},$$

with $\varepsilon_n \asymp n^{-\beta} \log^\gamma n$ and the constant $c_1 > 0$ as in Lemma 8. According to the latter lemma, we have $\mathbb{P}_{0,s_0}(F_n^c) \to 0$. Putting the above results together, we obtain on $E_n \cap F_n$ the inequality

$$\frac{\mathcal{N}_n}{\mathcal{D}_n} \leq \exp\left(\left(\frac{2\mathcal{K}^2}{\kappa^4} + \frac{1}{2} + \overline{C} \right) n\widetilde{\varepsilon}_n^2 - c_1 n\varepsilon_n^2 \right).$$

Taking $\varepsilon_n = \widetilde{M}\widetilde{\varepsilon}_n$, with $\widetilde{M} > 0$ large enough, gives a positive constant M for which on $E_n \cap F_n$ the inequality

$$\frac{\mathcal{N}_n}{\mathcal{D}_n} \leq \exp(-Mn\widetilde{\varepsilon}_n^2)$$

holds. It follows that

$$\mathbb{P}_{0,s_0}\left(\left\{ \frac{\mathcal{N}_n}{\mathcal{D}_n} > \varepsilon \right\} \cap E_n \cap F_n \right) = 0$$

for all large n. We then obtain from (A.3) that

$$\Pi_n\big(U^c_{s_0,\varepsilon_n}|\mathcal{X}_n\big) \xrightarrow{\mathbb{P}_{b_0,s_0}} 0.$$

The assertion of the theorem now follows from this and the dominated convergence theorem. The optimal choice for β in $\varepsilon_n \asymp n^{-\beta} \log^\gamma n$, as well as of α in $N_n \asymp n^\alpha$, is a consequence of a discussion in Remark 11. $\qquad\square$

A.3 Proofs of the remaining technical results

We use the following notation: M_ε will denote the smallest positive integer, such that $2^{M_\varepsilon}\varepsilon^2 \geq 4\mathcal{K}^2$. Note that this definition implies $2^{M_\varepsilon}\varepsilon^2 \leq 8\mathcal{K}^2$, so that $M_\varepsilon \asymp \log_2(1/\varepsilon)$ for $\varepsilon \to 0$. We also define sets $A_{j,\varepsilon} = \{s \in \mathcal{S} : 2^j\varepsilon^2 \leq \|s-s_0\|_2^2 < 2^{j+1}\varepsilon^2\}$ and $B_{j,\varepsilon} = \{s \in \mathcal{S} : \|s-s_0\|_2^2 < 2^{j+1}\varepsilon^2\}$ for $j = 0, 1, \ldots, M_\varepsilon$. The measure Q_n will be the uniform discrete probability measure on points z_i's, while $\|\cdot\|_{Q_n}$ will denote the $L_2(Q_n)$-norm. In this appendix, we use these sets for $\varepsilon = \widetilde{\varepsilon}_n \asymp n^{-\beta} \log^\gamma n$.

The next lemma verifies the prior mass condition in the proof of our main theorem. This corresponds, roughly speaking, to for example, condition (2.4) in Theorem 2.1 of Ghosal, Ghosh and van der Vaart (2000). This prior mass condition is crucial in establishing posterior contraction rates and we refer to Ghosal, Ghosh and van der Vaart (2000) for an additional discussion on it. Note that in Gugushvili and Spreij (2016) this condition is verified for another, somewhat artificial, prior.

Lemma 3. *Under the conditions of Theorem 3, the prior Π_n satisfies*

$$\Pi_n(V_{s_0,\widetilde{\varepsilon}_n}) \gtrsim e^{-\overline{C}n\widetilde{\varepsilon}_n^2} \tag{A.4}$$

for some constant $\overline{C} > 0$.

Proof. Since ξ_k's are independent, we have

$$\Pi_n(V_{s_0,\widetilde{\varepsilon}_n}) = \Pi_n\left(\bigcap_{k=1}^{N_n}\left\{\sup_{x \in B_k} |s(x) - s_0(x)| < \widetilde{\varepsilon}_n\right\}\right)$$

$$= \prod_{k=1}^{N_n} \Pi_n\left(\sup_{x \in B_k} |s(x) - s_0(x)| < \widetilde{\varepsilon}_n\right)$$

$$= \prod_{k=1}^{N_n} \Pi_n\left(\sup_{x \in B_k} |\xi_k - s_0(x)| < \widetilde{\varepsilon}_n\right).$$

Now, since s_0 is Hölder continuous and B_k is of length at most $2m_n/n$, we have

$$s_0(x) = s_0(a_{k-1}) + O\left(\frac{m_n}{n}\right)^\lambda, \quad x \in B_k,$$

where the order term is uniform in $x \in [0, 1]$. Remember also that by our conditions

$$\left(\frac{m_n}{n}\right)^\lambda \ll \widetilde{\varepsilon}_n. \tag{A.5}$$

Then for all n large enough and all k,

$$\Pi_n\left(\sup_{x \in B_k} |\xi_k - s_0(x)| < \widetilde{\varepsilon}_n\right) \geq \Pi_n\left(|\xi_k - s_0(a_{k-1})| < \frac{\widetilde{\varepsilon}_n}{2}\right)$$

$$\geq \text{const} \cdot \widetilde{\varepsilon}_n,$$

where const > 0 is some constant independent of k and n, and the last inequality comes from the fact that ξ_k has a density bounded away from zero on $[\kappa, \mathcal{K}]$. It then follows that

$$
\Pi_n(V_{s_0, \widetilde{\varepsilon}_n}) \geq (\text{const} \cdot \widetilde{\varepsilon}_n)^{N_n}
$$
$$
= e^{N_n \log(\text{const} \cdot \widetilde{\varepsilon}_n)}.
$$

We want to show existence of a constant $\overline{C} > 0$, such that for all n large enough,

$$
N_n \log(\text{const} \cdot \widetilde{\varepsilon}_n) \geq -\overline{C} n \widetilde{\varepsilon}_n^2.
$$

But this is immediate from our conditions on N_n (equivalently, m_n) and $\widetilde{\varepsilon}_n$, e.g. for $\overline{C} = 1$. The proof of the lemma is completed. $\square$

The following lemma is an analogue of Lemma A.2 in Gugushvili and Spreij (2016). The statement is slightly different, and so is the proof.

Lemma 4. *Let the conditions of Theorem 3 hold and let $s \in V_{s_0, \widetilde{\varepsilon}_n}$. Then, uniformly in s,*

$$
\frac{1}{2} \frac{1}{n} \sum_{i=1}^{n} \{\log(1 + f_s(z_i)) - f_s(z_i)\} \geq -\frac{2\mathcal{K}^2}{\kappa^4} \widetilde{\varepsilon}_n^2.
$$

Proof. As in the proof in Gugushvili and Spreij (2016), one derives

$$
\frac{1}{2} \frac{1}{n} \sum_{i=1}^{n} \{\log(1 + f_s(z_i)) - f_s(z_i)\} \geq -\frac{1}{2n} \sum_{i=1}^{n} f_s^2(z_i). \tag{A.6}
$$

Using the Cauchy–Schwarz inequality, the bounds on s and the fact that $s \in V_{s_0, \widetilde{\varepsilon}_n}$, one gets

$$
f_s^2(z_i) = \left(\frac{\int_z^{z+1/n} [s_0^2(u) - s^2(u)] \, du}{\int_z^{z+1/n} s^2(u) \, du} \right)^2 \leq \frac{\frac{1}{n} \int_z^{z+1/n} [s_0^2(u) - s^2(u)]^2 \, du}{\kappa^4 / n^2}
$$
$$
\leq \frac{\frac{4\mathcal{K}^2}{n} \int_z^{z+1/n} [s_0(u) - s(u)]^2 \, du}{\kappa^4 / n^2} \leq \frac{4\mathcal{K}^2 \widetilde{\varepsilon}_n^2}{\kappa^4}.
$$

The result follows by inserting the latter upper bound into (A.6). $\square$

Lemma 6 below is a more precise version of Lemma A.1 in Gugushvili and Spreij (2016). Its proof is similar in general terms, but differs substantially in the entropy estimates used, as well as some other arguments. In its proof and that of Lemma 7, we will need the following lemma.

Lemma 5. *Let $g : [0, \infty) \to \mathbb{R}$ be Hölder continuous of order $\lambda > 0$ and Hölder constant $L > 0$. Then*

$$
\left| \int_z^{z+h} g(u) \, du - h g(z) \right| \leq \frac{L h^{1+\lambda}}{1 + \lambda}.
$$

Furthermore, if $z_i = \frac{i}{n}$, for $i = 1, \ldots, n$, then

$$
\left| \int_0^1 g(u) \, du - \frac{1}{n} \sum_{i=1}^{n} g(z_i) \right| \leq \frac{L}{(1 + \lambda) n^\lambda}.
$$

Proof. For the first statement, we consider

$$
\left| \int_z^{z+h} g(u)\, \mathrm{d}u - h g(z) \right| = \left| \int_z^{z+h} \left(g(u) - g(z) \right) \mathrm{d}u \right|
$$

$$
\leq \int_z^{z+h} \left| g(u) - g(z) \right| \mathrm{d}u
$$

$$
\leq \int_z^{z+h} L(u - z)^\lambda \, \mathrm{d}u = \frac{L h^{1+\lambda}}{1 + \lambda}.
$$

For the second one we have, using the first part,

$$
\left| \int_0^1 g(u)\, \mathrm{d}u - \frac{1}{n} \sum_{i=1}^n g(z_i) \right| = \left| \sum_{i=1}^n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left(g(u) - g(z_i) \right) \mathrm{d}u \right|
$$

$$
\leq \sum_{i=1}^n \left| \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left(g(u) - g(z_i) \right) \mathrm{d}u \right|
$$

$$
\leq \frac{L}{(1 + \lambda) n^\lambda}.
$$

The proof is completed. $\qquad\square$

Lemma 6. *Let the conditions of Theorem* 3 *hold and assume* $b_0 = 0$. *Then*

$$
\mathbb{P}_{0,s_0} \left(\sup_{f_s \in \mathcal{F}_{s_0, \widetilde{\varepsilon}_n}} \left| \frac{1}{n} \sum_{i=1}^n \mathcal{W}_i f_s(z_i) \right| \geq \delta_n \right) \lesssim \frac{1}{n^\lambda \widetilde{\varepsilon}_n^2},
$$

where $\mathcal{F}_{s_0, \widetilde{\varepsilon}_n} = \{ f_s : \| s - s_0 \|_\infty < \widetilde{\varepsilon}_n \} = \{ f_s : s \in V_{s_0, \widetilde{\varepsilon}_n} \}$, *and* δ_n *is an arbitrary sequence of positive numbers, such that* $\delta_n \asymp \widetilde{\varepsilon}_n^2$. *In particular, as* $n \to \infty$, *the probability on the left-hand side of the above display converges to zero.*

Proof. Introduce

$$
g_s(z) = \frac{s_0^2(z) - s^2(z)}{s^2(z)}, \qquad \mathcal{G}_{s_0, \widetilde{\varepsilon}_n} = \{ g_s : \| s - s_0 \|_\infty < \widetilde{\varepsilon}_n \}.
$$

The function g_s approximates f_s in the following sense:

$$
\left| f_s(z_i) - g_s(z_i) \right| \leq \frac{2 \mathcal{K} L}{\kappa^2 n^\lambda}, \quad i = 1, \ldots, n, \tag{A.7}
$$

which can be seen as follows. Every interval $[z_i, z_i + 1/n)$ is contained in some bin B_k, so that s^2 is constant on $[z_i, z_i + 1/n)$. Hence, one obtains $f_s(z_i) = n \int_{z_i}^{z_i + 1/n} g_s(u)\, \mathrm{d}u$. It follows from Definition 1 and boundedness and Hölder continuity of s_0 in Assumption 1 that

$$
\left| g_s(u) - g_s(v) \right| \leq \frac{2 \mathcal{K} L}{\kappa^2} |u - v|^\lambda.
$$

Hence, equation (A.7) follows from Lemma 5. Note that the right-hand side of (A.7) is uniform in s.

Let

$$
\rho_n = \frac{2 \mathcal{K} L}{\kappa^2 n^{1+\lambda}} \sum_{i=1}^n |\mathcal{W}_i|.
$$

Then

$$\sup_{f_s \in \mathcal{F}_{s_0, \widetilde{\varepsilon}_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_i f_s(z_i) \right| \leq \sup_{g_s \in \mathcal{G}_{s_0, \widetilde{\varepsilon}_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_i g_s(z_i) \right| + \rho_n.$$

Hence,

$$\mathbb{P}_{0,s_0} \left(\sup_{f_s \in \mathcal{F}_{s_0, \widetilde{\varepsilon}_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_i f_s(z_i) \right| \geq \delta_n \right)$$

$$\leq \mathbb{P}_{0,s_0} \left(\sup_{g_s \in \mathcal{G}_{s_0, \widetilde{\varepsilon}_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_i g_s(z_i) \right| + \rho_n \geq \delta_n \right)$$

$$\leq \mathbb{P}_{0,s_0} \left(\sup_{g_s \in \mathcal{G}_{s_0, \widetilde{\varepsilon}_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_i g_s(z_i) \right| \geq \delta_n/2 \right) + \mathbb{P}_{0,s_0}(\rho_n \geq \delta_n/2).$$

The Markov inequality gives

$$\mathbb{P}_{0,s_0}(\rho_n \geq \delta_n/2) \lesssim \frac{1}{n^\lambda \delta_n} = \frac{1}{n^\lambda \widetilde{\varepsilon}_n^2}.$$

Therefore, in order to prove the lemma, it is enough to show that

$$\mathbb{P}_{0,s_0} \left(\sup_{g_s \in \mathcal{G}_{s_0, \widetilde{\varepsilon}_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_i g_s(z_i) \right| \geq \delta_n \right) \lesssim \frac{1}{n^\lambda \widetilde{\varepsilon}_n^2}. \tag{A.8}$$

We will apply Corollary 8.8 from van de Geer (2000) to show that (A.8) holds true, since the assertion of that corollary, inequality (8.30), implies that

$$\mathbb{P}_{0,s_0} \left(\sup_{g_s \in \mathcal{G}_{s_0, \widetilde{\varepsilon}_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_i g_s(z_i) \right| \geq \delta_n \right) \leq \tilde{c}_1 \exp(-\tilde{c}_2 n \widetilde{\varepsilon}_n^2), \tag{A.9}$$

for some positive constants $\tilde{c}_1$, $\tilde{c}_2$. The right-hand side of the above display is asymptotically much smaller than $1/(n^\lambda \widetilde{\varepsilon}_n^2)$.

Application of the mentioned corollary amounts to verification of formulae (8.23)–(8.29) in van de Geer (2000). Exactly as in Gugushvili and Spreij (2016), conditions (8.23)–(8.27), (8.29) can be satisfied by choosing $R_n = 2\mathcal{K}\widetilde{\varepsilon}_n/\kappa^2$, $K_1 = 3$, $\sigma_0^2 = 2\mathbb{E}_{0,s_0}[\mathcal{W}_i^2 e^{|\mathcal{W}_i|/3}]$, $K_2 = 2\mathcal{K}\widetilde{\varepsilon}_n/\kappa^2$, $C_1 = 3$, $K = 4K_1 K_2$, and $C_0 = 2C$, with C a universal constant as in Corollary 8.8 in van de Geer (2000). Finally, we need to verify formula (8.28) in van de Geer (2000),

$$\sqrt{n}\delta_n \geq C_0 \left(\int_{\delta_n/2^6}^{\sqrt{2}R_n\sigma_0} H_B^{1/2} \left(\frac{u}{\sqrt{2}\sigma_0}, \mathcal{G}_{s_0, \widetilde{\varepsilon}_n}, Q_n \right) du \vee \sqrt{2}R_n\sigma_0 \right). \tag{A.10}$$

Here $H_B(\delta, \mathcal{G}_{s_0, \widetilde{\varepsilon}_n}, Q_n)$ is the δ-entropy with bracketing of $\mathcal{G}_{s_0, \widetilde{\varepsilon}_n}$ for the $L_2(Q_n)$-metric (see Definition 2.2 in van de Geer (2000)).

First of all, note that

$$\sqrt{n}\delta_n \asymp \sqrt{n}\widetilde{\varepsilon}_n^2 \gg R_n \asymp \widetilde{\varepsilon}_n,$$

holds, since we have $n\widetilde{\varepsilon}_n^2 \to \infty$. Thus it suffices to show

$$\sqrt{n}\delta_n \geq C_0 \int_{\delta_n/2^6}^{\sqrt{2}R_n\sigma_0} H_B^{1/2} \left(\frac{u}{\sqrt{2}\sigma_0}, \mathcal{G}_{s_0, \widetilde{\varepsilon}_n}, Q_n \right) du.$$

In order to do this, we will upper bound the right-hand side by first upper bounding the integrand with a manageable and simple expression, and thereby we obtain a bound on the

integral itself. We will use H_∞, the entropy for the supremum norm (see Definition 2.3 in van de Geer (2000)) and the inequality $H_B(\delta, \mathcal{G}, Q) \leq H_\infty(\delta/2, \mathcal{G})$, valid for any collection of functions $\mathcal{G}$ and a probability measure Q, see Lemma 2.1 in van de Geer (2000), to obtain

$$\int_{\delta_n/2^6}^{\sqrt{2}R_n\sigma_0} H_B^{1/2}\left(\frac{u}{\sqrt{2}\sigma_0}, \mathcal{G}_{s_0,\widetilde{\varepsilon}_n}, Q_n\right) du$$

$$\leq \int_{\delta_n/2^6}^{\sqrt{2}R_n\sigma_0} H_\infty^{1/2}\left(\frac{u}{2\sqrt{2}\sigma_0}, \mathcal{G}_{s_0,\widetilde{\varepsilon}_n}\right) du$$

$$= 2\sqrt{2}\sigma_0 \int_0^{R_n/2-\delta_n/(2^7\sqrt{2}\sigma_0)} H_\infty^{1/2}\left(u + \frac{\delta_n}{2^7\sqrt{2}\sigma_0}, \mathcal{G}_{s_0,\widetilde{\varepsilon}_n}\right) du$$

$$\leq 2\sqrt{2}\sigma_0 \int_0^{R_n} H_\infty^{1/2}\left(u + \frac{\delta_n}{2^7\sqrt{2}\sigma_0}, \mathcal{G}_{s_0,\widetilde{\varepsilon}_n}\right) du.$$

We will now estimate the entropy in the last integral in the above display. Suppose $u > 0$ is fixed. For every $g \in \mathcal{G}_{s_0,\widetilde{\varepsilon}_n}$ construct an approximating function

$$\widetilde{g} = \sum_{k=1}^{N_n} u\left\lfloor \frac{g(a_{k-1})}{u} \right\rfloor 1_{B_k}.$$

The quality of approximation can be assessed as follows: we have

$$\|g - \widetilde{g}\|_\infty = \max_k \|(g - \widetilde{g})1_{B_k}\|_\infty$$

$$= \max_k \sup_{x \in B_k} \left| g(x) - g(a_{k-1}) + u\frac{g(a_{k-1})}{u} - u\left\lfloor \frac{g(a_{k-1})}{u} \right\rfloor \right|$$

$$\leq \max_k \sup_{x \in B_k} \left| g(x) - g(a_{k-1}) \right| + u.$$

Since s^2 is constant on B_k and s_0^2 is Hölder, while B_k is of length at most $2m_n/n$, we have

$$\sup_{x \in B_k} \left| g(x) - g(a_{k-1}) \right| = \sup_{x \in B_k} \left| \frac{s_0^2(x) - s_0^2(a_{k-1})}{s^2(a_{k-1})} \right|$$

$$\leq \frac{2\mathcal{K}L}{\kappa^2}\left(\frac{2m_n}{n}\right)^\lambda,$$

so that

$$\|g - \widetilde{g}\|_\infty \leq u + \frac{2\mathcal{K}L}{\kappa^2}\left(\frac{2m_n}{n}\right)^\lambda. \tag{A.11}$$

According to our assumptions,

$$\left(\frac{m_n}{n}\right)^\lambda \ll \delta_n, \tag{A.12}$$

and hence for any pair of positive constants A and B one eventually has

$$A\left(\frac{2m_n}{n}\right)^\lambda \leq B\delta_n.$$

This implies that for all n large enough,

$$H_\infty^{1/2}\left(u + \frac{\delta_n}{2^7\sqrt{2}\sigma_0}, \mathcal{G}_{s_0,\widetilde{\varepsilon}_n}\right) \leq H_\infty^{1/2}\left(u + \frac{2\mathcal{K}L}{\kappa^2}\left(\frac{2m_n}{n}\right)^\lambda, \mathcal{G}_{s_0,\widetilde{\varepsilon}_n}\right).$$

The entropy on the right-hand side of the above display can be estimated by bounding the corresponding covering number of $\mathcal{G}_{s_0,\widetilde{\varepsilon}_n}$, which can be achieved by counting the number of different $\widetilde{g}$'s. In fact, as we shall see below, although $\mathcal{G}_{s_0,\widetilde{\varepsilon}_n}$ is an infinite set, the number of different $\widetilde{g}$'s is finite and can by easily estimated from above.

First, recall that $\|g\|_\infty \leq (2\mathcal{K}/\kappa^2)\widetilde{\varepsilon}_n$ for $g \in \mathcal{G}_{s_0,\widetilde{\varepsilon}_n}$. This implies that there are at most

$$\left\lfloor \frac{4\mathcal{K}}{\kappa^2}\frac{\widetilde{\varepsilon}_n}{u} \right\rfloor + 1$$

possible values for $\widetilde{g}(a_0)$. Next, by the triangle inequality and (A.11),

$$\left|\widetilde{g}(a_k) - \widetilde{g}(a_{k-1})\right| \leq \left|\widetilde{g}(a_k) - g(a_k)\right| + \left|g(a_k) - g(a_{k-1})\right| + \left|\widetilde{g}(a_{k-1}) - g(a_{k-1})\right|$$

$$\leq 2u + \frac{4\mathcal{K}}{\kappa^2}L\left(\frac{2m_n}{n}\right)^\lambda + 2\|g\|_\infty.$$

Thus, once $\widetilde{g}(a_{k-1})$ has been determined, $\widetilde{g}(a_k)$ can take at most

$$\left\lfloor \frac{1}{u}\left(4u + \frac{8\mathcal{K}}{\kappa^2}L\left(\frac{2m_n}{n}\right)^\lambda + 4\|g\|_\infty\right)\right\rfloor + 1$$

values. Therefore, in total there can be at most

$$\left(\left\lfloor \frac{4\mathcal{K}}{\kappa^2}\frac{\widetilde{\varepsilon}_n}{u} \right\rfloor + 1\right)\left(\left\lfloor 4 + \frac{8\mathcal{K}}{\kappa^2}\frac{L}{u}\left(\frac{2m_n}{n}\right)^\lambda + \frac{8\mathcal{K}}{\kappa^2}\frac{\widetilde{\varepsilon}_n}{u} \right\rfloor + 1\right)^{N_n-1}$$

different $\widetilde{g}$'s, which yields an upper bound on the covering number of the set $\mathcal{G}_{s_0,\widetilde{\varepsilon}_n}$. Using (A.12), for large n, the logarithm of the above display is bounded by

$$N_n \log\left(\text{const} + \text{const}\frac{\widetilde{\varepsilon}_n}{u}\right)$$

for some constant const > 0 independent of n and u. This expression gives an upper bound on the entropy of the set $\mathcal{G}_{s_0,\widetilde{\varepsilon}_n}$. Inserting this upper bound into the entropy integral, we get that for all n large enough,

$$\int_0^{R_n} H_\infty^{1/2}\left(u + \frac{\delta_n}{2^7\sqrt{2}\sigma_0}, \mathcal{G}_{s_0,\widetilde{\varepsilon}_n}\right)du$$

$$\leq \sqrt{N_n}\int_0^{R_n} \log^{1/2}\left(\text{const} + \text{const}\frac{\widetilde{\varepsilon}_n}{u}\right)du$$

$$= \widetilde{\varepsilon}_n\sqrt{N_n}\int_0^{2\mathcal{K}/\kappa^2} \log^{1/2}\left(\text{const} + \text{const}\frac{1}{u}\right)du$$

$$\asymp \widetilde{\varepsilon}_n\sqrt{N_n},$$

since the integral in the second line is convergent, because the integrand is bounded by a constant times $u^{-1/2}$, and the latter function is integrable in the neighbourhood of zero.

Summarising the above intermediate calculations, we obtain that in order to prove (8.28) in van de Geer (2000), we need

$$\sqrt{n}\delta_n \gg \widetilde{\varepsilon}_n\sqrt{N_n} \tag{A.13}$$

and (A.12) to hold. Both are satisfied with our choice of $\widetilde{\varepsilon}_n$ and N_n (equivalently, m_n). Hence, all conditions of Corollary 8.8 in van de Geer (2000) are satisfied and (A.9) follows. This completes the proof. $\qquad\square$

The next lemma, to be used in the proof of Lemma 8, is an analogue of Lemma A.4 in Gugushvili and Spreij (2016). Its proof is also similar in structure, but differs in details.

Lemma 7. *Let the conditions of Theorem 3 hold and assume $b_0 = 0$. There exist two positive constants $\widetilde{c}_0$ and $\widetilde{C}_0$, depending on κ, $\mathcal{K}$ and L only, such that for all n large enough and all $s \in A_{j,\varepsilon_n}$, $j = 0, 1, \ldots, M_{\varepsilon_n}$, we have*

$$\sum_{i=1}^{n} \mathbb{E}_{0,s_0}[Z_{i,n,s}(Y_{i,n})] \leq -\frac{\widetilde{c}_0 \kappa^2}{\mathcal{K}^4} 2^j \varepsilon_n^2 n + \widetilde{C}_0 n^{1-\lambda}.$$

Proof. Below we use the following elementary inequality: for any fixed constant $C > 0$ there exists another constant $\widetilde{c}_0 > 0$, such that for $x \in (-1, C)$, $\log(1 + x) - x \leq -\widetilde{c}_0 x^2$ holds.

It follows from Assumption 1 that

$$\frac{\int_{[z_i,z_{i+1})} [s_0^2(u) - s^2(u)] \, du}{\int_{[z_i,z_{i+1})} s^2(u) \, du} \leq \frac{\mathcal{K}^2}{\kappa^2} - 1 =: C > 0.$$

Hence, for some $\widetilde{c}_0 > 0$,

$$\mathbb{E}_{0,s_0}[Z_{i,n,s}(Y_{i,n})] = \frac{1}{2} \log\left(1 + \frac{\int_{[z_i,z_{i+1})} [s_0^2(u) - s^2(u)] \, du}{\int_{[z_i,z_{i+1})} s^2(u) \, du}\right)$$

$$- \frac{1}{2} \frac{\int_{[z_i,z_{i+1})} [s_0^2(u) - s^2(u)] \, du}{\int_{[z_i,z_{i+1})} s^2(u) \, du}$$

$$\leq -\frac{\widetilde{c}_0}{2} \left\{ \frac{\int_{[z_i,z_{i+1})} [s_0^2(u) - s^2(u)] \, du}{\int_{[z_i,z_{i+1})} s^2(u) \, du} \right\}^2.$$

We now focus on the term in braces. Let $g(u) = \frac{(s^2(u) - s_0^2(u))^2}{s^4(u)}$. On bins, and hence on intervals $[z_i, z_{i+1})$, the function s is constant and positive, and by Hölder continuity and boundedness of s_0 also g is Hölder continuous. Application of Lemma 5 gives

$$\int_{[z_i,z_{i+1})} \frac{(s^2(u) - s_0^2(u))^2}{s^4(u)} \, du = \frac{1}{n} \frac{1}{s^4(z_i)} (s_0^2(z_i) - s^2(z_i))^2$$

$$+ O\left(\frac{1}{n^{1+\lambda}}\right). \tag{A.14}$$

Next, by a similar argument,

$$\frac{\int_{[z_i,z_{i+1})} [s^2(u) - s_0^2(u)] \, du}{\int_{[z_i,z_{i+1})} s^2(u) \, du} = \frac{s^2(z_i) - s_0^2(z_i)}{s^2(z_i)} + O\left(\frac{1}{n^{\lambda}}\right). \tag{A.15}$$

Combination of (A.14) and (A.15) and the bounds in Assumption 1 yields

$$\left\{ \frac{\int_{[z_i,z_{i+1})} [s^2(u) - s_0^2(u)] \, du}{\int_{[z_i,z_{i+1})} s^2(u) \, du} \right\}^2 - n \int_{[z_i,z_{i+1})} \frac{(s^2(u) - s_0^2(u))^2}{s^4(u)} \, du = O\left(\frac{1}{n^{\lambda}}\right).$$

Consequently, by summation,

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\int_{[z_i,z_{i+1})} [s^2(u) - s_0^2(u)] \, du}{\int_{[z_i,z_{i+1})} s^2(u) \, du} \right\}^2$$

$$- \int_0^1 \frac{(s^2(u) - s_0^2(u))^2}{s^4(u)} \, du = O\left(\frac{1}{n^{\lambda}}\right). \tag{A.16}$$

Hence, for all n large enough,

$$
\sum_{i=1}^{n} \mathbb{E}_{0,s_0}[Z_{i,n,s}(Y_{i,n})] \leq -\frac{\widetilde{c}_0 n}{2} \int_0^1 \frac{(s^2(u) - s_0^2(u))^2}{s^4(u)} \, du + O(n^{1-\lambda})
$$

$$
\leq -\frac{\widetilde{c}_0 \kappa^2}{\mathcal{K}^4} 2^j \varepsilon_n^2 n + \widetilde{C}_0 n^{1-\lambda}.
$$

Here the first inequality follows from (A.16) and the last inequality from the assumption $s \in A_{j,\varepsilon_n}$. The proof is completed. $\qquad\square$

The following lemma is an analogue of Lemma A.3 in Gugushvili and Spreij (2016). The proof also shares similar general arguments, but differs in details.

Lemma 8. *Let Assumption 1 be satisfied and assume also $b_0 = 0$. For a fixed and small enough constant $c_1 > 0$, for any sequence $\varepsilon_n \asymp n^{-\beta} \log^\gamma n$ of strictly positive numbers with $\gamma > 0$, and any sequence $m_n \asymp n^{1-\alpha}$ (equivalently, $N_n \asymp n^\alpha$) as in Theorem 3, there exist a constant $c_0 > 0$ (depending on κ, $\mathcal{K}$ and L) and a universal constant $C > 0$ for which the inequality*

$$
\mathbb{P}_{0,s_0}\left(\sup_{s \in U_{s_0,\varepsilon_n}^c} \prod_{i=1}^{n} \frac{p_{i,n,s}(Y_{i,n})}{p_{i,n,0}(Y_{i,n})} \geq \exp(-c_1 n \varepsilon_n^2) \right) \leq C M_{\varepsilon_n} \exp(-c_0 n \varepsilon_n^2) \tag{A.17}
$$

holds for all n large enough.

Proof. As in the proof of Lemma A.3 in Gugushvili and Spreij (2016), one has

$$
\mathbb{P}_{0,s_0}\left(\sup_{s \in U_{s_0,\varepsilon_n}^c} \prod_{i=1}^{n} \frac{p_{i,n,s}(Y_{i,n})}{p_{i,n,0}(Y_{i,n})} \geq \exp(-c_1 n \varepsilon_n^2) \right)
$$

$$
= \sum_{j=0}^{M_{\varepsilon_n}} \mathbb{P}_{0,s_0}\left(\sup_{s \in A_{j,\varepsilon_n}} \prod_{i=1}^{n} \frac{p_{i,n,s}(Y_{i,n})}{p_{i,n,0}(Y_{i,n})} \geq \exp(-c_1 n \varepsilon_n^2) \right). \tag{A.18}
$$

Furthermore, using Lemma 7 above, by arguments identical to those in the proof of Lemma A.3 in Gugushvili and Spreij (2016), we have for all n large enough

$$
\mathbb{P}_{0,s_0}\left(\sup_{s \in A_{j,\varepsilon_n}} \prod_{i=1}^{n} \frac{p_{i,n,s}(Y_{i,n})}{p_{i,n,0}(Y_{i,n})} \geq \exp(-c_1 n \varepsilon_n^2) \right)
$$

$$
\leq \mathbb{P}_{0,s_0}\left(\sup_{s \in B_{j,\varepsilon_n}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_i f_s(z_i) \right| \geq \delta_n \right), \tag{A.19}
$$

where

$$
\delta_n = \overline{\delta} 2^{j+1} \varepsilon_n^2 = \left(\frac{\widetilde{c}_0 \kappa^2}{\mathcal{K}^4} - \frac{\widetilde{C}_0}{2^j \varepsilon_n^2 n^\lambda} - \frac{c_1}{2^j} \right) 2^{j+1} \varepsilon_n^2. \tag{A.20}
$$

Note that $\overline{\delta} > 0$ for all n large enough, by choosing $0 < c_1 < \widetilde{c}_0 \kappa^2/(2\mathcal{K}^4)$ and $n^\lambda \varepsilon_n^2 \to \infty$ as $n \to \infty$. The former is a restriction on c_1 alluded to in the statement of the theorem.

To bound the right-hand side in (A.19), we will apply Corollary 8.8 from van de Geer (2000). To that end, we need to verify its conditions. Under our assumptions, we have

$$
\int_0^1 \frac{(s_0^2(u) - s^2(u))^2}{s^4(u)} \, du \leq \frac{4\mathcal{K}^2}{\kappa^4} 2^{j+1} \varepsilon_n^2,
$$

and the $O(n^{-\lambda})$ term in (A.16) is less than $2^{j+1}\varepsilon_n^2$ for all large n. Hence, (A.16) yields

$$\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\int_{[z_i,z_{i+1})}[s_0^2(u)-s^2(u)]\,\mathrm{d}u}{\int_{[z_i,z_{i+1})}s^2(u)\,\mathrm{d}u}\right\}^2 \leq \left(\frac{4\mathcal{K}^2}{\kappa^4}+1\right)2^{j+1}\varepsilon_n^2$$

for all n large enough and all $j=0,1,\ldots,M_{\varepsilon_n}$. Thus, taking

$$R_n = \left\{\frac{4\mathcal{K}^2}{\kappa^4}+1\right\}^{1/2}2^{(j+1)/2}\varepsilon_n$$

yields $\sup_{s\in B_{j,\varepsilon_n}}\|f_s\|_{Q_n} \leq R_n$. This verifies the unnumbered condition in Corollary (8.8) in van de Geer (2000).

With constants K_1, C, C_0 and C_1 chosen as in the proof of Lemma 6, $K_2 = 2\mathcal{K}^2/\kappa^2$ and $K = 4K_1K_2$, it is easy to see that conditions (8.23)–(8.8.27) and (8.29) in van de Geer (2000) will be verified.

Finally, we have to check (8.28) in van de Geer (2000), that amounts to checking the three inequalities $\delta_n \leq C_1 2R_n^2\sigma_0^2/K$, $\delta_n \leq 8\sqrt{2}R_n\sigma_0$, and

$$\sqrt{n}\delta_n \geq C_0\left(\int_{\delta_n/2^6}^{\sqrt{2}R_n\sigma_0} H_B^{1/2}\left(\frac{u}{\sqrt{2}\sigma_0}, \mathcal{F}_{s_0,j,\varepsilon_n}, Q_n\right)\mathrm{d}u \vee \sqrt{2}R_n\sigma_0\right), \tag{A.21}$$

where $\mathcal{F}_{s_0,j,\varepsilon_n} = \{f_s : s \in B_{j,\varepsilon_n}\}$, and $\sigma_0^2 = 2\mathbb{E}_{0,s_0}[\mathcal{W}_i^2 e^{|\mathcal{W}_i|/3}]$.

Both these inequalities follow by taking $\bar{\delta} > 0$ in (A.20) sufficiently small. This can be accomplished by taking $\tilde{c}_0$ and hence c_1 small enough, still obeying the inequality below that equation. Note that for such a small chosen $\tilde{c}_0$ the upper bound of Lemma 7, needed at the beginning of this proof, is still valid.

Now we move to verifying (A.21), the third inequality. This amounts to separately checking that for all n large enough and all $j=0,1,\ldots,M_{\varepsilon_n}$, the inequalities $n\delta_n^2 \geq C_0^2 2R_n^2\sigma_0^2$ and

$$n\delta_n^2 \geq C_0^2\left(\int_{\delta_n/2^6}^{\sqrt{2}R_n\sigma_0} H_B^{1/2}\left(\frac{u}{\sqrt{2}\sigma_0}, \mathcal{F}_{s_0,j,\varepsilon_n}, Q_n\right)\mathrm{d}u\right)^2 \tag{A.22}$$

hold. The first of these two inequalities is again straightforward, because $n\varepsilon_n^2 \to \infty$, so we move to the second one. As a preliminary step, we will show how the bracketing numbers of the set $\mathcal{F}_{s_0,j,\varepsilon_n}$ for the $L_2(Q_n)$-norm can be bounded by the bracketing numbers of the set B_{j,ε_n}. Suppose we have a bracket $[\ell, u]$ for functions s. Then it follows directly from the definition that $[f_u, f_\ell]$ is a bracket for functions f_s. Now we will compare the norms, in order to compare sizes of the brackets: let s_1, s_2 be two dispersion coefficients. Then, using (A.7), the c_2-inequality and applying Lemma 5, we obtain

$$\|f_{s_1} - f_{s_2}\|_{Q_n}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(f_{s_1}(z_i) - f_{s_2}(z_i))^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{s_0^2(z_i)-s_1^2(z_i)+O(n^{-\lambda})}{s_1^2(z_i)} - \frac{s_0^2(z_i)-s_2^2(z_i)+O(n^{-\lambda})}{s_2^2(z_i)}\right\}^2$$

$$\leq 2\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{s_0^2(z_i)-s_1^2(z_i)}{s_1^2(z_i)} - \frac{s_0^2(z_i)-s_2^2(z_i)}{s_2^2(z_i)}\right\}^2 + O(n^{-2\lambda})$$

$$= 2\int_0^1\left\{\frac{s_0^2(u)-s_1^2(u)}{s_1^2(u)} - \frac{s_0^2(u)-s_2^2(u)}{s_2^2(u)}\right\}^2\mathrm{d}u + O(n^{-\lambda})$$

$$= 2 \int_0^1 \frac{s_0^4(u)(s_1^2(u) - s_2^2(u))^2}{s_1^4(u)s_2^4(u)} \, du + O(n^{-\lambda})$$

$$\leq 2 \frac{\mathcal{K}^4}{\kappa^8} \|s_1^2 - s_2^2\|_2^2 + O(n^{-\lambda})$$

$$\leq 8 \frac{\mathcal{K}^6}{\kappa^8} \|s_1 - s_2\|_2^2 + O(n^{-\lambda}).$$

Taking square roots, we get from the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ with $a, b \geq 0$ that

$$\|f_{s_1} - f_{s_2}\|_{Q_n} \leq \sqrt{8} \frac{\mathcal{K}^3}{\kappa^4} \|s_1 - s_2\|_2 + O(n^{-\lambda/2}).$$

Suppose $u \geq \delta_n/2^6$, as in the entropy integral (A.21). Then for all n large enough, the remainder term in the above display satisfies $O(n^{-\lambda/2}) \leq u/2$. For this we need the condition $O(n^{-\lambda/2}) \leq \varepsilon_n^2$, that is,

$$4\beta \leq \lambda \tag{A.23}$$

to be satisfied, which holds under the stipulated assumptions on ε_n. Furthermore, if also

$$\|s_1 - s_2\|_2 \leq \frac{\kappa^4}{\sqrt{32}\mathcal{K}^3} u,$$

we get by the triangle inequality that $\|f_{s_1} - f_{s_2}\|_{Q_n} \leq u$. It follows that

$$H_B\left(\frac{u}{\sqrt{2}\sigma_0}, \mathcal{F}_{s_0, j, \varepsilon_n}, Q_n\right) \leq H_B\left(\frac{\kappa^4 u}{8\sqrt{2}\mathcal{K}^3\sigma_0}, B_{j, \varepsilon_n}, Q_n\right).$$

The entropy on the right-hand side, using $H_B(\delta, \mathcal{S}_n, Q_n) \leq H_\infty(\delta/2, \mathcal{S}_n)$, see Lemma 2.1 in van de Geer (2000), can be further bounded by

$$H_B\left(\frac{\kappa^2 u}{8\sqrt{2}\mathcal{K}^3\sigma_0}, \mathcal{S}_n, Q_n\right) \leq H_\infty\left(\frac{\kappa^2 u}{16\sqrt{2}\mathcal{K}^3\sigma_0}, \mathcal{S}_n\right). \tag{A.24}$$

Now recall that for all $s \in \mathcal{S}_n$, we have $\kappa \leq s \leq \mathcal{K}$. This and the fact that there are N_n bins in total imply that the minimal number of balls with radii v to cover the set $\mathcal{S}_n$ with respect to the supremum norm is bounded by

$$\left(1 + \text{const}\frac{1}{v}\right)^{N_n},$$

for some constant const > 0 independent of n. Hence, the entropy on the right-hand side of (A.24) is bounded by

$$N_n \log\left(1 + \text{const}\frac{1}{u}\right).$$

Now this bound shows that the required condition (A.21) will follow, if we show the inequality

$$\sqrt{n}\delta_n \gg \sqrt{N_n} \int_0^{R_n} \log^{1/2}\left(1 + \text{const}\frac{1}{u}\right) du.$$

The integral in this display (without loss of generality we take the const equal to 1) is of order $R_n \log^{1/2}(1 + 1/R_n)$, which can be seen as follows.

$$\int_0^{R_n} \log^{1/2}\left(1 + \frac{1}{u}\right) du$$

$$= R_n \sqrt{\log\left(1 + \frac{1}{R_n}\right)} + \frac{1}{2} \int_0^{R_n} \frac{1}{\sqrt{\log(1 + 1/u)}} \frac{1}{u+1} du,$$

where the latter integral is upper bounded by $\log(1 + R_n)/\sqrt{\log(1 + 1/R_n)}$. This upper bound is for small R_n of lower order than $R_n \sqrt{\log(1 + 1/R_n)}$. Hence, it is sufficient to show that

$$\sqrt{n}\delta_n \gg \sqrt{N_n} R_n \log^{1/2}\left(1 + \frac{1}{R_n}\right).$$

However, the latter relationship follows from

$$n\varepsilon_n^2 \gg N_n \log(\varepsilon_n^{-1}), \tag{A.25}$$

which is true under our choice of N_n and ε_n as given in Theorem 3.

Since we verified all the conditions in Corollary 8.8 in van de Geer (2000), as in the proof of Lemma A.3 in Gugushvili and Spreij (2016) we can apply it to the right-hand side of (A.19) to obtain

$$\mathbb{P}_{0,s_0}\left(\sup_{s \in A_{j,\varepsilon_n}} \prod_{i=1}^n \frac{p_{i,n,s}(Y_{i,n})}{p_{i,n,0}(Y_{i,n})} \geq \exp(-c_1 n \varepsilon_n^2)\right) \leq \exp(-c_0 n \varepsilon_n^2),$$

where $c_0 > 0$ can be expressed in terms of previous constants, cf. van de Geer (2000) and Gugushvili and Spreij (2016). This inequality and (A.18) yield the statement of the lemma. $\square$

Remark 11. The optimal choice for β in $\varepsilon \asymp n^{-\beta} \log^\gamma n$ in the statement of the lemma results from coupling the conditions (A.23) and (A.25) with our earlier restrictions (A.5), (A.12) and (A.13) on m_n, N_n, ε_n and $\widetilde{\varepsilon}_n$.

Indeed, putting $N_n \asymp n^\alpha$ gives from the coupling of these five conditions, in the above presented order, $4\beta \leq \lambda$, $1 - 2\beta \geq \alpha$ and $\gamma > 1$, $\beta \leq \alpha\lambda$, $\beta \leq \alpha\lambda/2$ and $\beta \leq (1 - \alpha)/2$. The third and the last conditions can be omitted, which leads to the conditions $4\beta \leq \lambda$, $\beta \leq (1 - \alpha)/2$ and $\beta \leq \alpha\lambda/2$ on α and β.

It turns out that the maximal possible β is $\beta = \frac{\lambda}{4}$, this attains the first condition and also satisfies the other two, for $\alpha \in [\frac{1}{2}, 1 - \frac{\lambda}{2}]$.

Acknowledgments

References

Aït-Sahalia, Y. and Jacod, J. (2014). *High-Frequency Financial Econometrics*. Princeton: Princeton University Press.

Allen, E. (2007). *Modeling with Itô Stochastic Differential Equations. Mathematical Modelling: Theory and Applications* **22**. Dordrecht: Springer. MR2292765

Aragon, Y. (2011). *Séries temporelles avec R – méthodes et cas. With a preface by Dominique Haughton. Pratique R*. Paris: Springer. MR3241814 https://doi.org/10.1007/978-2-8178-0208-4

Arjas, E. and Heikkinen, J. (1997). An algorithm for nonparametric Bayesian estimation of a Poisson intensity. *Computational Statistics* **12**, 385–402. MR1477272

Batz, P., Ruttor, A. and Opper, M. (2018). Approximate Bayes learning of stochastic differential equations. *Physical Review E* **98**, 022109. MR3862381 https://doi.org/10.1103/physreve.98.022109

Berger, J. O. and Wolpert, R. L. (1988). *The Likelihood Principle*, 2nd ed. Hayward, CA: Institute of Mathematical Statistics. MR0773665

Bezanson, J., Edelman, A., Karpinski, S. and Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review* **59**, 65–98. MR3605826 https://doi.org/10.1137/141000671

Board of Governors of the Federal Reserve System. Foreign Exchange Rate Series [DEXJPUS] and [DEXUSUK]. Retrieved from Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DEXJPUS and https://fred.stlouisfed.org/series/DEXUSUK, accessed November 2, 2016.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*, 2nd ed. *With 1 CD-ROM (Windows). Springer Texts in Statistics*. New York: Springer. MR1894099 https://doi.org/10.1007/b97391

Castillo, I. and Nickl, R. (2014). On the Bernstein–von Mises phenomenon for nonparametric Bayes procedures. *The Annals of Statistics* **42**, 1941–1969. MR3262473 https://doi.org/10.1214/14-AOS1246

Castillo, I. and Rousseau, J. (2015). A Bernstein–von Mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics* **43**, 2353–2383. MR3405597 https://doi.org/10.1214/15-AOS1336

De Gregorio, A. and Iacus, S. M. (2008). Least squares volatility change point estimation for partially observed diffusion processes. *Communications in Statistics Theory and Methods* **37**, 2342–2357. MR2446669 https://doi.org/10.1080/03610920801919692

Dette, H., Podolskij, M. and Vetter, M. (2006). Estimation of integrated volatility in continuous-time financial models with applications to goodness-of-fit testing. *Scandinavian Journal of Statistics* **33**, 259–278. MR2279642 https://doi.org/10.1111/j.1467-9469.2006.00479.x

Dimitriou-Fakalou, C. (2014). Gaussian pseudo-likelihood estimation for stationary processes on a lattice. *AStA Advances in Statistical Analysis* **98**, 21–34. MR3162969 https://doi.org/10.1007/s10182-013-0207-z

Elerian, O., Chib, S. and Shephard, N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica* **69**, 959–993. MR1839375 https://doi.org/10.1111/1468-0262.00226

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall. MR1383587

Faraway, J. (2016). Confidence bands for smoothness in nonparametric regression. *Stata Journal* **5**, 4–10. MR3478793 https://doi.org/10.1002/sta4.100

Florens-Zmirou, D. (1993). On estimating the diffusion coefficient from discrete observations. *Journal of Applied Probability* **30**, 790–804. MR1242012 https://doi.org/10.2307/3214513

Fuchs, C. (2013). *Inference for Diffusion Processes*. Heidelberg: Springer. MR3015023 https://doi.org/10.1007/978-3-642-25969-2

Gatheral, J. (2006). *The Volatility Surface: A Practitioner's Guide*. Hoboken, New Jersey: Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013). *Bayesian Data Analysis*, 3rd ed. *Chapman & Hall/CRC Texts in Statistical Science*. MR3235677

Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24**, 997–1016. MR3253850 https://doi.org/10.1007/s11222-013-9416-2

Genon-Catalot, V., Laredo, C. and Picard, D. (1992). Nonparametric estimation of the diffusion coefficient by wavelets methods. *Scandinavian Journal of Statistics* **19**, 317–335. MR1211787

Ghosal, S., Ghosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics* **28**, 500–531. MR1790007 https://doi.org/10.1214/aos/1016218228

Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *The Annals of Statistics* **35**, 192–223. MR2332274 https://doi.org/10.1214/009053606000001172

Giné, E. and Nickl, R. (2011). Rates of contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *The Annals of Statistics* **39**, 2883–2911. MR3012395 https://doi.org/10.1214/11-AOS924

Gobet, E., Hoffmann, M. and Reiß, M. (2004). Nonparametric estimation of scalar diffusions based on low frequency data. *The Annals of Statistics* **32**, 2223–2253. MR2102509 https://doi.org/10.1214/009053604000000797

Gugushvili, S. and Spreij, P. (2014a). Non-parametric Bayesian drift estimation for stochastic differential equations. *Lithuanian Mathematical Journal* **54**, 127–141. MR3212631 https://doi.org/10.1007/s10986-014-9232-1

Gugushvili, S. and Spreij, P. (2014b). Non-parametric Bayesian estimation of a dispersion coefficient of the stochastic differential equation. *ESAIM Probabilités Et Statistique* **18**, 332–341. MR3333993 https://doi.org/10.1051/ps/2013039

Gugushvili, S. and Spreij, P. (2016). Posterior contraction rate for non-parametric Bayesian estimation of the dispersion coefficient of a stochastic differential equation. *ESAIM Probabilités Et Statistique* **20**, 143–153. MR3528621 https://doi.org/10.1051/ps/2016008

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, NJ: Princeton University Press. MR1278033

Hamrick, J., Huang, Y., Kardaras, C. and Taqqu, M. S. (2011). Maximum penalized quasi-likelihood estimation of the diffusion function. *Quantitative Finance* **11**, 1675–1684. MR2850995 https://doi.org/10.1080/14697688.2011.615212

Hamrick, J. and Taqqu, M. S. (2009). Testing diffusion processes for non-stationarity. *Mathematical Methods of Operational Research* **69**, 509–551. MR2507762 https://doi.org/10.1007/s00186-008-0250-9

Heikkinen, J. and Arjas, E. (1998). Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics* **25**, 435–450. MR1650015 https://doi.org/10.1111/1467-9469.00114

Hoffmann, M. (1997). Minimax estimation of the diffusion coefficient through irregular samplings. *Statistics & Probability Letters* **32**, 11–24. MR1439493 https://doi.org/10.1016/S0167-7152(96)00052-1

Hoffmann, M. (1999a). Adaptive estimation in diffusion processes. *Stochastic Processes and Their Applications* **79**, 135–163. MR1670522 https://doi.org/10.1016/S0304-4149(98)00074-X

Hoffmann, M. (1999b). L_p estimation of the diffusion coefficient. *Bernoulli* **5**, 447–481. MR1693608 https://doi.org/10.2307/3318712

Höpfner, R. (2014). *Asymptotic Statistics. With a View to Stochastic Processes*. Berlin: De Gruyter Graduate. De Gruyter. MR3185373 https://doi.org/10.1515/9783110250282

Hualde, J. and Robinson, P. M. (2011). Gaussian pseudo-maximum likelihood estimation of fractional time series models. *The Annals of Statistics* **39**, 3152–3181. MR3012404 https://doi.org/10.1214/11-AOS931

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **60**, 271–293. MR1616041 https://doi.org/10.1111/1467-9868.00125

Iacus, S. M. (2008). *Simulation and Inference for Stochastic Differential Equations: With R Examples. Springer Series in Statistics*. New York: Springer. MR2410254 https://doi.org/10.1007/978-0-387-75839-8

Iacus, S. M. (2016). sde: Simulation and inference for stochastic differential equations. In *R Package Version 2.0.15*. https://CRAN.R-project.org/package=sde. MR2410254 https://doi.org/10.1007/978-0-387-75839-8

Ignatieva, K. and Platen, E. (2012). Estimating the diffusion coefficient function for a diversified world stock index. *Computational Statistics & Data Analysis* **56**, 1333–1349. MR2892345 https://doi.org/10.1016/j.csda.2011.10.004

Jacod, J. (2000). Non-parametric kernel estimation of the coefficient of a diffusion. *Scandinavian Journal of Statistics* **27**, 83–96. MR1774045 https://doi.org/10.1111/1467-9469.00180

Jacod, J. and Shiryaev, A. N. (2003). *Limit Theorems for Stochastic Processes*, 2nd ed. *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **288**. Berlin: Springer. MR1943877 https://doi.org/10.1007/978-3-662-05265-5

Kanaya, S. and Kristensen, D. (2016). Estimation of stochastic volatility models by nonparametric filtering. *Econometric Theory* **32**, 861–916. MR3530455 https://doi.org/10.1017/S0266466615000079

Karatzas, I. and Shreve, S. E. (1988). *Brownian Motion and Stochastic Calculus. Graduate Texts in Mathematics* **113**. New York: Springer. MR0917065 https://doi.org/10.1007/978-1-4684-0302-2

Kleijn, B. and van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics* **34**, 837–877. MR2283395 https://doi.org/10.1214/009053606000000029

Kristensen, D. (2010). Nonparametric filtering of the realized spot volatility: A kernel-based approach. *Econometric Theory* **26**, 60–93. MR2587103 https://doi.org/10.1017/S0266466609090616

Kutoyants, Yu. A. (2004). *Statistical Inference for Ergodic Diffusion Processes*. London: Springer. MR2144185 https://doi.org/10.1007/978-1-4471-3866-2

Lutz, B. (2010). *Pricing of Derivatives on Mean-Reverting Assets. Lecture Notes in Economics and Mathematical Systems* **630**. Berlin: Springer. MR2554106 https://doi.org/10.1007/978-3-642-02909-7

Mai, H. (2014). Efficient maximum likelihood estimation for Lévy-driven Ornstein–Uhlenbeck processes. *Bernoulli* **20**, 919–957. MR3178522 https://doi.org/10.3150/13-BEJ510

Malliavin, P. and Mancino, M. E. (2009). A Fourier transform method for nonparametric estimation of multivariate volatility. *The Annals of Statistics* **37**, 1983–2010. MR2533477 https://doi.org/10.1214/08-AOS633

Mishura, Y. (2015). The rate of convergence of option prices on the asset following a geometric Ornstein–Uhlenbeck process. *Lithuanian Mathematical Journal* **55**, 134–149. MR3323287 https://doi.org/10.1007/s10986-015-9270-3

Musiela, M. and Rutkowski, M. (2005). *Martingale Methods in Financial Modelling*, 2nd ed. *Stochastic Modelling and Applied Probability* **36**. Berlin: Springer. MR2107822

Nelson, D. B. (1990). ARCH models as diffusion approximations. *Journal of Econometrics* **45**, 7–38. MR1067229 https://doi.org/10.1016/0304-4076(90)90092-8

Nickl, R. and Söhl, J. (2017). Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *The Annals of Statistics* **45**, 1664–1693. MR3670192 https://doi.org/10.1214/16-AOS1504

Nickl, R. and Szabó, B. (2016). A sharp adaptive confidence ball for self-similar functions. *Stochastic Processes and Their Applications* **126**, 3913–3934. MR3565485 https://doi.org/10.1016/j.spa.2016.04.017

Papaspiliopoulos, O., Pokern, Y., Roberts, G. O. and Stuart, A. M. (2012). Nonparametric estimation of diffusions: A differential equations approach. *Biometrika* **99**, 511–531. MR2966767 https://doi.org/10.1093/biomet/ass034

Pfaff, B. (2008). *Analysis of Integrated and Cointegrated Time Series with R*, 2nd ed. New York: Springer. MR2450313 https://doi.org/10.1007/978-0-387-75967-8

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika* **88**, 603–621. MR1859397 https://doi.org/10.1093/biomet/88.3.603

Rohatgi, A. (2015). WebPlotDigitizer, Version 3.9. Available at. http://arohatgi.info/WebPlotDigitizer.

Sabel, T., Schmidt-Hieber, J. and Munk, A. (2015). Spot volatility estimation for high-frequency data: Adaptive estimation in practice. In *Modeling and Stochastic Learning for Forecasting in High Dimension* (A. Antoniadis, J.-M. Poggi and X. Brossat, eds.), *Lecture Notes in Statistics*, 213–241. Berlin: Springer. MR3588111

Scricciolo, C. (2003). Asymptotics for Bayesian histograms. Working Paper Series, 13/2003, Padova. http://paduaresearch.cab.unipd.it/7305.

Scricciolo, C. (2004). *Asymptotic Issues for Bayesian Histograms. Atti della XLII Riunione Scientifica della SIS*. Padova: CLEUP. http://hdl.handle.net/11565/40874.

Scricciolo, C. (2007). On rates of convergence for Bayesian density estimation. *Scandinavian Journal of Statistics* **34**, 626–642. MR2368802 https://doi.org/10.1111/j.1467-9469.2006.00540.x

Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics* **29**, 687–714. MR1865337 https://doi.org/10.1214/aos/1009210686

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. London: Chapman & Hall. MR0848134 https://doi.org/10.1007/978-1-4899-3324-9

Skorohod, A. V. (1964). *Sluchaĭnye protsessy s nezavisimymi prirashcheniyami. Random Processes with Independent Increments*. Moscow: Izdat. "Nauka". (Russian). MR0182056

Soulier, P. (1998). Nonparametric estimation of the diffusion coefficient of a diffusion process. *Stochastic Analysis and Applications* **16**, 185–200. MR1603904 https://doi.org/10.1080/07362999808809525

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **64**, 583–639. MR1979380 https://doi.org/10.1111/1467-9868.00353

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B, Statistical Methodology* **76**, 485–493. MR3210727 https://doi.org/10.1111/rssb.12062

Szabó, B., van der Vaart, A. W. and van Zanten, H. (2015a). Honest Bayesian confidence sets for the L_2-norm. *Journal of Statistical Planning and Inference* **166**, 36–51. MR3390132 https://doi.org/10.1016/j.jspi.2014.06.005

Szabó, B., van der Vaart, A. W. and van Zanten, J. H. (2015b). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *The Annals of Statistics* **43**, 1391–1428. MR3357861 https://doi.org/10.1214/14-AOS1270

Taleb, N. (1997). *Dynamic Hedging: Managing Vanilla and Exotic Options*. New York: Wiley.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. New York: Springer. MR2724359 https://doi.org/10.1007/b13794

van de Geer, S. A. (2000). *Applications of Empirical Process Theory. Cambridge Series in Statistical and Probabilistic Mathematics* **6**. Cambridge: Cambridge University Press. MR1739079

van der Meulen, F. and Schauer, M. (2017). Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals. *Electronic Journal of Statistics* **11**, 2358–2396. MR3656495 https://doi.org/10.1214/17-EJS1290

van der Meulen, F., Schauer, M. and van Zanten, H. (2014). Reversible jump MCMC for nonparametric drift estimation for diffusion processes. *Computational Statistics & Data Analysis* **71**, 615–632. MR3131993 https://doi.org/10.1016/j.csda.2013.03.002

van der Meulen, F. H. and van Zanten, J. H. (2013). Consistent nonparametric Bayesian inference for discretely observed scalar diffusions. *Bernoulli* **19**, 44–63. MR3019485 https://doi.org/10.3150/11-BEJ385

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing*. London: Chapman & Hall. MR1319818 https://doi.org/10.1007/978-1-4899-4493-1

Wang, Y. (2012). Model selection. In *Handbook of Computational Statistics. Springer Handbooks of Computational Statistics* (J. Gentle, W. Härdle and Y. Mori, eds.) 469–497. Berlin, Heidelberg: Springer. MR2985408 https://doi.org/10.1007/978-3-642-21551-3_16

Wasserman, L. (2006). *All of Nonparametric Statistics. Springer Texts in Statistics*. New York: Springer. MR2172729

Williams, D. (1991). *Probability with Martingales*. Cambridge: Cambridge University Press. MR1155402 https://doi.org/10.1017/CBO9780511813658

Wong, E. and Hajek, B. (1985). *Stochastic Processes in Engineering Systems. Springer Texts in Electrical Engineering*. New York: Springer. MR0787046 https://doi.org/10.1007/978-1-4612-5060-9

S. Gugushvili
Biometris
Wageningen University & Research
Postbus 16
6700 AA Wageningen
The Netherlands
E-mail: shota@yesdatasolutions.com
URL: https://gugushvili.github.io

M. Schauer
Mathematical Institute
Leiden University
P.O. Box 9512
2300 RA Leiden
The Netherlands
E-mail: m.r.schauer@math.leidenuniv.nl
URL: https://pub.math.leidenuniv.nl/%7eschauermr/

F. van der Meulen
Delft Institute of Applied Mathematics
Faculty of Electrical Engineering
    Mathematics and Computer Science
Delft University of Technology
Mekelweg 4
2628 CD Delft
The Netherlands
E-mail: f.h.vandermeulen@tudelft.nl
URL: http://diamweb.ewi.tudelft.nl/%7emeulen/

P. Spreij
Korteweg-de Vries Institute for Mathematics
University of Amsterdam
P.O. Box 94248
1090 GE Amsterdam
The Netherlands
E-mail: spreij@uva.nl
URL: https://staff.fnwi.uva.nl/p.j.c.spreij/