いかに研究成果を世界に発信するか

ー判別分析の4つの問題と新事実ー

新村秀一
成蹊大学　経済学部

How to Spread Our Research to the World

ーFour Problems of the Discriminant Analysisー

Shuichi Shinmura
Economic Department, Seikei Univ.

要旨：

1948年に大学を卒業し、4年間大阪府立成人病センターで、「心電図の自動診断システム」の診断論理を線形判別関数（LDF）と2次判別関数（QDF)で行った。しかし医師の開発した枝分かれ論理にかなわなかった。これが「最適化線形判別関数」を研究する動機である。

2010年に基礎研究を終了したが、なかなか理解が得られなかった。

そこで、2014年にSASユーザー会と科研費の研究集会をのぞく日本での学会発表と論文投稿を停止し、フリージャーナルと採択率の厳しい国際会議にフルペーパーを発表し、年4件の英語論文を年末からResearch GateにUPすることにした。本発表では、SASユーザー会のユーザーや研究者にその概略を報告する。その後で、2015年夏に発表した「判別分析の4つの問題」を紹介する。

キーワード：Research Gate, Research Map, ISI, 新しい国際会議,

判別分析の問題と新知見

1. はじめに

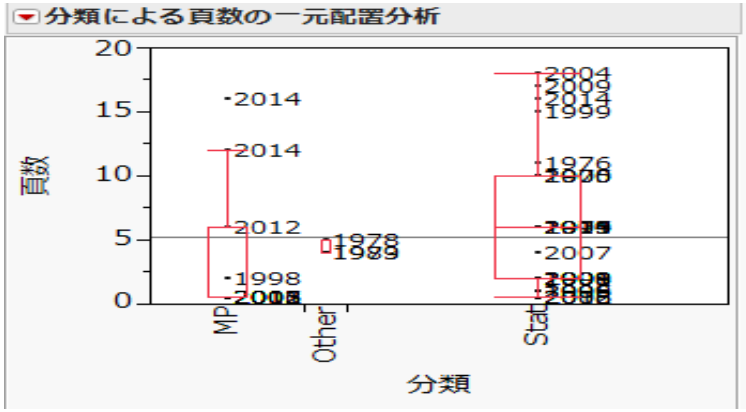
- 大学卒業後（1948年）大阪府立成人病センターで、「心電図の自動診断システム」の診断論理の開発を線形判別関数（LDF）と2次判別関数（QDF）で行った。
  - しかし医師の開発した「枝分かれ論理」にかなわなかった。
  - 当初は自分の技量がないと考えたが、数多くの医学診断でSASやJMPで実証研究するうちに、判別超平面上の近辺に多くのケースがある場合（医学診断、各種格付けや合否判定）、確率分布を基礎とする判別分析に問題があると疑うようになった。
  - これが「誤分類数最小化（MNM)基準による**最適化線形判別関数（OLDF)**」の研究動機
- 1997年から2010年で基礎研究を終了し、2012年まで合否判定データを用いて応用研究を行ったが、日本では理解が得られなかった。
  - 当初は、JMPが**OLDF**を取り込んでくれればと目標を定めたが、当面は予定がないとの決定。
- そこで、2013年末に世界へ情報発信しようと考え、Googleで自分の数少ない英語論文を検索したが、ほどんと検索に引っかかってこない。
- 2014年にSASユーザー会と科研費の研究集会をのぞく、日本での学会発表と論文投稿を停止し、フリージャーナルと採択率の厳しい国際会議にフルペーパーを発表し、年4件程の英語論文を目標とし達成した。

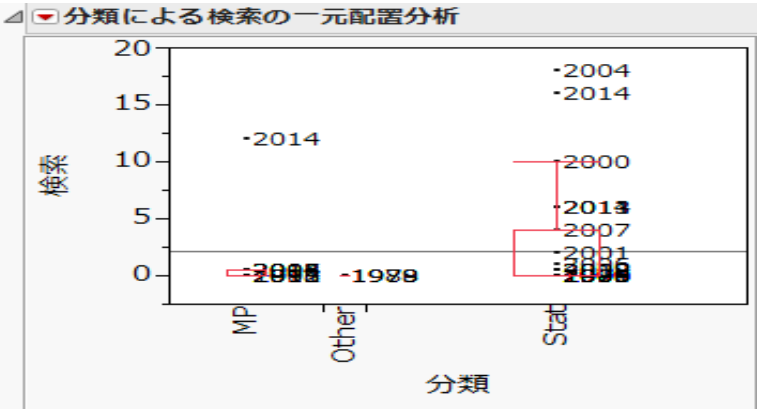
- 2014年末に、検索でResearch Gateを知り、そこにUP。
  - 2015年5月22日で、ダウンロード数が1,404件、閲覧数が1,381、引用数が59、CVの閲覧数が571件であり一応手ごたえを得た。
  - 5000件のダウンロードを当面の世界普及の目標と考えている。
- 本発表では、SASユーザー会の仲間にその概略を報告し、成果発表と普及の参考になればと考える。
- 大学の研究者にはResearch Mapという研究者DBがある。
  - 民間企業の研究者は、例え日本語の研究論文であってもResearch GateにUPすれば、研究成果を日本中に普及できる。
- それに加えて、2015年5月のSpringerの論文と夏に発表した国際統計会議（ISI）の内容「判別分析の4つの問題」を紹介。

2. 2013年末のGoogleでの調査

SN	年	学会誌	分類	頁数	検索	SN	年	学会誌	分類	頁数	検索
1	2014	**ICORE**	MP	12	12	21	2004	VIP	Stat	18	18
2	2014	**SOIC**	Stat	16	16	22	2003	EuroInforms	MP	0.5	0
3	2014	**Analytics**	Stat	6	6	23	2003	ISI	Stat	2	0
4	2014	**Springer**	MP	16	0.5	24	2002	IFORS	MP	0.5	0.5
5	2013	ISI	Stat	6	6	25	2001	ISI	Stat	2	0
6	2013	EURO	MP	0.5	0	26	2001	EMS	Stat	2	2
7	2012	DEA	MP	6	0	27	2000	Taruto	Stat	10	10
8	2012	Informs	MP	0.5	0.5	28	2000	Compstat	Stat	2	0
9	2012	ERCIM	Stat	0.5	0.5	29	2000	Data Mining	Stat	10	0
10	2011	ISI	Stat	6	6	30	1999	AppliedStatistics	Stat	6	0
11	2011	OR53	MP	0.5	0	31	1999	ARS	Stat	15	0
12	2009	IPSI	Stat	17	0.5	32	1999	**ISI**	Stat	2	0
13	2008	IASC	Stat	0.5	0.5	33	1999	Taruto	Stat	1	0
14	2007	ISI	Stat	4	4	34	1998	SMABS	MP	2	0
15	2006	EuroXX	MP	0.5	0.5	35	1998	Compstat	Stat	2	0
16	2006	Compstat	Stat	0.5	0	36	1983	**MEDINFO**	Other	4	0
17	2006	Matrix Comp	Stat	1	1	37	1979	**COMPSAC**	Stat	6	0
18	2005	ISI	Stat	6	0	38	1978	**MEDINFO**	Other	5	0
19	2005	IFORS	MP	0.5	0.5	39	1979	**ICCS**	Other	4	0
20	2005	EMS	Stat	1	0	40	1976	**Dijion**	Stat	11	0
						41	1975	NERC	Stat	10	0

2.1 無残な結果





- **表1**は2013年末の41件の調査結果。研究成果を広く世界に発信するために「どの国際会議で発表すべきだったのか？」という疑問をもちGoogleで検索。
- MEDINFOは発表者が多いにもかかわらず1978年の時点で5頁である。インターネットの世界には、4頁以下の論文が満ち溢れているが、とても著者以外は内容が分かるとは思えない。
  - すなわち志高く4頁以上要求される国際会議しか参加しないという基準をとらなかったことは愚かであった。
- しかし1999年にタルトで開催されたISIの関連国際会議にISIの発表後に参加した。
  - 僅か1頁のAbstractであったが、その後10頁のOLDFの論文を書籍に発表する機会を得たのは怪我の功名である。
- 以上を振り返れば、その後の研究もMEDINFO等で発表を継続すべきだった。
  - 1984年から1996年までの空白は、企業人であったため簡単に国際会議に参加することは容易でなく、仕事も忙しく、研究費を稼ぐため出版に傾倒していたようです。

2.2 新しい国際会議とフリージャーナル

Googleの検索結果を踏まえて、2013年末に

- 6頁以上のFull Paperが PDFで公開できる国際会議に参加。
  - これによって、国際会議はISIとICORESにする。
- お金のかからないフリージャーナルのSOICに年1本以上の論文を載せる。
- 国内の統計とOR関連は、年4回以上の会議で発表してきたが、2014年3月に定年を迎え、自費で発表する負担削減のため停止。
  - そこでSASユーザー会と、旅費のでる研究集会だけに参加発表する。

(1) ICORES (International Conference on Operations Research and Enterprise Systems)

- INSTICCという組織に幾つかの学会が参加し、論文の投稿、査読、管理、参加登録、採択された論文の管理が一つのシステムのPRIMORISで行える。
  - FullとShortペーパーだけがDigital Libraryとしてロイターなどのインデックスに登録。
- 採択率20%のFull Paperセッション(12頁、40分)、Short Paper（6頁、20分）、PosterセッションとDoctorセッション。
  - 発表は、Fullペーパー、Shortペーパー、博士ペーパーの順に発表します。
  - また、複数の類似の国際会議が1部屋あるいは2部屋で並行して行われ、参加は自由。
  - Posterは同時開催。
  - Proceedingsは理論のFullが15編、Shortが29編、その後に応用のFull4編と Shortが10編のみがセッションと無関係に並んでいます
  - ISBNとDOIつきでCrossRef (http://www.crossref.org/)のThomson Reuters Conference Proceedings Citation Index (ISI)に登録）。
  - また年会費なしでINSTICCの会員証がもらえる。
  - しばらくしてSpringerの叢書に執筆依頼。
    - 4割以上を新しい内容に変えた16頁の論文の提出。
    - タイトルとAbstractは書き直し、本文は40%以上新しい内容にする。
    - すぐに別の論文2編が世の中に出ることが分かり、宝くじにでも当たった気分です。

- ICORES2015は「線形分離可能なデータの判別」をテーマにしたが不採択通知。
- レフリーコメントとして。
  - 判別はOverlapデータの判別が重要で、線形分離を取り上げるのは問題
  - ICORES2014の小規模改善である。
  - 幾つもの予見が書いてあるが、これが許されるのは革新的な研究だけである。
  - 英文が雑で当学会の水準に満たない、などとあります。
- 最初はあきらめていましたが、投稿システムに「異議申し立て」の機能があります。

- そこで「４）は前回条件付採択に味を占め、この通知がきてからNative Checkし、Camera Readyの最終校を考えていた」とわびた。
- しかし、線形分離可能な判別はハードマージン最大化SVM（H-SVM）で明確に定義されたが、SVMの研究者は行っていない。また、線形分離可能と判別する2群はオーバーラップしていないと同意義であり、これができるのはH-SVMと改定IP-OLDF　だけである。
- また線形分離可能な判別は、結果が明確であるので研究に意義がある

- と反論した。やがてShort Paperという訂正が着たが、早まってSOICに投稿していたし、Short Paperには興味がないと断った。
- Springerの方は2014年6月に投稿したが音なしで、依頼されながら不採択かとやきもきしたが、2015年3月に校正が来て、5月25日に刊行。
- ICORES2016には、十分気をつけて望むつもりである。

(2)ISIとDATA Analytics

- 奇数年はISI、遇数年はDATA Analytics
- 6頁のFull Paperあるいは4頁のShort Paperとポスター。
  - Full PaperとShort Paperは会議後PDFで公開され、Full Paperの中から組織委員が24頁までの依頼論文を書く機会がある。
  - 会議終了後に首を長くして待っていましたが着ませんでした。
    - 論文には多くの表を載せ、Camera Readyの最終校では幾つかを図に置き換える指示に従わなかった。
  - DATA Analytics2015でFull Paperに選ばれました。他の論文の査読も引き受けたので、24頁の論文依頼の可能性があったが、リオデジャネイロのISI2015に参加。
    - ISI　にはその後の楽しみはないが、日本の学術誌のレフリーから問題にされる内容を過去に発表し、今回はその最後の締めである”Four Problems of the Discriminunt Analysis”を統計会議の頂点で発表したほうが良いという判断。
    - DATA Analyticsの4人の差読者のうちの2名のコメントに納得できない。

(3) 電子ジャーナル

- 電子ジャーナルは、紙媒体を主体とした学術誌と対比し、インターネット上に自分の研究成果を残せるメリットがあります。
- インドの電子ジャーナルからは良くメールが来ます。歴史もありインパクト・ファクターもあるということですが、有償。
  - 2014年以降年間200万円ほどかかる研究費の自己負担を減らして100万円以下に減らすことにしたので対象から省く。
  - その後幾つかのフリージャーナルから案内が来て投稿しましたが、最後の段階で掲載料を10万円以上要求され急遽取りやめ。
- SOIC(STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING)というフリージャーナルが、1)掲載料を取らない（ただしTex作成料を支払った）、2)頁数の制限のないなどの特徴がある。
  - 現在のところShinmura(2014c)の引用数は5件とトップであることが分かり、これまでの日本の学術誌より効果的と考えています。

3. Research GateとResearch Map

- 2014年は英語論文3編（Springerは2015年にずれ込む）が達成できたが、研究成果を世界に発信できるか決め手を欠いていた。
- 秋ごろたまたまGoogleでResearch Gateを見つけてた。
  - クリックすると、丸山ワクチンに関する論文1件が登録されているだけである。
  - Wikipediaで調べるとBill Gatesらが寄付して、世界の研究者が多く登録している研究者のソシアルネットワーク型のDBであるらしい。
  - ソシアルネットワークには興味がないが、取り合えずインターネット上で検索できる自分の英語論文の「MNM基準のOLDF」に限定しUP。
- ー日本の研究者DBとしてREEDと呼ばれるものがありResearch Mapに移行
  - 折角論文と口頭発表などに分類していたのに、“MISC.”にまとめられ更新なかった。
  - 2015年3月に、学長名で整備する指示があり、5月まででほぼ完備した
  - そこで成蹊大学経済学部論集のDEAに関するものと、ゼミ生のレポートを掲載。
  - OR学会や計算機統計学会の一部に下線が引かれてあり、これらはPDFをダウンロードできる。

3.1 Research Gateの紹介

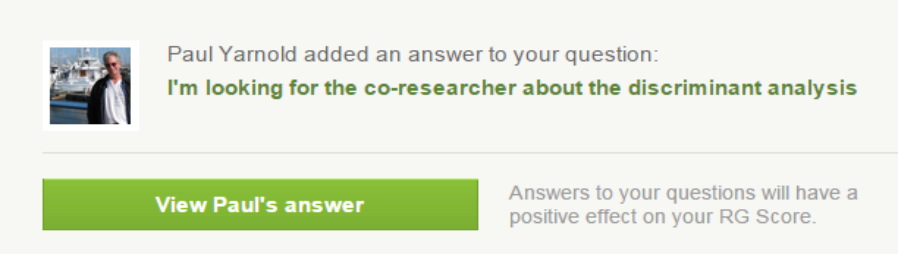
0.59
Your RG Score

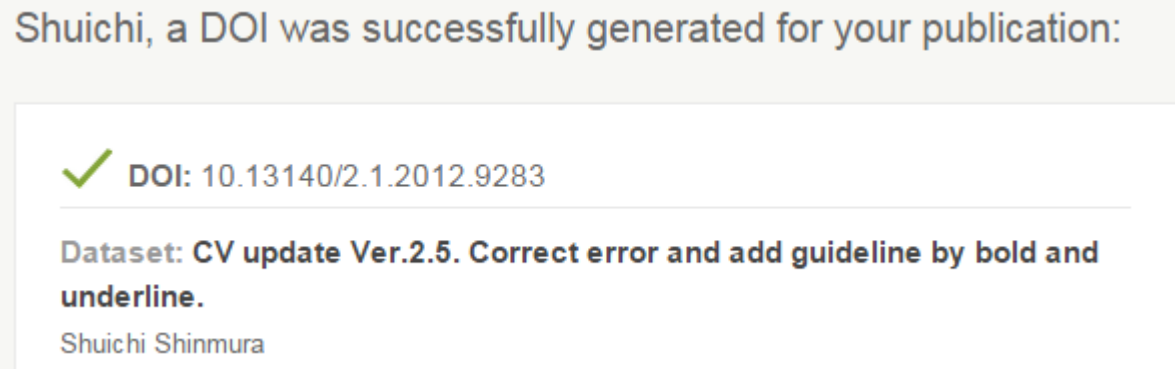
-1.11
Score change

See your score

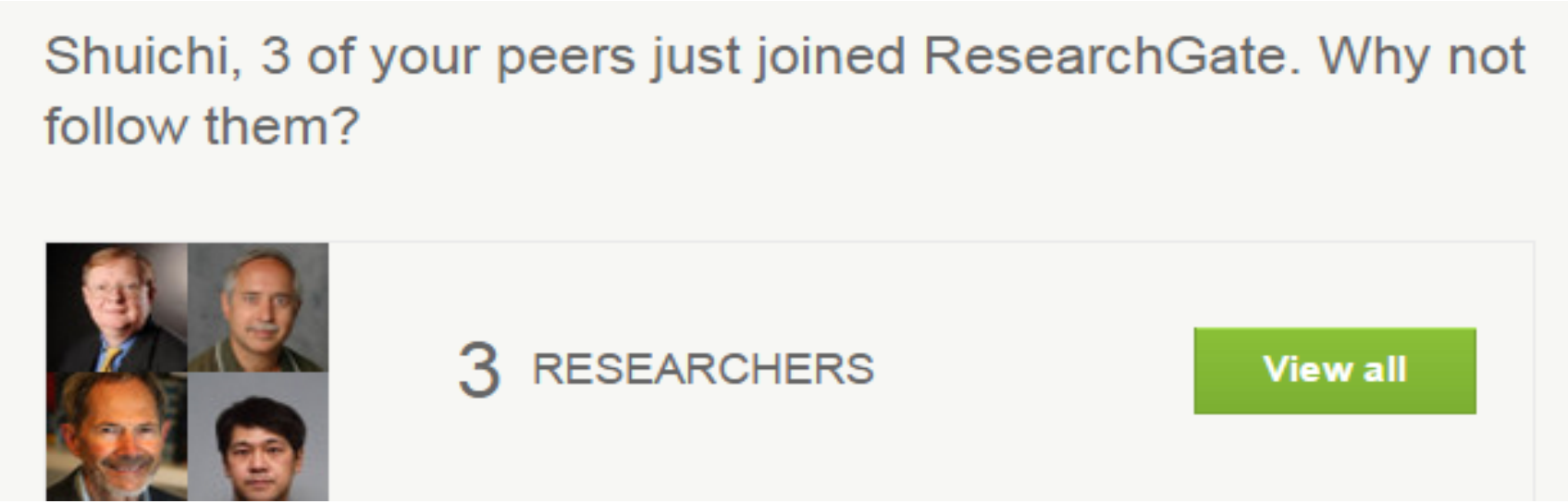
Personal stats	Total	Last week	
Publication views	30	28	See views by country / institution
Full-text downloads	57	31	View details
Dataset downloads	36	29	View details
Citations	18	6	See who cited your research
Profile views	17	5	See views by country / institution

- 2014年9月19日(金)の最初の通知。
- RG Score, Publication View, Full-text download, dataset download, Citation, Profile Views, Impact Factorなどの情報が毎週末に配信される。

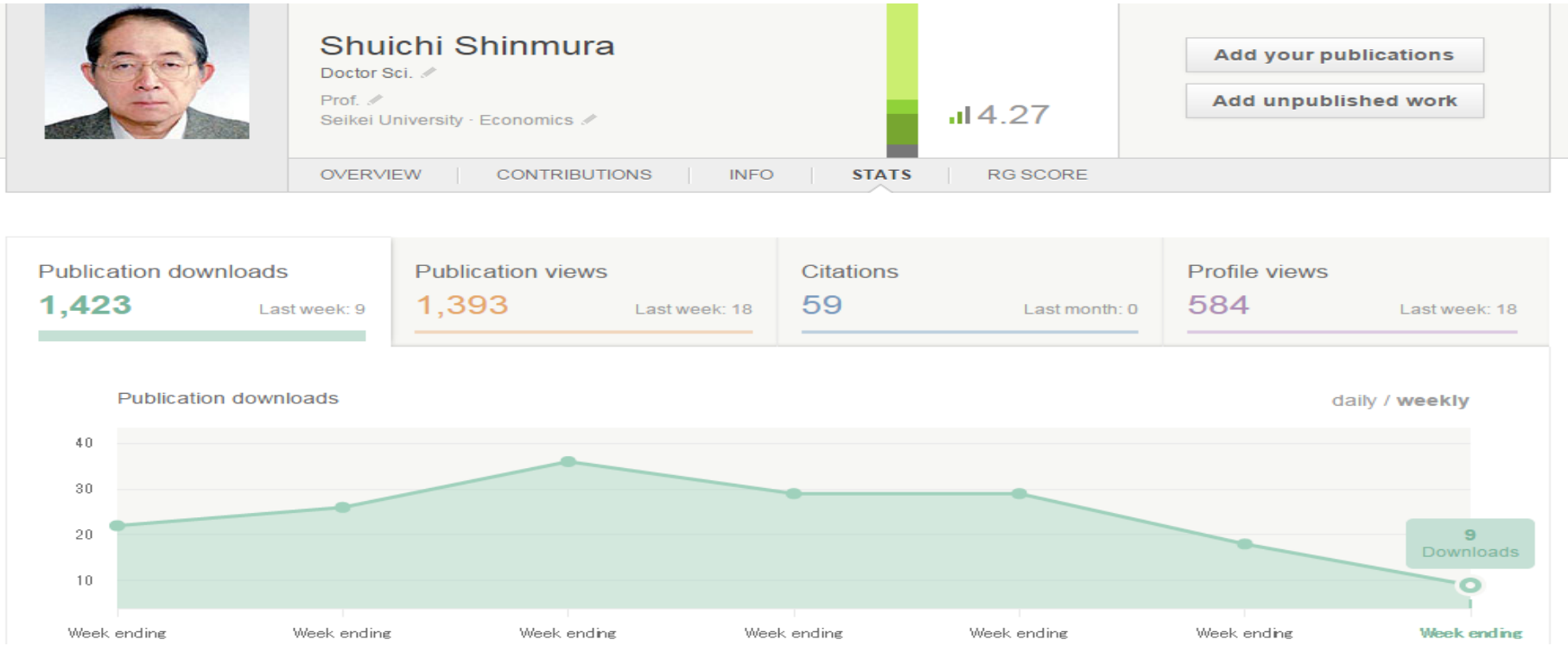




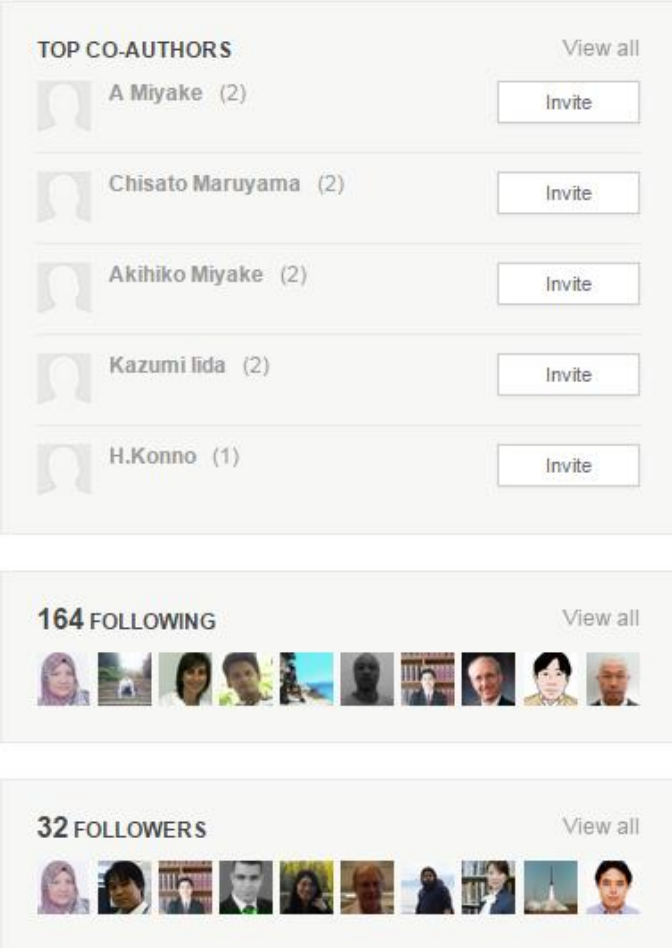
- 同日、最初のFollowerのPaulに連絡を促す通知
- 質問を行い、Paulの回答。彼はRGのスパースター
- CV2.5にDOIが振られる。



- 私の研究を4人の先達が調べているという案内。
- 初めて、写真を見た感激



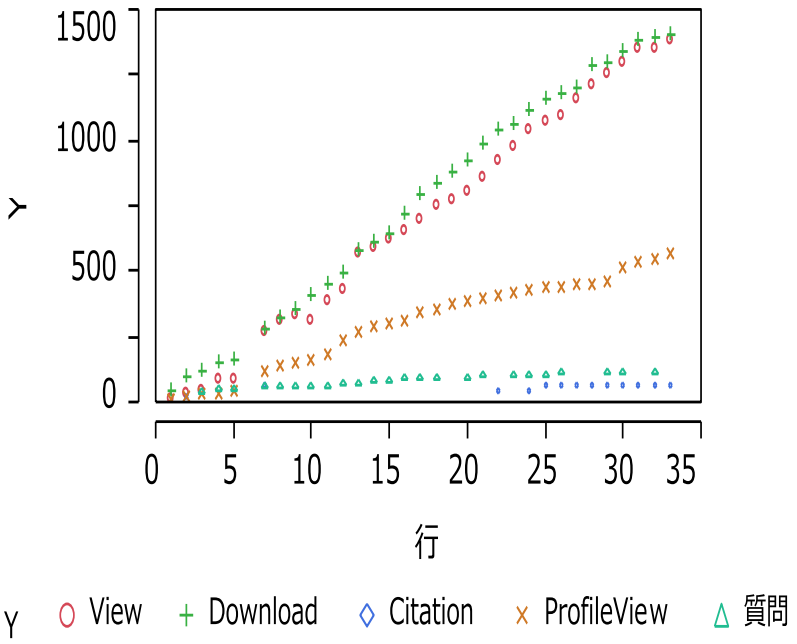
- 2015年5月22日(金)の状況



- 上は、退職されている共同研究者に招待状の督促。
- 下は、私のフォロワー

- 気候学の財城先生の登録1ヵ月後のHP.
- 国際誌や国内誌の論文はPDFで公開され、DOIも負荷。
- ラッキーな学会誌との遭遇はうらやましい。
- しかし、RGの情報はない。

月日	RGスコア	View	Download	引用	経歴	質問	保証
9月19日	1.7	8	48		11		
10月 3日	1.95	48	119		32	28	
11月14日	3.07	265	282	59	123	51	
11月28日	3.07	333	354		151	53	4
2月27日	3.49	923	1040	41	408		42
3月 6日	3.46	974	1064		416	94	
5月15日	4.18	1354	1392	59	548	109	
5月22日	4.16	1381	1404	59	571		



3.2 Research Map

- 2015年3月に学長名でResearch Mapの整備の指示。
  - 5月末までかかり、ようやく見れ内容。
- Miscに
  - 成蹊大学論集のDEAに関する論文と学生のレポート
  - JSCSの論文と学会発表
- 講演
  - OR学会の発表

4.判別分析の4つの問題と判別関数の評価

- 1972年から4年間行った心電図の診断論理の開発で判別分析は「枝分かれ論理」にかなわなかった。
- そのため1998年からMNM基準による改定IP-OLDFを開発し、「k-fold cross validation for small sample」法で改定IP-OLDFは他の線形判別関数と比較し検証標本で平均誤分類確率が一番小さいことが分かった。
- 特にLDF、QDFは判別超平面上の近辺に多くのケースがくる医学診断や各種各ずけや試験の合否判定に著しく誤分類確率が大きく利用してはいけないという結論になる。
- この他次の**重大なつの問題**を抱えている。

問題1：判別超平面上のケースの帰属 判別超平面上のケースをclass1/class2のいずれに判別するかは決定不能であるにもかかわらず，多くの研究論文は理由なくclass1に含めてきた．この問題は、改定IP-OLDF以外は正しい誤分類確率が分からないことを示す。

問題2：線形分離可能なデータを認識できない H-SVMは2つのSV間にケースが来ないことで**線形分離可能を明確に定義**した．しかし「線形分離可能なデータの判別の問題点の研究」が少なかった．H-SVMと改定IP-OLDF以外の判別関数は理論的に線形分離可能なデータを認識できない．特にFisherのLDFとQDFは判別成績が悪い．またOver lappingを定義できない。

3) **問題3：分散共分散行列の深刻な問題** QDFに用いられている一般化逆行列の手法は、一方の群の変数値のみが一定値で他方の群がバラツク場合class2をclass1に誤判別する問題が分かった[34].

4) **問題4：判別分析は推測統計学と無縁な点** Fisherは誤分類確率や判別係数の標準誤差を定式化していないのに、判別分析を重回帰分析と同じく推測統計学手法と間違った理解が行われている。これは「k-fold cross validation for small sample」法で解決できる。

5. まとめ

- 筆者の一生の研究テーマは、SASとJMPと数理計画法のLINGOという世界最高水準のソフトを駆使することで漸く解決できた。
- 後は、世界の研究者と利用者に認知させ、統計ソフトにインストールし普及する最後の仕事が実現できるか否かである。
- 最近の研究結果はResearch Gateに、2013年までは計算機統計学会、OR学会誌、統計学会等に発表してあり、Research Mapからダウンロードできます。できない場合は、連絡ください。また共同研究、講演、講習会などに対応できます。

文献:

S. Shinmura(2015). Four Serious Problems and New Facts of the Discriminant Analysis. ICORES 2014, 15-30, Springer.