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

平成　２９　年　　６　月　１６　日現在

機関番号：　１４４０１
研究種目：　基盤研究(A)（一般）
研究期間：　2013～2016
課題番号：　２５２４００３６

研究課題名（和文）超高次元データ空間における統計的推定・シミュレーション原理の開発と応用展開

研究課題名（英文）Development and Application of Statistical Estimation and Simulation for Super High Dimensional Data Space

研究代表者

鷲尾　隆（Washio, Takashi）

大阪大学・産業科学研究所・教授

研究者番号：００１９２８１５

交付決定額（研究期間全体）：（直接経費）　　35,200,000円

研究成果の概要（和文）：本研究では、(1)超高次元にロバストな統計的推定・シナリオ生成の一般的原理、(2)超高次元データからの統計的推定手法、(3)超高次元状態空間における確率的シナリオ生成手法、(4)開発推定手法・シミュレーション手法の応用、(5)新たな国際的研究コミュニティーの構築を目指した。
この結果、類似性尺度や密度の評価、ロバスト推定、シナリオ探索、検索・クラスタリング、分類、異常検知、希少シナリオ高効率生成、高頻出パターン導出規則抽出の手法を開発した。そしてこれらを適用した生体高分子の運動シミュレーション手法を開発した。研究活動においては、国際会議２件、国際ワークショップ・セミナー７回を開催した。

研究成果の概要（英文）：In this study, we aimed to develop (1) generic and robust principles of statistical estimation and scenario generation against super high dimensionality, (2) statistical estimation methods using super high dimensional data, (3) probabilistic scenario generation methods for super high dimensional space, (4) an application of these developed methods and simulation techniques, and (5) an international research community.
Throughout this project, we developed techniques of similarity measure, density evaluation, robust estimation, scenario search, retrieval and clustering, classification, anomaly detection, rare scenario generation, and frequent pattern derivation. We also organized two international conferences and seven international workshops/seminars.

研究分野：　機械学習

キーワード：　超高次元データ　機械学習　データマイニング　人工知能　次元の呪い　シミュレーション　希少事象

１．研究開始当初の背景

近年のユビキタスセンシングやネットワーク、実験測定、並列計算の技術進歩により、様々な社会・科学分野でビッグデータの利用や大規模確率シミュレーションが行われている。特に各観測事例やシミュレーション状態がデータ・状態空間の多くの次元に亘って分布し、見かけの表現次元のみならず「本質次元」の高い超高次元ベクトルで表される大規模・複雑な対象が増えている。例えば、ネットクチコミ拡散は種々の非線形効果と確率的揺らぎにより広範な状態空間に分布する拡散過程である。また、地球上各点の同時大気状態も多数のカオス挙動を含み広範な状態空間に分布する。このような高い本質次元のデータ・状態の下での高精度、ロバストな統計的推定やシナリオ生成は、大規模データ解析、シミュレーションの基礎である。しかし、一定の性能を大数の法則で達成するには$O(r^{d/4+1})$ (dは本質次元数,r>>1は定数)の大量データ収集や状態計算が必要となり[1]、数十次元以上ではこの「次元の呪い」効果で性能が大きく損なわれる[2,3]。

また世界的に特異値分解等の次元圧縮法が研究され[4]、特殊な高次元データ分布ではパラメータ適応による推定手法[5]も提案されているが、何れも高本質次元問題には適用困難である。これに対し研究代表者等は科研（特定）情報爆発公募研究(課題番号18049052,19024048)、科研基盤(B)(課題番号22300054)を通じて次元の呪い効果を分析し[6]、その結果、同一中心の２つの超球間の体積が内側超球体積の$(1+\varepsilon/r)^{d-1}$ 倍であることから、殆どの超高次元データ・状態が中心から$[r,r+\varepsilon]$の距離に分布する「球面集中効果」、各次元に独立同一分布する $p(x)^d$ 分布のように超高次元空間の局所に確率が集中する「確率密度集中効果」、各次元に一定の標準偏差σで分布するデータや状態の存在体積が$(2\sigma)^d$ に比例するように超高次元データ・状態が広大な体積内に分布する「スパース化効果」等を特徴付けし、前者２効果を打ち消す人工的歪みをデータ・状態分布に与える高精度、ロバストな推定法を提案した[7,8]。３つ目の効果には国際共同研究によりランダム部分標本推定を多数重ねる高精度、ロバストなアンサンブル推定法を提案した[9]。これらが一定性能を達するのに必要なデータ数・状態数は事実上 $O(d^{\alpha})(\alpha>0)$に抑えられ大きなブレークスルーとなった。これらは一流国際会議 SDM,ICDM に採択され、更に[9]は ICDM ベスト論文の１つとして国際ジャーナル特集号への掲載が決まった。

しかしこれらの成果では、αが大きくなり数千を超える本質次元で十分な推定が得られない場合がある。これに対し、本研究提案当時、国立情報学研究所の M.Houle 教授等が計算幾何学を基に、次元に依らない空間の性質により球面集中とスパース化にロバストな異常データ検出原理を ICDM に発表しベスト論文賞を取った[10]。また、研究代表者鷲尾と当時オーストラリア・モナシュ大の K.M.Ting 准教授との国際共同研究でも[9]を拡張しスパース化に対する新たな推定原理を得てた。一方、大規模確率シミュレーションも統計的推定と多くの原理を共有し「次元の呪い」問題を抱えるが、統計数理研究所の伊庭准教授は拡張 MCMC 手法で確率密度集中に対抗した特徴的ランダム行列の生成に成功し Physical Review に発表した[11]。

本研究は、これら研究者が協力して上記新原理を深く共通レベルで考察し、本質次元が数千以上の問題解決を行う可能性を見出したことを背景として開始された。

２．研究の目的

近年のビッグデータや大規模シミュレーション出力結果の多くでは、その中の各事例やシナリオが超高次元の属性や変数を含む。しかし、「次元の呪い」によるデータ解析・シナリオ生成の性能低下問題の研究は世界的に遅れている。我々は科研(特定)情報爆発公募や科研基盤(B)を通じ、人工データ分布導入やアンサンブル推定で「本質次元」が数千程度まで統計的推定性能を向上した[4,5,6]。

本研究では更に最新の計算幾何学、モンテカルロ計算と融合し、シナリオ生成も対象に加えより高次元の統計的推定とシミュレーション技術へ拡張を目指した。そして、その技術を生体高分子のシミュレーション、機械学習等、重要大自由度問題に適用展開することを目指した。データマイニング・機械学習の研究代表者に計算幾何学とモンテカルロシミュレーションの研究分担者を加え、４年で以下の研究項目について、原理確立と応用展開を目指した。

(1)超高次元に一層ロバストな統計的推定・シナリオ生成の一般的原理の探求

上記提案者等の新たな原理の関係を整理し、従来困難だった「次元の呪い」問題のブレークスルーを実現する一般的原理を探求・確立する。

(2)上記一般的原理に基づく超高次元データからの統計的推定手法の開発

高本質次元データからの推定、クラスタリング等の高精度、ロバスト手法を確立する。

(3)上記一般的原理に基づく超高次元状態空間における確率的シナリオ生成手法の開発

高本質次元状態空間での確率的シナリオ生成の高精度、ロバストな手法を確立する。

(4)開発推定手法・シミュレーション手法の応用展開

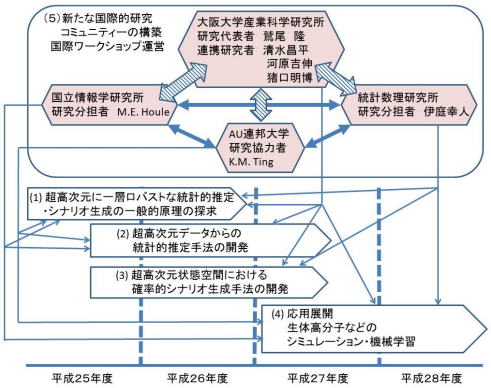
生体高分子など高本室次元シミュレーション等、理科学で重要な大自由度系のデータ解析・シミュレーションにを適用し、スケーラブルで有用な手法ツールを提供する。

(5)新たな国際的研究コミュニティーの構築

提案者等が国際ワークショップ運営し、当該分野の世界的研究活性化と成果普及をリードする。

３．研究の方法

４年間で高本質次元空間での高性能な統計的推定・シナリオ生成の一般的原理確立と手法開発、実問題展開、国際コミュニティー構築を行うため、下図に示すようにデータマイニング、機械学習の研究代表者・連携研究者と、計算幾何学と確率シミュレーションの研究分担者２名、海外研究協力者１名で体制を組んだ。



研究体制図

大阪大学の研究代表者（鷲尾）が４年間を通じ全体統括を行った。(1)超高次元に一層ロバストな統計的推定・シナリオ生成の一般的原理の探求には、データマイニング、機械学習の知見に加えて計算幾何学、確率シミュレーションの知識が必要なことから、データマイニング、機械学習が専門の研究代表者鷲尾とこれまで同分野で国際共同研究をして来た研究協力者であるオーストラリア連邦大学の Ming 准教授、計算幾何学が専門で国立情報学研究所の研究分担者 Houle 教授、確率シミュレーションが専門で統計数理研究所の研究分担者伊庭教授が、これまでの成果を踏まえて密接に協力して取り組んだ。

(2)一般的原理に基づく超高次元データからの統計的推定手法の開発とその応用展開である(4)生体高分子などのシミュレーション、機械学習については、これまでデータマイニングや機械学習における超高次元データからの統計的推定法を研究して来た大阪大学の研究代表者鷲尾と同研究室の連携研究者３名、オーストラリア・モナシュ大学の Ming 准教授、国立情報学研究所の研究分担者 Houle 教授が中心となり取り組んだ。

更に(3)一般的原理に基づく超高次元状態空間における確率的シナリオ生成手法の開発とその応用展開(4)生体高分子などのシミュレーション、機械学習については、超高次元データからの統計的推定法を研究して来た大阪大学の研究代表者鷲尾と同研究室の連携研究者３名、超高次元状態空間での確率シミュレーションを研究して来た統計数理研究所の研究分担者伊庭教授が中心となって取り組んだ。

並行してこれら一連の研究活動を核に、参加者全員で当該分野の(5)新たな国際的研究コミュニティーの構築に取り組み、国際会議及び国際ワークショップの開催に取り組んだ。

４．研究成果

以下、各研究項目について、得られた成果の概要と対応する主な発表論文をまとめる。

(1) 超高次元に一層ロバストな統計的推定・シナリオ生成の一般的原理の探求

過去の提案者等の提案原理の関係を整理し、従来困難だった「次元の呪い」問題のブレークスルーを実現する一般的原理を様々な方向から探求した。まず、ランダムサンプリングよって次元の呪いの影響を軽減しつつ、高本質次元空間上の事例の類似性やデータ密度を機械学習手法に反映する基礎原理を探求した。その結果、種々の類似性尺度や密度評価方法を確立することができた[雑誌論文⑬,学会発表⑧,⑩,⑪]。また、密度推定を行う上で外れ値などの外乱の影響を効果的に排除する基礎原理として Density Power Divergence を導入する方法を確立した[雑誌論文⑩,⑪]。さらに、高本質次元空間でのモンテカルロ・シミュレションによって目的シナリオを効果的に探索する基礎原理の探求を行った[学会発表⑫]。並行して、非常に多数のデータ変数間の決定関係を解析する統計的因果推論に関しても、新しい原理の見通しを得た[雑誌論文⑫, 学会発表⑨]。

(2)上記一般的原理に基づく超高次元データからの統計的推定手法の開発

上記(1)の基礎原理をさらに拡張し、高本質次元データからの検索・クラスタリング、分類、異常検知などの各種統計的推定、機械学習の手法を確立した。検索・クラスタリングについては、高本質次元データに適した類似性尺度を用いた手法の開発を行った[雑誌論文③,⑥,⑦,学会発表⑥,⑦]。分類については、半空間 mass という高本質次元データの密度分布推定に基づく手法の開発を行った[雑誌論文②,⑤,学会発表③]。また、異常検知については、k-近傍探索やデータ mass を使った高本質次元データに効果的な手法を開発し、その性質を明らかにした[雑誌論文①,④,学会発表②]。

(3)上記一般的原理に基づく超高次元状態空間における確率的シナリオ生成手法の開発

高本質次元状態空間での希少な確率的シナリオの高効率生成について、Multi-canonical MCMC 手法の拡張を行った[雑誌論文⑧]。また、非常に大規模で高本質次元データから高頻出なパターンやシナリオ導出規則を抽出する手法について、統計的サンプリングによるものを開発した[学会発表①,⑤]。

(4)開発推定手法・シミュレーション手法の応用展開
以上、(1),(2),(3)で得られた原理の一部を適用して、高本質次元の自由度を有する生体高分子の確率動力学的シミュレーションを行い、生体反応の希少シナリオを生成するスケーラブルな手法を開発した[雑誌論文⑨,学会発表④]。

(5)新たな国際的研究コミュニティーの構築
研究代表者、研究分担者、連携研究者、研究協力者が中心となり、２回の国際会議
SISAP2015: The 8th International Conf. on Similarity Search and Applications
SISAP2016: The 9th International Conf. on Similarity Search and Applications
とデータ本質次元に関する２回の国際ワークショップ・セミナー
NII Seminar Series on Dimensionality and Scalability, 2014
Dimensionality and Scalability II: Hands-On Intrinsic Dimensionality, 2015
さらに確率的シナリオシミュレーションに関する５回の国際ワークショップ・セミナー
Symposium on Rare Event Sampling and Related Topics, 2014
Symposium on Rare Event Sampling and Related Topics II, 2015
Symposium on Rare Event Sampling and Related Topics III, 2015
Workshop: Topics in Advanced Monte Carlo Methods, 2016
Simulations Encounter with Data Science -- Data Assimilation, Emulators, Rare Events and Design, 2017
を開催し、当該研究分野の国際研究コミュニティを創成した。今後も、さらなる国際会議やワークショップ・セミナーを、東アジアと欧米、オセアニア各地域の研究者持ち回りで開催していく予定である。

参考文献

[1] B.W.Silverman, Density estimation for statistics and data analysis, Chapman and Hall (1986)
[2] C.M.Bishop, Pattern recognition and machine learning, Springer (2006)
[3] C.Snyder, et al., Obstacles to high-dimensional particle filtering, Monthly Weather Rev., 136(12), pp.4629-4640 (2008)
[4] S.Michael et al., Using linear algebra for Intelligent Information Retrieval, Tech. Rep.: UT-CS-94-270, Univ. of Tennessee (1994)
[5] H.Liu et al., Sparse non-parametric density estimation in high dimensions using the rodeo, Proc. of Artif. Intel. and Stat. (AISTATS07) (2007)
[6] 鷲尾隆，情報爆発時代の高次元データマイニング，電子情報通信学会誌,94(8), pp.679-683 (2011)
[7] K.Kido, H.Kuwajima and T.Washio, A range query approach for high dimensional euclidean space based on EDM estimation, Proc. of SIAM Data Mining Conf. (SDM08), pp.387-398 (2008)
[8] N.V.Phuong, T.Washio and T.Higuchi, A new particle filter for high-dimensional state-space models based on intensive and extensive proposal distribution, Int. J. Knowledge Eng. and Soft Data Paradigms, 2(4), pp.284-311 (2010)
[9] K.M.Ting, T.Washio, J.Wells and T.Liu, Density estimation based on mass, Proc. of IEEE Int. Conf. on Data Mining (ICDM11), pp.715-724 (2011)
[10] T.de Vries, S.Chawla and M.E.Houle, Finding local anomalies in very high dimensional space, Proc. of IEEE Int. Conf. on Data Mining (ICDM10), pp.128-137 (2010)
[11] N.Saito, Y.Iba and K.Hukushima, Multicanonical sampling of rare events in random matrices, Phys. Rev. E, 82, pp.031142 (2010)

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 13 件）

① Kai Ming Ting, Takashi Washio, Jonathan R. Wells and Sunil Aryal, Defying the gravity of learning curve: a characteristic of nearest neighbor anomaly detectors, Machine Learning, Vol.106, pp.55-91, 2017, DOI:10.1007/s10994-016-5586-4(査読有)

② Guillaume Casanova, Elias Englmeier, Michael E. Houle, Peer Kroger, Michael Nett and Arthur Zimek, Dimensional Testing for Reverse k-Nearest Neighbor Search, The VLDB Endowment, Vol.10, No.7, pp.769-780, 2017, DOI: 10.14778/3067421.3067426 (査読有)

③ Michael E. Houle, Xiguo Ma, Vincent Oria and Jichao Sun, Efficient Similarity Search within User-Specified Projective Subspaces, Information Systems, Vol.59, pp.2-14, 2016, DOI: 10.1016/j.is.2016.01.008 (査読有)

④ Guilherme O. Campos, Arthur Zimek, Jorg Sander, Ricardo J. G.B. Campello, Barbora Micenkova, Erich Schubert,

Ira Assent and Michael E. Houle, On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study, Data Mining and Knowledge Discovery, Vol.30, No.4, pp.891-927, 2016, DOI: 10.1007/s10618-015-0444-8 **(査読有)**

⑤ Bo Chen, Kai Ming Ting, Takashi Washio and Gholamreza Haffari, Half-space mass: a maximally robust and efficient data depth method, Machine Learning, Vol.100, pp. 677–699, 2015, DOI: 10.1007/s10994-015-5524-x **(査読有)**

⑥ Michael E. Houle, Xiguo Ma and Vincent Oria, Effective and Efficient Algorithms for Flexible Aggregate Similarity Search in High Dimensional Spaces, IEEE Transactions on Knowledge and Data Engineering, Vol.27, No.12, pp.3258-3273, 2015, DOI: 10.1109/TKDE.2015.2475740 **(査読有)**

⑦ Michael E. Houle and Michael Nett, Rank-based similarity search: Reducing the dimensional dependence, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.37, No.1, pp.136-150, 2015, DOI: 10.1109/TPAMI.2014.2343223 **(査読有)**

⑧ Yukito Iba, Nen Saito and Akimasa Kitajima, Multicanonical MCMC for Sampling Rare Events: An Illustrative Review, Annals of the Institute of Statistical Mathematics, Vol.66, No.3, pp.611-645, 2014, DOI: 10.1007/s10463-014-0460-2 **(査読有)**

⑨ **伊達 幸人, 藤崎 弘士, 松永 康佑, 生体高分子の揺らぎとダイナミクスーシミュレーションと実験の統計解析, 統計数理**, Vol.62, No.2, pp.163-170, 2014**(査読無)**

⑩ Yasuhiro Sogawa, Tsuyoshi Ueno, Yoshinobu Kawahara and Takashi Washio, Active learning for noisy oracle via density power divergence, Neural Networks, Vol.46, pp.133-143, 2013, DOI: 10.1016/j.neunet.2013.05.007 **(査読有)**

⑪ **十河 泰弘, 植野 剛, 河原 吉伸, 鷲尾 隆**, Density Power Divergence **を用いたロバスト能動回帰学習, 人工知能学会論文誌**, Vol.28, No.1, pp.13-21, 2013, DOI: 10.1527/tjsai.28.13**（査読有）**

⑫ Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvarinen and Takashi Washio, ParceLiNGAM: A Causal Ordering Method Robust Against Latent Confounders, Neural Computation, Vol.26, No.1, pp.57-83, 2014, DOI: 10.1162/NECO_a_00533 **(査読有)**

⑬ Jonathan R. Wells, Kai Ming Ting and Takashi Washio, LiNearN: A new approach to nearest neighbour density estimator, Pattern Recognition, Vol.47, No.8, pp.2702–2720, 2014, DOI: 10.1016/j.patcog.2014.01.013 **(査読有)**

〔学会発表〕（計 20 件）

① **馬場 祥人, 杉山 麿人, 鷲尾 隆, サンプリングを用いた精度保証つき頻出パターンマイニング, 第 30 回人工知能学会全国大会**, 2016 **年** 06 **月** 06 **日**～2016 **年** 06 **月** 09 **日, 北九州市国際会議場, 福岡県, (国内学会)**

② Takashi Washio, Defying the Gravity of Learning Curves: Are More Samples Better for Nearest Neighbor Anomaly Detectors?, The 9th International Conference on Similarity Search and Applications (SISAP2016), 2016 **年** 10 **月** 24 **日**～2016 **年** 10 **月** 26 **日**, Tokyo, Japan **（招待講演）（国際会議）**

③ Bo Chen, Kai Ming Ting, Takashi Washio and Gholamreza Haffari, Half-space Mass: A maximally robust and efficient data depth method, Machine Learning and Knowledge Discovery in Databases(ECML /PKDD2015), 2015 **年** 09 **月** 07 **日**～2015 **年** 09 **月** 11 **日**, Porto, Portugal **（国際会議）**

④ **鷲尾　隆, ビッグデータに基づくモデリング―生体・医療への適用を例として―**, SAS **ユーザー総会** 2015, 2015 **年** 08 **月** 06 **日**～2015 **年** 08 **月** 06 **日, 東京大学伊藤国際学術研究センター, 東京都（招待講演）, (国内学会)**

⑤ **馬場 祥人, 杉山 麿人, 鷲尾 隆, サンプリングを用いた高速頻出パターンマイニング, 第９回人工知能学会全国大会**, 2015 **年** 05 **月** 30 **日**～2015 **年** 06 **月** 02 **日, 函館未来大学, 北海道, (国内学会)**

⑥ Michael E. Houle, Xiguo Ma and Vincent Oria, Flexible Aggregate Similarity Search in High-Dimensional Data Sets, The 8th International Conference on Similarity Search and Applications (SISAP2015)（国際学会）,2015 年 10 月 12 日～2015 年 10 月 14 日, Glasgow, United Kingdom（国際会議）

⑦ Laurent Amsaleg, Oussama Chelly, Teddy Furon, Stéphane Girard, Michael E. Houle, Ken-ichi Kawarabayash and Michael Nett, Estimating Local Intrinsic Dimensionality, The 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2015), 2015 年 08 月 10 日～2015 年 08 月 13 日, Sydney, Australia（国際会議）

⑧ Sunil Aryal, Kai Ming Ting, Jonathan R. Wells and Takashi Washio, Improving iForest with relative mass, The 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2014), 2014 年 05 月 14 日, Tainan, Taiwan（国際会議）

⑨ Patrick Blöbaum, Shohei Shimizu and Takashi Washio, A performance comparison of generative and discriminative models in causal and anticausal problems, Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS2014), 2014 年 04 月 22 日, Reykjavík, Iceland（国際会議）

⑩ Sunil Aryal, Kai Ming Ting, Gholamreza Haffari and Takashi Washio, mp-dissimilarity: A data dependent dissimilarity measure, IEEE International Conference on Data Mining (ICDM2014), 2014 年 12 月 17 日, Shenzhen, China（国際会議）

⑪ Michael E. Houle, Xiguo Ma, Vincent Oria and Jichao Sun, Efficient algorithms for similarity search in axis aligned subspaces, 7th Int. Conf. on Similarity Search and Applications (SISAP2014), 2014 年 10 月 29 日～2014 年 10 月 31 日, Los Cabos, Mexico（国際会議）

⑫ 伊庭幸人, 高柳慎一, 粒子モンテカルロ法による時間逆転シミュレーション, IBIS2014, 2014 年 11 月 18 日, 名古屋工業大学, 愛知県名古屋市, (国内学会)

〔図書〕（計 1 件）
鷲尾 隆, エヌ・ティ・エス社, ビッグデータ・マネージメント　データサイエンティストのためのデータ利活用技術と事例　第 2 章第 4 節ビッグデータからのモデリング手法, 2014 年 11 月

〔その他〕
ホームページ等
http://www.ar.sanken.osaka-u.ac.jp/

６．研究組織

(1)研究代表者
鷲尾　隆（WASHIO, Takashi）
大阪大学・産業科学研究所・教授
研究者番号：00192815

(2)研究分担者
伊庭　幸人（IBA, Yukito）
統計数理研究所・モデリング研究系・教授
研究者番号： 30213200

マイケル フール（Michael E.Houle）
国立情報学研究所・大学共同利用機関等の部局等・客員教授
研究者番号：90399270

(3)連携研究者
清水 昌平（SHIMIZU, Shohei）
滋賀大学・データサイエンス学部・准教授
研究者番号：10509871

河原 吉伸（KAWAHARA, Yoshinobu）
大阪大学・産業科学研究所・准教授
研究者番号：00514796

猪口 明博（INOGUCHI, Akihiro）
関西学院大学・理工学部・准教授
研究者番号：70452456

(4)研究協力者
Kai Ming Ting
Federation University Australia・Faculty of Science and Technology・Professor