Bootstrapping least-squares estimates in biochemical reaction networks

Daniel F. Linder\textsuperscript{a} and Grzegorz A. Rempala\textsuperscript{b,}\textsuperscript{*}

\textsuperscript{a}Department of Biostatistics, Georgia Southern University, Statesboro, GA 30458, USA;
\textsuperscript{b} Division of Biostatistics and Mathematical Biosciences Institute, The Ohio State University, Columbus, OH 43210, USA

(Received 27 June 2014; accepted 18 March 2015)

The paper proposes new computational methods of computing confidence bounds for the least-squares estimates (LSEs) of rate constants in mass action biochemical reaction network and stochastic epidemic models. Such LSEs are obtained by fitting the set of deterministic ordinary differential equations (ODEs), corresponding to the large-volume limit of a reaction network, to network’s partially observed trajectory treated as a continuous-time, pure jump Markov process. In the large-volume limit the LSEs are asymptotically Gaussian, but their limiting covariance structure is complicated since it is described by a set of nonlinear ODEs which are often ill-conditioned and numerically unstable. The current paper considers two bootstrap Monte-Carlo procedures, based on the diffusion and linear noise approximations for pure jump processes, which allow one to avoid solving the limiting covariance ODEs. The results are illustrated with both in-silico and real data examples from the LINE 1 gene retrotranscription model and compared with those obtained using other methods.

Keywords: reaction network; density-dependent Markov jump process; diffusion approximation; network reverse engineering; least-squares estimation; bootstrap monte-carlo method

AMS Classification: 62F; 60J28; 92C40; 92C45

1. Introduction

The rapid advances in modern molecular quantification technology, such as fluorescence microscopy, cellular flow cytometry and high-throughput RNA-seq (see, e.g. [15,26,27]), provide us with an unprecedented opportunity to discover the basic operating principles of molecular systems biology. However, progress seems to largely depend on our ability to bring to bear proper statistical methods capable of separating the biological signal of interest from the vast amount of experimental noise. In order to be effective, such methods must be tailored to the specific experimental tools and data collection mechanisms which, despite continuing improvements, still often provide incomplete and highly noisy information.
In this paper, we are concerned with statistical methods for biochemical reaction rates estimation from the data consisting of one or more time course trajectories of (partial) network species concentrations measured on a pre-defined time grid via, for example, RT-PCR or flow cytometry assays [4,21]. Although this particular problem has been discussed extensively in the computational biology and statistical literature ([9,12] and references therein), its asymptotic analysis was often dismissed due to a small number of longitudinal data points available from typical laboratory experiments. However, as pointed out recently in [11] in many practical situations, it is reasonable to assume that the size of biological system (e.g. cell culture) under consideration is sufficiently large to allow for its large-volume approximation, the technique which is often used in modern population dynamics models, like, for example, epidemic spread studies [1]. The volume limit idea should be particularly appealing to molecular biology experimentalists as it circumvents the problem of small longitudinal sample sizes due to instrumental limitations and cost considerations. The same approach is also applicable (although this is not the focus of current paper) to stochastic epidemic models [1].

Using the large-volume argument Rempala [16] has shown that the popular method of parameter estimation based on the least-squares criterion yields consistent and asymptotically efficient results when fitting the deterministic system of ordinary differential equations (ODEs) to the stochastic data arriving from large collections of molecules (e.g. in the RT-PCR experiments) under the law of mass action kinetics. However, as also pointed out in [16], the practical obstacle to large-volume approximation for the least-square estimates (LSEs) is in the fact that their joined asymptotic covariance structure is given very non-explicitly by a solution to a set of nonlinear differential equations. Consequently, due to often numerically ill-posed inverse problems, it is quite difficult to obtain the asymptotic confidence bounds for the LSEs, even in relatively simple models.

In the current paper, we aim at improving the practical applicability of the results discussed in [16]. First, we consider a slightly more general version of the LSE, the so-called penalized LSE or PLSE. Whereas in many modern statistical analysis problems the penalized methods are used to improve identifiability and decrease variability in the model predictions, in our case the biggest potential gain in their usage is the stabilization of the variance matrix (for the fixed volume) in a way analogous to the ridge model in classical regression analysis [10]. Furthermore, based on the consistency of PLSEs, we derive and illustrate with examples two large-volume parametric bootstrap (Monte-Carlo) procedures which allow for consistent estimation of the covariance structure, possibly under limited and only partially observable trajectory data. In order to improve efficiency in re-computing model trajectories, we propose the usage of diffusion and closely related linear noise approximations (LNAs) [12] with the bootstrap algorithm. We show that the consistency of our procedures follows from the classical large-volume limit theorem for the density-dependent jump process (see, e.g. [7, Chapter 10]). Finally, in order to apply our methods to practical experimental data, we also propose a procedure based on the profile likelihood bootstrap algorithm which allows for a more realistic selection of the effective volume size of the modelled system. In essence, we treat the volume size as an estimable parameter of the model and fit it via the maximal likelihood method.

Beyond this introductory section, the paper is organized as follows. In Section 2, we introduce the concept of a stochastic biochemical network under mass action dynamics, along with the necessary notation. In Section 3, we discuss the least-squares estimation problem for mass action biochemical networks and prove basic consistency and normality results for penalized least-squares estimates (PLSEs). These PLSE results are analogous to those in [16] but form the foundation for our latter analysis and are provided for completeness. As they are consequences of a general law of large numbers and a central limit theorem (CLT) for density-dependent stochastic processes, we provide brief review of the relevant facts about such processes in Appendix 1. In Section 4, our main results on the bootstrap algorithms for estimating LSE distribution are
discussed, along with the formulation of the required large-volume consistency and diffusion approximation results. The proofs are deferred to Appendix 2. Finally, in Section 5, we illustrate some of the concepts discussed earlier with the analysis of the LINE 1 retrotransposition model for both synthetic and real data, comparing the latter analysis to the one presented in [16] relying on simple asymptotic LSE bounds. Our conclusions and a brief summary are given in Section 6.

2. Jump processes and mass action kinetics

A reaction system or reaction network is typically defined as a set of constituent chemical species reacting according to a finite set of equations, each describing the system state change after the occurrence of a particular reaction. For instance, the chemical reaction

$$A + B \xrightarrow{\lambda} C,$$

is interpreted as ‘a molecule of species A combines with a molecule of species B to create a molecule of species C at rate $\lambda$’. In this paper, the reaction system is treated as a continuous-time Markov chain, with its state described by a $d$-dimensional vector $X(t)$ keeping track of each species molecular counts at time $t$. The $k$th reaction is determined by the reaction rate function $\lambda_k(x)$ along with two vectors, $\nu_k$ and $\nu'_k$ which give, respectively, the number of molecules of each species consumed and produced in the reaction. Both $\nu_k$ and $\nu'_k$ are the elements of $\mathbb{Z}^d_{\geq 0}$ and the corresponding reaction rate is a non-negative scalar. Under the assumptions that the system is well-stirred and thermally equilibrated, the reaction rates take on the classical mass action form and the waiting times between reactions are exponentially distributed (see, e.g. [8]). Intuitively, the mass action form of a $k$th reaction rate function represents the number of different ways we can choose molecular substrates needed for the reaction $k$ to occur [7, Chapter 10]. Defining $|\nu_k| = \sum_i \nu_{ik}$, the $k$th reaction rate function has the form

$$\lambda_k^{(n)}(x) = n \kappa_k \prod_i \nu_{ik}^{!n_{ik}} (\frac{\nu_{ik}}{n})^{\nu_{ik}}.$$

Here, the parameter $n$ denotes the effective volume of the system. Whereas in many classical chemistry applications $n$ is simply the physical volume of a container (e.g. cell) multiplied by Avagadro’s number, in biological systems, where often the physical units of the parameters are not readily available, it is convenient to consider $n$ simply as a scaling parameter proportionate to the overall system volume [2,3].

The exponential waiting time distribution, along with the probability of a particular reaction justifies the system time-evolution equation

$$X(t) = X(0) + \sum_k Y_k \left( \int_0^t \lambda_k^{(n)}(X(s)) \ ds \right) (\nu'_k - \nu_k).$$

(1)

Denoting by $c = n^{-1}x$ the species concentration per unit volume and observing the behaviour of the mass action rate functions in the large-volume limit ($n \to \infty$) one obtains the approximation

$$\lambda_k^{(n)}(x) \approx n \kappa_k \prod_j c_j^{\nu_{jk}} =: n\lambda_k(c).$$

(2)

When $\lambda_k^{(n)}(x)$ is replaced with $n\lambda_k(x/n)$ in Equation (1), the process $X(t)$ becomes a member of a general family of the so-called density-dependent Markov jump processes (DDMJP) described in the appendix. As shown in the appendix (cf. Theorem A.1) the approximation (2) ties together the
stochastic jump process with its large-volume limit which turns out to be given by the solution of an integral version of the classical law of mass action ODE (the so-called reaction rate equation, cf., e.g. [24]),

\[ \dot{c}(t) = \sum_k \kappa_k \prod_i c_i^{\nu_k}(t)(v_k^i - v_k). \] (3)

The above ODE is parametrized by a vector of parameters, say \( \beta \), involving the rate constants \( \kappa_k \) as well as possibly the initial conditions \( c(0) \). In what follows, we write \( \hat{c}^\beta(t) \) for any such solution to either Equation (3) or to its integral version (see also Appendix 1).

3. Penalized least squares

Let the counts of \( d \) species observed at discrete time points, \( t_i, (i = 1, \ldots, T < \infty) \) be denoted by \( X_i^{(n,\beta)} \in \mathbb{Z}_{\geq 0}^d \) where for now we assume that this observed data arrived from a single trajectory of DDMJP for which the effective system volume \( n \) was known. The extension to multiple trajectories is straightforward, but not considered here except for a very simple case of independent observations (see Corollary 3.5 of Theorem 3.3). The case when \( n \) is unknown is discussed in Section 5. Following Rempala [16], we denote the scaled molecular counts \( C_n(t_i) = C_n^\beta(t_i) = X_i^{(n,\beta)}(t_i)/n \) and assume that \( \beta^0 \in \Theta \subset \mathbb{R}^p \) is the true parameter value we intend to estimate. Throughout the paper, the PLSE is defined as any solution (not necessarily unique) to the optimization problem

\[ \hat{\beta}_n = \arg\min_{\beta \in \Theta} \sum_i |C_n(t_i) - \hat{c}^\beta(t_i)|^2 + 2\gamma_n h(\beta), \] (4)

where \( \gamma_n \geq 0 \), \( h \) is a penalty function and \( \Theta \) is an open set in \( \mathbb{R}^p \) containing \( \beta^0 \). Common forms of \( h(x) \) include \( |x|_2^2 = \sum_j x_j^2 \) for the ridge regression, \( |x|_1 = \sum_j |x_j| \) for the lasso [23], and \( |x|_2^2 + |x|_1 \) for the elastic net [29]. Note that in our setting we will typically have \( \beta \geq 0 \) component-wise, in which case the lasso penalty becomes (right) differentiable as it reduces to \( h(\beta) = \sum_j \beta_j \).

For any matrix \( A \), define \( \|A\| = \sqrt{\text{Tr}AA^\top} \). The following two results extending those in [16] will be useful in the following sections. The first one establishes the strong law of large numbers (SLLN) and the consistency of \( \hat{\beta}_n \) as \( n \to \infty \).

**THEOREM 3.1 SLLN**  Let \( \partial_k c^\beta \) denote the derivative of \( c^\beta \in \mathbb{R}^d \geq 0 \) with respect to the \( k \)-th coordinate of \( \beta \in \mathbb{R}^p \) and set \( \partial c^\beta = [\partial_k c^\beta] \in \mathbb{R}^{p \times d} \). Assume that \( \|\partial c^\beta\| < \infty \) and that the SLLN for DDMJP holds (cf. Theorem A.1 in the appendix), the subderivatives of the penalty function \( h \) are bounded and \( \gamma_n \to 0 \) as \( n \to \infty \). If

\[ \forall \beta_1, \beta_2 \in \Theta, c^{\beta_1}(t_i) = c^{\beta_2}(t_i) \quad \text{for } i = 1, \ldots, T \quad \text{implies} \quad \beta_1 = \beta_2 \] (5)

and for some \( \varepsilon > 0 \)

\[ \exists k \sum_i \partial_k c^\beta(t_i)(c^\beta(t_i) - c^{\beta_0}(t_i)) \neq 0 \quad \text{for } \beta \neq \beta_0, |\beta - \beta_0| < \varepsilon \] (6)

then the penalized LSE \( \hat{\beta}_n \) given by Equation (4) are strongly consistent, that is,

\[ \hat{\beta}_n \to \beta^0 \quad \text{as } n \to \infty \quad \text{a.s.} \]
Proof Let $\theta \in \mathbb{R}^{d}$ denote the vector of subderivatives (the sub-gradient) of $h(\beta)$. By the so-called KKT necessary optimality conditions (see, e.g. [20], Chapter 3), any solution $\hat{\beta}_n$ of the optimization problem (4) has to satisfy $-\theta \sum_i (C_n(t_i) - c^{\hat{\beta}}(t_i)) (C_n(t_i) - c^\beta(t_i)) = 2\gamma_n \theta$ with $C_n(t_i)$ and $c^\beta(t_i) \in \mathbb{R}^{d}$ denoting, respectively, the observations from $d$ chemical species concentrations measured at $t_i$ and the corresponding deterministic solution of Equation (3) evaluated at $t_i$ and $\hat{\beta}$. This yields $\sum_i \partial c^{\hat{\beta}}(t_i)(C_n(t_i) - c^\beta(t_i)) = \gamma_n \theta$. Adding and subtracting the deterministic limit gives $\sum_i \partial c^{\hat{\beta}}(t_i)(C_n(t_i) - c^{\theta_0}(t_i)) + \gamma_n \theta = \sum_i \partial c^{\hat{\beta}}(t_i)(C_n(t_i) - c^{\theta_0}(t_i))$.

Under the SLLN for DDMJP cited in the appendix (cf. Theorem A.1), the condition $\gamma_n \to 0$ and the boundedness of $\partial c^{\beta}(t_i)$ imply the RHS above converges a.s. to 0 and hence so does the LHS. Consequently, $\lim_{n \to \infty} \sum_i \partial c^{\hat{\beta}}(t_i)(c^{\hat{\beta}}(t_i) - c^{\theta_0}(t_i)) = 0$. Under the assumptions of the trajectories identifiability (5) and non-degeneracy (6), the asymptotic inversion gives $\lim_{n \to \infty} \hat{\beta}_n = \beta^0$ a.s. 

Remark 3.2 Note that the conditions (5) and (6) are necessary to guarantee the LSE consistency. The selection of points $\{t_i\}$ is therefore important to ensure good statistical properties of LSEs, as also seen in our next theorem below. The required boundedness of $h$ subderivatives is, on the other hand, a relatively mild condition and satisfied, e.g. for $h(x) = |x|_1$ and $h(x) = |x|_2$. Indeed, the latter is simply differentiable, and for the former, recall that the subderivative of $|x|$ is $-1, 1$, or any number in $[-1, 1]$, respectively, according to $x < 0, x > 0$ or $x = 0$.

The second result of this section is the CLT for the scaled PLSEs $\sqrt{n}(\hat{\beta} - \beta^0)$. Similarly to the SLLN above, this one is also a consequence of the appropriate result for DDMJP.

Theorem 3.3 CLT Suppose that the assumptions of the previous theorem hold and additionally $\gamma_n = o(n^{-1/2})$ and $h$ is twice differentiable in $\Theta \in \mathbb{R}^d$. Furthermore, assume that the hypothesis of CLT for DDMJP holds (cf. Theorem A.2 in the appendix). Then for any $s \in \mathbb{R}^d$

$$\sup_x |P(\sqrt{n}(s^T \hat{\beta} - s^T \beta^0) \leq x) - P(s^T \mathcal{Z}_n \leq x)| \to 0 \quad \text{as} \quad n \to \infty,$$

where $\mathcal{Z}_n \sim \text{MVN}(0, \Sigma_{\hat{\beta}})$ is a multivariate zero-centred Gaussian distribution with the covariance matrix

$$\Sigma_{\hat{\beta}} = \text{Var} \left( \sum_i (B_{\hat{\beta}} + \gamma_n H(\hat{\beta}))^{-1} \partial c^{\hat{\beta}}(t_i) \sqrt{n} (C_n(t_i) - c^{\theta_0}(t_i)) \right)$$

$$= \sum_i (B_{\hat{\beta}} + \gamma_n H(\hat{\beta}))^{-1} \partial c^{\hat{\beta}}(t_i) \text{Var}(V_i)[(B_{\hat{\beta}} + \gamma_n H(\hat{\beta}))^{-1} \partial c^{\hat{\beta}}(t_i)]^T$$

$$+ \sum_{i \neq j} [(B_{\hat{\beta}} + \gamma_n H(\hat{\beta}))^{-1} \partial c^{\hat{\beta}}(t_i) \text{Cov}(V_i, V_j)[(B_{\hat{\beta}} + \gamma_n H(\hat{\beta}))^{-1} \partial c^{\hat{\beta}}(t_i)]^T$$

$$+ \sum_{i \neq j} (B_{\hat{\beta}} + \gamma_n H(\hat{\beta}))^{-1} \partial c^{\hat{\beta}}(t_i) \text{Cov}(V_i, V_j)[(B_{\hat{\beta}} + \gamma_n H(\hat{\beta}))^{-1} \partial c^{\hat{\beta}}(t_i)]^T].$$

(8)

Here, $B_{\hat{\beta}} = \sum_i \partial c^{\hat{\beta}}(t_i)[\partial c^{\theta_0}(t_i)]^T$, $H$ is the Hessian matrix of the penalty function $h$ and $\text{Cov}(V_i, V_j) = \text{Cov}(V(t_i), V(t_j))$ is given by Equation (23) of Theorem A.2 in the appendix.
Proof In the notation of the previous proof, under the differentiability assumptions on the penalty function \( h \), in Equation (7) we may take \( \theta = \nabla h \), the gradient vector of the penalty. Expanding the LHS of Equation (7) about \( \hat{\beta}^0 \) gives

\[
\sum_i \partial \hat{\beta}(t_i)(C_n(t_i) - c^{\beta^0}(t_i)) + \gamma_n \nabla h(\hat{\beta})
\]

\[
= \sum_i \partial \hat{\beta}(t_i)(c^{\hat{\beta}}(t_i) - c^{\beta^0}(t_i)) + \gamma_n H(\beta^0)(\hat{\beta} - \beta^0) + \gamma_n \nabla h(\beta^0) + O_p(|\hat{\beta} - \beta^0|)(\hat{\beta} - \beta^0)
\]

\[
= \sum_i \partial \hat{\beta}(t_i)[\partial c^{\beta}(t_i)]^\top(\hat{\beta} - \beta^0) + \gamma_n H(\beta^0)(\hat{\beta} - \beta^0) + O_p(|\hat{\beta} - \beta^0|)(\hat{\beta} - \beta^0) + \gamma_n \nabla h(\beta^0)
\]

\[
= (B_{\hat{\beta}} + \gamma_n H(\beta^0))(\hat{\beta} - \beta^0) + O_p(|\hat{\beta} - \beta^0|)(\hat{\beta} - \beta^0) + \gamma_n \nabla h(\beta^0).
\]

In view of the theorem assumptions, the higher order terms scaled by \( \sqrt{n} \) tend to 0, and therefore

\[
\sqrt{n}(\hat{\beta} - \beta^0) = (B_{\hat{\beta}} + \gamma_n H(\beta^0))^{-1} \sum_i \partial c^{\hat{\beta}}(t_i) \sqrt{n}(C_n(t_i) - c^{\beta^0}(t_i))
\]

\[
- \gamma_n \sqrt{n}(B_{\hat{\beta}} + \gamma_n H(\beta^0))^{-1} \nabla h(\beta^0) + o_p(1)
\]

\[
= (B_{\hat{\beta}} + \gamma_n H(\beta^0))^{-1} \sum_i \partial c^{\hat{\beta}}(t_i) \sqrt{n}(C_n(t_i) - c^{\beta^0}(t_i)) + o_p(1), \tag{9}
\]

since the condition \( \gamma_n = o(n^{-1/2}) \) implies that the term in the second line above converges to 0. From Theorem A.2 in the appendix, it follows that the \( d \)-vectors \( \sqrt{n}(C_n(t_i) - c^{\beta^0}(t_i)) \) \((i = 1, \ldots, T)\) converge weakly to the \( d \)-variate normal variables whose covariance structure is given by \( \text{Cov}(V_i, V_j) \) described in Equation (A5). Therefore, the relation (9), along with the continuous mapping and Slutsky’s theorem arguments (see, e.g. [22], p. 133), implies that \( \sqrt{n}(\hat{\beta} - \beta^0) \Rightarrow \text{MVN}(0, \Sigma_{\beta^0}) \) where \( \Rightarrow \) denotes the usual weak convergence of random vectors. Since \( \text{MVN}(0, \Sigma_{n,\beta}) \Rightarrow \text{MVN}(0, \Sigma_{\beta^0}) \) by Slutsky’s theorem, the hypothesis follows via the Cramer–Wold device (see, e.g. [22], p. 56). \[\blacksquare\]

Remark 3.4 Note that the relation (9) indicates that for finite \( n \) the penalty term may decrease the variability of the LSE, by stabilizing the inverse of \( B_{\hat{\beta}} \) matrix. Indeed, for instance for \( h(x) = |x|_2 \) this effect is similar to that observed in the ridge regression model. This is further illustrated in the numerical example in Section 5 (cf. Table 1).

In Theorems 3.1 and 3.3, we have assumed that data come from a single trajectory of the jump process. However, the results are obviously immediately extendable to the case of multiple independent trajectories. This special case of independent observations is of interest, for instance, in RT-PCR experiments with multiple cell-culture dishes and therefore deserves a special attention.

Corollary 3.5 Assume that the trajectory data \( X^{(n,\beta)}_i \in \mathbb{Z}_{\geq 0} \) \((i = 1, \ldots, T < \infty)\) are mutually independent, then \( \text{Cov}(Z_i, Z_j) = 0 \) for \( i \neq j \) and, assuming that remaining conditions in Theorem 3.3 hold true, its hypothesis also holds with

\[
\Sigma_{\beta} = \sum_i (B_{\hat{\beta}} + \gamma_n H(\hat{\beta}))^{-1} \partial c^{\hat{\beta}}(t_i) \text{Var}(V_i)[(B_{\hat{\beta}} + \gamma_n H(\hat{\beta}))^{-1} \partial c^{\hat{\beta}}(t_i)]^\top.
\]

As already indicated in Section 1, the complicated, non-explicit covariance structure of the collection of random variables \( \sqrt{n}(\hat{\beta})(C_n(t_i) - c^{\beta^0}(t_i)) \) \((i = 1, \ldots, T)\) limits the usefulness of
Table 1. Penalized vs. ordinary LSEs for L1.

| Coefficient | LSE | PLSE |
|-------------|-----|------|
|             | Mean | Variance | Mean | Variance |
| $\beta_1 = 5$ | 5.84 | 55.54 | 3.95 | 5.51 |
| $\beta_2 = 1$ | 1.82 | 47.27 | 0.96 | 10.05 |
| $\beta_3 = 0.1$ | 0.10 | 0.0115 | 0.13 | 0.0049 |
| $\beta_1 = 5$ | 5.16 | 4.09 | 4.74 | 2.24 |
| $\beta_2 = 1$ | 1.02 | 4.18 | 0.94 | 3.56 |
| $\beta_3 = 0.1$ | 0.10 | 0.0013 | 0.11 | 0.0010 |
| $\beta_1 = 5$ | 5.04 | 1.18 | 4.95 | 1.01 |
| $\beta_2 = 1$ | 0.99 | 0.67 | 0.97 | 0.66 |
| $\beta_3 = 0.1$ | 0.10 | 0.0005 | 0.10 | 0.0004 |
| $\beta_1 = 5$ | 5.02 | 0.89 | 5.00 | 0.88 |
| $\beta_2 = 1$ | 1.00 | 0.25 | 0.99 | 0.25 |
| $\beta_3 = 0.1$ | 0.10 | 0.0003 | 0.10 | 0.0004 |

The comparison of the lasso-penalized and the ordinary LSE estimates in terms of their mean and variance illustrates the reduction in variation due to the presence of penalty term. For smaller system sizes the advantage of the penalized estimates for variance reduction are especially apparent.

the relation (9) in obtaining asymptotic confidence bounds for the penalized LSE. Although for some simple reaction networks the explicit analytic expression for Equation (8) may sometimes exist, in most cases of practical interest it may be only numerically evaluated (cf. Section 5). As such evaluations become more unstable with increasing system complexity (see, e.g. [25]), the methods of approximating the LSEs asymptotic distribution without the need to explicitly evaluate the formula (8) are naturally of great interest. This issue is discussed next.

4. Bootstrap algorithms

We present here two resampling algorithms for estimating the distribution of the PLSEs discussed in the previous section. As seen below, this may be done either directly, by considering the appropriate resampling distributions themselves, or indirectly, by considering the resampling estimates of the asymptotic covariance in Theorem 3.3.

4.1. Diffusion approximation

For the purpose of bootstrapping trajectories, it is useful to have an easy to simulate approximation to the concentration process $C_n(t) = X(t)/n$ where $X(t)$ is given by Equation (1) (cf. also Equation (A2) in the appendix). To this end, consider the following Itô diffusion process $D_n(t)$ (see, e.g. [2]):

$$D_n(t) = D_n(0) + \sum_k \frac{1}{\sqrt{n}} \int_0^t \sqrt{\lambda_k(D_n(s))} dB_k(s)(v'_k - v_k) + \int_0^t F(D_n(s)) ds,$$  \hspace{1cm} (10)

where $B_k(t)$ are independent standard Brownian motions and

$$F(x) = \sum_k \lambda_k(x)(v'_k - v_k)$$  \hspace{1cm} (11)

is the drift function. Note that under the law of mass action approximation (2), $F(x)$ is asymptotically ($n \to \infty$) equal to the RHS of the reaction rate equation (3). As shown in Lemma A.3
in the appendix, the process defined by Equation (10) approximates \( C_n(t) \) in the sense that both share the same deterministic large-volume limit given by the solution \( c(t) \) of the reaction rate equation (3). Note that there is an obvious computational benefit in considering the process \( D_n \) instead of the pure jump process \( X(t) \) which requires an exact (hence expensive) simulation of the trajectory (1) [8]. For large systems, the simulation of \( D_n \) is typically substantially less computationally intense and may be performed via one of the many standard numerical methods available in the literature [28, Chapter 5].

The considerations above lead to the diffusion bootstrap algorithm (Algorithm 1) for approximating either directly the distribution of the penalized LSE \( \hat{\beta} \) or the covariance matrix \( \Sigma_\hat{\beta} \). The formal justification of Algorithm 1 is given in the following consistency result (Theorem 4.1) stating that as the system volume increases, the bootstrap samples have the asymptotically correct empirical distribution and, consequently, may be used to estimate the covariance structure of the least squares and penalized least-squares estimates with vanishing penalty. The proof is given in Appendix 2.

**Algorithm 1 (Diffusion Bootstrap)**

**Require:** The trajectory data, \( C_n(t_i) \) at time points \( t_1, t_2, \ldots, t_T \), the parameterized ODE \( c^\beta(t) \), the number \( (M) \) of bootstrap samples, and the value of \( \gamma_n \) (chosen possibly by the cross-validation method).

1. Compute the penalized LSE, \( \hat{\beta} = \arg\min_{\beta \in \Theta} \sum_i |C_n(t_i) - c^\beta(t_i)|^2 + 2\gamma_n h(\beta) \).
2. Simulate a trajectory, \( D^*_{n}(t) \), from the diffusion approximation under \( \beta = \hat{\beta} \).
3. Recompute and store the penalized LSE, \( \hat{\beta}^* = \arg\min_{\beta \in \Theta} \sum_i |D^*_{n}(t_i) - c^\beta(t_i)|^2 + \gamma_n h(\beta) \), increment counter.
4. If counter \( \leq M \) then go to step 2. Else go to step 5.
5. Use the \( M \) values of \( \hat{\beta}^* \) to estimate the covariance matrix \( \Sigma_\hat{\beta} \) or directly the distribution of \( \hat{\beta} \).

**Theorem 4.1** Assume \( C_n(t) = X(t)/n \) is a scaled trajectory from the pure jump process (1) with \( \beta = \beta^0 \). Let \( \hat{\beta} = \arg\min_{\beta \in \Theta} \sum_i |C_n(t_i) - c^\beta(t_i)|^2 + \gamma_n h(\beta) \) be associated with the probability space, \( (\Omega, \mathcal{F}, P) \). Let \( D^*_{n}(t) \) denote the diffusion process parametrized by the LSE \( \beta = \hat{\beta} \), and let \( \hat{\beta}^* = \arg\min_{\beta \in \Theta} \sum_i |D^*_{n}(t_i) - c^\beta(t_i)|^2 + \gamma_n h(\beta) \) be the penalized least-squares estimate associated with the conditional probability space, \( (\Omega^*, \mathcal{F}^*, P^*_n) \). Then, under the assumptions of Theorem 3.3,

\[
\forall x \in \mathbb{R}^p \quad \sup_x |P_n(\sqrt{n}s^T(\beta^* - \hat{\beta}) \leq x) - P(\sqrt{n}s^T(\hat{\beta} - \beta^0) \leq x)| \to 0 \quad \text{a.s.} \quad n \to \infty.
\]

Although in the above theorem we have assumed that the penalty function is differentiable, it should be noted that this is not necessary. Provided that the subderivatives of the penalty term are bounded, the term vanishes asymptotically and the theorem hypothesis still holds.

**Remark 4.2** From the proof of the theorem, it is evident that the hypothesis of Theorem 4.1 extends to the case of independent observations of Corollary 3.5.

### 4.2. Linear noise approximation

An alternative approximation of Equation (1), which is often numerically simpler than \( D_n \), may be obtained by directly appealing to the limit theorem for DDMJP, that is, Theorem A.2 in the
Substituting $\ln$ where $\frac{\partial}{\partial t}$ another heuristic derivation cf., e.g., [28], Chapter 8). Since the proof is again deferred to the appendix. For the record, we present the second resampling algorithm below as Algorithm 2.

For the record, we present the second resampling algorithm below as Algorithm 2.

The usage of $L_n(t)$ in place of $D_n(t)$ leads to another resampling algorithm similar to the diffusion bootstrap, but potentially exhibiting better numerical stability due to the natural decoupling of the deterministic and stochastic components of the resamples following from Equation (12). For the record, we present the second resampling algorithm below as Algorithm 2.

Formally, the validity of Algorithm 2 is derived from the following analogue of Theorem 4.1. The proof is again deferred to the appendix.

Formally, the validity of Algorithm 2 is derived from the following analogue of Theorem 4.1. The proof is again deferred to the appendix.
**Algorithm 2 (LNA Bootstrap)**

**Require:** The trajectory data, $C_n(t_i)$ at time points $t_1, t_2, \ldots, t_T$, the parameterized ODE $c^\beta(t)$, the number $(M)$ of bootstrap samples, and the value of $\gamma_n$ (chosen possibly by the cross-validation method).

1. Compute the penalized LSE, $\hat{\beta} = \arg\min_{\beta \in \Theta} \sum_i |C_n(t_i) - c^\beta(t_i)|^2 + \gamma_n h(\beta)$.
2. Simulate a trajectory, $L^*_n(t_i)$, from the linear noise approximation under $\beta = \hat{\beta}$.
3. Recompute and store the penalized LSE, $\hat{\beta}^* = \arg\min_{\beta \in \Theta} \sum_i |L^*_n(t_i) - c^\beta(t_i)|^2 + \gamma_n h(\beta)$, increment counter.
4. If counter $\leq M$ then go to step 2. Else go to step 5.
5. Use the $M$ values of $\hat{\beta}^*$ to estimate the covariance matrix $\Sigma_\beta$ or directly the distribution of $\hat{\beta}$.

**Theorem 4.3** Let $C_n(t)$ and $\hat{\beta}$ be as in Theorem 4.1. Let $L^*_n(t)$ denote the LNA process conditional on $\hat{\beta}$, and let $\hat{\beta}^* = \arg\min_{\beta \in \Theta} \sum_i |L^*_n(t_i) - c^\beta(t_i)|^2 + \gamma_n h(\beta)$ be the penalized least-squares estimate associated with the conditional probability space, $(\Omega^*, \mathcal{F}^*, P_n^*)$. Then, under the assumptions of Theorem 3.3,

$$\forall s \in \mathbb{R}^d \sup_x |P^*_n(\sqrt{n}s^\top (\hat{\beta}^* - \hat{\beta}) \leq x) - P(\sqrt{n}s^\top (\hat{\beta} - \beta^0) \leq x)| \to 0 \quad \text{a.s. } n \to \infty.$$

Note that Remark 4.2 following Theorem 4.1 also applies to Theorem 4.3.

**4.3. Unobserved species and the EM algorithm**

In most practical circumstances, some of the species in the chemical model will be unobservable due to technical or other experimental limitations, like, for instance, the inability to simultaneously record genomic and proteomic data [16]. Let us briefly outline how the inference problem may be handled in such cases. Following the set-up of Section 3, denote by $x'_j$ the limiting $(n \to \infty)$ scaled concentration of chemical species $j$ at time point $t_i$ (assumed as observed) and let $x' = (x'_1, \ldots, x'_j)$ and $x = (x^1, \ldots, x'^r) \in \mathbb{R}^{dT}$. Define in a similar fashion vector $y \in \mathbb{R}^{d-rT}$ corresponding to the unobserved species, so the full species vector is then $[x, y] = ([x^\top, y^\top]^\top$. Under either LNA or the diffusion approximation assumption, the joint limited likelihood of the full data is multivariate normal with its mean $c^\beta = [c^\beta_x, c^\beta_y] \in \mathbb{R}^{dT}$ being the vector of the ODE solutions to Equation (3) for all $d$ species and its precision matrix

$$\Sigma^{-1}(\beta) = \begin{bmatrix} \tilde{\Sigma}_{11}(\beta) & \tilde{\Sigma}_{12}(\beta) \\ \tilde{\Sigma}_{21}(\beta) & \tilde{\Sigma}_{22}(\beta) \end{bmatrix},$$

where the partitioning patterns reflects the covariance structure between species vectors $x$ and $y$. In this notation, the expression for the limiting log-likelihood function $\ell(x, y|\beta)$ of the full data satisfies

$$\ell(x, y|\beta) \propto -\log |\Sigma(\beta)| - ([x, y] - c^\beta)^\top \Sigma^{-1}(\beta) ([x, y] - c^\beta)$$

$$= -\log |\Sigma(\beta)| - (x - c^\beta_x)^\top \tilde{\Sigma}_{11}(\beta) (x - c^\beta_x) - (y - c^\beta_y)^\top \tilde{\Sigma}_{22}(\beta) (y - c^\beta_y)$$

$$- (x - c^\beta_x)^\top \tilde{\Sigma}_{12}(\beta) (y - c^\beta_y) - (y - c^\beta_y)^\top \tilde{\Sigma}_{21}(\beta) (x - c^\beta_x).$$

With the help of the above, the asymptotic inference on $\beta$ may be performed via the usual EM algorithm [5]. In particular, it follows that the algorithm’s E-step has a closed form expression
Journal of Biological Dynamics

135

(see, e.g. [14] Chapter 2) and may be maximized numerically as soon as the Gaussian covariance structure is estimated (e.g. via Algorithm 1 or 2).

5. Example: L1 system

As an example, consider a retrotranscription model of LINE1 or L1 gene introduced and studied earlier in [18]. The system equations given below describe the first stage of the retrotranscription process in which the L1 transcript ($X_1$) production is activated by aggregated amounts of transcription facilitating proteins ($X_2$) that interact with the regulatory region of the gene. The set of chemical equations in the system with effective size $n$ is

$$X_2 \xrightarrow{\kappa_1} X_1 + X_2 \quad X_1 \xrightarrow{\kappa_2} \phi,$$

$$\phi \xrightarrow{nk_3} X_2 \quad X_2 \xrightarrow{\kappa_4} \phi. \quad (15)$$

In the above system, the so-called equal dispersal condition (EDC) is often assumed, where $\kappa_2 = \kappa_4$. Under EDC the re-parameterized mass action ODE solution $c^\beta$ in Equation (13) has the closed form

$$c_1^\beta(t) = \beta_1 \beta_3 [1 + \beta_5 (1 + t \beta_2) e^{-\beta_2 t}],$$

$$c_2^\beta(t) = \beta_3 (1 + \beta_5 e^{-\beta_2 t}), \quad (16)$$

where

$$\beta_1 = \frac{\kappa_1}{\kappa_2}, \quad \beta_2 = \kappa_2, \quad \beta_3 = \frac{\kappa_3}{\kappa_4}, \quad \beta_4 = \beta_2,$$

and $\beta_5$ gives the fold change from the initial concentrations [16]. Note that in this parametrization the regularity conditions (5) and (6) are satisfied for L1 system and that the diffusion approximation (10) specializes to

$$D_{n1}(t) = D_{n1}(0) + \frac{1}{\sqrt{n}} \left( \int_0^t \sqrt{\beta_1 \beta_2 D_{n2}(s)} \, dB_1(s) - \int_0^t \sqrt{\beta_2 D_{n1}(s)} \, dB_2(s) \right)$$

$$+ \beta_1 \beta_2 \int_0^t D_{n2}(s) \, ds - \beta_2 \int_0^t D_{n1}(s) \, ds,$$

$$D_{n2}(t) = D_{n2}(0) + \frac{1}{\sqrt{n}} \left( \sqrt{\beta_3 \beta_4 B_3(t)} - \int_0^t \sqrt{\beta_2 D_{n2}(s)} \, dB_4(s) \right) + \beta_3 \beta_2 t - \beta_2 \int_0^t D_{n2}(s) \, ds, \quad (17)$$

where $B_1, B_2, B_3, B_4$ are independent standard Brownian motions. The LNA for L1 may be derived analogously.

5.1. Asymptotic LSE distribution for L1

In order to illustrate the bootstrap methods from Section 4, we compare the bootstrap estimates of the LSE distributions to their true asymptotic ones in the L1 model under EDC. The mass action ODE for L1 is given by Equation (16). Upon normalizing to the initial concentrations, $\beta_5 = 1$ and the drift function and its gradient are

$$F(x_1, x_2) = (\beta_1 \beta_2 x_2 - \beta_2 x_1, \beta_2 \beta_3 - \beta_2 x_2),$$
\[ \frac{\partial F(x_1, x_2)}{\partial F(x_1, x_2)} = \begin{bmatrix} -\beta_2 & \beta_1 \beta_2 \\ 0 & -\beta_2 \end{bmatrix}. \]

The matrices \( \Phi \) and \( G \) in the corresponding CLT covariance calculations (Theorem A.2) are, respectively,

\[ \Phi(t, s) = \begin{bmatrix} e^{\beta_2(s-t)} & -\beta_1 \beta_2 (s-t) e^{\beta_2(s-t)} \\ 0 & e^{\beta_2(s-t)} \end{bmatrix} \]

and

\[ G(x_1(s), x_2(s)) = \begin{bmatrix} \beta_1 \beta_2 x_2(s) + \beta_2 x_1(s) & 0 \\ 0 & \beta_2 \beta_3 + \beta_2 x_2(s) \end{bmatrix}. \]

Recall that the asymptotic covariance formula for the components \( Z_i \) of the scaled limiting process in Theorem 3.2 is

\[ \text{Cov}(Z_i, Z_j) = \int_{t_i \wedge t_j}^{t_i \wedge t_j} \Phi(t_i, s) G(c^\beta(s)) (\Phi(t_j, s))^\top ds \]

which may be numerically evaluated along with the remaining terms in the asymptotic variance \( \Sigma_{\hat{\beta}} \) expression (8). For comparison, we generate the L1 system trajectories with \( \beta^0 = (\beta_1, \beta_2, \beta_3) = (5, 1, 1) \), \( \beta_4 = \beta_2 \) and \( \beta_5 = 1 \) and assume the process to be observed at the time points from 0.03 to 1.00 in the alternating increments of 0.03 and 0.04. For example, the first 5 time points are 0.03, 0.07, 0.10, 0.13, 0.17. The resulting asymptotic covariance matrix for the LSE (with no penalty, \( \gamma_n = 0 \)) obtained with the help of numerical integration is shown in Figure 1 along with its bootstrap estimates for varying \( n \). For the bootstrap, the diffusion approximation with corresponding \( \hat{\beta}^0 \) was applied using the Euler–Maruyama method with step size, \( dt = 0.01 \). The number of bootstrap samples was taken as \( M = 10,000 \), for each of the system sizes \( n = 10^2, 10^3, 10^4 \). To additionally graphically illustrate the recovery of the asymptotic covariance structure by the bootstrap algorithms, we show in Figure 2 the theoretical (upper panel) and bootstrap approximated (lower panel) bivariate marginal distributions of the LSE of \( \beta^0 \), denoted \( \hat{\beta} \).

Figures 1 and 2 indicate that the empirical distribution of the bootstrapped LSEs closely approximates that of the theoretical distribution illustrating in practice the application of our Algorithms 1 and 2. Note that obtaining the bootstrap samples only requires the ability to simulate from the diffusion process.

\[ \Sigma_{\beta^0} = \begin{pmatrix} 27.62 & -3.48 & -0.46 \\ -3.48 & 41.96 & 0.12 \\ -0.46 & 0.12 & 0.01 \end{pmatrix}, \quad \hat{\Sigma}_{\beta^0}^{n=100} = \begin{pmatrix} 30.41 & -4.27 & -0.48 \\ -4.27 & 41.32 & 0.10 \\ -0.48 & 0.10 & 0.01 \end{pmatrix} \]

\[ \hat{\Sigma}_{\beta^0}^{n=1,000} = \begin{pmatrix} 27.26 & -4.27 & -0.46 \\ -4.27 & 42.80 & 0.13 \\ -0.46 & 0.13 & 0.01 \end{pmatrix}, \quad \hat{\Sigma}_{\beta^0}^{n=10,000} = \begin{pmatrix} 25.93 & -3.65 & -0.43 \\ -3.65 & 42.20 & 0.12 \\ -0.43 & 0.12 & 0.01 \end{pmatrix} \]

Figure 1. Bootstrap estimate of the LSE covariance for the L1 system. True asymptotic LSE covariance matrix vs. diffusion bootstrap covariance matrices for L1 systems with varying \( n \) (via Algorithm 1). To facilitate comparisons, the penalty was removed, i.e. \( \gamma_n = 0 \) for all cases considered.
5.2. Penalized LSE for L1

In most practical cases, the data-generating process cannot be measured with arbitrary precision, due to various data collection limitations or errors arising during data processing. In general, such processing errors (see, e.g. [19]) add variation to the LSEs and may compound the problems with formula (8) which in practice needs to be evaluated using estimated values of the model parameters in a fixed volume and hence is sensitive to the LSEs lack of precision. The penalty term $\gamma_n h(\beta)$ in Equation (4) is intended to improve the LSE estimates precision by giving up some small amount of accuracy (recovering it in the limit, due to $\gamma_n \to 0$). Since, as already indicated in Section 3, the penalty term also contributes to stabilizing the inverse matrix in Equation (8), the use of penalized formulation (4) for LSEs with fixed sample size may bring considerable benefits.

To illustrate this point, we compare now the bootstrap estimates for the lasso-penalized and ordinary ($\gamma_n = 0$) LSE in the L1 model with various system sizes. The values of the parameters $\beta^0$ are as above. Since the lasso penalty is right-differentiable in our setting, as $\beta \geq 0$ coordinate-wise, all the forms of penalty discussed in Section 3 are smooth and hence roughly equivalent (see, [10]). To account for the observational errors, we also introduce an independent mean 0 Gaussian noise at each observation grid time point. The variance of this noise is set to 5% of the maximum concentration of mRNA L1 transcripts ($X_1$) which is equivalent to approximately 25% of the maximum concentration of the transcription facilitating proteins ($X_2$). The choice of penalty rate is $\gamma_n = n^{-3/4}$ for $n = 10, 10^2, 10^3$, and $10^4$, so as to satisfy the LSE consistency requirements, although in a real data analysis one would typically choose it by some form of a
cross-validation procedure (cf. [6]). The results of the analysis for $M = 10,000$ resamples are summarized in Table 1. As seen from the table, for the system sizes of $n = 10$ and $n = 100$ the penalized LSE enjoys, respectively, a 10-fold and 2-fold decrease in variance as compared with the ordinary LSE of $\beta_1$, with only modest increase in bias. For the penalized estimate of $\beta_2$, there is nearly a 5-fold decrease in variance for $n = 10$ but only a moderate decrease for $n \geq 100$, while the variability of the $\beta_3$ estimates remaining roughly the same for the LSE and penalized LSE for $n \geq 100$. From the inspection of table entries it should be clear, that with the noisy data the penalized LSE does better in terms of the variance for small system sizes, whereas for larger $n$ the performance of ordinary and penalized LSE is very similar.

5.3. RT-PCR Data for L1

In order to look at a more realistic biological example, in addition to the above synthetic data, we also consider the real time-PCR data from one of the time course experiments analysing L1 system in HeLa cell cultures [13] under external chemical stress, as described in [17]. The data consist of the L1 transcripts values ($X_1$) only, as the protein ($X_2$) is unobservable. The L1 transcript measurements from the experiment are presented in Table 2 taken from Rempala [16] where the LSE-based asymptotic confidence bound for this particular data set was obtained by approximating the overall system size and using the asymptotic distribution of the LSE. The system size approximation in [16] was based on a relatively crude fluorescence conversion method introduced in [17]. We shall illustrate now how the bootstrap approach may be used to help identify most likely system size, without using a (possibly unreliable) method based on fluorescent signal conversion. Note that, unlike in previous examples, the time-course L1 transcript measurements reported in Table 2 were obtained from genetically identical but spatially separated L1 cell cultures (different tissue culture plates) and hence represent not a single one, but rather multiple trajectories of the stochastic system (15), each obtained at different time point. Nonetheless, due to the genetic homogeneity of tissue cultures, the initial conditions for the observed trajectories and the respective experimental system sizes $n$ may be assumed to be approximately equal and Theorems 4.1 and 4.3 (and Algorithms 1 and 2) are still applicable in view of Remark 4.2.

As in our earlier analysis, we took $\beta_4 = \beta_2$ and $\beta_5 = 1$ reducing Equation (16) to the following one-dimensional model:

$$c_1^{\hat{\beta}}(t) = \beta_1 \beta_3 [1 + (1 + t \beta_2) e^{-\beta_2 t}].$$

(18)

Since with the concentration $c_2^{\hat{\beta}}$ unobserved this model is no longer identifiable, we also fix $\beta_3 = 1$ and hence reduce the initial L1 model to only two parameters ($\beta_1, \beta_2$). Application of the LNA diffusion approximation and the EM algorithm to the limiting LSE problem, as discussed in Section 4.3 (with no penalty, i.e. $\gamma_n = 0$), gave the following values of the estimates, which are very similar to those obtained in [16]

$$\hat{\beta}_1 = 6.18, \quad \hat{\beta}_2 = 0.027.$$ 

Based on these numerical values, the bootstrap algorithm (Algorithm 2) was applied to calculate the confidence bounds for the observed longitudinal trajectory of the L1 transcripts. The
Figure 3. Bootstrap likelihood profiling. The LNA bootstrap ($M = 5000$) plots approximating for several different $n$ values the residuals distributions for the model (18) fitted to data in Table 2 and comparing them with the actual observed residuals (grey vertical lines). As seen from the plots, around the value $n = 5000$ the mode of the residuals distribution best aligns with the observed value. Note that the alignment is very poor for $n = 1000$, which is the estimated volume used in [16].

bounds depend on the unknown system size $n$, which was estimated via the method of likelihood profiling as follows. First, for the grid of $n$ values (from $n = 500$ to 9000 with 500 step size), the bootstrap approximation based on $M = 5000$ resamples was obtained using Algorithm 2 (note that with Algorithm 2 the same set of bootstrap resamples may be used with different $n$). Next, the actual value of the observed error was compared with the modes of the error distribution for varying $n$. The best mode-to-data alignment (over the gridded values of $n$) was obtained at $n = 5000$ and hence this value was taken as the estimated effective system size. The process is illustrated in Figure 3. In Figure 4, the 95% confidence envelopes of the original analysis in [16] (where the effective system size was taken as $n = 1000$) are compared with the ones obtained from the likelihood profiling ($n = 5000$). As may be seen from the plots the new larger estimate of the system size has the effect of shrinking the uncertainty around the observed trajectory due to generally tighter confidence bounds about observed data points. illustrating the advantage of the bootstrap-based analysis for data in Table 2.

6. Summary and discussion

We proposed here and justified with the appropriate large-volume limit results some new ways of practically estimating the covariance structure and, in general, improving the performance of LSEs in biochemical network models. We discussed three possible approaches based on (a) the penalized least-squares criterion, (b) the fast bootstrap algorithms for covariance estimation and (c) the bootstrap-based profiling method for the determination of effective system size.
With regard to (a), we have shown that under the vanishing penalty on parameters size, the LSEs continue to be consistent in the limit, but may additionally exhibit much better precision in the small-volume settings. Concerning (b), we have argued that the usage of bootstrap Monte-Carlo algorithms may allow to avoid the often difficult calculations of the LSEs asymptotic covariance structure. To improve efficiency, we proposed the usage of diffusion-based bootstrap methods and shown them to be consistent in the large-volume limit. Two types of diffusion approximation were considered: (i) the classical chemical Langevin equation approximation and (ii) the LNA method directly derived from the CLT for jump processes. The latter one, although slightly more biased in our simulation studies, seems particularly attractive as it allows one to generate resamples which are independent of the system size (volume) \( n \). With regard to (c), this aspect of the LNA-based bootstrap may be used to develop an efficient algorithm for estimating the effective volume of the biochemical model, in essence, by analysing the size of the model fit fluctuations for different volume sizes. We have illustrated the practical applicability of this approach to reducing model uncertainty for partially observed, longitudinal RT-PCR data from the LINE 1 retrotranscription network.

Overall, our current work seems to support the conclusion that the bootstrap methods, although not without limitations, like, for example, the requirement to be able to simulate from the entire reaction network, seem to offer a viable alternative to a direct asymptotic approximation based on the estimated covariance structure. One hopes that the results presented here will encourage further study of the general application of resampling to biochemical network inference problems, both in its theoretical and applied aspects, in particular in the setting specifically not considered here when both \( T, n \to \infty \).

Acknowledgments

Research supported in part by US NIH grant R01CA-152158 (DFL, GAR) and US NSF grant DMS-1106485 (GAR)

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. See Section 5.3.
This is done out of convenience to eliminate superfluous parameters. Note that any such renormalization possibly effects the units of system volume $n$.

In the particular case, given the large system size $n$, the penalty would have only negligible effect.

References

[1] H. Andersson and T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis*, Vol. 4, Springer, New York, NY, 2000.

[2] D.F. Anderson and T.G. Kurtz, *Continuous Time Markov Chain Models for Chemical Reaction Networks*, in *Design and Analysis of Biomolecular Circuits*, H. Koeppl, G. Setti, M. di Bernardo, and D. Denosmore, eds., Springer, New York, NY, 2011, 3–42.

[3] K. Ball, T.G. Kurtz, L. Popovic, and G. Rempala, *Asymptotic analysis of multiscale approximations to reaction networks*, Ann. Appl. Probab. 16(4) (2006), pp. 1925–1961.

[4] M.W. Chevalier and H. El-Samad, *A data-integrated method for analyzing stochastic biochemical networks*, J. Chem. Phys. 135(21) (2011), 214110.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the em algorithm*, J. R. Statist. Soc. Ser. B (Methodol.) 39(1) (1977), pp. 1–38.

[6] B. Efron and G. Gong, *A leisurely look at the bootstrap, the jackknife, and cross-validation*, Amer. Statist. 37(1) (1983), pp. 36–48.

[7] S.N. Ethier and T.G. Kurtz, *Markov Processes: Characterization and Convergence*, Vol. 282, Wiley, New York, NY, 2009.

[8] D.T. Gillespie, *A rigorous derivation of the chemical master equation*, Phys. A: Statist. Mech. Appl. 188(1) (1992), pp. 404–425.

[9] A. Golightly and D.J. Wilkinson, *Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo*, Interface Focus 1(6) (2011), pp. 807–820.

[10] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*, Vol. 2, Springer, New York, NY, 2009.

[11] M. Komorowski, M.J. Costa, D.A. Rand, and M. P.H. Stumpf, *Sensitivity, robustness, and identifiability in stochastic chemical kinetics models*, Proc. Natl. Acad. Sci. USA 108(21) (2011), pp. 8645–8650.

[12] M. Komorowski, B. Finkenstädt, C.V. Harper, and D.A. Rand, *Bayesian inference of biochemical kinetic parameters using the linear noise approximation*, BMC Bioinformatics 10(1) (2009), p. 343.

[13] J.R. Masters, *Hela cells 50 years on: The good, the bad and the ugly*, Nat. Rev. Cancer 2(4) (2002), pp. 315–319.

[14] G.J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Vol. 382, Wiley-Interscience, New York, NY, 2007.

[15] O.D. Perez, P.O. Krutzik, and G.P. Nolan, *Flow cytometric analysis of kinase signaling cascades*, Methods Mol. Biol. 263(2004), pp. 67–94.

[16] G.A. Rempala, *Least squares estimation in stochastic biochemical networks*, Bull. Math. Biol. 74(8) (2012), pp. 1938–1955.

[17] G.A. Rempala, K.S. Ramos, and T. Kalbfleisch, *A stochastic model of gene transcription: an application to l1 retrotransposition events*, J. Theoret. Biol. 242(1) (2006), pp. 101–116.

[18] G.A. Rempala, K.S. Ramos, T. Kalbfleisch, and I. Teneng, *Validation of a mathematical model of gene transcription in aggregated cellular systems: application to l1 retrotransposition*, J. Comput. Biol. 14(3) (2007), pp. 339–349.

[19] A. Roberts, C. Trapnell, J. Donaghey, J.L. Rinn, and L. Pachter, *Improving RNA-Seq expression estimates by correcting for fragment bias*, Genome Biol. 12(3) (2011), p. R22.

[20] A. Ruszczynski, *Nonlinear Optimization*, Vol. 13, Princeton University Press, Princeton, NJ, 2011.

[21] L. Salwinski and D. Eisenberg, *In silico simulation of biological network dynamics*, Nat. Biotechnol. 22(8) (2004), pp. 1017–1019.

[22] J. Shao, *Mathematical Statistics*, 2nd ed., Springer, New York, NY, 2003.

[23] R. Tibshirani, *Regression shrinkage and selection via the lasso: a retrospective*, J. R. Statist. Soc: Ser. B (Statist. Methodol.) 73(3) (2011), pp. 273–282.

[24] N.G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, Vol. 1, Elsevier North-Holland, Amsterdam, 1992.

[25] M.D. Wang, M.J. Schnitzer, H. Yin, R. Landick, J. Gelles, and S.M. Block, *Force and velocity measured for single molecules of RNA polymerase*, Science 282(5390) (1998), pp. 902–907.

[26] D.A. Wheeler, M. Srinivasan, M. Egholm, Y. Shen, L. Chen, A. McGuire, W. He, Y.-J. Chen, V. Makhijani, G.T. Roth, X. Gomes, K. Tartaro, F. Niazi, C.L. Turcotte, G.P. Irzyk, J.R. Lupski, C. Chilaut, X.-z. Song, Y. Liu, Y. Yuan, L. Nazareth, X. Qin, D.M. Muzny, M. Margulies, G.M. Weinstock, R.A. Gibbs, and J.M. Rothberg, *The complete genome of an individual by massively parallel DNA sequencing*, Nature 452(7189) (2008), pp. 872–876.

[27] D.J. Wilkinson, *Stochastic modelling for quantitative description of heterogeneous biological systems*, Nat. Rev. Genet. 10(2) (2009), pp. 122–133.

[28] D.J. Wilkinson, *Stochastic Modelling for Systems Biology*, Vol. 44, CRC Press, Boca Raton, FL, 2011.

[29] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. R. Statist. Soc: Ser. B (Statist. Methodol.) 67(2) (2005), pp. 301–320.
Appendix 1. Density-dependent Markov jump process

In the following, suppose the state \( \{X_n(t) : t \geq 0\} \) is a continuous-time Markov process on the positive orthant of \( d \)-dimensional lattice, \( \mathbb{Z}_d^+ \), with jump intensities \( \lambda_k^{(n)} = n\lambda_k(n^{-1}x) \) for \( x \in \mathbb{Z}_d^+ \), \( v_k' - v_k \in \mathbb{Z}_d^+ \) and the transition probabilities \( P(X_n(t + h) = x + v_k' - v_k | X_n(t) = x) = h\lambda_k(n^{-1}x) + o(h) \) and \( P(X_n(t + h) = x | X_n(t) = x) = 1 - \lim_{n \to \infty} \sum_k \lambda_k(n^{-1}x) + o(h) \). It is assumed that there are a finite number of vectors, \( v_k' - v_k \in \mathbb{Z}_d^+ \) such that \( \sup_k \lambda_k(x) > 0 \) and that \( \lambda_k(x) \) is a continuous function of \( x \) for each \( k \). Such processes are known as the DDMJPs since their transition rates depend on the density of the process (see, e.g., [16]). Note that under the approximation (2) the process given in Equation (1) is a special case of the DDMJP. If \( Y = \{Y(t) : t \geq 0\} \) is a collection of independent unit Poisson processes then

\[
X_n(t) = X_n(0) + \sum_k Y_k \left( n \int_0^t \lambda_k(n^{-1}X_n(s)) \, ds \right) (v_k' - v_k). \tag{A1}
\]

Define the centred Poisson process \( \tilde{Y}_k(nt) = Y_k(nt) - nt \) and \( C_n(t) = X_n(t)/n \). The corresponding drift function is given by Equation (11), that is, \( F(x) = \sum_k \lambda_k(x)(v_k' - v_k) \), where for notational simplicity we no longer emphasize the trajectory or rate dependance on any parameters. Then, Equation (A1) becomes

\[
C_n(t) = C_n(0) + n^{-1} \sum_k \tilde{Y}_k \left( n \int_0^t \lambda_k(C_n(s)) \, ds \right) (v_k' - v_k) + \int_0^t F(C_n(s)) \, ds. \tag{A2}
\]

The following result due to Kurtz [1, Chapter 5] establishes SLLN for \( C_n(t) \).

**Theorem A.1 (SLLN)** Let \( \lim_{n \to \infty} C_n(0) = c_0 \) and suppose that for any compact \( K \subseteq \mathbb{R}^d \) there exists a constant \( M_K \) such that \( |F(x) - F(y)| \leq M_K |x - y|, \forall x, y \in K \). Then

\[
\lim_{n \to \infty} \sup_{t \geq 0} |C_n(t) - c(t)| = 0 \quad \text{a.s.,}
\]

where \( c(t) \) is the deterministic solution to the integral equation,

\[
c(t) = c_0(t) + \int_0^t F(c(s)) \, ds.
\]

For the inferential purposes, it is important to know that the scaled fluctuation of the DDMJP around the deterministic ODE solution, \( V_n = \sqrt{n}(C_n(t) - c(t)) \) is Gaussian. This fact follows from the following CLT due to Kurtz [1, Chapter 5]. Let \( \partial F \in \mathbb{R}^d \times \mathbb{R}^d \) such that \( [\partial F]_{ij} = [\partial F]_{ji} \).

**Theorem A.2 (CLT)** Suppose \( \partial F \) is continuous and \( \lim_{n \to \infty} V_n(0) = v_0 \) is non-random. Then, \( V_n = \sqrt{n}(C_n(t) - c(t)) \Rightarrow V \), where the process \( V \) is given by

\[
V(t) = v_0 + \sum_k B_k \left( \int_0^t \lambda_k(c(s)) \, ds \right) (v_k' - v_k) + \int_0^t \partial F(c(s)) V(s) \, ds, \tag{A3}
\]

or, in a distributionally equivalent Itô diffusion form, by

\[
V(t) = v_0 + \sum_k \int_0^t \sqrt{\lambda_k(c(s))} \, dB_k(s)(v_k' - v_k) + \int_0^t \partial F(c(s)) V(s) \, ds. \tag{A4}
\]

The process \( V \) is a Gaussian vector process with covariance

\[
\text{Cov}(V(t), V(r)) = \int_0^t \Phi(t, s) G(c(s)) \Phi(r, s) \, ds, \tag{A5}
\]

where \( \Phi \in \mathbb{R}^{d \times d} \) satisfies

\[
\frac{\partial}{\partial s} \Phi(t, s) = \Phi(t, s) \partial F(c(s)) \quad \Phi(s, s) = I
\]

and \( G(x) = \sum_{\lambda_k(x)} (v_k' - v_k)(v_k' - v_k)^\top \lambda_k(x) \).
Appendix 2. Diffusion approximation and the bootstrap

Recall the Itô diffusion process $D_n$ defined by Equation (10). The following result indicates that $D_n$ has the same consistency property as the scaled pure jump process (A2) and may be therefore used as its approximation.

**Lemma A.3** Suppose that $\lim_{n \to \infty} D_n(0) \to c(0)$ a constant, and that for each compact set $K \in \mathbb{R}^d$ we have a Lipschitz constant, $M_K$, such that $|F(x) - F(y)| \leq M_K|x - y|$, $\forall x, y \in K$, then

$$ \lim_{n \to \infty} \sup_{s \leq t} |D_n(s) - c(s)| = 0 $$

almost surely, where $c(\cdot)$ is the solution to $c(t) = c(0) \pm \int_0^t F(c(s)) \, ds$ and $F(x) = \sum_k (v_k' - v_k) \lambda_k(x)$, $x \in \mathbb{R}^d$.

**Proof** By the definition of $D_n(t)$ given in Equation (10)

$$ |D_n(s) - c(s)| = \left| D_n(0) - c(0) + \frac{1}{\sqrt{n}} \int_0^s \sqrt{\lambda_k(D_n(r))} \, dB_k(r)(v_k' - v_k) + \left( \int_0^s (F(D_n(r)) - F(c(r))) \, dr \right) \right| $$

$$ \leq |D_n(0) - d(0)| + \frac{1}{\sqrt{n}} \sum_k |v_k' - v_k| \left| \int_0^s \sqrt{\lambda_k(D_n(z))} \, dB_k(z) \right| + \int_0^s M_K |D_n(r) - c(r)| \, dr $$

implying,

$$ |D_n(s) - c(s)| \leq |D_n(0) - c(0)| + \frac{1}{\sqrt{n}} \sum_k |v_k' - v_k| \sqrt{\lambda_k} \sup_{0 \leq r_1 \leq r_2 \leq s} |B_k(r_1) - B_k(r_2)| + \int_0^s M_K |D_n(r) - c(r)| \, dr, $$

where $\tilde{\lambda}_k = \sup_{s \in K} \lambda_k(s)$. For almost all $\omega$, $B_k(\omega, r)$, is a continuous process, and thus achieves its maximum and minimum on every compact set, implying that $\sup_{0 \leq r_1 \leq r_2 \leq s} |B_k(\omega, r_1) - B_k(\omega, r_2)|$ is bounded, except on a set of measure 0. (Note: We have suppressed the inherent sample space notation, $\omega$, except when needed, as is standard for brevity). Then, by Gronwall’s inequality (cf. [1], Chapter 5),

$$ |D_n(s) - c(s)| \leq (|D_n(0) - c(0)| + \frac{1}{\sqrt{n}} \sum_k |v_k' - v_k| \sqrt{\lambda_k} \sup_{0 \leq r_1 \leq r_2 \leq s} |B_k(r_1) - B_k(r_2)|) \, e^{M_K s} $$

yielding

$$ \sup_{s \leq t} |D_n(s) - c(s)| \leq (|D_n(0) - c(0)| + \frac{1}{\sqrt{n}} \sum_k |v_k' - v_k| \sqrt{\lambda_k} \sup_{0 \leq r_1 \leq r_2 \leq s} |B_k(r_1) - B_k(r_2)|) \, e^{M_K t}. $$

As $n \to \infty$, the first term of the RHS converges almost surely to 0 by the theorem’s hypothesis. The second term involves a finite weighted sum of the expressions of type $\sup_{0 \leq r_1 \leq r_2 \leq s} |B_k(\omega, r_1) - B_k(\omega, r_2)|$ which are bounded for almost all $\omega$, implying that $(1/\sqrt{n}) \sup_{0 \leq r_1 \leq r_2 \leq s} |B_k(\omega, r_1) - B_k(\omega, r_2)|$ converges to 0 a.s. $\blacksquare$

The following result shows that the properly centred and scaled diffusion process converges in distribution to the same limit as the centred and scaled pure jump process.

**Lemma A.4** With the assumption $c_n(0) = \sqrt{n}(D_n(0) - c(0)) \to c_0$ a constant, and those of Theorem 2, the process, $V'_n(t) = \sqrt{n}(D_n(t) - c(t))$ converges in distribution to the process, $V'(t)$ which is distributionally equivalent to $V(t)$ defined in Theorem A.2.

**Proof** By definition,

$$ V'_n(t) = \sqrt{n}(D_n(t) - c(t)) = c_n(0) + \sum_k \int_0^t \sqrt{\lambda_k(D_n(s))} \, dB_k(s)(v_k' - v_k) $$

$$ + \int_0^t \varphi(c(s)) V'_n(s) \, ds + \int_0^t O(|D_n(s) - c(s)|) V'_n(s) \, ds. $$
Define, as before, 
\[ U_n(t) := \sum_k \int_0^t \sqrt{\kappa_k(D_n(s))} \, dB_k(s)(v_k' - v_k) + \int_0^t O(|D_n(s) - c(s)|)V_n'(s) \, ds \]
\[ = -c_n(0) + V_n'(t) - \int_0^t \partial F(c(s))V_n'(s) \, ds. \]

Since the process \( D_n(s) \) is non-anticipating and converges in probability pointwise to \( c(s) \) by Lemma A.3, the dominated convergence theorem (DCT) for stochastic integrals implies that \( \lim_{n \to \infty} U_n(t) \) can be computed by interchanging the limit with the first integral, and that the classical DCT can be applied to the second. This gives
\[ U_n(t) \Rightarrow \sum_k \int_0^t \lim_{n \to \infty} \sqrt{\kappa_k(D_n(s))} \, dB_k(s)(v_k' - v_k) + \int_0^t \lim_{n \to \infty} O(|D_n(s) - c(s)|)V_n'(s) \, ds \]
\[ = \sum_k \int_0^t \sqrt{\kappa_k(c(s))} \, dB_k(s)(v_k' - v_k) = U(t). \]

Again, defining \( \Phi \) to be the matrix function solution to
\[ \frac{\partial}{\partial s} \Phi(t, s) = -\Phi(t, s)\partial F(c(s)), \quad \Phi(s, s) = I, \]
we have
\[ \int_0^T \Phi(t, s) \, dU_n(s) = \int_0^T \Phi(t, s) \, dV_n'(s) - \int_0^T \Phi(t, s)\partial F(c(s))V_n'(s) \, ds \]
\[ = \Phi(t, t)V_n'(t) - \Phi(t, 0)c_n(0) - \int_0^T \Phi'(t, s)V_n'(s) \, ds + \int_0^T \Phi(t, s)\partial F(c(s))V_n'(s) \, ds \]
\[ = V_n'(t) - \Phi(t, 0)c_n(0). \]

Thus,
\[ V_n'(t) = \Phi(t, 0)c_n(0) + \int_0^T \Phi(t, s) \, dU_n(s), \]
and \( U_n \Rightarrow U \) implies \( V_n' = h(U_n) \Rightarrow h(U) = \Phi(t, 0)c_0 + \int_0^T \Phi(t, s) \, dU(s) = V' \) again by the continuous mapping theorem. This is clearly a Gaussian vector process by definition, with its covariance function given by
\[ E(V'(t)V'(r)') = E \left[ \int_0^T \Phi(t, s) \, dU(s) \, \left( \int_0^T \Phi(r, s) \, dU(s) \right)' \right] \]
\[ = E \lim_{|s_1 - s_2| \to 0} \left( \sum_i \Phi(t, s_i)(U(s_{i+1}) - U(s_i)) \left( \sum_i \Phi(r, s_i)(U(s_{i+1}) - U(s_i)) \right)' \right) \]
\[ = E \lim_{|s_1 - s_2| \to 0} \sum_i \Phi(t, s_i) \sum_k \sqrt{\kappa_k(c(s_i))}(B_k(s_{i+1}) - B_k(s_i))(v_k' - v_k) \]
\[ \times \left( \sum_i \Phi(r, s_i) \sum_k \sqrt{\kappa_k(c(s_i))}(B_k(s_{i+1}) - B_k(s_i))(v_k' - v_k)' \right) \]
\[ = \lim_{|s_1 - s_2| \to 0} E \left( \sum_{i,j} \Phi(t, s_i) \sum_{k,j} (v_k' - v_k)(v_j' - v_j)' \sqrt{\kappa_k(c(s_i))} \sqrt{\kappa_j(c(s_i))} \right. \]
\[ \times (B_k(s_{i+1}) - B_k(s_i))(B_l(s_{j+1}) - B_l(s_j)) \Phi(r, s) \right)'. \]

Since the Brownian motions \( B_k \) and \( B_l \) are independent, as well as the increments \((B_k(s_{i+1}) - B_k(s_i))\) are independent of \((B_k(s_{j+1}) - B_k(s_j))\), we have
\[ E(V'(t)V'(r)') = \lim_{|s_1 - s_2| \to 0} \sum_i \Phi(t, s_i) \sum_k \lambda(c(s_i))(v_k' - v_k)(v_k' - v_k)' E(B_k(s_{i+1}) - B_k(s_i))^2 \Phi(r, s) \]
and therefore $V^*$ and $V$ have the same distribution.

The following observation regarding the LNA process will be useful at this point.

Remark A.5 In view of the definition of the LNA process $L_n(t)$ given by Equation (12), it is easy to see that the results of Lemmas A.3 and A.4 hold true when $C_n(t)$ is replaced by $L_n(t)$.

With the above lemmas, we are finally in a position to prove the bootstrap consistency result of Theorem 4.1.

Proof of Theorem 4.1

Let $\Rightarrow^*$ indicate that the sequence of probability distributions associated with the conditional (bootstrap) space converges to the distribution of a multivariate normal random vector with covariance $\Sigma_{\hat{\beta}}$. The notation $a.s. (\omega)$ and $a.s. (\omega^*)$ indicates the property holds for almost every $\omega$ and almost every $\omega^*$, respectively. In view of the assumed differentiability of the penalty function $h$, any minimizing solution, $\hat{\beta}^*$, satisfying $\hat{\beta}^* = \arg\min_{\beta \in \Theta} \sum_i |D_n^\beta(t_i) - c \beta(t_i)|^2 + 2\gamma_n h(\beta)$ must also necessarily satisfy $\sum_i \partial c \hat{\beta}^*(t_i)(D_n^\beta(t_i) - c \hat{\beta}(t_i)) + \gamma_n \nabla h(\hat{\beta}^*) = 0$. Expansion about $\hat{\beta}$ gives

$$
\sum_i \partial c \hat{\beta}^*(t_i) \sqrt{\hat{\nabla}}(D_n^\beta(t_i) - c \hat{\beta}(t_i)) = \sum_i \partial c \hat{\beta}^*(t_i) \sqrt{\hat{\nabla}}(\hat{\beta}^* - \hat{\beta}) + \sqrt{\gamma_n} \nabla h(\hat{\beta}^*) + O(|\hat{\beta}^* - \hat{\beta}|)$
$$

where $\hat{H}(\hat{\beta})$ is the Hessian of the penalty term evaluated at $\hat{\beta}$. Lemma A.3 gives $O(|\hat{\beta}^* - \hat{\beta}|) \rightarrow 0$ a.s.($\omega^*$) implying that $O(|\hat{\beta}^* - \hat{\beta}|) \sqrt{\hat{\nabla}}(\hat{\beta}^* - \hat{\beta})$ converges to 0 in view of Lemma A.4, validating the removal in the limit of the higher order terms. Setting $B_{\hat{\beta}^*} = \sum_i \partial c \hat{\beta}^*(t_i) |\partial c \hat{\beta}(t_i)| \sqrt{\hat{\nabla}} + (\gamma_n/2)H(\hat{\beta}^*)$ we arrive at

$$
\sqrt{\hat{\nabla}}(\hat{\beta}^* - \hat{\beta}) = \sum_i B_{\hat{\beta}^*}^{-1} \partial c \hat{\beta}^*(t_i) \sqrt{\hat{\nabla}}(D_n^\beta(t_i) - c \hat{\beta}(t_i)).
$$

By the consistency for the diffusion approximation (Lemma A.3), $\hat{\beta}^* \rightarrow \hat{\beta}$ a.s.($\omega^*$), implying that $\partial c \hat{\beta}^*(t_i) \rightarrow \partial c \hat{\beta}(t_i)$ and $B_{\hat{\beta}^*}^{-1} \rightarrow B_{\hat{\beta}}^{-1}$ a.s.($\omega^*$). Since $\sqrt{\hat{\nabla}}(D_n^\beta(t_i) - c \hat{\beta}(t_i))$ converges to the same distribution as the Gaussian vector process with mean 0 and covariance function

$$
\Sigma_{\hat{\beta}}(t, r) = \int_0^{t \wedge r} \Phi(\hat{\beta}(s), \Phi(\hat{\beta}(r), s)^\top) \, ds,
$$

where

$$
\frac{\partial}{\partial s} \Phi(\hat{\beta}(s), s) = -\Phi(\hat{\beta}(s), s) \partial F(\hat{\beta}(s)), \quad \Phi(\beta, s, s) = I,
$$

we have

$$
\forall s \in \mathbb{R}^p \sup_x |P_n^*(\sqrt{\hat{\nabla}}(\hat{\beta}^* - \hat{\beta})) \leq x) - P(s^\top Z_n \leq x)| \rightarrow 0 \quad a.s.
$$

(A6)

where $Z_n \sim MVN(0, \Sigma_{\hat{\beta}})$ is the zero-centred multivariate normal measure with covariance

$$
\Sigma_{\hat{\beta}} = \text{Var} \left( \sum_i B_{\hat{\beta}}^{-1} \partial c \hat{\beta}(t_i) \sqrt{\hat{\nabla}}(D_n^\beta(t_i) - c \hat{\beta}(t_i)) \right)
$$

$$
= \sum_i B_{\hat{\beta}}^{-1} \partial c \hat{\beta}(t_i) \Sigma_{\hat{\beta}}(t_i, t_i) [B_{\hat{\beta}}^{-1} \partial c \hat{\beta}(t_i)]^\top
$$
\[
+ \sum_{i \neq j}[B^{-1}_{\hat{\beta}} \partial \hat{c}^2(t_i) \Sigma^2_{\hat{\beta}}(t_i, t_j)[B^{-1}_{\hat{\beta}} \partial \hat{c}^2(t_j)]^\top \\
+ \sum_{i \neq j} B^{-1}_{\hat{\beta}} \partial \hat{c}^2(t_j) \Sigma^2_{\hat{\beta}}(t_i, t_j)[B^{-1}_{\hat{\beta}} \partial \hat{c}^2(t_i)]^\top.
\]

By consistency of the penalized LSE for the pure jump process, \(C_n(t), \hat{\beta} \rightarrow \beta^0\) a.s.(\(\omega\)), implying that \(\partial \hat{c} \beta(t_i) \rightarrow \partial c \beta^0(t_i)\) and \(B^{-1}_{\hat{\beta}} \rightarrow B^{-1}_{\beta^0}\) a.s.(\(\omega\)). An appeal to Slutsky’s theorem gives

\[
\text{MVN}(0, \Sigma_{\hat{\beta}}) \Rightarrow \text{MVN}(0, \Sigma_{\beta^0}). \quad (A7)
\]

Asymptotic normality of the penalized LSE for the pure jump process implies

\[
\sqrt{n}(\hat{\beta} - \beta^0) \Rightarrow \text{MVN}(0, \Sigma_{\beta^0}). \quad (A8)
\]

Application of the Cramer–Wold device to Equations (A7) and (A8), along with Equation (A6) imply

\[
\forall s \in \mathbb{R}^p \sup_x |P^*_n(\sqrt{n} s^\top (\hat{\beta}^* - \hat{\beta}) \leq x) - P_n(\sqrt{n} s^\top (\hat{\beta} - \beta^0) \leq x)| \rightarrow 0 \quad \text{a.s.,}
\]

which in turn implies

\[
\sqrt{n}(\hat{\beta}^* - \hat{\beta}) \Rightarrow^* \text{MVN}(0, \Sigma_{\beta^0}).
\]

**Proof of Theorem 4.3**

In view of Remark A.5, proof of Theorem 4.1 may be virtually repeated with \(L_n(t)\) in place of \(C_n(t)\) to obtain the hypothesis.