Statistical success of MAMAs: a guided tour through Francis' contributions

Catherine Matias
with the help of Dasha & Oleg Loukianov
under the benevolent tutelage of Francis

CNRS - Laboratoire de Probabilités et Modèles Aléatoires, Paris
catherine.matias@math.cnrs.fr
http://cmatias.perso.math.cnrs.fr/

Outline

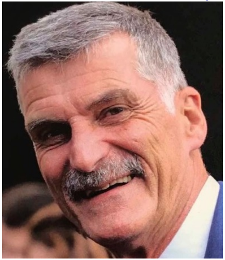
When and where it all started

Nearest-neighbour one-dimensional random walk in random environment

Estimator's construction and link with a branching process with immigration in random environment (BPIRE)

More than 10 years ago, in Évry...

MAMAs (marches aléatoires en milieu aléatoire) met Statistics



- Francis was a driving force to bridge the gap between probability and statistics
- In the context of MAMAs, only the work of [Adelman & Enriquez (04)] touched on this issue
- We started as an initial group of 5 authors . . .

Biophysical context for Statistics in MAMAs: DNA unzipping

- MAMA's introduced by [Chernov(67)] to model DNA replication
- By the end of 90's, various DNA unzipping experiments appeared

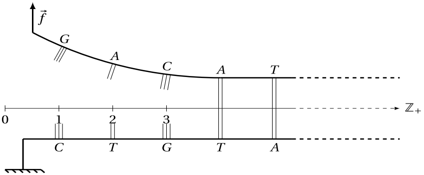


Fig. 1. DNA unzipping.

Goals

- DNA sequencing (exploratory),
- Study the structural properties of the molecule.

Statistics for MAMAs

General idea

Assume you observe a single (arbitrarily long) trajectory from a (one-dimensional) MAMA, what can you say about the underlying environment?

Overview of the results

Stat set.	Estim type	Proba set.	Results	Tools
Param.	MLE	transient, ballistic [Comets *et al.*(14), Falconnet *et al.*(14b)]	consist.+AN+eff.	BPIRE
		trans. ball.+ Markov env [Andreoletti *et al.*(15)]	consist.+AN+eff.	+ link with HMM
		transient, sub-ballistic [Falconnet *et al.*(14a)]	consistency+AN	BPIRE
		recurrent, finite unknown support [Comets *et al.*(16)]	consistency (+ AN ongoing)	loc. in the infinite valley (+law env. seen from particle)
Non param.	Moment-based	general state space [Adelman & Enriquez (04)]	consist.	reinforced RW
		rec. or right-transient [Diel & Lerasle(18), Havet *et al.*(19)]	consist. + risk bounds	BPIRE

Statistics for MAMAs

General idea

Assume you observe a single (arbitrarily long) trajectory from a (one-dimensional) MAMA, what can you say about the underlying environment?

Overview of the results

Stat set.	Estim type	Proba set.	Results	Tools
Param.	MLE	transient, ballistic [Comets *et al.*(14), Falconnet *et al.*(14b)]	consist.+AN+eff.	BPIRE
		trans. ball.+ Markov env [Andreoletti *et al.*(15)]	consist.+AN+eff.	+ link with HMM
		transient, sub-ballistic [Falconnet *et al.*(14a)]	consistency+AN	BPIRE
		recurrent, finite unknown support [Comets *et al.*(16)]	consistency (+ AN ongoing)	loc. in the infinite valley (+law env. seen from particle)
Non param.	Moment-based	general state space [Adelman & Enriquez (04)]	consist.	reinforced RW
		rec. or right-transient [Diel & Lerasle(18), Havet *et al.*(19)]	consist. + risk bounds	BPIRE

Outline

When and where it all started

Nearest-neighbour one-dimensional random walk in random environment

Estimator's construction and link with a branching process with immigration in random environment (BPIRE)

Model description I

Random environment on $\mathbb{Z}$

- $\omega = \{\omega_x\}_{x\in\mathbb{Z}}$ i.i.d. with $\omega_x \in (0,1)$ and $\omega_x \sim \nu$,
- $\nu \in \mathcal{M}$ a class of distributions on $(0,1)$
- $\mathbf{P}_\nu = \nu^{\otimes\mathbb{Z}}$ law on $(0,1)^{\mathbb{Z}}$ of ω and $\mathbf{E}_\nu$ expectation,

Markov process conditional on the environment

For fixed ω, let $X = \{X_t\}_{t\in\mathbb{N}}$ be the Markov chain on $\mathbb{Z}$ starting at $X_0 = 0$ and with transitions

$$P_\omega(X_{t+1} = y | X_t = x) = \begin{cases} \omega_x & \text{if } y = x+1, \\ 1-\omega_x & \text{if } y = x-1, \\ 0 & \text{otherwise.} \end{cases}$$

P_ω is the measure on the path space of X given ω (quenched law) and E_ω corresponding expectation.

Model description II

Random walk in random environment (MAMA)

The (unconditional) law of X is the annealed law

$$\mathbb{P}(\cdot) = \int P_\omega(\cdot)\mathrm{d}\mathbf{P}_\nu(\omega),$$

with $\mathbb{E}$ for the corresponding expectation.

Note that X is not a Markov process.

Limiting behaviour of X

Let

$$\rho_x = \frac{1-\omega_x}{\omega_x}, \qquad x \in \mathbb{Z}.$$

[Solomon(75)] proved the classification:

(a) Recurrent case: If $\mathbf{E}_\nu(\log \rho_0) = 0$, then

$$-\infty = \liminf_{t\to\infty} X_t < \limsup_{t\to\infty} X_t = +\infty, \quad \mathbb{P}\text{-almost surely.}$$

(b) Transient case: if $\mathbf{E}_\nu(\log \rho_0) < 0$, then

$$\lim_{t\to\infty} X_t = +\infty, \quad \mathbb{P}\text{-almost surely.}$$

If we moreover let $T_n = \inf\{t \in \mathbb{N} : X_t = n\}$, then

(b1) Ballistic case: if $\mathbf{E}_\nu(\rho_0) < 1$, then, $\mathbb{P}$-almost surely, $T_n/n \to c$, $\mathbb{P}$-a.s.

(b2) Sub-ballistic case: If $\mathbf{E}_\nu(\rho_0) \geq 1$, then $T_n/n \to +\infty$ $\mathbb{P}$-almost surely, when $n \to \infty$

Statistical estimation of the law of the environment

Goal and context

- Goal: Estimate the distribution ν relying on the observation of a trajectory $X_{[0,T_n]}$.
- In a much more general setting, [Adelman & Enriquez (04)] provide a link between the MAMA and the environment, leading to moment estimators for the distribution ν.
- Drawback of [Adelman & Enriquez (04)]: In a parametric setting, estimate some moments first and then invert a function to recover the parameters θ. May induce a loss of efficiency.
- In a series of papers, Francis and his co-authors have focused on maximum likelihood estimation (MLE).
- Later, a similar (to [Adelman & Enriquez (04)]) moment approach has been successfully used to estimate the law of the environment in a non-parametric setting [Diel & Lerasle(18), Havet *et al.*(19)].

Outline

When and where it all started

Nearest-neighbour one-dimensional random walk in random environment

Estimator's construction and link with a branching process with immigration in random environment (BPIRE)

Estimator's construction I

Recall that $T_n = \inf\{t \in \mathbb{N} : X_t = n\}$. We let

$$L_x^n := \sum_{s=0}^{T_n-1} \mathbf{1}\{(X_s, X_{s+1}) = (x, x-1)\}$$

and $R_x^n := \sum_{s=0}^{T_n-1} \mathbf{1}\{(X_s, X_{s+1}) = (x, x+1)\}$, the number of left steps (resp. right steps) from site x. Then,

$$P_\omega(X_{[0,T_n]}) = \prod_{x\in\mathbb{Z}} \omega_x^{R_x^n}(1-\omega_x)^{L_x^n}$$

$$\text{and } \mathbb{P}(X_{[0,T_n]}) = \prod_{x\in\mathbb{Z}} \int_0^1 a^{R_x^n}(1-a)^{L_x^n}\mathrm{d}\nu(a).$$

Note that

- Only the visited sites contribute in this product.

Estimator's construction II

- Moreover, $L_n^n = 0$ and

$$R_x^n = \begin{cases} L_{x+1}^n & \text{if } x < 0 \\ L_{x+1}^n + 1 & \text{if } x \in [0, n-1] \end{cases}$$

Hence, the likelihood function $\nu \mapsto \ell_n(\nu)$ is defined as

$$\ell_n(\nu) = \sum_{x \leq n-1} \log \int_0^1 a^{L_{x+1}^n + \mathbf{1}\{x \geq 0\}} (1-a)^{L_x^n} \mathrm{d}\nu(a).$$

(the sequence $(L_x^n)_{x \leq n}$ is an exhaustive stat).

Underlying BPIRE

From [Kesten *et al.*(75)], under the annealed law $\mathbb{P}$, the sequence $L_n^n, L_{n-1}^n, \ldots, L_0^n$ has the same distribution as a BPIRE denoted $Z_0, \ldots, Z_n$, and defined by

$$Z_0 = 0, \quad \text{and for } k = 0, \ldots, n-1, \quad Z_{k+1} = \sum_{i=0}^{Z_k} \xi'_{k+1,i},$$

with $\{\xi'_{k,i}\}_{k \in \mathbb{N}^*; i \in \mathbb{N}}$ independent and

$$\forall m \in \mathbb{N}, \quad P_\omega(\xi'_{k,i} = m) = (1-\omega_k)^m \omega_k,$$

Under annealed law $\mathbb{P}$, $\{Z_n\}_{n \in \mathbb{N}}$ is a homogeneous Markov chain with transition kernel

$$Q_\nu(x,y) = \binom{x+y}{x} \int_0^1 a^{x+1}(1-a)^y \mathrm{d}\nu(a).$$

Back to the estimators I

Parametric setting: MLE

In a parametric setting, $\nu = \nu_\theta$ and we have an equality in $\mathbb{P}$-distribution (for some explicit function ϕ_θ)

$$\ell_n(\nu_\theta) \stackrel{dist.}{=} \sum_{k \le n-1} \phi_\theta(Z_k, Z_{k+1})$$

Then we let $\widehat{\theta}_n \in \mathrm{Argmax}_{\theta\in\Theta}\, \ell_n(\nu_\theta)$ and the right-hand side is the likelihood of

- a Markov chain (when environment is iid)
- a hidden Markov chain (HMM, when environment is Markov)

Back to the estimators II

Challenges in the parametric setting: transient case

In the transient regime, the chain is positive recurrent aperiodic with unique stationary distribution π_ν characterized by $\pi_\nu(k) = \mathbf{E}_\nu[S(1-S)^k]$, where

$$S := (1 + \rho_1 + \rho_1\rho_2 + \cdots + \rho_1 \dots \rho_n + \dots)^{-1} \in (0,1).$$

Tools:

- LLN for $n^{-1}\sum_k h(Z_k, Z_{k+1})$
- TCL for $n^{-1/2}\sum_k h(Z_k, Z_{k+1}) - \mathbb{E}[h(Z_k, Z_{k+1})|\mathcal{F}_{k-1}]$

Back to the estimators III

Challenges in the parametric setting: recurrent case

In the recurrent regime, BPIRE explodes and is useless. What can we do instead?

- [Comets *et al.*(16)] restrict to ν with finite unknown support, things can be written explicitely. Assume $\nu(\cdot) = \sum_{i=1}^{d} p_i \delta_{a_i}(\cdot)$. Now the parameter is $(a_i, p_i)_{1 \le i \le d}$
- Consistency relies on Sinai's localisation of the walk in the infinite valley [Sinai(83)] and result of [Gantert *et al.*(10)] about the convergence of centered (with respect to the bottom of the valley) local times.
- AN requires limit theorems for normalised functionals of the type $\sum_{t=1}^{n} f(w_{X_t}, X_{t+1} - X_t)$. Here w_{X_t} is the zero coordinate of the environment seen from the particle.

Back to the estimators IV

Non parametric setting: moment estimator

[Diel & Lerasle(18)] estimate the moment quantity

$$m^{\alpha,\beta} = \int_0^1 a^\alpha (1-a)^\beta \mathrm{d}\nu(a) = \mathbf{E}_\nu[\omega_0^\alpha (1-\omega_0)^{1-\beta}],$$

through its empirical counterpart (for some explicit function $h_{\alpha,\beta}$),

$$\hat{m}_n^{\alpha,\beta} = \sum_{k=1}^{n} h_{\alpha,\beta}(Z_k, Z_{k+1}).$$

(recurrent or transient to the right cases). This is later used to construct an estimator of the cdf of ν [Diel & Lerasle(18)] and of its density [Havet *et al.*(19)].

Back to the estimators V

Challenges in non parametric setting

- Cdf and density of ν can be consistently estimated from the above moments, indifferently in recurrent or right-transient case. Rates of convergence are provided.
- Tool: Concentration inequalities for $\sum_{k=1}^{n} h(Z_k, Z_{k+1})$
- However no minimax rate is known in this context: are those rates of convergence optimal?

Some remaining challenges

- In the parametric, transient, sub-ballistic case, under Temkin's model (finite and unknown support of size 2) Fisher's information is infinite. Actual rate of convergence of MLE might be faster than $1/\sqrt{n}$;
- In the recurrent and parametric case
  - prove AN of MLE when ν has finite support, (*preliminary results by Comets, Loukianova, Loukianov*)
  - build a MLE when ν is absolutely continuous;
- In the Markov environment case, [Andreoletti *et al.*(15)] worked under a (technical) assumption that ν is reversible: go beyond that;
- In the non parametric setting, what are the minimax rates of convergence? And can they be achieved?

My personal recordings on this tremendous time

- Francis has been a driving force in the development of statistical results for estimating the environment law of a MAMA;
- Our group meetings mixing probabilists and statisticians were enriching for all of us as well as for our disciplines;
- Following Francis heritage, we hope many researchers will follow this fruitful path of mixing our sub-disciplines
- Quoting one of my favourite colleagues at LPSM: "We are all probabilists" and I would add "we also are all statisticians".

Thank you !

And if you have questions, Dasha & Oleg will be happy to answer them ;)

My personal recordings on this tremendous time

- Francis has been a driving force in the development of statistical results for estimating the environment law of a MAMA;
- Our group meetings mixing probabilists and statisticians were enriching for all of us as well as for our disciplines;
- Following Francis heritage, we hope many researchers will follow this fruitful path of mixing our sub-disciplines
- Quoting one of my favourite colleagues at LPSM: "We are all probabilists" and I would add "we also are all statisticians".

Thank you !

And if you have questions, Dasha & Oleg will be happy to answer them ;)

References I

Omer Adelman and Nathanaël Enriquez.
Random walks in random environment: what a single trajectory tells.
Israel J. Math., 142:205–220, 2004.

Andreoletti, Pierre, Loukianova, Dasha, and Matias, Catherine.
Hidden markov model for parameter estimation of a random walk in a markov environment.
ESAIM: PS, 19:605–625, 2015.

A.A. Chernov.
Replication of a multicomponent chain by the lightning mechanism.
Biofizika, 12:297–301, 1967.

References II

F. Comets, M. Falconnet, D. Loukianova, and O. Loukianov.
Maximum likelihood estimator consistency for recurrent random walk in a parametric random environment with finite support.
Stochastic Processes and their Applications, 126:3578–3604, 2016.

Francis Comets, Mikael Falconnet, Oleg Loukianov, Dasha Loukianova, and Catherine Matias.
Maximum likelihood estimator consistency for a ballistic random walk in a parametric random environment.
Stochastic Process. Appl., 124(1):268–288, 2014.

Roland Diel and Matthieu Lerasle.
Non parametric estimation for random walks in random environment.
Stochastic Process. Appl., 128(1):132–155, 2018.

References III

M. Falconnet, A. Gloter, and D. Loukianova.
Mle in the context of a sub-balistic random walk in a parametric random environment.
Mathematical Methods of Statistics, 23:159–175, 2014.

M. Falconnet, D. Loukianova, and C. Matias.
Asymptotic normality and efficiency of the maximum likelihood estimator for the parameter of a ballistic random walk in a random environment.
Math. Methods Statist., 23(1):1–19, 2014.

Nina Gantert, Yuval Peres, and Zhan Shi.
The infinite valley for a recurrent random walk in random environment.
Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 46(2):525 – 536, 2010.

References IV

A. Havet, M. Lerasle, and É Moulines.
Density estimation for rwre.
Math. Meth. Stat., 28:18–38, 2019.

H. Kesten, M. V. Kozlov, and F. Spitzer.
A limit law for random walk in a random environment.
Compositio Math., 30:145–168, 1975.

Ya. G. Sinai.
The limiting behavior of a one-dimensional random walk in a random medium.
Theory of Probability & Its Applications, 27(2):256–268, 1983.

Fred Solomon.
Random walks in a random environment.
Ann. Probab., 3:1–31, 1975.

Outline

Three examples

Simulations

Examples of environment distributions I

Example 1: Finite and known support

- Fix $a_1 < a_2 \in (0,1)$ and let $\nu_p = p\delta_{a_1} + (1-p)\delta_{a_2}$, where δ_a is the Dirac mass located at value a.
- Unknown parameter $p \in \Theta \subset (0,1)$ (namely $\theta = p$)
- Assume that a_1, a_2 and Θ are such that the process is transient and ballistic.

Then, the assumptions are satisfied and one can estimate p consistently and efficiently.

- May be generalised to $k > 2$ fixed and known support points and $\theta = (p_1, \dots, p_{K-1})$.

Examples of environment distributions II

Example 2: Two unknown support points (Temkin's model)

- $\nu_\theta = p\delta_{a_1} + (1-p)\delta_{a_2}$ and unknown parameter $\theta = (p, a_1, a_2) \in \Theta$, where Θ is a compact subset of

$$(0,1) \times \{(a_1, a_2) \in (0,1)^2 \,:\, a_1 < a_2\}$$

  such that the process is transient and ballistic.

Then, the assumptions are satisfied and one can estimate θ consistently. Moreover, if $\mathbf{E}^\theta(\rho_0^3) < 1$, the MLE estimator is asymptotically normal and efficient.

Examples of environment distributions III

Example 3: Beta distribution

- $\mathrm{d}\nu(a) = \frac{1}{\mathrm{B}(\alpha,\beta)} a^{\alpha-1}(1-a)^{\beta-1}\mathrm{d}a$,
- Unknown parameter $\theta = (\alpha, \beta) \in \Theta$ where Θ is a compact subset of
$$\{(\alpha, \beta) \in (0, +\infty)^2 \,:\, \alpha > \beta + 1\}.$$
- As $\mathbf{E}^{\theta}(\rho_0) = \beta/(\alpha - 1)$, the constraint $\alpha > \beta + 1$ ensures that the process is transient and ballistic.

Then, the assumptions are satisfied and one can estimate θ consistently and efficiently.

Outline

Simulations protocol

- Three models corresponding to the previous 3 examples, with $\theta^\star$ as in Table 1.
- In each model, $1,000$ repeats of the following procedure
  - Generate a random environment according to distribution $\nu_{\theta^\star}$ on the set of sites $\{-10^4, \dots, 10^4\}$.
  - Run a random walk in this environment and stop it successively at the hitting times T_n, with $n \in \{10^3 k; 1 \leq k \leq 10\}$.
  - For each value of n,
    - Estimate $\theta^\star$ with MLE and [Adelman & Enriquez (04)]'s procedure
    - Estimate the Fisher information matrix $\Sigma_{\theta^\star}$ and compute a confidence interval for $\theta^\star$

Simulation	Fixed parameter	Estimated parameter
Example 1	$(a_1, a_2) = (0.4, 0.7)$	$p^\star = 0.3$
Example 2	-	$(a_1^\star, a_2^\star, p^\star) = (0.4, 0.7, 0.3)$
Example 3	-	$(\alpha^\star, \beta^\star) = (5, 1)$

Table: Parameter values for each experiment.

Lengths of the walks

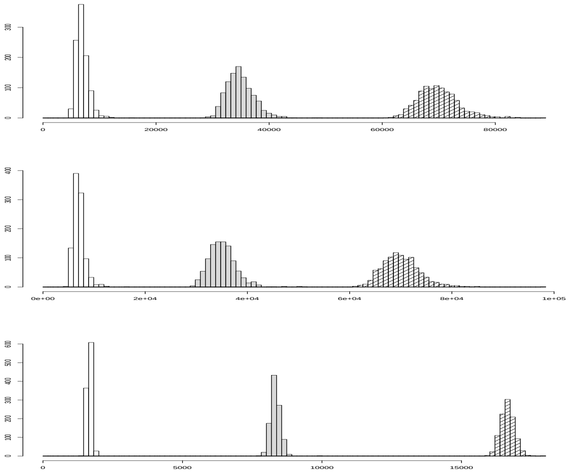
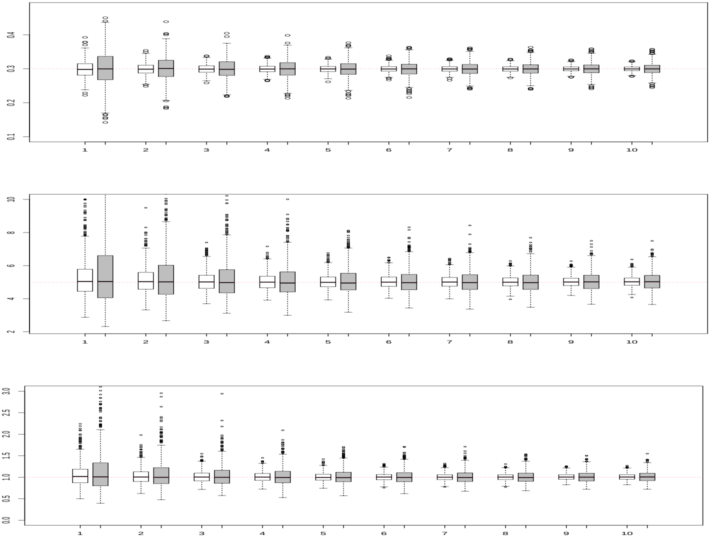
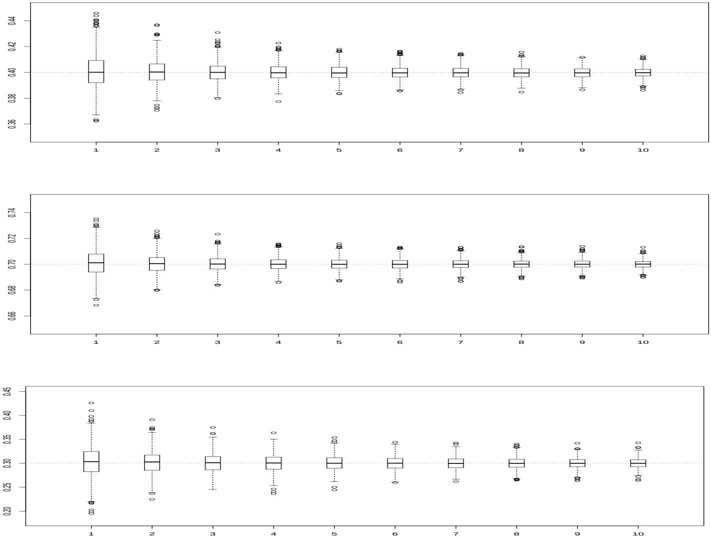


Figure: Histograms of the hitting times T_n for values n equal to 1 000 (white), 5 000 (grey) and 10 000 (hatched). Top panel: Example 1; middle panel: Example 2; bottom panel: Example 3.

Boxplots of MLE (white) and [Adelman & Enriquez (04)]'s estimate (grey) - Ex. 1 ($\hat{p}$) and 3 ($\hat{\alpha}, \hat{\beta}$)



Boxplots of MLE - Ex. 2 $(\hat{p}, \hat{a}_1, \hat{a}_2)$



Estimation of the Fisher information matrix - Ex 1

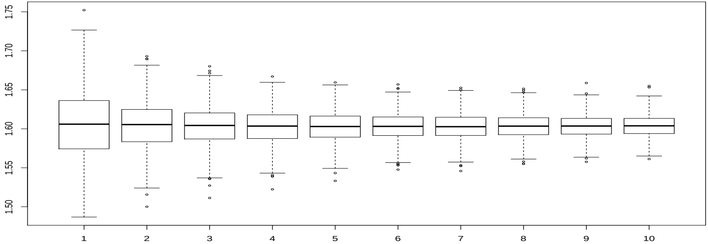


Figure: Boxplot of the estimator $\hat{\Sigma}_n$ obtained from 1000 iterations and for values n ranging in $\{10^3 k : 1 \leq k \leq 10\}$ in the case of Ex. 1.

Estimation of the Fisher information matrix - Ex 2

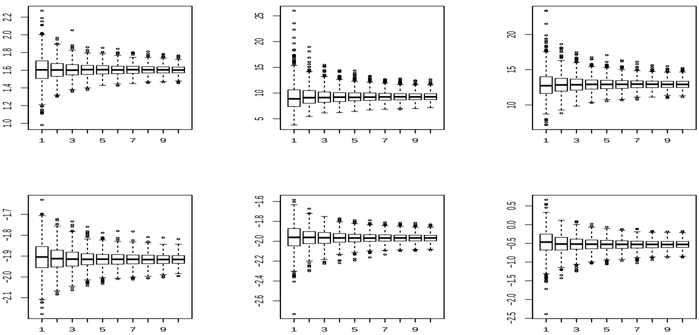


Figure: Boxplots of the values of the matrix $\hat{\Sigma}_n$ obtained from 1000 iterations and for values n ranging in $\{10^3 k : 1 \leq k \leq 10\}$ in the case of Ex. 2. The parameter is ordered as $\theta = (\theta_1, \theta_2, \theta_3) = (p, a_1, a_2)$ and the figure displays the values: $\hat{\Sigma}_n(1,1); \hat{\Sigma}_n(2,2); \hat{\Sigma}_n(3,3); \hat{\Sigma}_n(1,2); \hat{\Sigma}_n(1,3)$ and $\hat{\Sigma}_n(2,3)$, from left to right and top to bottom.

Estimation of the Fisher information matrix - Ex 3

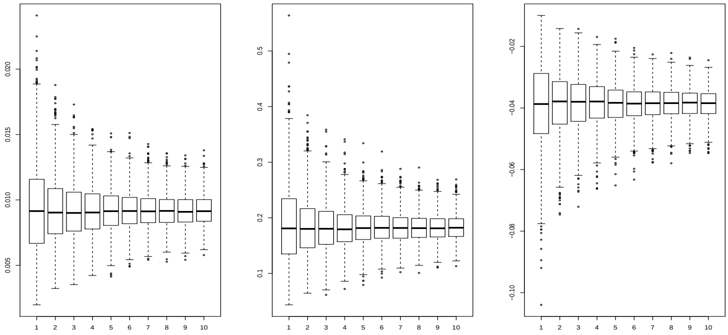


Figure: Boxplots of the values of the matrix $\hat{\Sigma}_n$ obtained from 1000 iterations and for values n ranging in $\{10^3 k : 1 \leq k \leq 10\}$ in the case of Ex. 3. The parameter is ordered as $\theta = (\theta_1, \theta_2) = (\alpha, \beta)$ and the figure displays the values: $\hat{\Sigma}_n(1,1); \hat{\Sigma}_n(2,2)$ and $\hat{\Sigma}_n(1,2)$, from left to right.

Empirical coverages of confidence regions

n	Ex. 1			Ex. 2			Ex. 3		
	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
1000	0.994	0.952	0.899	0.992	0.953	0.909	0.977	0.942	0.901
2000	0.989	0.952	0.903	0.994	0.953	0.910	0.978	0.928	0.884
3000	0.988	0.942	0.901	0.990	0.938	0.886	0.981	0.940	0.889
4000	0.991	0.944	0.896	0.991	0.951	0.894	0.988	0.945	0.900
5000	0.990	0.942	0.896	0.993	0.942	0.891	0.986	0.941	0.883
6000	0.983	0.948	0.901	0.987	0.951	0.888	0.988	0.937	0.897
7000	0.986	0.950	0.900	0.992	0.951	0.900	0.986	0.942	0.898
8000	0.987	0.956	0.898	0.988	0.950	0.903	0.981	0.946	0.903
9000	0.990	0.959	0.913	0.990	0.949	0.893	0.985	0.939	0.901
10000	0.987	0.954	0.908	0.990	0.949	0.899	0.983	0.944	0.892

Table: Empirical coverages of $(1-\gamma)$ asymptotic level confidence regions, for $\gamma \in \{0.01, 0.05, 0.1\}$ and relying on 1000 iterations.

Conclusions from the simulations

- Good performances of $\widehat{\theta}_n$ on simulated data
  - Unbiased estimator (like [Adelman & Enriquez (04)]'s one)
  - Less spread out than [Adelman & Enriquez (04)]'s one (in fact efficient)
  - Easier to compute (Ex. 2 [Adelman & Enriquez (04)]'s estimate is out of reach)
- Confidence regions build from $\widehat{\theta}_n$ have accurate empirical coverage.