Multivariate Normal Approximation for the Stochastic Simulation Algorithm: limit theorem and applications

Vincent Picard, Anne Siegel, Jérémie Bourdon

Multivariate Normal Approximation for the Stochastic Simulation Algorithm: limit theorem and applications

Vincent Picard[1]

Université de Rennes 1
IRISA UMR 6074
INRIA Dyliss
Rennes, France

Anne Siegel[2]

CNRS
IRISA UMR 6074
INRIA Dyliss
Rennes, France

Jérémie Bourdon[3]

Université de Nantes
LINA UMR 6241 – Computational Biology (ComBi) group
INRIA Dyliss
Nantes, France

Abstract

Stochastic approaches in systems biology are being used increasingly to model the heterogeneity and the intrinsic stochasticity of living systems, especially at the single-cell level. The *stochastic simulation algorithm* – also known as the *Gillespie algorithm* – is currently the most widely used method to simulate the time course of a system of bio-chemical reactions in a stochastic way.
In this article, we present a *central limit theorem* for the Gillespie stochastic trajectories when the living system has reached a *steady-state*, that is when the internal bio-molecules concentrations are assumed to be at equilibrium. It appears that the stochastic behavior in steady-state is entirely characterized by the stoichiometry matrix of the system and a single vector of reaction probabilities.
We propose several applications of this result such as deriving multivariate confidence regions for the time course of the system and a constraints-based approach which extends the flux balance analysis framework to the stochastic case.

Keywords: Stochastic Simulation, Multivariate statistics, Random walks

[1] Email: vincent.picard@irisa.fr
[2] Email: anne.siegel@irisa.fr
[3] Email: jeremie.bourdon@univ-nantes.fr

This paper is electronically published in
Electronic Notes in Theoretical Computer Science
URL: www.elsevier.com/locate/entcs

1 Introduction

The quantitative analysis of systems of coupled chemical reactions also known as *reaction networks* is a major center of interest in systems biology. Two main mathematical frameworks have been proposed to investigate their kinetic behaviour [12]. On the one hand the *ordinary differential equations* (ODEs) provide deterministic trajectories for the average quantities of molecules at the population level. On the other hand the *chemical master equation* gives a probabilistic description of the trajectories at the single-cell level. In both cases and especially in the latter one the solutions rapidly become non analytical or intractable as the size and the complexity of the networks increase. As a consequence, computational methods have been thoroughly developed during the last decades to provide insight into the dynamics of reaction networks.

In the world of ODEs, one can use numerical analysis algorithms to compute approximate trajectories of the average quantities of molecules. The main issue of this approach is that a perfect knowledge of the system is required to derive an appropriate system of ODEs in the first place. In other words, all the molecular species, reactions, kinetic laws (law of mass actions, Michaelis-Menten, ...) and their parameters must be known. This level of knowledge is far beyond the scope of the current experimental possibilities and so the quantitative analysis of large metabolic networks cannot be obtained from this approach. A successful method, with respect to the aforementioned difficulties, is to consider the particular class of the ODEs solutions where the speeds of the internal molecular concentrations are equal to zero. In that case, the network is in a simpler state referred to as the *steady-state* where the reactions are balanced and the internal metabolite concentrations are constant. This method, named *flux balance analysis* (FBA) [16], is a *constraint-based approach* that takes advantage from the computational advances in *linear programming* to explore the fluxomic capabilities of reaction networks. Its strength is mainly due to the little required knowledge about the network: basically, the only needed information is the *stoichiometry* matrix. Consequently, FBA has lead to numerous applications (*E. coli*, ...) and extensions such as flux coupling analysis.

While ODEs have become widespread in the systems biology community they fail to provide a model at the individual scale that can exhibit stochastic behaviours. Yet, multiple biological examples have been presented to demonstrate that life is inherently stochastic [1,5]. Moreover, the current development of single-cell biology techniques [11] (fluoresence, microscopic imaging, mass spectrometry) provide more and more experimental data at the individual scale. As a consequence, there is currently an important need for probabilistic methods in systems biology [19]. For instance, consider the reaction network on Figure 1 together with a (fictitious) experimental time serie that represents the concentrations of molecules for a *single* cell. A frequent question in systems biology, named *model validation*, is to decide whether the proposed set of reactions are consistent with the observed data or not. At the single-cell level stochastic fluctuations exist, especially when the quantities of molecules are small, so ODEs models cannot help much since they are inherently deterministic and cannot allow one to discriminate between incorrect trajectories and normal stochastic fluctuations.

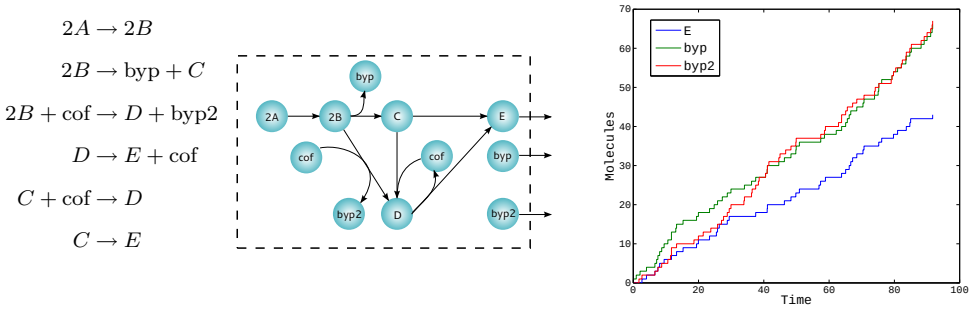


Fig. 1. An example of reaction network (slighty modified from [15]) with experimental data.

The probabilistic counterpart of ODEs is the chemical master equation which was believed to be computationally intractable until D. T. Gillespie popularized a simple yet efficient kinetic Monte Carlo algorithm [7] to simulate exact probabilistic trajectories. The algorithm, referred to as the *stochastic simulation algorithm* (SSA) has become the cornerstone of the probabilistic kinetic methods in systems biology [20], multiple biological applications have been presented [14,1] and improved [6] versions have been proposed in the literature. Hence, the SSA and its variations can be seen as the probabilistic counterpart of the numerical analysis algorithms for ODEs. However, using the SSA leads to the same problems as the ODEs approach: a perfect knowledge of the system is required including probabilistic kinetic parameters. These parameters are even more difficult to infer since numerous individual trajectories are necessary. Another known problem is that the algorithm becomes computationally too expensive when the simulation time is too long or when the number of reaction events explodes. To deal with this issue, approximated versions [2,9] of the SSA have been proposed as well as approximations of the chemical master equation [8,18].

The aim of this work is to demonstrate that the steady-state analysis that has been used to derive FBA from ODEs is also completely within the reach of the probabilistic methods. We derive the consequences of applying the same simplifying assumption (that is, the system as the system has reached a steady-state in which the quantities of internal chemical species are constant) to the SSA. The consequence is the existence of a multivariate central limit theorem (CLT) for the trajectories where the limiting distribution is specified by the stoichiometry matrix and *reaction probabilities* which are the analogous of FBA fluxes. Thus our approach needs as much information as FBA, that is to say mainly the stoichiometry, but is inherently *stochastic*. In the article, we derive the CLT for the stochastic trajectories of a reaction network in steady-state. Then, we present multiple theoretical and practical applications of this result. For instance, we derive *confidence regions* for the aforementioned model validation problem (Figure 1) and we propose a constraints-based approach, similar to FBA, to integrate experimental data.

2 The SSA in Steady-State

We study systems of coupled chemical reactions known as reaction networks. A *reaction network* consists of n molecular species $X_1, \ldots, X_n$ that are involved in m

chemical reactions

$$R_i : \quad a_{i,1}X_1 + \cdots + a_{i,n}X_n \ \rightarrow \ b_{i,1}X_1 + \cdots + b_{i,n}X_n \quad (1 \leq i \leq m). \tag{1}$$

The parameters $a_{i,j}$, $b_{i,j} \in \mathbb{N}$ are the *stoichiometry coefficients* of the reaction network. The number $a_{i,j}$ represents the quantity of X_j molecules consumed by the reaction R_i and the number $b_{i,j}$ represents the quantity of X_j molecules produced by the reaction R_j. The global effect of the reactions on the molecular quantities is often summarized by the *stoichiometry matrix* $S = (s_{i,j})_{1\leq i\leq m,\, 1\leq j\leq n}$ where $s_{i,j} = b_{i,j} - a_{i,j}$. In our notations, each row of the stoichiometry matrix represents the effect of one reaction on the molecular quantities.

In this work we study the *dynamics of reaction networks*, that is, the time evolution of the molecular quantities of a reaction network. We denote by x_t^i the number of X_i molecules at time $t \geq 0$ and $\boldsymbol{x}_t \in \mathbb{R}^n$ the column vector of all quantities at time t. The *trajectory* of the system is the family $(\boldsymbol{x}_t)_{t\in\mathbb{R}}$. Let $\boldsymbol{e}_i = (0, \cdots, 0, 1^{(i)}, 0, \cdots, 0)^\top$ be the $\mathbb{R}^m$-canonic basis vectors and assume that we know the initial state $\boldsymbol{x}_0$ of the system. We consider that the stochastic process $\boldsymbol{x}_t$ is a *jump process*, that is, there exists an increasing sequence of random events $(t_k)_{k\in\mathbb{N}}$ such that $\boldsymbol{x}_t$ is constant on each interval $[t_k, t_{k+1}[$ and $t_0 = 0$. At each time t_k ($k > 0$), one reaction randomly occurs and the value of $\boldsymbol{x}$ changes according to the occurring reaction, thus one can write

$$\forall k \in \mathbb{N}, \quad \boldsymbol{x}_{t_{k+1}} = \boldsymbol{x}_{t_k} + S^\top \boldsymbol{e}_{\mu_{k+1}}, \tag{2}$$

where $(\mu_k)_{k\in\mathbb{N}^*}$ is a random variable representing the index of the reaction that occurs at time t_k and $S^\top$ is the transpose of the stoichiometry matrix. The *inter-reaction times* τ_k are also random variables defined as $\tau_k = t_k - t_{k-1}$ for $k \in \mathbb{N}^*$. Notice that the stochastic process is entirely determined by $\boldsymbol{x}_0$ and the distribution of $(\mu_k, \tau_k)_{k\in\mathbb{N}^*}$. In the article, we mainly focus on the stochastic process $(\boldsymbol{y}_k)_{k\in\mathbb{N}} = (\boldsymbol{x}_{t_k})_{k\in\mathbb{N}}$ referred to as the *embedded process* which represents the succession of changes in the chemical species quantities.

2.1 The Stochastic Simulation Algorithm

The *SSA* [7] is a kinetic Monte Carlo methods that implements a choice for (μ_{k+1}, τ_{k+1}) based only on the current molecular quantities $\boldsymbol{x}_{t_k}$. This assumption is justified when considering homogeneous well stirred systems in thermal equilibrium. The fundamental idea behind the SSA is that the *reaction waiting time* for each *possible combination* of the R_i reactants is independent and randomly exponential with a parameter c_i named the *stochastic kinetic rate*. This fundamental assumption leads to a simulation of correct probabilistic trajectories with regard to the chemical master equation [7]. In the SSA, it is important to notice that the exponential assumption concerns each possible instance of a chemical rule, so the waiting time for any instance of a particular chemical reaction is exponential with parameter $h_i = c_i \times \sharp\{\text{reactant combinations}\}$. This value referred to as the *propensity* of the reaction increases with the available quantities of reactants.

2.2 The Steady-State

Inspired from the success of steady-state analysis for ODEs, that is to say FBA approaches, we aim to investigate steady-state analysis in the context of probabilistic dynamics. First, let us now introduce a formal definition of the *steady-state conditions* for reaction networks that form the initial assumptions of our analysis.

Definition 1 *A reaction network follows the* steady-state conditions *if all τ_k and μ_k are mutually independent, $(\tau_k)_k$ are identically distributed, and $(\mu_k)_k$ are identically distributed.*

If the steady-state conditions hold then the stochastic process $(\boldsymbol{x}_t)$ is entirely determined by the initial state $\boldsymbol{x}_0$, the stoichiometry matrix S and both the distributions of $(\tau_k)_{k\in\mathbb{N}}$ and $(\mu_k)_{k\in\mathbb{N}}$. The *reaction probabilities* column vector $\boldsymbol{p} = (\Pr(\mu_k = i))_{1\leq i\leq m}$ describes the distribution of $(\mu_k)_{k\in\mathbb{N}}$ at steady-state. Without loss of generality, we assume that p_i is positive for every i (otherwise, R_i never occurs and can be removed from the system).

Importantly, the steady-state conditions hold for an execution of the SSA as soon as the propensities are constant.

Proposition 2.1 *In the SSA, if all propensities h_i are constant at each execution of the iteration, then the steady-state condition holds.*

Indeed, according to the SSA when the propensities h_i are constant, $(\mu_k, \tau_k)_{k\in\mathbb{N}^*}$ are sampled independently using a Bernoulli distribution with reaction probabilities $p_i = \frac{h_i}{\sum_{l=1}^m h_l}$ and the inter-reactions times are exponentially distributed with parameter $\sum_{i=1}^m h_i$. To define our framework, we will consider the strong hypothesis that the system has constant propensities. Notice that this excludes numerous oscillating systems such as the Lotka-Volterra dynamics.

2.3 Central Limit Theorem for the Embedded Process

We focus on the *embedded process* $(\boldsymbol{y}_k) = (\boldsymbol{x}_{t_k})$ which describes the succession of changes in the molecular quantities. It is straightforward to establish from (2) that

$$\forall k \in \mathbb{N}, \quad \boldsymbol{y}_k = \boldsymbol{y}_0 + S^\top \left(\sum_{l=1}^{k} \boldsymbol{e}_{\mu_l} \right). \tag{3}$$

The stochastic process $\boldsymbol{q}_k = \sum_{l=1}^k \boldsymbol{e}_{\mu_l}$ counts all the occurrences of each reaction until time t_k, so we refer to $(\boldsymbol{q}_k)$ as the *reaction counting process* (RCP). In this section we demonstrate that, under the steady-state conditions, the embedded process is a random walk that admits a central limit theorem. We proceed in two steps, first we prove this result on the RCP (which corresponds to the case $S = \mathrm{Id}_n$) and then we use an affine transformation to obtain the general result.

We remark that if the steady-state conditions hold then the RCP is a *random walk* [4,10] in $\mathbb{N}^m$, since it has independent and identically distributed increments $\boldsymbol{e}_{\mu_l}$. In other words, q_{k+1} is obtained from q_k by randomly selecting a dimension according to the probabilities $\boldsymbol{p}$ and then moving forward this direction. It is well known that this type of Markovian processes [4,3] admits a central limit theorem (CLT). Formally, the result can be obtained using the classical multivariate

CLT [17].

Proposition 2.2 *Under the steady-state conditions, the RCP* $\boldsymbol{q}_k = \sum_{l=1}^{k} \boldsymbol{e}_{\mu_l}$ *converges to a multivariate Gaussian distribution with covariance matrix* $V(\boldsymbol{p}) = diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top$*:*

$$\frac{1}{\sqrt{k}} \left(\boldsymbol{q}_k - k\boldsymbol{p}\right) \xrightarrow[k\to+\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, V(\boldsymbol{p})\right). \tag{4}$$

Now that we have demonstrated that the RCP converges to a normal distribution, we can notice that by virtue of equation (3), the embedded process $\boldsymbol{y}_k$ is simply an affine transformation of the RCP. Thus, the embedded process is also a random walk where the possible steps are the affine transformation of the canonic basis vectors (see Figure 2 for an example). Due to the stability of normal distributions with regard to affine transformations, the embedded process also tends to a normal distribution and we obtain a multivariate CLT for the embedded process.

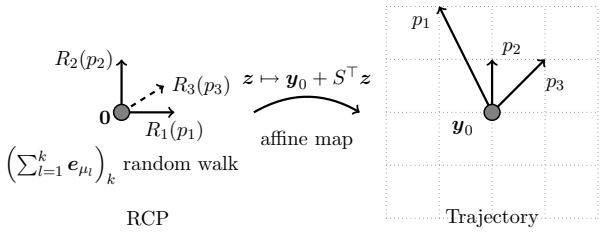


Fig. 2. Illustration of the random walk behaviour on the example system $\{X \to 2Y; \varnothing \to Y; \varnothing \to X + Y\}$.

Proposition 2.3 (Central limit theorem for the embedded process) *Under the steady-state conditions, the the embedded process* $\boldsymbol{y}_k = \boldsymbol{x}_{t_k}$ *converges in distribution to a multivariate Gaussian distribution with covariance matrix* $W(S, \boldsymbol{p}) = S^\top \left(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\right) S$*:*

$$\frac{1}{\sqrt{k}} \left(\boldsymbol{y}_k - \left(\boldsymbol{y}_0 + kS^\top \boldsymbol{p}\right)\right) \xrightarrow[k\to+\infty]{\mathcal{D}} \mathcal{N}\left(\mathbf{0}, W(S, \boldsymbol{p})\right). \tag{5}$$

In other words, the $\boldsymbol{y}_k$ distribution asymptotically tends to a multivariate Gaussian distribution $\mathcal{N}\left(\boldsymbol{y}_0 + kS^\top \boldsymbol{p}, kW(S, \boldsymbol{p})\right)$. Our contribution includes the analytical expressions for the mean and the variance-covariance matrix which depend *only* on S and $\boldsymbol{p}$. In the appendix, we also provide the reader with a complete characterization of the degenerated cases for the limiting distribution depending on the form of the stoichiometry matrix. The rest of the article presents theoretical and practical applications of this asymptotic result.

3 Confidence ellipsoids and accumulation speeds

In this section we introduce other convergence results that are consequences of the CLT. The first result is the introduction of α-confidence ellipsoids that are likely sets of possible values for $\boldsymbol{y}_k$ with asymptotic probability $1 - \alpha$. We present an illustrative application to the model validation problem. The second result is a convergence proposition which demonstrates the relation between the steady-state

reaction probabilities and the practically relevant notion of ratios of accumulation speeds.

3.1 Confidence ellipsoids

The value of a real random variable with known probability distribution is often estimated using confidence intervals. A confidence ellipsoid is a generalization of a confidence interval when the random variable is a Gaussian vector. Let us start by giving a formal definition of *confidence ellipsoids*.

Proposition 3.1 (confidence ellipsoid) *Let S be a $m \times n$ stoichiometry matrix, $\boldsymbol{p}$ a positive probability vector, $\alpha \in]0,1]$ a tolerance error and $\boldsymbol{y}_0$ initial quantities. Consider t_α the unique solution to equation $(2\pi)^{-\frac{n}{2}} \int_{\boldsymbol{x} \in B_n(\mathbf{0},t_\alpha)} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right) d\boldsymbol{x} = 1-\alpha$ where $B_n(\mathbf{0},t_\alpha)$ is the $\mathbb{R}^n$ centred ball of radius t_α and $\|\cdot\|$ is the Euclidean norm. Assume that $\ker W(S,\boldsymbol{p}) = \{\mathbf{0}\}$ and consider $V \in GL_n(\mathbb{R})$ such that $W(S,\boldsymbol{p}) = VV^\top$. Then the subset*

$$\mathcal{E}(S,\boldsymbol{p},\boldsymbol{y}_0,\alpha,k) = \left\{ \boldsymbol{z} \in \mathbb{R}^n \quad | \quad \left\| \frac{1}{\sqrt{k}} V^{-1} \left(\boldsymbol{z} - \boldsymbol{y}_0 - kS^\top \boldsymbol{p} \right) \right\| \leq t_\alpha \right\} \tag{6}$$

*does not depend on the particular choice of V and is a non degenerated $\mathbb{R}^n$-*ellipsoid *called the α-*confidence ellipsoid.

Remark that a well-suited V can be calculated using the Choleski decomposition [13] or the spectral decomposition. The idea behind the definition is to determine the appropriate affine map that transforms $\boldsymbol{y}_k$ into a centred reduced normal random variable. The soundness of the approach is demonstrated by the following proposition.

Proposition 3.2 *The embedded process $\boldsymbol{y}_k$ belongs to the confidence ellipsoid $\mathcal{E}(S,\boldsymbol{p},\boldsymbol{y}_0,\alpha,k)$ with a probability that tends to $1-\alpha$ when k increases*

$$Pr\left(\boldsymbol{y}_k \in \mathcal{E}(S,\boldsymbol{p},\boldsymbol{y}_0,\alpha,k)\right) \underset{k \to +\infty}{\longrightarrow} 1-\alpha. \tag{7}$$

Contrary to ODE or FBA approaches, our method does take fluctuations and inter-species correlations into account: the confidence ellipsoid is centred on the expected value of $\boldsymbol{y}_k$ but its dimensions are calculated using the variance-covariance matrix. Moreover, the advantages of the multivariate approach is now clearly illustrated. Imagine one uses a non multivariate method to calculate the asymptotic α-confidence intervals for each specie $(y_k^a)_{1 \leq a \leq n}$, then he would obtain a multi-dimensional rectangle for the possible values of $\boldsymbol{p}$ which contains the confidence ellipsoid.In other words, the multivariate rectangle has an asymptotic probability larger than $1-\alpha$. Conversely, it is straightforward to derive a confidence interval for y_k^a from the confidence ellipsoid by using $S' = (s_{i,a})_{1 \leq i \leq m}$ (a-th column of S) instead of S, that is to calculate the projection of the ellipsoid on the $(\mathbf{0}, \boldsymbol{e}_a)$ axis.

Illustation on a model validation example. Let us illustrate our results on the example of a metabolic pathway initially proposed in [15]. The system is slightly modified to distinguish the by-products of reactions 2 and 3, leading to the reaction network depicted in Figure 1. We assume for instance that $\boldsymbol{y}_0 = \mathbf{0}$ and $\boldsymbol{p} =$

$(0.3, 0.2, 0.1, 0.2, 0.1, 0.1)^\top$. This value allows us to equilibrate the production and the consumption of metabolites B, C, and D. Focusing on the outputs of the system, we consider the reduced stoichiometry matrix S' where only the output columns $(E, byp, byp2)$ are kept. Now let us consider that we can measure the quantities of E, byp and $byp2$ in 3 different individual cells after $k = 100$ reactions: $\boldsymbol{o}_1 = (40, 15, 5)^\top$, $\boldsymbol{o}_2 = (23, 19, 11)^\top$ and $\boldsymbol{o}_3 = (35, 25, 15)^\top$.

The problem of *model validation* is to decide which of these cells are consistent with the given reaction network and $\boldsymbol{p}$. To address this issue, we calculate the equation of a confidence ellipsoid

$$\mathcal{E}(\alpha, 100): \quad err(\boldsymbol{z}) = \left\| \frac{1}{\sqrt{100}} \begin{pmatrix} 2.1822 & 0 & 0 \\ 0.7559 & 2.6458 & 0 \\ 0.7071 & 0.7071 & 3.5355 \end{pmatrix} \left(\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} - 100 \begin{pmatrix} 0.3 \\ 0.2 \\ 0.1 \end{pmatrix} \right) \right\| \leq 3 \quad (8)$$

corresponding to a tolerance error $\alpha \simeq 2.9\%$. The reader can refer to the appendix for more details about the derivation. Applied to the datasets at hand, we assume that k is large enough so that the asymptotic regime is reached. Computing the quadratic errors for the data at hand $err(\boldsymbol{o}_1) \simeq 2.66$, $err(\boldsymbol{o}_2) \simeq 1.73$ and $err(\boldsymbol{o}_3) \simeq 3.2$ yields that both $\boldsymbol{o}_1$ and $\boldsymbol{o}_2$ belong to the confidence ellipsoid $\mathcal{E}(\alpha, 100)$ whereas $\boldsymbol{o}_3$ does not. We conclude that, with probability $1-\alpha \simeq 97\%$, data about cells 1 and 2 are consistent with the model prediction. On the contrary, data about cell 3 are not consistent with the model and deserve a careful study. Importantly, one may remark that the expected value $E(\boldsymbol{y}_{100}) = (30, 20, 10)^\top$ was easy to compute but provides no relevant information to check the consistency of $\boldsymbol{o}_1$, $\boldsymbol{o}_2$ and $\boldsymbol{o}_3$ alone. Our method is relevant because it makes also use of the variances and covariances.

3.2 Ratios of accumulation speeds

Our second convergence result concerns the *ratios of accumulation speeds* between two output species X_a and X_b, defined as

$$\rho_{a,b}(k) = \frac{(y_k^a - y_0^a)/t_k}{(y_k^b - y_0^b)/t_k} = \frac{y_k^a - y_0^a}{y_k^b - y_0^b} \qquad (k > 0). \quad (9)$$

This quantity is highly interesting for multiple reasons. First, it gives information about the production rates of the system. Second, it is easy to measure experimentally. Indeed, by virtue of Proposition 2.3, the quantities of outputs in steady state are linear in average and $\rho_{a,b}$ is simply the ratio of the slopes corresponding to the average production of X_a and X_b. Moreover, the knowledge of the exact reaction times t_k is not necessary. Third, it is by nature a relative quantity while many biological experiments (western blots, Southern blots and other electrophoresis-based techniques) initially provide *relative* quantitative data between species (however absolute quantitative data can be obtained using a reference chemical specie whose absolute quantity is known). We now introduce a proposition demonstrating that $\rho_{a,b}$ is also theoretically very interesting.

Proposition 3.3 *For all $a, b \in \{1, \ldots, n\}$, if $(S^\top \boldsymbol{p})_b \neq 0$ then the ratio of accumulation speeds $\rho_{a,b}(k)$ between X_a and X_b converges in probability to $\bar{\rho}_{a,b} = (S^\top \boldsymbol{p})_a / (S^\top \boldsymbol{p})_b$: $\forall \varepsilon > 0, \lim_{k \to +\infty} Pr\left(\left|\rho_{a,b}(k) - (S^\top \boldsymbol{p})_a/(S^\top \boldsymbol{p})_b\right| > \varepsilon\right) = 0$.*

The proposition can be viewed as a prediction of $\rho_{a,b}(k)$ since the probability that $\rho_{a,b}(k)$ belongs to any positive-length interval that contains $\bar{\rho}_{a,b}$ tends to 1. In other words, the proposition states that $\rho_{a,b}$ is a *consistent estimator* of a ratio involving the reaction probabilities $\boldsymbol{p}$. Thus, the proposition establishes a relation between the measurable quantity $\rho_{a,b}$ and the parameters of the steady-state.

4 From observations to constraints-based analysis

The previous applications consisted in deriving properties about $\boldsymbol{y}_k$ based on perfect knowledge of $\boldsymbol{p}$. However, in most of the biological applications, the reaction probabilities are unknown and one has to rely on experimental results to infer the model parameters. This is the main motivation of the following applications that consists in deriving *constraints* on the reaction probabilities $\boldsymbol{p}$ from experimental data. As the vector $\boldsymbol{p}$ is only a description of the system dynamics at steady-state, one can only use the experimental data obtained when the system is in steady-state regime. Thus the time $t_0 = 0$ refers to the start of the steady state regime that is when the reactant quantities are assumed to be constant. One advantage of our analysis is that qualitative asymptotic results about $\boldsymbol{y}_k$ have already been established in the previous sections. For instance Proposition 2.3 states that the expectancies, variances and covariances grows linearly with k. What we do not know are the slopes of these linear growths. In this section, we assume that these steady-state slopes can be experimentally measured and we derive constraints on $\boldsymbol{p}$ based on these observations (Table 1).

	Observation	Matricial Constraint	Algebraic Constraints	Type
(1a)	$E(y_k^a) = y_0^a$	$(S^\top \boldsymbol{p})_a = 0$	$\sum_{i=1}^m s_{ia} p_i = 0$	linear
(1b)	$E(y_k^a) \leq y_0^a + k\gamma$	$(S^\top \boldsymbol{p})_a \leq \gamma$	$\sum_{i=1}^m s_{ia} p_i \leq \gamma$	linear
(1c)	$y_0^a + k\gamma \leq E(y_k^a)$	$\gamma \leq (S^\top \boldsymbol{p})_a$	$\gamma \leq \sum_{i=1}^m s_{ia} p_i$	linear
(2a)	$Var(y_k^a) \leq k\gamma$	$\left(S^\top(\mathrm{diag}\boldsymbol{p} - \boldsymbol{p}\boldsymbol{p}^\top)S\right)_{aa} \leq \gamma$	$\sum_{i=1}^m s_{ia}^2 p_i - \sum_{1\leq i,j\leq m} s_{ia}s_{ja}p_i p_j \leq \gamma$	quadratic
(2b)	$k\gamma \leq Var(y_k^a)$	$\gamma \leq \left(S^\top(\mathrm{diag}\boldsymbol{p} - \boldsymbol{p}\boldsymbol{p}^\top)S\right)_{aa}$	$\gamma \leq \sum_{i=1}^m s_{ia}^2 p_i - \sum_{1\leq i,j\leq m} s_{ia}s_{ja}p_i p_j$	quadratic
(3a)	$Cov(y_k^a, y_k^b) \leq k\gamma$	$\left(S^\top(\mathrm{diag}\boldsymbol{p} - \boldsymbol{p}\boldsymbol{p}^\top)S\right)_{ab} \leq \gamma$	$\sum_{i=1}^m s_{ia}s_{ib}p_i - \sum_{1\leq i,j\leq m} s_{ia}s_{jb}p_i p_j \leq \gamma$	quadratic
(3b)	$k\gamma \leq Cov(y_k^a, y_k^b)$	$\gamma \leq \left(S^\top(\mathrm{diag}\boldsymbol{p} - \boldsymbol{p}\boldsymbol{p}^\top)S\right)_{ab}$	$\gamma \leq \sum_{i=1}^m s_{ia}s_{ib}p_i - \sum_{1\leq i,j\leq m} s_{ia}s_{jb}p_i p_j$	quadratic
(4a)	$\lim_k \rho_{a,b}(k) = \gamma$	$(S^\top \boldsymbol{p})_a = \gamma(S^\top \boldsymbol{p})_b$	$\sum_{i=1}^m (s_{ia} - \gamma s_{ib})p_i = 0$	linear
(4b)	$\lim_k \rho_{a,b}(k) \leq \gamma$	$(S^\top \boldsymbol{p})_a \leq \gamma(S^\top \boldsymbol{p})_b$	$\sum_{i=1}^m (s_{ia} - \gamma s_{ib})p_i \leq 0$	linear
(4c)	$\gamma \leq \lim_k \rho_{a,b}(k)$	$(S^\top \boldsymbol{p})_a \geq \gamma(S^\top \boldsymbol{p})_b$	$\sum_{i=1}^m (s_{ia} - \gamma s_{ib})p_i \geq 0$	linear

Table 1
Translation table from biological observations to constraints on the reaction probabilities.

Direct application of Proposition 2.3 lead us to the matrix constraints (1) (2) and (3) of Table 1 while constraints (4) are direct consequences of Proposition 3.3. The algebraic constraints are a rewriting of the matrix constraints as simple algebraic expressions. Remark that the constraint (1a) simply states that the reaction probabilities must be balanced to maintain an average constant quantity of X_a molecules. Notice that γ in observations (1bc) and (2abcd) may be difficult to measure on realistic experimental time series since one needs to know the exact reaction times (t_k). However, this is not the case of observations (1a) and (4abc). The linear

constraints $\sum_{i=1}^{m} p_i = 1$ and $0 \leq p_i \leq 1$ $(i = 1 \dots m)$ should also be added to the set of constraints.

A toy example. We propose a toy example (Figure 3) to illustrate our constraints-based approach. The system contains one input reaction that produces a metabolite A. The metabolite is then transformed into other metabolites (B, C, D, E) that are used to produce four different outputs (O_1, O_2, O_3, O_4). In the model, the input reaction has no reactant meaning that the input metabolites quantities are assumed to be constant. We assume the following assumptions on the system : (H_0) a steady-state has been reached where the quantities of internal species (A, B, C, D, E) remain constant and the outputs O_1, O_2, O_3 and O_4 are accumulating, (H_1) the variability in the accumulation of O_1 is bounded : $Var(y_k^{O_1}) \leq k \cdot 0.2$, (H_2) the covariance between accumulations of O_1 and O_2 satisfies $Cov(y_k^{O_1}, y_k^{O_2}) \leq -0.01k$, (H_3) the speed of accumulation of O_3 is more than half the speed of accumulation of O_2. The proposed numerical values (0.2, -0.01 and $1/2$) are purely arbitrary and are chosen for illustration purposes.

$R1: B \to O_1$
$R2: C \to O_2$
$R3: D \to O_3 + E$
$R4: E \to O_4$
$R5: A \to B$
$R6: A \to C$
$R7: A \to D$
$R8: \varnothing \to A$

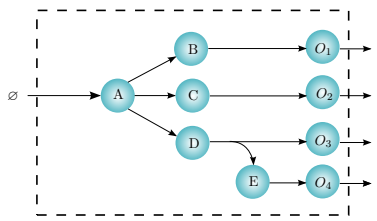


Fig. 3. A toy example to illustrate the constraints-based approach.

The hypothesis H_0 led us to the following set of constraints for the steady-state reaction probabilities (left column). These constraints are linearly independent so the system has 2 degrees of freedom and we focus on the possible values of (p_1, p_2) keeping in mind that the other components of $\boldsymbol{p}$ can be calculated using the following system (right column).

$$\begin{aligned} &\forall i, 0 \leq p_i \leq 1, \\ &\sum_{i=1}^{8} p_i = 1 \\ p_8 &= p_5 + p_6 + p_7 \ (A \text{ in steady-state}) \\ p_5 &= p_1 \ (B \text{ in steady-state}) \\ p_6 &= p_2 \ (C \text{ in steady-state}) \\ p_7 &= p_3 + p_4 \ (D \text{ in steady-state}) \\ p_3 &= p_4 \ (E \text{ in steady-state}) \end{aligned} \qquad \begin{aligned} p_3 &= 1/6 - p_1/2 - p_2/2 \\ p_4 &= 1/6 - p_1/2 - p_2/2 \\ p_5 &= p_1 \\ p_6 &= p_2 \\ p_7 &= 1/3 - p_1 - p_2 \\ p_8 &= 1/3 \end{aligned}$$

Hence in most of the cases, if one has m reactions and n' balanced metabolites, the number of degrees of freedom will be $m - n' - 1$. Notice that the (H_0) hypothesis also includes that the reaction probabilities belong to $[0, 1]$, so we must only consider the values of (p_1, p_2) such that the above expressions for $p_i (i = 3, \dots, 8)$ are in the correct range, that is (H_0) : $p_1 + p_2 \leq 1/3$. Now we translate the hypothesis into constraints using the translation table: (H_1) : $p_1 - p_1^2 \leq 0.2$, (H_2) : $-p_1 p_2 \leq -0.01$ and (H_3) : $p_3 \geq p_2/2 \Leftrightarrow p_2 \leq 1/6 - p_1/2$. The possible values of (p_1, p_2) subjected to these constraints are depicted on Figure 4. The association of the four constraints gives rise to a small set $\mathcal{S}$ of possible values for (p_1, p_2), thus we obtain a good idea bout the steady-state parameters.

This constraints-based approach allows a derivation of new insights on the model. Indeed, we have derived a set $\mathcal{S}$ of possible values for (p_1, p_2). Then, by virtue of Proposition 3.3 we know that the ratios of accumulation speeds between O_1 and O_2 will converge to p_1/p_2. In Figure 4 we represent the extremal values of p_1/p_2 for two sets of constraints: the first does not include (H_3) while the second one does. Under hypothesis (H_0), (H_1) and (H_2) alone the possible values for p_1/p_2 belong to the interval $[0.11, 7.6]$. If the hypothesis H_3 – concerning the accumulating speeds between O_2 and O_3 – is added, then the area of the solution set is clearly reduced and the possible values for p_1/p_2 is restricted to $[0.6, 6.5]$. Thus, we have derived from the hypothesis a range of possible values for the ratio of accumulating speeds between O_1 and O_2.

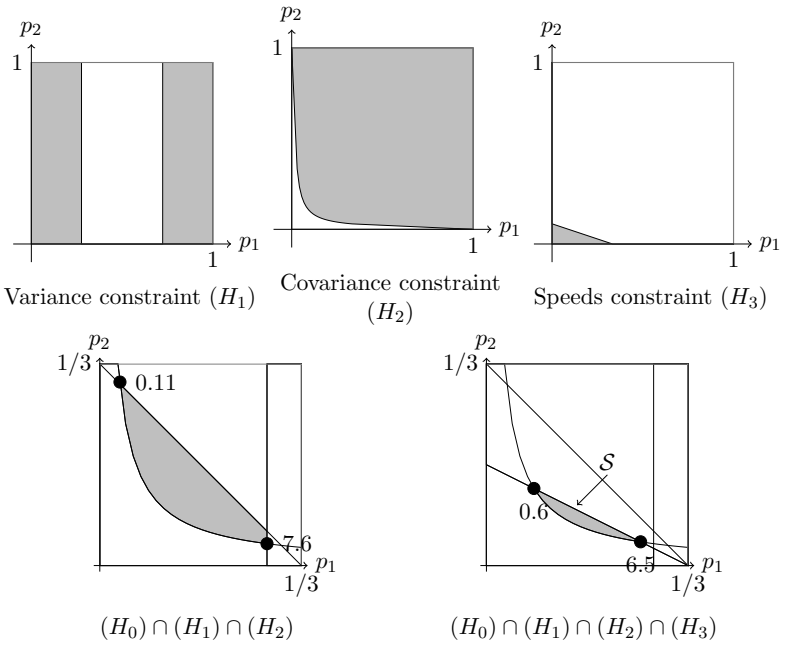


Fig. 4. The gray regions correspond to the sets of all valid (p_1, p_2) with respect to the steady-state hypothesis (H_0) and the given observations (H_1), (H_2) and (H_3). The black dots correspond to the extreme possible values of p_1/p_2.

5 Conclusion

In this work, we have studied the asymptotic distribution of the molecular quantities of a reaction network working in a steady-state regime, when these quantities are calculated using the SSA. We provided analytical expressions for the mean and the variance-covariance matrix which depend on S and $\boldsymbol{p}$. We presented several theoretical and practical consequences of this theorem: the possibility to derive confidence ellipsoids, the model validation problem, the convergence of the ratios of accumulation speeds. Toy examples illustrates our results.

A very interesting aspect of our work is that the constraints-based approach can be consider as a probabilistic counterpart of FBA. In FBA, the fluxes $\boldsymbol{f}$ quantify the occurrence rates of each reaction in steady-state with the main hypothesis that the fluxes are balanced such that the internal metabolites concentrations remain

constant. Thus, FBA-based methods are *constraints-based approaches* where $\boldsymbol{f}$ is assumed to belong to the so-called *steady-state cone*. In our approach, internal chemical species are also considered to be balanced. While FBA only concentrates on fluxes at the population level, our original approach at the cell level not only integrates the metabolite balance constraints but can also make use of the second moments (variances and co-variances) of the outputs. Hence, we can integrate the intrinsic variability of productions at the cell level. However, the additional constraints are no longer linear but quadratic (and not necessarily positive quadratic) and cannot be solved (for instance) by the classical Dantzig simplex algorithm. Thus, the constraints-based approach opens perspectives to apply efficient constraint resolution or optimization techniques.

References

[1] Adam Arkin, John Ross, and Harley H McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected escherichia coli cells. *Genetics*, 149(4):1633–1648, 1998.

[2] Yang Cao, Daniel T Gillespie, and Linda R Petzold. Efficient step size selection for the tau-leaping simulation method. *The Journal of chemical physics*, 124:044109, 2006.

[3] Stewart N Ethier and Thomas G Kurtz. *Markov processes: characterization and convergence*, volume 282. John Wiley & Sons, 2009.

[4] William Feller. *Introduction to Probability Theory and Its Applications, Vol. II POD.* John Wiley & sons, 1974.

[5] Naama Geva-Zatorsky, Nitzan Rosenfeld, Shalev Itzkovitz, Ron Milo, Alex Sigal, Erez Dekel, Talia Yarnitzky, Yuvalal Liron, Paz Polak, Galit Lahav, et al. Oscillations and variability in the p53 system. *Molecular systems biology*, 2(1), 2006.

[6] Michael A Gibson and Jehoshua Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The journal of physical chemistry A*, 104(9):1876–1889, 2000.

[7] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

[8] Daniel T Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113:297, 2000.

[9] Daniel T Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115:1716, 2001.

[10] Allan Gut. *Stopped random walks: Limit theorems and applications*. Springer, 2009.

[11] Matthias Heinemann and Renato Zenobi. Single cell metabolomics. *Current Opinion in Biotechnology*, 22(1):26–31, 2011.

[12] Volkhard Helms. *Principles of computational cell biology*. Wiley, 2008.

[13] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[14] HarleyH. McAdams and Adam Arkin. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*, 94(3):814–819, 1997.

[15] Jason A Papin, Nathan D Price, Sharon J Wiback, David A Fell, and Bernhard O Palsson. Metabolic pathways in the post-genome era. *Trends in biochemical sciences*, 28(5):250–258, 2003.

[16] Karthik Raman and Nagasuma Chandra. Flux balance analysis of biological systems: applications and challenges. *Briefings in bioinformatics*, 10(4):435–449, 2009.

[17] A. W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[18] EWJ Wallace, DT Gillespie, KR Sanft, and LR Petzold. Linear noise approximation is valid over limited times for any chemical system that is sufficiently large. *Systems Biology, IET*, 6(4):102–115, 2012.

[19] Darren J Wilkinson. Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2):122–133, 2009.

[20] Darren J Wilkinson. *Stochastic modelling for systems biology*, volume 44. CRC press, 2012.

Degenerate Cases of the Limiting Distribution

In the main matter section, we have seen that the embedded process $(\boldsymbol{y}_k)_k$ under the steady-state conditions is a random walk entirely determined by S and the reaction probabilities $\boldsymbol{p}$. When normalized, it asymptotically follows a Gaussian distribution with variance-covariance matrix $W(S, \boldsymbol{p})$. However this does not mean that the process actually spreads in all directions. Another formulation of this fact is to notice that $W(S, \boldsymbol{p})$ is symmetric and positive but not necessarily *definite*. The study of the degenerate cases is interesting for at least two reasons: it provides a better understanding of the dynamics of a given system and it allows considering a reduced equivalent and simpler system S' where $W(S', \boldsymbol{p})$ is definite. We prove that there are only two causes of degeneracy. The first one is the well known possible existence of *P-invariants* [20] that restricts the dynamics to an affine subspace. The second one corresponds to the degeneracy of the underlying reaction counting process: in some cases the set of reachable points after a given number of random steps is included in an affine hyperplane.

.1 P-invariants

A solution $\boldsymbol{z}$ to the linear equation $S\boldsymbol{z} = \boldsymbol{0}$ is called a *P-invariant*. P-invariants are meaningful since they correspond to *conservation laws* of the network [20]. Indeed, if $\boldsymbol{z}$ is a P-invariant then its coordinates are also the coordinates in the *dual basis* of a conserved *linear form* φ, that is to say $\forall t, \varphi(\boldsymbol{x}_t) = \varphi(\boldsymbol{x}_0)$. The main consequence of the relationship $\varphi \neq 0$ is that the system trajectory is included in the affine hyperplane defined by the equation $\varphi(\boldsymbol{z}) = \varphi(\boldsymbol{x}_0)$. This necessarily leads to a degenerate case for the Gaussian limiting distribution. More generally if $\dim \ker S = k$ then the trajectory is included in an affine subspace of dimension $n - k$.

To perform a *P-invariant elimination*, it is always possible to consider an equivalent reaction network S' without non-null P-invariants by removing certain columns of S. Indeed, let us consider a non-null P-invariant $\boldsymbol{z}$ and assume without loss of generality that $z_1 \neq 0$, then the molecular quantity x_t^1 is a function of $x_t^2, \cdots, x_t^n$:

$$\forall t, x_t^1 = \frac{1}{z_1}\left(z_1 x_0^1 - \sum_{i=2}^{n} z_i (x_0^i - x_t^i)\right). \tag{.1}$$

Hence, one can always remove the first column of S to decrease by 1 the dimension of the P-invariants space. By repeating successively this procedure, we eventually obtain a reaction network without non-null P-invariant.

.2 RCP invariants

The second source of degeneracy is a particular property of certain random walks where the set of reachable points after a given number of steps is included in an affine hyperplane. For instance, this is the case of the aforementioned RCP $\boldsymbol{q}_k = \left(\sum_{l=1}^{k} \boldsymbol{e}_{\mu_l}\right)_k$ (which is equal to $\boldsymbol{y}_k$ when $S = \mathrm{Id}_n$ and $\boldsymbol{x}_0 = \boldsymbol{0}$). Indeed, it is straightforward to prove that for all probability distributions on $(\mu_k)_k$, $\boldsymbol{y}_k$ belongs to the affine hyperplane defined by the equation $\sum_{i=1}^{n} \boldsymbol{z}_i = k$. This is because the

system trajectory steps forward in one direction at each reaction (see Fig .1 for a 2D example). Thus the degeneracy of the embedded process can be inherited from the degeneracy of the RCP. Similarly to P-invariant elimination, knowing the equation of these affine hyperplanes allows constructing a reduced equivalent and non-degenerated S' by eliminating one of the molecular species.

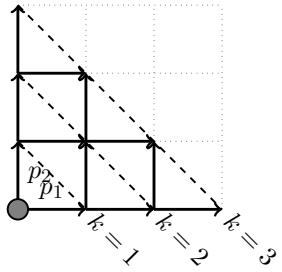


Fig. .1. Illustration of a degenerate random walk on the example system $\{\varnothing \to X; \varnothing \to Y\}$. After k reactions, the system lies necessarily in the affine hyperplane defined by equation $X + Y = k$.

.3 Characterization

Let us now introduce a general theorem that characterizes the degenerate cases of the limiting Gaussian distribution. The theorem takes into account both sources of degeneracy. Our proof relyies on the orthogonal reduction of symmetric matrices.

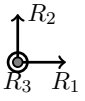
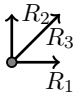
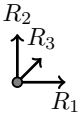
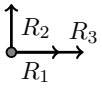
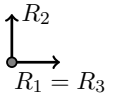
Proposition .1 *Let S be the stoichiometry matrix of a system. Let $\boldsymbol{p}$ be a positive reaction probability vector. Consider $\boldsymbol{u} = (1, \cdots, 1)^\top \in \mathbb{R}^n$. Then one and only one of the following cases occurs.*

(i) *If S has non null P-invariants, then* $\dim \ker W(S, \boldsymbol{p}) > 0$.

(ii) *If S is injective and $S\boldsymbol{z} = \boldsymbol{u}$, then has a unique solution $\boldsymbol{\eta}$ then* $\ker W(S, \boldsymbol{p}) = span(\boldsymbol{\eta})$.

(iii) *If S is injective and $S\boldsymbol{z} = \boldsymbol{u}$ has no solution, then* $\ker W(S, \boldsymbol{p}) = \{\boldsymbol{0}\}$.

The first case occurs when S is not injective, which is equivalent to $S^\top$ being not surjective. Equation efeq:affinemap implies that $(\boldsymbol{y}_k)$ is in the image of $S^\top$, so when $S^\top$ is not surjective we necessarily obtain a degenerate case. The second case is a non trivial condition that corresponds to the second source of degeneracy. When a solution $\boldsymbol{\eta}$ exists, $\boldsymbol{y}_k$ is included in the affine hyperplane parallel to the hyperplane $\sum_{i=1}^n \eta_i z_i = 0$ and passing through $\boldsymbol{x}_0 + kS^\top \boldsymbol{p}$. The last case is the regular one. The following table depicts some examples of simple reaction systems that illustrates the three cases.

System							
S injective?	yes	no	yes	yes	yes	yes	yes
Solution to $S\boldsymbol{z} = \boldsymbol{u}$?	yes	no	no	no	yes	no	yes
degenerate ?	yes	yes	no	no	yes	no	yes

Derivation of the ellipsoid equation (8)

We fix $t_\alpha = 3$ corresponding to a tolerance error of

$$1 - \frac{1}{(2\pi)^{\frac{3}{2}}} \int_{\boldsymbol{x} \in B_3(\mathbf{0},3)} \exp\left(-\frac{\|\boldsymbol{x}\|^2}{2}\right) d\boldsymbol{x} \simeq 0.0292909 \simeq 2.9\% \tag{.2}$$

and we calculate the corresponding confidence ellipsoid. According to the previous the results, the limiting variance-covariance matrix is

$$W(S', \boldsymbol{p}) = \begin{pmatrix} 0.21 & -0.06 & -0.03 \\ -0.06 & 0.16 & -0.02 \\ -0.03 & -0.02 & 0.09 \end{pmatrix} \tag{.3}$$

and its Choleski decomposition is

$$V = \begin{pmatrix} 0.4583 & 0 & 0 \\ -0.1309 & 0.378 & 0 \\ -0.0655 & -0.0756 & 0.2828 \end{pmatrix}. \tag{.4}$$

As the distribution is not degenerated, we can determine the equation of the α-confidence ellipsoid for the embedded process :

$$\mathcal{E}(\alpha, 100) : \underbrace{\left\| \frac{1}{\sqrt{100}} \underbrace{\begin{pmatrix} 2.1822 & 0 & 0 \\ 0.7559 & 2.6458 & 0 \\ 0.7071 & 0.7071 & 3.5355 \end{pmatrix}}_{V^{-1}} \left(\begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} - 100 \begin{pmatrix} 0.3 \\ 0.2 \\ 0.1 \end{pmatrix} \right) \right\|}_{err(\boldsymbol{z})} \leq t_\alpha. \tag{.5}$$