様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 １０ 日現在

機関番号：１４４０１
研究種目：研究活動スタート支援
研究期間：2019～2022
課題番号：１９Ｋ２３１９３

研究課題名（和文）Parsimonious statistical modelling for high-dimensional problems

研究課題名（英文）Parsimonious statistical modelling for high-dimensional problems

研究代表者
ＰＯＩＧＮＡＲＤ　ＢＥＮＪＡＭＩＮ（POIGNARD, BENJAMIN）

大阪大学・大学院経済学研究科・准教授

研究者番号：４０８４５２５２

交付決定額（研究期間全体）：（直接経費）　　1,600,000円

研究成果の概要（和文）：本研究では、高次元多変量モデルにおける少数のパラメータと柔軟な表現力を持つパラメータ数のバランスに対して、非零の値が推定されるパラメータが自動的に疎になるよう、新しい推定法の開発と理論的性質の解明を行った。多変量GARCHモデル、因子分析とコピュラモデルにスパース推定を適用した。また実データによる予測力の評価、シミュレーションによる検証を通して提案手法の有用性を明らかにした。

研究成果の学術的意義や社会的意義
The curse of dimensionality is the main drawback inherent to most multivariate models due to the explosive number of parameters. The research main purpose was to fix this curse, provide methods to efficiently model high-dimensional vectors and improve the prediction performances.

研究成果の概要（英文）：The research was devoted to the sparse modelling of multivariate models and to the development of statistical methods to fix the curse of dimensionality. The sparse approach aimed to improve the precision of the M-estimators and to improve the prediction performances. Three multivariate models were under consideration: multivariate stochastic volatility models (financial econometrics literature); factor models; copula models. For each of these models, we specified a sparsity-based estimation framework, derived the corresponding theoretical properties (finite/large sample properties) and illustrated the relevance of the proposed method through numerical experiments. In particular, the specification of a suitable M-estimation criterion was key to allow for fast-solving implementation methods. We could apply the sparse modelling to high-dimensional random vectors (e.g., financial data) and obtain better out-of-sample performances compared to non-sparse methods.

研究分野：Econometrics/Statistics

キーワード：Copulas　Factor models　Financial econometrics　Sparsity

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

High dimensionality in parametric models relates to the explosive number of parameters to estimate. This is a problem inherent to most multivariate econometric models, where the modelling of a large vector of variables often induces a significantly large number of parameters to consider: this is the so-called “curse of dimensionality”. Copula modelling, factor models, or, in financial time series, the specification of variance-covariance matrix processes within the family of Multivariate GARCH (MGARCH) or Multivariate Stochastic Volatility (MSV) are illustrative cases of such a curse. Sparsity is a way to tackle this problem as it aims to find the optimal balance between a richly parameterized model and, yet, parsimonious enough to avoid over-fitting issues, among others. In parametric models, sparsity refers to the assumption that the vector of parameters has zero entries. At the estimation stage, a penalty function is applied to the loss function to foster sparsity in the vector of parameters. This is a penalized M-estimation problem. An alternative approach to sparsity is the sure independent screening (SIS) viewpoint, which aims to discover a majority of inactive features for describing a target variable.

２．研究の目的

The research centers on four topics: sparse variance-covariance matrix processes; sparse approximate factor models; sparse copula models; feature screening methods.

(1) Sparse variance-covariance matrix processes: The specification of covariance processes for large-dimensional vectors that provides the optimal trade-off between parsimony and a sufficiently parameterized model to capture complex relationships while allowing for fast-solving optimization procedures under the positive-definite constraint is the main purpose of this research topic. The variance-covariance processes under consideration are based on the MGARCH and MSV families. The analysis of the asymptotic properties of the M-estimator of the sparsity-based variance-covariance processes is another important topics of interest.

(2) Sparse approximate factor models: Factor models are a way to summarize the information from large data sets through a small number of variables called factors. The variance-covariance matrix of the vector of observations is deduced from the summation of the cross-product of the factor loading matrix and the variance-covariance of the idiosyncratic errors. The latter is not required to be diagonal, which enables the idiosyncratic errors to be cross-sectionally correlated. The objective of this topic is to model sparsity with respect to the variance-covariance of the idiosyncratic errors and to provide the theoretical properties (finite sample) of the corresponding penalized M-estimator.

(3) Sparse copula models: Copulas are a standard way of modelling the joint distribution of a random vector. They are flexible in that they allow a separate modelling between the dependence structure and the marginal distributions of the vector components. Fully parametric copula-based models can be estimated by assuming parametric models for both the copula and the marginals and then performing maximum likelihood estimation. As an alternative, the empirical cumulative distribution of each margin (the pseudo-observations) can be plugged at the maximization step of the likelihood function. Within the latter semi-parametric framework, the copula model complexity is under consideration: the copula parametrization may require the estimation of many parameters, as in the Gaussian copula, conditional copulas, or mixture of copulas. The study of the finite-sample properties of the M-estimator in the presence of pseudo-observations is the main focus for this topic.

(4) Feature selection: An inactive feature refers to the situation where the distribution of the target variable does not depend on this feature, given the other features. To screen out this set of inactive features, and without assuming a specific underlying relationship between the features and the target variable (e.g., linear model), a measure of importance is applied between each feature and the target variable to quantify their dependence. The purpose of this research topic is to propose novel screening methods to identify this latter set of inactive features and refines the techniques to account for the redundancy problem.

３．研究の方法

Research topic (1): Within the MGARCH family, rather than modelling a GARCH dynamic, which can be estimated by quasi maximum likelihood only, we rely on MARCH type processes, in which case we can apply the Ordinary Least Squares (OLS) method. Indeed, the latter uses the autoregressive representation on the squares of the observed process. This idea is applied to different variance-covariance dynamics (Cholesky GARCH, Vector GARCH). A penalized OLS loss function can be specified at the estimation stage, and fast-solving algorithm can be developed. Moreover, the theoretical properties are easier to derive for the latter loss function (convex; no third order term). The MSV family offers an alternative specification for variance-covariance processes. But their estimation techniques such as Bayesian MCMC typically suffer from the curse of dimensionality. Specifying the MSV model as a multivariate state space model, we can carry out a two-step penalized procedure based on OLS loss functions.

Research topic (2): We propose a two-step estimation: we first obtain an estimator of the factor loading matrix under suitable identifiability constraints; given this estimator, we assume sparsity with respect to the variance-covariance matrix of the idiosyncratic errors and thus consider a penalized M-estimation criterion, where the non-penalized loss function is a Gaussian QML and, alternatively, a least squares loss.

Research topic (3): We specify a general penalized M-estimation criterion and consider the finite sample properties of the corresponding sparse M-estimator. These properties are established under regularity conditions for both the non-penalized loss function, which should satisfy the restricted strong convexity property, and the penalty function, which should satisfy the so-called “amenable” condition.

Research topic (4): We devise new association measures to perform feature screening (mixture of Kendall’s tau and Spearman’s rho, Maximum Mean Discrepancy). Moreover, to improve the performances of the screening procedure (in terms of correct identification of the inactive features), we consider the so-called redundancy problem (correlated features are selected while they are inactive) using the minimum redundancy maximum relevance (mRMR). Using the continuous version of the mRMR problem, we can perform feature screening while tackling the redundancy issue through a penalized M-estimation criterion, which is an OLS loss penalized by convex/non-convex penalty functions. The continuous version of the mRMR problem with penalization refines the standard screening approach by including the feature-feature relationship in the selection process.

４．研究成果

Research topic (1): We derive the asymptotic properties of the penalized M-estimators of the MARCH process and of the proposed two-step MSV model. We provide the conditions for which the sparsity-based estimator satisfies the oracle property. The corresponding variance-covariance matrix processes are applied to high-dimensional vector of observations (over a hundred variables), where the penalized OLS loss function can be efficiently optimized. The prediction accuracy of the variance-covariance models is tested (out-of-sample analysis using the Diebold-Mariano test and Model Confidence Set) through portfolio allocation experiments (global minimum variance) based on financial data. The empirical results show that: the sparse MARCH processes provide better out-of-sample performances than alternative standard MGARCH models (e.g., scalar DCC); the two-step MSV model based on the state space representation outperforms MGARCH dynamics; the sparse approach provides a gain in terms of prediction performances over non-sparse dynamics.

Research topic (2): We show some finite sample consistency results for the two-step sparse approximate factor model estimators and provide the conditions for support recovery, that is the correct identification of the non-zero coefficients of the variance-covariance matrix of the idiosyncratic errors. The conditions for consistency and recovery highly depend on the regularity of the loss function under consideration. Our approach is compared to alternative estimation techniques (POET, PML estimators).

Research topic (3): After deriving the finite sample error bounds (l1 and l2 norms) for general

loss functions and penalty functions (convex and non-convex), we verify that the sparse M-estimators deduced from the semi-parametric Gaussian copula, Archimedean copulas and mixture of copulas satisfy these bounds. The "informativeness" of the theoretical upper bounds on the errors highly depend on the regularity of the loss function through the restricted strong convexity coefficients and on the level of non-convexity of the penalty function (SCAD and MCP).

Research topic (4): We establish the SIS property based on different association measures. As for the screening approach based on the sparse mRMR with HSIC as the association measure, we prove the large sample properties of the corresponding penalized estimator. The real data experiments based on UCI machine learning data sets suggest that the sparse mRMR performs better in terms of classification accuracy, i.e. correct classification, than screening methods which do not take the redundancy issue into account.

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　7件／うち国際共著　7件／うちオープンアクセス　4件）

１．著者名 Poignard Benjamin、Yamada Makoto	４．巻 108
２．論文標題 Sparse Hilbert-Schmidt Independence Criterion regression	５．発行年 2020年
３．雑誌名 Proceedings of Machine Learning Research, AISTATS 2020	６．最初と最後の頁 538～548
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Poignard Benjamin、Terada Yoshikazu	４．巻 14
２．論文標題 Statistical analysis of sparse approximate factor models	５．発行年 2020年
３．雑誌名 Electronic Journal of Statistics	６．最初と最後の頁 3315～3365
掲載論文のDOI（デジタルオブジェクト識別子） 10.1214/20-EJS1745	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Freidling Tobias、Poignard Benjamin、Climente-Gonzalez Hector、Yamada Makoto	４．巻 139
２．論文標題 Post-selection inference with HSIC-Lasso	５．発行年 2021年
３．雑誌名 Proceedings of Machine Learning Research, ICML 2021	６．最初と最後の頁 3439～3448
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

１．著者名 Poignard Benjamin、Fermanian Jean-David	４．巻 40
２．論文標題 High-dimensional penalized arch processes	５．発行年 2021年
３．雑誌名 Econometric Reviews	６．最初と最後の頁 86～107
掲載論文のDOI（デジタルオブジェクト識別子） 10.1080/07474938.2020.1761153	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Poignard Benjamin、Asai Manabu	4．巻 44
2．論文標題 High‐dimensional sparse multivariate stochastic volatility models	5．発行年 2022年
3．雑誌名 Journal of Time Series Analysis	6．最初と最後の頁 4～22
掲載論文のDOI（デジタルオブジェクト識別子） 10.1111/jtsa.12647	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

1．著者名 Poignard Benjamin、Naylor Peter、Climente-Gonzalez Hector、Yamada Makoto	4．巻 151
2．論文標題 Feature screening with kernel knockoffs	5．発行年 2022年
3．雑誌名 Proceedings of Machine Learning Research, AISTATS 2022	6．最初と最後の頁 1935～1974
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 該当する

1．著者名 Poignard Benjamin、Fermanian Jean-David	4．巻 74
2．論文標題 The finite sample properties of sparse M-estimators with pseudo-observations	5．発行年 2022年
3．雑誌名 Annals of the Institute of Statistical Mathematics	6．最初と最後の頁 1～31
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10463-021-00785-4	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計9件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 Benjamin POIGNARD
2．発表標題 Feature screening with kernel knockoffs
3．学会等名 AISTATS 2022 conference
4．発表年 2022年

１．発表者名
Benjamin POIGNARD

２．発表標題
Sparse Factor Models: Asymptotic Properties

３．学会等名
CFE-CMStatistics 2021 Conference

４．発表年
2021年

１．発表者名
Benjamin POIGNARD

２．発表標題
Sparse Factor Models: Asymptotic Properties

３．学会等名
Ecodep Conference 2021

４．発表年
2021年

１．発表者名
Benjamin POIGNARD

２．発表標題
Sparse Hilbert-Schmidt Independence Regression Criterion

３．学会等名
AISTATS 2020 conference

４．発表年
2020年

１．発表者名
Benjamin POIGNARD

２．発表標題
High-dimensional Sparse Multivariate Stochastic Volatility models

３．学会等名
Seminar presentation at the 4th Asian Quantitative Finance Seminar

４．発表年
2020年

１．発表者名
Benjamin POIGNARD

２．発表標題
High-dimensional Sparse Multivariate Stochastic Volatility models

３．学会等名
CFE-CMStatistics 2020 Conference

４．発表年
2020年

１．発表者名
Benjamin POIGNARD

２．発表標題
Sparse Hilbert-Schmidt Independence Criterion regression

３．学会等名
Riken AIP - 9th AIP Open Seminar

４．発表年
2021年

１．発表者名
Benjamin POIGNARD

２．発表標題
Statistical analysis of sparse approximate factor models

３．学会等名
SETA International Conference

４．発表年
2019年

１．発表者名
Benjamin POIGNARD

２．発表標題
The finite sample properties of sparse M-estimators with seudo-observations

３．学会等名
EcoSta

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関