様式Ｆ－１９

科学研究費助成事業（学術研究助成基金助成金）研究成果報告書

平成　25 年　6 月　5 日現在

機関番号：14301
研究種目：若手研究(B)
研究期間：2011　～　2012
課題番号：23700172
研究課題名（和文）　カーネル法に基づく複雑な時系列の分析手法の開発
研究課題名（英文）　New approaches for the analysis of complex time-series using kernel methods.

研究代表者
クトゥリ　マルコ　（Cuturi Marco）
京都大学・情報学研究科・准教授
研究者番号：80597344

研究成果の概要（和文）：　時系列は現在、ますます複雑化しています。それぞれのデータポイントは、構造化オブジェクト（例えば、画像またはグラフ）または非常に高い次元の特徴ベクトルを記述することが一般である。我々のプロジェクトの目的は、カーネル法と最適化法を介して複雑なデータの時系列を処理するための新しい方法を開発することである。

研究成果の概要（英文）：Time series are now increasingly complex. Each observation may describe a structured object (an image or a graph for instance) or alternatively a very high dimensional feature vector. The goal of our project is to develop new methods to handle time-series of complex data through kernel methods and optimization methods.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
交付決定額	3,300,000	990,000	4,290,000

研究分野：　知能情報学
科研費の分科・細目：　機械学習
キーワード：Machine Learning, Kernel Methods

１．研究開始当初の背景

The subject of analyzing time series through kernel methods was relatively underdevelopped at the time this project started. Most techniques would only consider distances for multivariate time-series, and few would consider their positive definiteness. In particular, the interplay between the way kernels on single observations could be used to form kernels on time-series of observations had not been explored, apart from our previous work on the Global Alignment Kernel. Additionally, some numerical issues relative to the use of kernels for time-series were also hindering their application to large scale problems, since they were both too slow and too unstable to be of practical use.

２．研究の目的

The goal of this research was to develop kernel methods that could accommodate times-series of large-dimensional data, provide good performances and be computationally effective. In practical terms, our goal was to develop theoretical foundations and practical implementations (computer code) of positive definite kernel functions. These functions, to be attractive to practitioners, must be easily parameterized and intuitive, fast to compute and numerically stable (i.e. their performance does not vary too quickly with parameter setting). We wanted to distribute that software on our website and provide utilization guidelines. Our goal was also to apply modern optimization techniques to provide a new outlook on a long-standing problem in time-series analysis: cointegration. With colleagues, we conjectured that the detection of cointegrated relationships between the different components of a multivariate time series could be studied under the light of semidefinite programming. One of the goal of this project was to investigate further this connection and propose algorithms that could handle this issue in an innovative way.

３．研究の方法

Our contributions have relied on different tools, among which (1) dynamic programming and its generalization from a (min,sum) algebra to a (sum,product) algebra to produce a *soft-minimum*; (2) Bayesian linear regression, to propose closed forms for kernels that average the likelihood under a vector autoregressive model of two time series; (3) semidefinite programming, to propose exact minimizers of problems that are non-convex when studying the properties of time-series.

４．研究成果

We have proposed two novel kernels for time series: triangular global alignment kernels as well as autoregressive kernels. Both offer state-of-the-art performance and were shown to outperform other alternative.

Triangular Kernels

Global alignment kernels were proposed a few years ago, and have been successful in comparing timer series of structured data by reinterpreting the widely used *Dynamic Time Warping* family of distances for time-series.

Triangular global alignment kernels were proposed by noticing that a substantial acceleration at a very low (if nonexistent) cost in performance could be proposed to compute global alignment kernels more efficiently. We have also studied in that work conditions for the numerical stability of global alignment kernels as well as further elements of their theoretical validity by defining the class of "geometrically divisible kernels". Geometrically divisible kernels are positive definite kernels which can be written as an infinite expansion of the powers of another base kernel. This concept may prove useful in the future to study other kernels on structures.

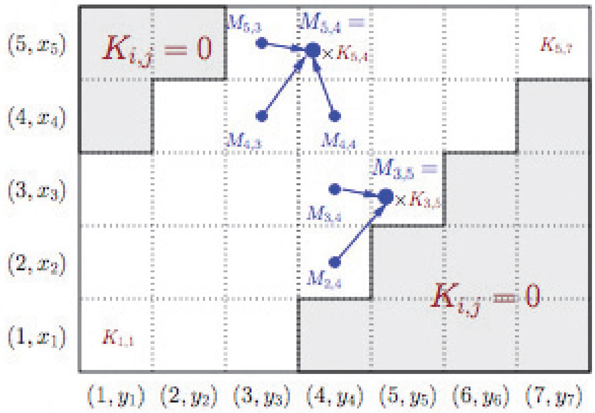

Figure 1: Rather than considering the set of all possible alignments to compare two string $x_1 \ldots x_5$ and $y_1 \ldots y_7$, the triangular global alignment kernels only considers alignments that lie not too far from the diagonal by constraining the set of possible alignments.

Autoregressive Kernels

Autoregressive kernels build upon the idea that, two be similar, it suffices that two time series have simultaneaously high likelihoods for a wide choice of parameters taken in a family of parameterized distributions,

$$k(\mathbf{x}, \mathbf{x}') = \int_{\theta \in \Theta} p_\theta(\mathbf{x})\, p_\theta(\mathbf{x}')\, \omega(d\theta),$$

where, in that work, we focus more explicitly on the Vector Autoregressive Models (VAR), hence the name of the kernels we propose.

AR kernels are extremely fast to compute when considered on high-dimensional multivariate time series and perform comparably to many other kernel functions. A possible extension to time series in reproducing kernel Hilbert spaces considered in that paper might be, despite good experimental results,
false from a theoretical point of view. Indeed, a key result in our paper (infinite divisibility of the inverse generalized variance kernel between distributions) needs to be reconsidered under the light of recent work by S. Sra ("A new metric on the manifold of kernel matrices with application to matrix geometric means", NIPS 2012) and this has delayed the publication of our work so far. We are, however, confident that we will find a proper formulation to correct for that problem in the next months.

Experimental results which illustrate the interest of this approach are provided in Fig.2 below. In that figure we can see that the autoregressive kernels perform at least as well as many other alternatives over many different parameter settings. The version of our kernel that is not kernelized (represented in dark blue in the figures below) is markedly faster than any of the other kernels considered in this benchmark.

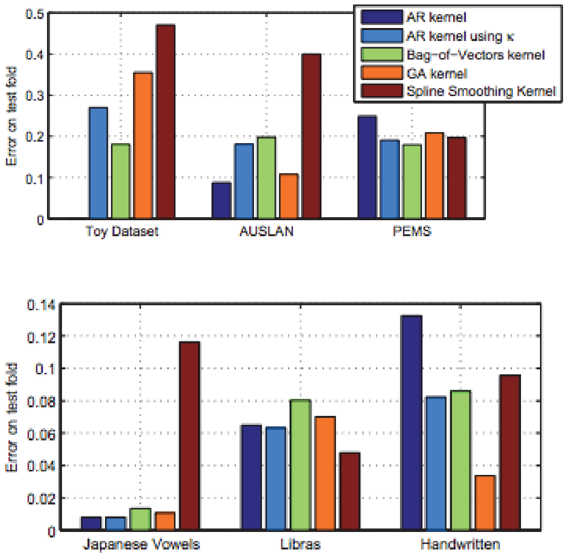

Figure 2: Experimental error rate for the autoregressive kernels on different data sets compared to other choices.

Mean Reversion with a Variance Threshold

We have proposed a new framework to study equilibria (cointegration) between different components of a multivariate time-series, using semi-definite programming.

The problem we have tried to tackle was that of avoiding the detection of cointegrated relationships in a stochastic process that have **a very low variance**. Indeed, from an econometric point of view, it is sufficient to find a vector y such that $(y^T x_t)$ is stationary to obtain a cointegrated relationship.

However, we argue that in many practical settings such a relationship is useless. In particular, in the context of statistical arbitrage of financial assets, finding a perfectly cointegrated relationships is not preferable, since trading with a mean-reversion strategy on such baskests usually incurrs a high transaction cost.

Our method now allows to find a direction y such that $(y^T x_t)$ is stationary but at the same time ensure that the variance of var$(y^T x_t)$ is not too small. Our approach makes use of the $\mathcal{S}$-lemma, that is the ability to solve for non-convex quadratic problems in variable y by solving a semidefinite program in a positive definite matrix Y and then recover an approximate solution by considering the first eigenvalue of Y as a possible solution for y.

This approach has direct applications in finance, but we expect that it can be extended to anomaly detection by estimating mean-reverting functions that can act as alarm functionals when studying a high-dimensional time varying system.

This approach is illustrated in Fig.3 below, where we able to detect *meaningful* cointegrated relationships (that have a sufficient variance) between different assets.

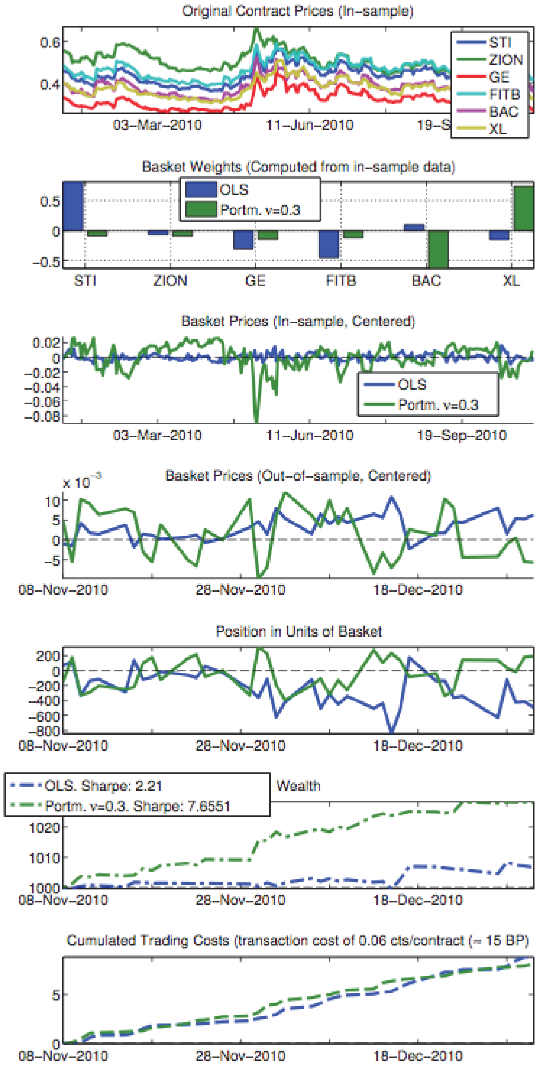

Figure 3: Comparison of a classic cointegration technique (OLS) with our approach (Portmanteau) when selecting a basket of mean-reverting assets. The resulting portfolio is more volatile (third figure from top) and selects different weights (second). This selection is beneficial to record higher gains with lower transaction costs (bottom).

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　２件）

M. Cuturi, A.Doucet, *Autoregressive Kernels for Time Series*, Journal of Machine Learning Research. This article has been accepted pending minor modifications, which we need to take care of to finalize our submission. A mistake in one of our proofs has delayed the final publication of this paper.

M. Cuturi, A. d'Aspremont, Mean Reversion with a Variance Threshold, International Conference on Machine Learning 2013, JMLR W&CP 28(3):271-279, 2013.

〔学会発表〕（計　２件）

M. Cuturi, Fast Global Alignment Kernels, Proceedings of the International Conference on Machine Learning 2011.

〔図書〕（計　０件）

〔産業財産権〕
○出願状況（計　０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　０件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://www.iip.ist.i.kyoto-u.ac.jp/member/cuturi/GA.html

http://www.iip.ist.i.kyoto-u.ac.jp/member/cuturi/AR.html

６．研究組織
(1)研究代表者
クトゥリ　マルコ　（Cuturi Marco　）
研究者番号：80597344