# ROBUST MODEL SELECTION FOR STOCHASTIC PROCESSES

*J. E. García*[1a] *V. A. González-López*[1b] *and M. L. L.Viola*[2]

[1]University of Campinas, Brazil.
[a]jg@ime.unicamp.br, [b]veronica@ime.unicamp.br,
[2]Federal University of São Carlos, Brazil.

## ABSTRACT

In this paper we address the problem of model selection for the set of finite memory stochastic processes with finite alphabet, when the data is contaminated. We consider $m$ independent samples, with most of them being realizations of the same stochastic process with law $Q$, which is the one we want to retrieve. We devise a model selection procedure such that for a sample size large enough, the selected process is the one with law $Q$. Our model selection strategy is based on estimating relative entropies to select a subset of samples that are realizations of the same law. Although the procedure is valid for any family of finite order Markov models, we will focus on the family of variable length Markov chain models, which include the fixed order Markov chain model family. We define the asymptotic breakdown point $\gamma$ for a model selection procedure, and we show the value $\gamma$ for our procedure. This means that if the proportion of contaminated samples is smaller than $\gamma$, then, as the sample size grows our procedure selects a model for the process with law $Q$.

## 1. INTRODUCTION

In this paper we propose a robust strategy to select models from samples coming from a process which is contaminated and it is a discrete time stochastic process, on a finite alphabet. We will only consider the family of variable length Markov chain models, from now on VLMC (see [4, 1, 2, 5]) because it includes the fixed order Markov chain models and the independent case. For VLMC model selection we will use the version of the CTM algorithm introduced by [2], which is based on the Bayesian Information Criterion (BIC). It has been shown by [3] that a small Bernoulli random perturbation on a sample produced by a VLMC will effectively transform the process to an infinity memory process. They also show a variation of the original context algorithm given by [4] which can recover the VLMC model of the original chain, provided that the noise is small enough.

In this work we consider a different kind of contamination, we have a set of $m$ independent samples, with most of them being from the same stochastic process with law $Q$, whose model we want to recover. The approach of this paper can be applied yet in the case in which we have only one sample produced by the concatenation of realizations of a mixture process which is the process $Q$ plus a contaminant process. We define the asymptotic breakdown point $\gamma$ for the model selection problem and we show the value of $\gamma$ for our procedure.

Our procedure can be applied when the data is coming from a mixture of stochastic processes, for example in the problem of classification of languages according to their rhythmic features, using speech samples. The usual procedure to deal with this topic has been choose a subset of the original sample which seems best represent each language. Instead, if we apply this kind of robust procedure can be taken the complete dataset, see [6].

## 2. PRELIMINARIES

Let $(X_t)$ be a discrete time stochastic process on a finite alphabet $A$ with cardinal $|A|$. Denote the string (concatenation of elements from $A$) $a_k a_{k+1} \ldots a_r$ by $a_k^r$, where $a_i \in A$, $k \leq i \leq r$. If the stochastic process $(X_t)$ has probability law $Q$, and if $x_1^n$ is a $n$ realization of that process, we denote $Q(x_1^n) = Prob(X_1^n = x_1^n)$. The transition probability from the sequence $x_1^n$ to the symbol $a \in A$ is $Q(a|x_1^n) = Prob(X_{n+1} = a|X_1^n = x_1^n)$. Given a string $s = a_k a_{k+1} \ldots a_r$, we denote its length as $l(s) = r - k + 1$. The empty string is denoted by $\emptyset$ and $l(\emptyset) = 0$. We say that the string $v$ is a postfix of a string $s$ when there exists a string $u$ such that $s = uv$. When $s \neq v$, $v$ is a proper postfix of $s$.

**Definition 1** *A set $\mathcal{T}$ of strings is called a tree if satisfies the following rules*

1. *no $s_1 \in \mathcal{T}$ is a postfix of any other $s_2 \in \mathcal{T}$,*
2. *no $s_1 \in \mathcal{T}$ can be replaced by a proper postfix without violating rule 1.*

We denote by $d(\mathcal{T}) = \max\big(l(s), s \in \mathcal{T}\big)$ the depth of the tree $\mathcal{T}$.

**Definition 2** *Let $(X_t)$ be a finite order stationary ergodic stochastic process on a finite alphabet $A$ with probability law $Q$. We will say that the tree $\mathcal{T}$ is a context tree for $(X_t)$ if for any $n \geq d(\mathcal{T})$ and for any sequence of symbols in $A$, $x_1^n$, there exist a postfix $s \in \mathcal{T}$ such that*

$$Q(a|x_1^n) = Q(a|s), \ \ \forall a \in A, \tag{1}$$

*and no proper postfix of $s$ satisfies equation (1). In that case $s$ is called a context for the process $Q$.*

**Definition 3** *We will say that the stochastic process $(X_t)$ is a variable length Markov chain compatible with the context tree $\mathcal{T}$ if it verify definition 2.*

Each model in the family of variable length Markov chain models, is identified by its context tree. For more details see [4, 1]. There are diverse methodologies for the selection and estimation of context trees, see for example [1, 2, 4, 5]. The context tree maximization CTM algorithm proposed by [2] is based on the BIC criterion and it will be used in this work for the statistical estimation of context trees.

For a given value $D$ with $n > D$, if $s$ is some string $l(s) < D$, $a \in A$ we denote by $N_n(s,a)$ the number of occurrences of the string $s$ followed by $a$ in the sample $x_1^n$, $N_n(s,a) = \big|\{i : D < i \leq n, x_{i-D}^{i-1} = s, x_i = a\}\big|$. The number of occurrences of $s$ in the sample $x_1^n$ is denoted by $N_n(s)$ and $N_n(s) = \big|\{i : D < i \leq n, x_{i-D}^{i-1} = s\}\big|$. We denote by $\mathcal{K}(x_1^n, D)$ the family of feasible context trees, where a feasible context tree $\mathcal{T}$ is such that $d(\mathcal{T}) \leq D$ and $N_n(s) \geq 1$ for all $s \in \mathcal{T}$ and for each string $s'$ with $N_n(s') \geq 1$ it has a postfix $s \in \mathcal{T}$. Now we can define the context tree estimator

$$\hat{\mathcal{T}}(x_1^n) = \arg \max_{\mathcal{T} \in \mathcal{K}(x_1^n, D)} \prod_{s \in \mathcal{T}} \tilde{P}_s(x_1^n) \qquad (2)$$

where $\tilde{P}_s(x_1^n) = n^{-\frac{(|A|-1)}{2}} \tilde{P}_{\text{ML},s}(x_1^n)$. $\tilde{P}_{\text{ML},s}(x_1^n) = \prod_{a \in A} \left(\frac{N_n(s,a)}{N_n(s)}\right)^{N_n(s,a)}$ if $N_n(s) \geq 1$ and $\tilde{P}_{\text{ML},s}(x_1^n) = 1$ if $N_n(s) = 0$.

For fixed $n$ is considered $D = D(n) = \log(n)$. For a finite memory Markov process, $\hat{\mathcal{T}}(x_1^n)$ converges eventually almost surely to the true $\mathcal{T}$ of the law $Q$. The algorithm in [2] allows to compute these estimators in $O(n)$ time, and to compute them on-line for all $i \leq n$ in $o(n\log(n))$ time. According to the corollary 2.12 in [2] the empirical probabilities $\hat{Q}_{\hat{\mathcal{T}}}(a|s) = \frac{N_n(s,a)}{N_n(s)}$, $a \in A, s \in \hat{\mathcal{T}}$ converges to the true conditional probabilities $Q(a|s), a \in A, s \in \mathcal{T}$ almost surely as $n \to \infty$.

In order to simplify the notation we avoid the reference to the context tree $\mathcal{T}(or\hat{\mathcal{T}})$ when the underlying context tree is understood and we adopt the notation

$$\hat{Q} = \widehat{CTM}((x_t)_{t=1}^n)$$

to emphasize that the estimation uses the CTM algorithm.

## 3. RELATIVE ENTROPY

**Definition 4** *Given two probability mass functions $P(\cdot)$ and $Q(\cdot)$, the relative entropy is*

$$D(P||Q) = \sum_{x \in \chi} P(x) \log\big(\frac{P(x)}{Q(x)}\big).$$

**Remark 1** *Let $P(\cdot)$, $Q(\cdot)$ be two probability functions. Then, $D(P||Q) \geq 0$. The equality occurs if and only if $P(x) = Q(x), \forall x \in \chi$.*

**Definition 5** *Let $\mathcal{T}_P$ and $\mathcal{T}_Q$ be two context trees following the definition 2 with probability law $P$ and $Q$ respectively. $\mathcal{T}_{PQ}$ is defined by all the strings from $\mathcal{T}_P$ and $\mathcal{T}_Q$, such that $\mathcal{T}_{PQ}$ satisfy the definition 1.*

From the previous definition,
$\mathcal{T}_{PQ} = \{s \in \mathcal{T}_P \cup \mathcal{T}_Q : \nexists s' \in \mathcal{T}_P \cup \mathcal{T}_Q$ postfix of $s\}$. From Theorem 3 (see [6]), using $\mathcal{T}_{PQ}$ it is possible to express the entropy between two processes through its conditional entropies as $D\big(P||Q\big) = \sum_{s \in \mathcal{T}_{PQ}} P(s) D\big(P(\cdot|s)||Q(\cdot|s)\big)$.

**Remark 2** *For $s \in \mathcal{T}_{PQ}$, we observe that $P(\cdot|s)$ is the usual probability when $s \in \mathcal{T}_P$. If $s \notin \mathcal{T}_P$, $\exists s_1 \in \mathcal{T}_P$ and $x$ some string, such that $s = xs_1$ and $P(\cdot|s) = P(\cdot|s_1)$.*

## 4. ASYMPTOTIC BREAKDOWN POINT

**Assumption 1** *For a family $\mathcal{F}$ of stochastic processes, consider a collection $\{(X_{i,t}), i = 1, \ldots, m\}$ of $m$ independent finite memory stationary processes belonging to $\mathcal{F}$, where $(X_{i,t})$ has probability law $Q_i$. If $\mathcal{J}_{Q_i} =$ $\{j \in \{1, ..., m\} : (X_{jt}) \sim Q_i\}$, suppose that exists $i_0$ such that $\forall i \neq i_0$, $|\mathcal{J}_{Q_{i_0}}| > |\mathcal{J}_{Q_i}|$, with $i, i_0 \in \{1, \ldots, m\}$, $Q_{i_0}$ will be called as majority law of $\mathcal{F}$. Denote by $\mathcal{C}_n^m = \{(x_{1,t})_{t=1}^n, (x_{2,t})_{t=1}^n \ldots, (x_{m,t})_{t=1}^n\}$ a collection of $m$ samples of size $n$ from $(X_{i,t}), i = 1, \ldots, m$.*

Let $\mathcal{S}$ be a strategy of estimation, i.e.

$$\mathcal{S} : \Omega_{\mathcal{F}} \to \mathcal{F}$$

where $\Omega_{\mathcal{F}}$ is the sample space of processes in $\mathcal{F}$ and $\mathcal{S}(\mathcal{C}_n^m)$ denotes the value of the estimator from the sample collection $\mathcal{C}_n^m$.

**Remark 3** *Under the Assumption 1, $\mathcal{S}(\mathcal{C}_n^m)$ indicates some strategy to select a sample (from $\mathcal{C}_n^m$ ) or some set of samples (the best ones) to make the estimation of the majority law $Q_{i_0}$.*

We define now the asymptotic breakdown point of the model estimator $\mathcal{S}(\mathcal{C}_n^m)$.

**Definition 6** *Under the Assumption 1, the model estimator $\mathcal{S}(\mathcal{C}_n^m)$ has an asymptotic breakdown point equal to $\gamma$ for the family $\mathcal{F}$, if $\gamma$ is the smallest value into $(0, 1]$ such that, if $\frac{|\mathcal{J}_{Q_{i_0}}|}{m} < \gamma$ then,*

$$\lim_{n \to \infty} \mathcal{S}(\mathcal{C}_n^m) \neq Q_{i_0}, \textit{almost surely.}$$

## 5. ESTIMATORS

Given the collection of samples $\mathcal{C}_n^m$, for each $i \in \{1, ..., m\}$ denote $\hat{Q}_i = \widehat{CTM}((x_{i,t})_{t=1}^n)$ the model estimated from the sample $(x_{i,t})_{t=1}^n$ using the algorithm introduced in [2]. For each $i, j \in \{1, ..., m\}$, denote by $\hat{d}_{(i||j)}(\mathcal{C}_n^m)$ the relative entropy between $\hat{Q}_i$ and $\hat{Q}_j$, i.e. $\hat{d}_{(i||j)}(\mathcal{C}_n^m) = D\Big(\hat{Q}_i||\hat{Q}_j\Big)$. Define then,

$$\overline{d}_{(i,j)}(\mathcal{C}_n^m) = \frac{\hat{d}_{(i||j)}(\mathcal{C}_n^m) + \hat{d}_{(j||i)}(\mathcal{C}_n^m)}{2}$$

and

$$\hat{V}_j(\mathcal{C}_n^m) = \frac{1}{m}\sum_{i=1}^{m} \overline{d}_{(j,i)}(\mathcal{C}_n^m).$$

We will refer to $\overline{d}_{(i,j)}(\mathcal{C}_n^m)$ as being the Symmetrized Relative Entropy (SRE) between the samples $i$ and $j$ from $\mathcal{C}_n^m$. We will also say that $\hat{V}_j(\mathcal{C}_n^m)$ is the mean SRE between the sample $j$ and the other samples in $\mathcal{C}_n^m$.
Now, sort in increasing order the set $\{\hat{V}_j(\mathcal{C}_n^m),\ j = 1, ..., m\}$ and call $j_i^*(\mathcal{C}_n^m)$ the index of the sample in the $i$th position on the ordered set, i.e.

$$j_1^*(\mathcal{C}_n^m) = \arg\min_{j=1,\ldots,m}\left\{\hat{V}_j(\mathcal{C}_n^m)\right\},$$

also

$$j_m^*(\mathcal{C}_n^m) = \arg\max_{j=1,\ldots,m}\left\{\hat{V}_j(\mathcal{C}_n^m)\right\}.$$

**Remark 4** *To evaluate $D\left(\hat{Q}_i || \hat{Q}_j\right)$ it is used Theorem 3 (see [6]), replacing the true probabilities by its empirical estimators and taking by the set of strings, the common tree given by definition 5 using the estimated trees from $\hat{Q}_i$ and $\hat{Q}_j$.*

**Theorem 1** *Under the Assumption 1, if the estimator $\mathcal{S}(\mathcal{C}_n^m)$ is defined as being $\hat{Q}_{j_i^*(\mathcal{C}_n^m)}$ for some natural number $i < m/2$, then, $\mathcal{S}(\mathcal{C}_n^m)$ has asymptotic breakdown point equal to $\frac{1}{2}$.*

(See details of the proof in [6]).
In terms of quality of estimation, we can use Theorem 1 in order to propose a better strategy that can take advantage of the best samples detected by $\left\{\hat{V}_j(\mathcal{C}_n^m)\right\}$ to construct a more powerful estimator for the majority law $Q_{i_0}$.

**Definition 7** *Under the Assumption 1, we define the $\alpha$-trimmed CTM model estimator for $Q$ as being*

$$\hat{Q}^\alpha = CTM\left(\left(x_{j_i^*(\mathcal{C}_n^m),t}\right)_{t=1}^n, i = 1, \ldots, [(1-\alpha)m]\right),$$

*for $\alpha$ such that $[(1-\alpha)m] \geq 1$. Where $[(1-\alpha)m]$ is the integer part of $(1-\alpha)m$.*

**Remark 5** *$\hat{Q}^\alpha$ computes the CTM estimator assuming the selected samples as independent, this means that to compute the occurrences of each string $s$ followed by $a \in A$ will be necessary compute $\hat{Q}(a|s) = \frac{N_n^\alpha(s,a)}{N_n^\alpha(s)}$ with $N_n^\alpha(s) = \sum_{i=1}^{[(1-\alpha)m]} N_n^i(s)$ and $N_n^\alpha(s,a) = \sum_{i=1}^{[(1-\alpha)m]} N_n^i(s,a)$ where $N_n^i$ are the occurrences computed from the sample $\left(x_{j_i^*(\mathcal{C}_n^m),t}\right)_{t=1}^n$. Where each string $s$ comes from the set of feasible trees, with the same restriction as was assumed for equation(2).*

**Theorem 2** *Under the Assumption 1, for $\alpha$ such that $[(1-\alpha)m] \geq 1$. The estimator $\mathcal{S}(\mathcal{C}_n^m)$ defined by $\hat{Q}^\alpha$ has*

*(i) an asymptotic breakdown point equal to $\alpha$, when $\alpha \in (0, \frac{1}{2})$ and*

*(ii) an asymptotic breakdown point equal to $\frac{1}{2}$ when $\alpha \in [\frac{1}{2}, 1]$.*

(See details of the proof in [6]).

**Remark 6** *If $\alpha = (1 - \frac{1}{m})$ the estimator is given by $\hat{Q}_{j_1^*(\mathcal{C}_n^m)}$, i.e. the most representative empirical law, because the sample $(x_{j_1^*(\mathcal{C}_n^m),t})_{t=1}^n$ would be considered the most representative in terms of the mean SRE.*

## 6. CONCLUSION

In this paper we introduce a strategy of robust estimation to estimate the majority law from a collection of samples coming from VLMC processes. That strategy takes advantage from the convergence "almost surely" guaranteed by the CTM algorithm, but it is not restricted to this algorithm and can be applied using other algorithms of estimation. From a practical point of view, the strategy takes advantage also from the structure of trees (of VLMC), because the structure of tree allows to express the relative entropy between two processes in terms of the conditional probabilities. Using a very convenient structure of tree, that is a composition between the trees of the two processes (from [6]) the strategy can be formulated as a precise calculus between the empirical probability laws. The strategy achieves the best level of robustness, that is at most 50% of contamination. In addition, the strategy reveals how to improve the estimation, doing to grow the number of samples used for it, with the selection of the best samples to do the estimation.

## 7. ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support for this research provided by CNPq's projects 485999/2007-2 and 476501/2009-1 and USP project "Mathematics, computation, language and the brain".

## 8. REFERENCES

[1] P. Buhlmann and A. Wyner, *Ann. Statist.* **27**, 480 (1999).

[2] I. Csiszár and Z. Talata, *IEEE Trans. Inform. Theory,* **52**, 1007 (2006).

[3] P. Collet, A. Galves and F. Leonardi, *Electronic Journal of Probability.* **13**, 1345 (2008).

[4] J. Rissanen, *IEEE Trans. Inform. Theory,* **29**(5) 656 (1983).

[5] A. Galves, C. Galves, J. Garcia, N. L. Garcia and F. Leonardi, *Annals of Applied Statistics,* **6**(1) 186 (2012).

[6] J. E. García, V. A. González-López and M.L.L. Viola, *Robust model selection for finite memory stochastic processes.* Submitted.