CK-002

秘匿信号領域における LASSO モデリングのための ランダムユニタリ変換の秘匿性強化
Security improvement for LASSO modeling on encryption domain

坂東幸浩†，仲地孝之‡
Yukihiro BANDOH†, Takayuki NAKACHI‡

1 はじめに

近年、エッジ／クラウドコンピューティングはビッグデータ解析の計算リソースとして急速に普及している [1]。その解析対象は、音声・映像等のメディア信号から商品取引情報等の経済データ、臨床結果等の医療データまで多岐に渡る [2]。

しかし、取得データの個人特定に繋がる可能性のあるデータは、プライバシー保護の観点から、エッジ／クラウドコンピューティングの利用は制限される。例えば、臨床検査結果、購買履歴、移動経路などがあげられる。こうしたデータに対しては、データを取得した組織・機関に閉じて利用される。大量のユーザを抱え、所望の規模のデータを取得可能な場合は、問題ない。しかし、医療機関における臨床データのように、各機関で取得可能なデータ数が限られている場合、各機関に閉じた分析では、十分な分析精度を得られない場合がある [3]。問題の原因は、取得されたデータが分散しており、集約できない点にある。

こうした問題を解決する方法の一つとして、データを暗号化した状態で計算可能な秘密計算が研究されている。秘密計算は、一般にマルチパーティプロトコルや準同型暗号に基づき実行される [4] [5] 。しかし、秘密計算は、除算の困難性、計算効率および計算精度に課題がある。このため、その適用は、ソーティング処理や幾つかの統計処理に限定されており、十分な普及にはいたっていない。

これに対して、ランダムユニタリ変換に基づく秘匿計算が提案されている [6] 。この秘匿計算は、準同型暗号やマルチパーティプロトコルと比較して高速な演算が可能であり、さらに、スパース信号表現 [7]、画像圧縮 [8] 等の広く普及した信号処理アルゴリズムと併用可能である。こうした特性を活かして、秘匿領域における信号処理アルゴリズム（例：秘匿領域におけるスパース表現のための辞書学習 [9]、秘匿領域における画像信号圧縮 [10] 等）が研究されてきた。上述の研究が単一拠点内に閉じた秘匿化であったのに対し、ランダムユニタリ変換による秘匿化の機能拡張として、分散した拠点において情報を秘匿化する分散秘匿化が検討されている。[11] では、LASSO[12] による分析モデル構築を対象とし、分散秘匿化が LASSO 解を保全する理論的保証を与えている。つまり、各拠点において個別に秘匿化されたデータを集約し、LASSO 解を求めたとしても、秘匿化前の原データと同一の LASSO 解が導出可能であることを保証している。この結果、分散秘匿化されたデータに対する直接的な分析が可能となった。

しかし、ランダムユニタリ変換による分散秘匿化を多様な条件下で利用する場合、秘匿化強度に関する問題が残っていた。ランダムユニタリ変換の秘匿性強度は秘匿化対象のデータ数に依存する。このため、各拠点において取得できるデータ数が少なくなると、同変換の秘匿性強度が低下するのである。拠点毎に十分なデータを取得できないという課題を解決するために導入した分散秘匿化において、少データ数の場合の秘匿性強度の低下は、看過できない問題となる。

そこで、本稿では、データ数の減少に伴い秘匿性強度が低下するランダムユニタリ変換に対して、取得したデータ数によらず秘匿化強度を保持することを目的とし、秘匿化強度の強化法を提案する。提案法は、高次元空間への秘匿信号の埋め込みを通して秘匿性の強化を図る。この埋め込みは、ランダムユニタリ変換の次元拡大と摂動情報の付加から構成される。さらに、高次元空間へ埋め込まれた秘匿信号領域において、秘匿化前の原信号の LASSO 解を求めるためには、コスト関数を適切に設定する必要があることを示す。提案法により、データの機密性は確保した上で、集約したデータを利用した大規模な分析が可能となるため、分散取得されたきたプライバシー保護が必要なデータに対しても、分析精度の向上が実現される。

2 秘匿化信号に対するスパースモデリング

提案法の説明に先立ち、提案法のベースとなるランダムユニタリ変換を用いた秘匿信号領域でのスパースモデリングについて概説する。

まず、対象とするスパースモデリングの定式化について述べる。観測ベクトル $\boldsymbol{y} = (y_0, \cdots, y_{n-1})^T \in \mathbb{R}^n$ を p 本の特徴ベクトル $\boldsymbol{x}_j = (x_{0,j}, \cdots, x_{n-1,j})^T \in \mathbb{R}^n$, $(j = 0, \cdots, p-1)$ の線形和で表現することを考える。特徴ベク

†日本電信電話株式会社 NTT メディアインテリジェンス研究所
‡日本電信電話株式会社 NTT 未来ねっと研究所

トル $\boldsymbol{x}_j$ の重み係数を w_i とし、重み係数ベクトルを $\boldsymbol{w} = (w_0, \cdots, w_{p-1})^T \in \mathbb{R}^p$ とすると、LASSO[12] と呼ばれる定式化では、重み係数ベクトルを以下の最小化問題の解として求解する。

$$\begin{aligned}&\min_{\boldsymbol{w}\in\mathbb{R}^p} L(\boldsymbol{w}),\\ &L(\boldsymbol{w}) \triangleq \frac{1}{2}||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_1\\ &\qquad = \frac{1}{2}\sum_{i=0}^{n-1}\left(y_i - \sum_{j=0}^{p-1} x_{i,j}w_j\right)^2 + \lambda\sum_{j=0}^{p-1}|w_j| \quad (1)\end{aligned}$$

ここで、$\boldsymbol{X}$ は $\boldsymbol{x}_j$ を第 j 列とする行列 (特徴行列と呼ぶ) $\boldsymbol{X} = (\boldsymbol{x}_0, \cdots, \boldsymbol{x}_{p-1}) \in \mathbb{R}^{n\times p}$ であり、λ は解のスパース性を調整するパラメータである。なお、以下の議論では、$\sum_{j=0}^{p-1} x_{i,j} = 0$, $\sum_{j=0}^{p-1} x_{i,j}^2 = 1$ であることを仮定する。

次に、ランダムユニタリ変換を用いた情報の秘匿化について説明する。鍵 ζ（定義は後述）によって生成されるランダムユニタリ行列 $\boldsymbol{Q}_{\zeta,n} \in \mathbb{R}^{n\times n}$ を用いて、観測ベクトル $\boldsymbol{y}$、特徴行列 $\boldsymbol{X}$ を以下のように秘匿信号 $\hat{\boldsymbol{y}}$, $\hat{\boldsymbol{X}}$ へ変換する。

$$\hat{\boldsymbol{y}} \triangleq \boldsymbol{Q}_{\zeta,n}\boldsymbol{y} \quad (2)$$

$$\hat{\boldsymbol{X}} \triangleq \boldsymbol{Q}_{\zeta,n}\boldsymbol{X} \quad (3)$$

このとき、$\boldsymbol{Q}_{\zeta,n}$ は以下の関係を満たす：

$$\boldsymbol{Q}_{\zeta,n}^*\boldsymbol{Q}_{\zeta,n} = \boldsymbol{I} \quad (4)$$

ここで、$[\cdot]^*$ はエルミート転置、$\boldsymbol{I}$ は単位行列を表す。ランダムユニタリ行列の生成には、擬似乱数行列に対してグラムシュミットの直交化を適用する方法 [6] を用いる。なお、擬似乱数行列を生成する際の乱数シードが鍵 ζ となる。

最後に、ランダムユニタリ変換により秘匿化された信号から式 (1) の解を求める。この求解は、秘匿化された信号に対して以下のコストを最小化することで実現できる。

$$\hat{L}(\boldsymbol{w}) \triangleq \frac{1}{2}||\hat{\boldsymbol{y}}-\hat{\boldsymbol{X}}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_1 \quad (5)$$

実は、式 (4) の性質を考慮すると、次の関係が導かれる [11]。

$$\arg\min_{\boldsymbol{w}} \hat{L}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}} L(\boldsymbol{w})$$

上式は、秘匿化された信号に対して求めた LASSO 解は、秘匿化前の信号に対する解と一致することを示している。

3 ランダムユニタリ変換の秘匿性強化

ランダムユニタリ変換の秘匿化強度、即ち、秘匿化信号からの原信号を復元する際の不確定性はランダムユニタリ変換のサイズ n に依存する。n は秘匿信号を埋め込む空間の次元数を規定するためである。この n は、$\boldsymbol{y}$ の次元数および $\boldsymbol{X}$ の行数、即ち、観測データのサンプル数に対応する。このため、観測データのサンプルが少なくなると、ランダムユニタリ変換による秘匿化強度が低下することになる。そこで、n が小さな場合に、秘匿化強度を強化するために、$\boldsymbol{y}$ および $\boldsymbol{X}$ を各々、$\tilde{n}$ 次元ベクトルおよび $\tilde{n}\times p$ 行列に拡張 $(\tilde{n} > n)$ し、大きなサイズのランダムユニタリ変換を用いるアプローチをとる。しかし、次元拡張前のサイズ n を攻撃者が既知の場合、ランダムユニタリ行列のサイズを増加させるだけでは秘匿化強度の強化に繋がらない。サイズ $\tilde{n}$ のランダムユニタリ変換により秘匿化された信号は、$\tilde{n}$ 次元空間内の超球面（原点を中心とし、秘匿化前の原信号のユークリッドノルムを半径とする）上に射影される。見かけ上、秘匿信号は $\tilde{n}$ 次元空間に埋め込まれる。しかし、次元拡張前のサイズ n を既知の攻撃者であれば、原信号の存在範囲を n 次元部分空間内の超球面上に絞り込める。このため、原信号を復元する際の不確定性は、次元拡張前の n 次元空間内に制限されてしまう。

そこで、秘匿化信号 $\tilde{\boldsymbol{y}} \in \mathbb{R}^{\tilde{n}}$, $\tilde{\boldsymbol{X}} \in \mathbb{R}^{\tilde{n}\times p}$ を次式の変換により生成する。

$$\tilde{\boldsymbol{y}} = \boldsymbol{Q}_{\zeta,\tilde{n}}\boldsymbol{S}\boldsymbol{y} + \boldsymbol{\psi} \quad (6)$$

$$\tilde{\boldsymbol{X}} = \boldsymbol{Q}_{\zeta,\tilde{n}}\boldsymbol{S}\boldsymbol{X} + \boldsymbol{\Phi} \quad (7)$$

ここで、$\boldsymbol{Q}_{\zeta,\tilde{n}} \in \mathbb{R}^{\tilde{n}\times\tilde{n}}$ は、$\tilde{n}\times\tilde{n}$ サイズのランダムユニタリ行列である。$\boldsymbol{S} \in \mathbb{R}^{\tilde{n}\times n}$ は、ベクトルの次元を n から $\tilde{n}$ へ拡張する変換である。$\boldsymbol{S}$ は、1 および 0 を要素とし、各列は一つだけ 1 を含み、かつ、各行は高々一つの 1 を含むように構成される。このとき、$\boldsymbol{Q}_{\zeta,\tilde{n}}\boldsymbol{S}$ は、$\boldsymbol{Q}_{\zeta,\tilde{n}}$ の n 本の列ベクトルにより構成される。$\boldsymbol{\psi} \in \mathbb{R}^{\tilde{n}}$ は、$\boldsymbol{Q}_{\zeta,\tilde{n}}\boldsymbol{S}$ に含まれない $\boldsymbol{Q}_{\zeta,\tilde{n}}$ の列ベクトルとし、$\boldsymbol{\Phi} \in \mathbb{R}^{\tilde{n}\times p}$ は $\boldsymbol{Q}_{\zeta,\tilde{n}}\boldsymbol{S}$ および $\boldsymbol{\psi}$ に含まれない $\boldsymbol{Q}_{\zeta,\tilde{n}}$ の p 本の列ベクトルにより構成される。摂動情報として、$\boldsymbol{\psi}$, $\boldsymbol{\Phi}$ を加えることで、次元拡張前のサイズ n を既知の攻撃者に対しても、原信号の存在する n 次元部分空間内の超球面を秘匿可能となる。

式 (6)(7) による秘匿化信号領域において、原信号の LASSO 解を算出するために、適切なコスト関数を設定する必要がある。式 (6)(7) の秘匿化信号に対する式 (5) のコスト関数の最小解は、原信号の LASSO 解と一致しない為である。そこで、次のコスト関数を導入する。

$$\tilde{L}(\boldsymbol{w}) \triangleq \frac{1}{2}||\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{X}}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_1 - \frac{1}{2}||\boldsymbol{w}||_2^2 \quad (8)$$

このとき、$\boldsymbol{\psi}$ および $\boldsymbol{\Phi}$ が次の関係

$$\boldsymbol{\psi}^T(\boldsymbol{Q}_{\zeta,\tilde{n}}\boldsymbol{S}) = \boldsymbol{0}_n \quad (9)$$

$$\boldsymbol{\Phi}^T(\boldsymbol{Q}_{\zeta,\tilde{n}}\boldsymbol{S}) = \boldsymbol{0}_{p\times n} \quad (10)$$

$$\boldsymbol{\psi}^T\boldsymbol{\Phi} = \boldsymbol{0}_p \quad (11)$$

を満たす (なお、$\boldsymbol{0}_n$、$\boldsymbol{0}_p$ および $\boldsymbol{0}_{p\times n}$ は、各々、全ての要素を 0 とする n 次元ベクトル、p 次元ベクトルおよび $p\times n$

行列である) ことに注意すると、次式が得られる。

$$||\tilde{\boldsymbol{y}} - \tilde{\boldsymbol{X}}\boldsymbol{w}||_2^2 = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + ||\boldsymbol{w}||_2^2 + ||\boldsymbol{\psi}||_2^2 \qquad (12)$$

さらに、上式を用いることで、コスト関数を以下のように変形できる。

$$\begin{aligned}\tilde{L}(\boldsymbol{w}) &= ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_1 + \frac{1}{2}||\boldsymbol{\psi}||_2^2 \\ &= L(\boldsymbol{w}) + \frac{1}{2}||\boldsymbol{\psi}||_2^2\end{aligned}$$

上式を用いて、以下の関係を導出できる。

$$\arg\min_{\boldsymbol{w}} \tilde{L}(\boldsymbol{w}) = \arg\min_{\boldsymbol{w}} L(\boldsymbol{w})$$

つまり、式 (6)(7) により秘匿化された信号に対して、$\arg\min_{\boldsymbol{w}} \tilde{L}(\boldsymbol{w})$ の解を求めれば、秘匿化前の信号に対する LASSO 解を導出できることを上式は示している。

4 実験

提案法による秘匿性強化の効果を検証するために、糖尿病の臨床データを用いて以下のような実験を行った。用いた糖尿病データ [13] は、442 人の患者のデータから構成され、各患者に対して 10 項目の検査結果と検査から 1 年後の疾病進行度をデータとして含む。10 項目の検査結果から疾病進行度を予測する予測モデルを LASSO を用いて構築する。上述の検査結果（特徴行列を構成）および疾病進行度（観測ベクトルを構成）が秘匿化対象となる。秘匿化対象サンプル数の減少に伴う秘匿化強度の変化を測定するため、秘匿化対象サンプル数を $n = 352, 88, 44, 22, 11$ の 5 種類に設定した。秘匿化強度は、秘匿化時と異なるランダムユニタリ行列を用いて復号された信号に対して、秘匿化前の原信号との相関係数に基づき評価した。両信号の相関係数は信号間の類似度を表しており、相関係数が低ければ、正しく復号できなかったことを示すためである。サンプル数毎に 100 種類のランダムユニタリ行列を用いて復号を試みた。

表 1 に上記相関係数の絶対値の平均値 (以下、相関係数と略記) を示す。同表 (a)(b) は各々、特徴行列および観測ベクトルに対する結果である。提案法における次元拡大率 $(\tilde{n}/n)$ は 4,8,16,32 とした。なお、$\tilde{n}/n = 1$ の列の値は、次元拡大を伴わないナイーブなランダムユニタリ変換に対する結果である。同表によれば、サンプル数の減少に伴い、相関係数が増加していることが確認できる。一般的に、相関係数が 0.2 未満であれば、ほぼ相関は無いとみなせる。しかし、同表 (b) の $\tilde{n}/n = 1$ の列（ナイーブなランダムユニタリ変換の結果）では、$n = 11$ のように、サンプル数が極端に少なくなると、相関係数が 0.2 を超えてしまっている。一方、提案法を用いることで、相関係数の増加を抑制できている。例えば、同表 (a)(b) の対角セルでは、ほぼ、

表 1: 秘匿時と異なるランダムユニタリ行列による復号信号と原信号の類似度（n は秘匿化対象としたサンプル数を表す。$\frac{\tilde{n}}{n}$ は提案法による次元拡大率を表す。なお、$\frac{\tilde{n}}{n} = 1$ の列は、ナイーブなランダムユニタリ変換の結果を表す。）

(a) 特徴行列

n \ $\frac{\tilde{n}}{n}$	1	4	8	16	32
352	0.022	0.008	0.006	0.004	0.003
88	0.039	0.017	0.01	0.007	0.005
44	0.069	0.02	0.014	0.009	0.007
22	0.088	0.033	0.021	0.013	0.009
11	0.142	0.033	0.029	0.019	0.013

(b) 観測ベクトル

n \ $\frac{\tilde{n}}{n}$	1	4	8	16	32
352	0.039	0.025	0.018	0.01	0.009
88	0.107	0.04	0.032	0.018	0.016
44	0.126	0.061	0.035	0.031	0.019
22	0.166	0.085	0.068	0.037	0.036
11	0.256	0.131	0.089	0.059	0.039

一定の類似度となっている。この結果は、サンプルが減少しても、その減少率に相当する程度に次元拡大率を設定することで、相関係数の増加を抑制できることを示している。

提案手法により、データのプライバシーを保護した上で、拠点毎に分散取得された少数データを集約して、分析を行うことが可能となる。そこで、データを集約して分析を行う効果を検証するため、糖尿病データを K 個のサブセットに分割し、各サブセットが拠点毎に観測されるデータとみなして、以下の実験を行った。なお、拠点数は $K = 16, 32$ とした。各サブセット内のデータを学習データと検証データに分離し、以下の 2 種類の予測モデルを比較した。一つ目の予測モデルは、秘匿化して集約した全拠点の学習データを用いて構築した。秘匿化は式 (6)(7) に基づき実施し、予測モデルは、集約した秘匿データに対する式 (8) の最小化を通して構築した。同予測モデルを用いて、各拠点の検証データに対して、予測を実施した。上記予測モデルを統合予測モデルと呼ぶ。二つ目の予測モデルは、自拠点内の学習データのみを用いて構築した。予測モデルは式 (1) に基づき構築した。拠点毎に構築した予測モデルを用いて、各拠点の検証データに対して、予測を実施した。上記予測モデルを独立予測モデルと呼ぶ。

表 2 に統合予測モデルおよび独立予測モデルにより得られる予測誤差を示す。あわせて、次式の尺度を用いて、統

表 2: 予測誤差

拠点数	予測誤差: 統合予測モデル	予測誤差: 独立予測モデル	予測誤差 低減率 [%]
16	2377435	3194945	25.6
32	2377435	5539466	57.1

合予測モデルによる予測誤差低減率も評価した。

$$\text{予測誤差低減率} = \frac{\text{独立予測モデルの予測誤差} - \text{統合予測モデルの予測誤差}}{\text{独立予測モデルの予測誤差}}$$

同表の結果から、統合予測モデルは独立予測モデルに比べて予測誤差を低減できており、各拠点のデータを集約して予測することにより、予測精度の向上に繋がることを確認できた。従来、個人情報等を含むために拠点内に閉じた利用に限定されていたデータであっても、提案技術により、分散取得されたデータを統合した状態での分析が可能となり、分析性能の向上を実現できることを、本実験結果は示している。

5 まとめ

本稿では、秘匿対象サンプル数の低減に伴い、ランダムユニタリ変換の秘匿化強度が低下する問題に対して、同秘匿化強度の強化法について検討した。高次元空間への秘匿信号の埋め込みを通して、秘匿対象サンプル数によらず、高い秘匿化強度を実現できることを確認した。あわせて、秘匿化前の原信号に対する LASSO 解を秘匿信号領域において求めるための方策について明らかにした。この結果、プライバシー保護の必要なデータが分散取得され、かつ、取得拠点内のデータが少数であったとしても、提案技術により、データの機密性は確保した上で他拠点と共有可能となり、分析の高精度化が可能となる。

参考文献

[1] G. Premsankar, M. Di Francesco, and T. Taleb, "Edge computing for the internet of things: A case study," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1275–1284, 2018.

[2] R. Xu, S. Y. Nikouei, Y. Chen, A. Polunchenko, S. Song, C. Deng, and T. R. Faughnan, "Real-time human objects tracking for smart surveillance at the edge," *Proc. IEEE Int. Conf. Commun.*, pp. 1–6, 2018.

[3] Y. Xiao, Y. Jia, C. Liu, X. Cheng, J. Yu, and W. Lv, "Edge computing security: State of the art and challenges," *Proc. IEEE*, vol. 107, no. 8, pp. 1608–1631, 2019.

[4] M. Barni, G. Droandi, and R. Lazzeretti, "Privacy protection in biometric-based recognition systems: A marriage between cryptography and signal processing," *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 66–76, 2015.

[5] Z. Brakerski, "Fundamentals of fully homomorphic encryption - A survey," *Electronic Colloquium on Computational Complexity*, report no. 125, 2018.

[6] I. Nakamura, Y. Tonomura, and H. Kiya, "Unitary transform-based tempalte protection and its application to l2-norm minimization problems," *IEICE Trans. Inf. & Syst.*, vol. E99-D, no. 1, p. 60 68, 2016.

[7] M. Elad, "Sparse and redundant representation modeling - what next?," *IEEE Trans. Signal Process. Lett.*, vol. 19, no. 12, pp. 922–928, 2012.

[8] ITU-T and ISO/IEC JTC 1, *Information technology JPEG 2000 image coding system: Core coding system, ITU-T Rec.T.800 and ISO/IEC 15444-1:2004, Edition 3.0*, June 2019.

[9] T. Nakachi, Y. Bandoh, and H. Kiya, "Secure overcomplete dictionary learning for sparse representation," *IEICE Trans. Inf. & Syst.*, vol. E103-D, no. 1, pp. 50–58, 2020.

[10] T. Chuman, K. Iida, W. Sirichotedumrong, and H. Kiya, "Encryption-then-compression systems using grayscale-based image encryption for JPEG images," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 6, pp. 1515–1525, 2019.

[11] Y. Bandoh, T. Nakachi, and H. Kiya, "Sparse modeling on distributed encryption data," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2020.

[12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[13] *https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html*.