経済産業省委託

令和３年度政府戦略分野に係る国際標準開発活動

【戦 17】キャッシュレス取引のセキュリティ性に関わる生体認証精度評価を容易とする精度評価方法に関する国際標準化

成 果 報 告 書

第Ⅱ部

－3 年間のまとめ－

令和４年２月

一般社団法人日本自動認識システム協会

はじめに

少ないサンプル数で実現する新しい生体認証精度評価方法により、国内外での生体認証の用途拡大や普及促進への貢献していくため、3 つの事業（経済産業省受託事業、国際標準化活動、JAISA 自主事業）を連携して活動を実施している。

これまで、生体認証の精度評価方法は ISO/IEC 19795 シリーズの国際標準に基づいて行われてきた。この標準に従うと 19795-1 に記載された「３の法則」に従い、要求される精度の逆数の 3 倍のサンプルによる試験を行うことが必要で、近年の高精度の生体認証装置やシステムのハードウエアまたはソフトウエアに何らかの改良改訂が行われる度に必要な精度の再評価の負担が極めて大きいという問題があった。そこで、より少ないサンプル数で従来の評価試験を代替する新標準を研究開発し、生体認証装置・システムの開発効率を改善することに適用評価と規格開発を進めた。

本検討委員会は JAISA 精度評価技術グループと生体認証装置・システム精度評価に関わるメンバーで、2019 年度（平成 31 年度）から 2021 年度（令和 3 年度）の 3 年間に①新しい精度評価方法を検討する委員会の設立と運営、②新しい精度評価方法の詳細検討ならびに実証データ収集と適用性確認、③国際的ロビー活動、④国際標準素案の作成支援、⑤産業界への普及活動を実施してきた。

1 年目の活動では指紋データを用い、2 つの極値統計モデルで同じ推定結果が導き出せるか、サンプル削減効果がどの程度得られるかを検討、新しい精度評価方法の国際標準化を提案、New Work Item (ISO/IEC5152)として登録されました。2 年目には、指紋以外の生体情報（顔、歩容、音声等）での適用評価を進め、どこまで少ないサンプルでその程度の精度評価ができるかを検証した。3 年目には評価結果の信頼区間を検討、国際標準化段階も委員会段階へ進み、2 年後には国際標準ドキュメントが発行される予定である。

本報告は、この 3 年間の活動で得られた知見をまとめたものである。生体認証の精度評価に極値統計技術を用いて、稀に起こる現象「他人受入」の発生確率を求める難易度の高い性能評価方法に関するものである。今後、この方法を普及啓発するテキストとしていきたく、技術的な違和感、わかりにくい説明箇所など、皆様からのご意見・ご要望をいただければ幸いです。

2022 年 2 月 10 日

精度評価方法に関する国際標準化検討委員会
委員長　鷲見 和彦

目　　次

第1章 バイオメトリクス精度評価の課題

近年の生体認証システムの大幅な精度向上に伴い、精度評価に必要とされる生体サンプルの数は非常に多くなっている。生体認証システムにおけるテクノロジーテスト(照合アルゴリズムの性能を計測するための試験)の代表的な精度評価指標には FMR(False match rate)と FNMR(False non-match rate)があり、世界的に広く利用されている。FMR は他人を誤って受け入れる率、FNMR は本人を誤って棄却する率として直感的に理解しやすい指標であるため、複数の生体認証システムを比較検討する際には、重要な判断基準の一つとなっている。

一般に、生体認証システムにおける照合アルゴリズムは、利用者が事前に身体的特徴や行動的特徴から作成される登録データ(テンプレート)と、認証時にカメラ等のセンサーで取得した照合データを比較することにより、類似度を計算しスコア値として出力するように設計されている。類似度スコアは、登録データと照合データが似ていればいるほど高い数値を示し、似ていなければ低い値を示す。本人と他人の識別には、この類似度スコアに閾値処理を実施し、閾値以下の場合は他人、閾値以上の場合は本人、と判定することが一般的である 。この閾値を上下に変更することにより、生体認証装置の判定を厳しく(FMR を低く)したり緩く(FNMR を低く)したりすることができる。判定を厳しくすればセキュリティ重視の設定となり、判定を緩くすれば利便性重視の設定となるが、一般に FMR と FNMR はトレードオフの関係にある。

実際の生体認証システムのユースケースでは、他人を誤って受け入れることは、本人を誤って棄却するよりも深刻な問題と捉えられる場合が多い。本人を棄却した場合は、リトライや代替手段(パスワード等)によって受け入れることが可能になるが、他人を受け入れた場合は一度でも発生するとセキュリティ上の大きなリスクになるからである。そのため、多くの生体認証システムでは如何に低い FMR で運用ができるかという点を重視しており、製品仕様にも FMR または FAR(他人受入率)の低さをアピールポイントとしているケースが多い。今日では他人受入率は 0.0001%(100 万分の 1)以下を謳う製品も多く存在する。

生体認証システムの開発の過程では、照合アルゴリズムの精度向上のために、多くの被験者を集めて大規模な生体サンプルデータベースを構築する必要がある。FMR の計測には他人対スコア(non-mated score)、FNMR の計測には本人対スコア(mated score)がそれぞれ必要であり、n 名の被験者を集めた場合、他人対スコアは n(n-1)組、本人対スコアは n 組得られる（1 名あたり登録・照合データ各 1 組を収集した場合)。そのため、FMR 計測のためのデータは、FNMR 計測のためのデータより多く集めることが可能である。しかしながら、先述のように他人受入率を 100 万分の 1 に設定した生体認証システムでは、単純計算でも、他人対スコアを最低でも 100 万組集め、他人受入が 1 件しか発生しなかったことを示さなければならない。

ISO/IEC 19795-1(バイオメトリック性能の試験および報告:Biometric performance testing and reporting)では、精度計測に必要とされる生体サンプルの規模について、統計的知見からより厳密なルー

ルを定義している。一つは 3 のルール(the rule of three)と呼ばれるもので、n 回の他人対試行で 1 回も他人受入が発生しなかった場合、95%の信頼度レベルで FMR≈3/n と推定できるというものである。もう一つは 30 のルール(the rule of 30)と呼ばれるもので、実測の FMR 値±30%の範囲に信頼度レベル 90%で真の FMR 値が存在することを主張するためには、最低 30 組の他人対スコアが必要というものである。

前述のとおり、今日の生体認証システムは他人受入率数百万分の 1 というレベルの性能を持っており、事実、30 のルールを満たすように 30 件ものの他人受入の事例を得ることは非常に困難である。そのため、多くのケースで 3 のルールを適用することとなるが、百万分の 1 を主張するためには 3 百万件の他人対スコアが必要となる。これは、2,450 名の被験者に相当する。

ISO/IEC 19795-1 に準拠した精度評価を実施するためには、生体認証システムのベンダは、新しい生体認証装置を開発する度にこの規模の被験者を募集してデータを収集する必要がある。そのためにかかる費用や時間はベンダにとって大きな負担となり、開発サイクルの長期化、製品・サービスの高価格化など生体認証システムの普及の阻害要因となっている。また、小規模のベンチャー企業などでは、十分に精度評価にコストをかけることができず、その結果 ISO 規格準拠を謳えないことが事実上の参入障壁となり、業界の活性化の妨げとなっている。

他方、生体認証システムを導入しようと考えている企業や、生体認証精度評価を実施する第三者機関やなどでは、ベンダが標榜する FMR/FNMR が正しいのかどうかを確認するためにはベンダと同じレベルの被験者を集めなければ評価ができない。小規模のデータベースではいずれの生体認証システムも他人受入が発生することは無く、ユーザー数が将来増えていったときに発生するであろう他人受入のリスクを自ら判断することができないという問題がある。

そこで、今回、少数の生体サンプルで従来と同等の精度評価を実施するための統計学的手法を、ISO/IEC の国際規格として導入することを提案することとした。具体的には、従来の 3 のルールが他人対分布を二項分布に基づく推定モデルで近似していたことに対し、極値統計学に基づく新しいモデルを用いて、発生頻度が極端に低い部分の確率密度関数を推定するというアプローチをとった。FMR の推定には、一般極値分布(GEV)モデルおよび一般化パレート分布(GP)モデルを採用し、他人対スコアの統計的外挿を実施した。本手法の適用方法を具体的に示し、多様な生体認証モダリティに活用できる形で国際標準とすることにより、上述の業界共通課題を解決しようと試みた。

以下、第 2 章では極値統計の基礎と精度評価への適用、第 3 章では新精度評価方法（評価の手順と報告）第 4 章では適用評価例について報告する。

第2章 極値統計の基礎と精度評価への適用

本章では 2 つの極値統計モデルの特徴と極値統計適用の前提条件を示し、既存精度評価方法との違いを説明する。

2.1. 極値統計の基礎とGEVとGPの特徴

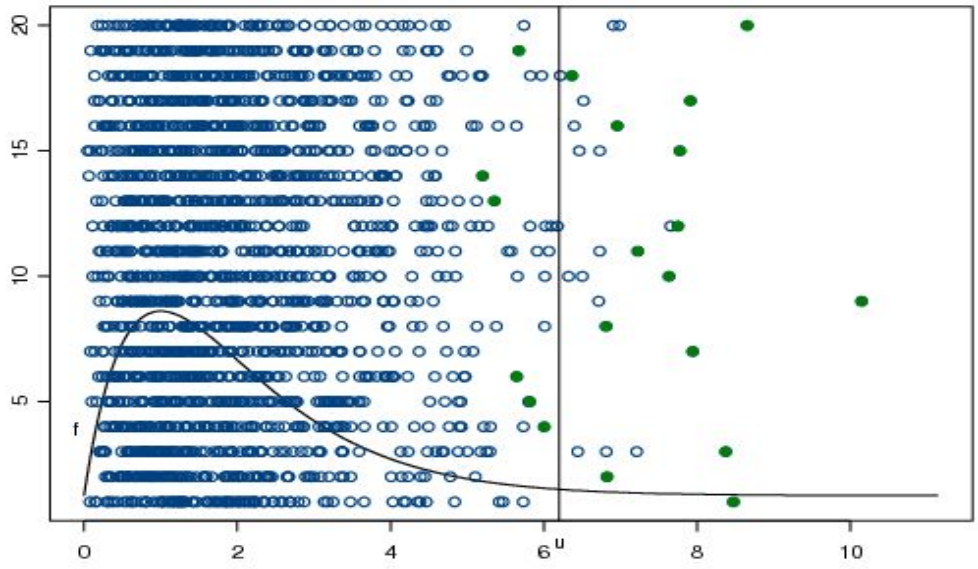
（1）極値統計とは

自然災害のように、稀であっても甚大な影響を及ぼす現象は数多い．稀な現象を扱うのが極値統計学であり、その興味の対象は、異常に大きな（小さな）値を取るデータ（極値データと呼ぶ）の出現の仕方、すなわち、その確率分布である。

統計における推測は、過去に得られたデータに数理統計の理論を適用することで、将来起こりうることを予測することである。母集団分布の端（裾）に対する推測を行う極値統計では、データのうち、一部の極端に大きい（小さい）ものだけを用い、それに極値理論を適用することで、これまでに起きたことのない大きなものを含む、極端な事象の起こり方を推測する。過去に起こったことのない事象の予測は、（統計的）外挿（extrapolation）或いは、補外と呼ばれ、極値統計では極めて重要である。

（2）極値統計で用いられるデータ

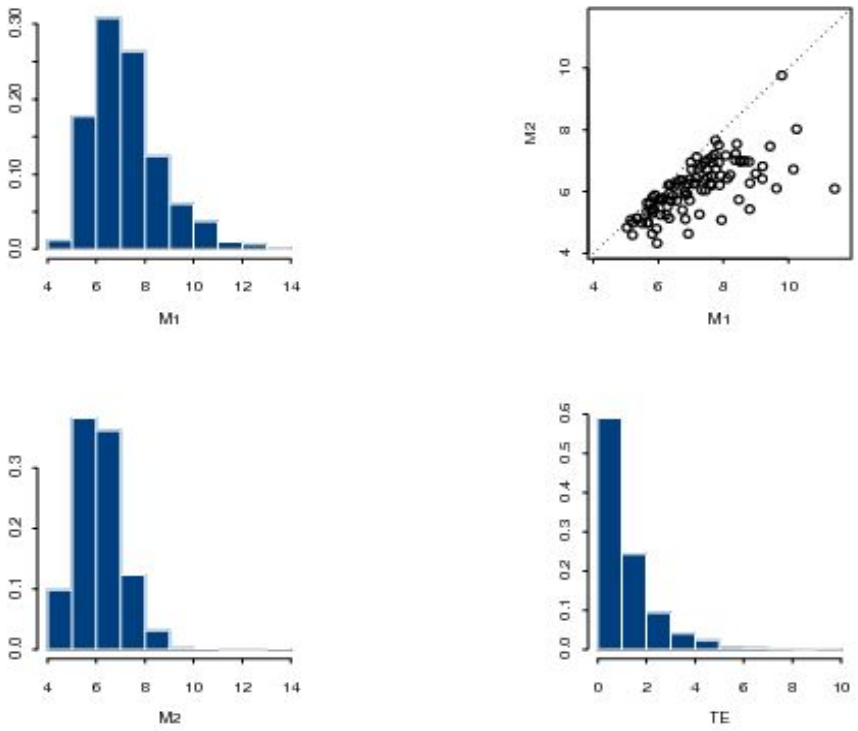
例えば大雨による洪水対策のために必要となるデータは「各年での最大日降水量」、或いは、「ある大きな値（閾値）を超えるすべての日降水量」であり、前者の最大を「各年度の上位 5 番目までの日降水量」とすることも考えられる。このような極値統計で用いるデータをシミュレーション実験により説明する。



図表 2.1.1　極値統計で用いられるデータの例

この図表 2.1.1 は、密度関数 f をもつ（母集団）分布Fからの 2000 個の乱数を、100 個ずつ 20 個のブロック毎（縦軸）に分けたものを表し、各ブロックの 100 個の乱数のうち、一番右の緑丸がそのブロックの最大のものである。この乱数をデータと考え、各ブロックから最大データ（緑丸）、それから（例えば）上位 5 個までのデータを取り出す。また、十分大きな閾値 u を決めそれを超えるすべてのデータ（exceedances）を取り出し、それからの超過（u を引いた値）を求める。この閾値を引いたデータを閾値超過（threshold excess）データと呼ぶ。

続いて、同じ分布Fからの乱数を 100 個の乱数からなるブロックを 1000 個（10^5個の乱数を 1000 個ずつ 100 個のブロックに分けたもの）に増やして、ブロック最大 M1、ブロック第 2 位 M2、これらの組合せ（上位 2 個(r=2)）、および閾値 u=6.2 としたときの閾値超過した乱数 TE を取り出したものが次の図表 2.1.2 である。



図表 2.1.2.　閾値超過した乱数 TE を取り出した図

図表 2.1.2.は、左側の上と下はそれぞれ相対頻度で描いた乱数M１とM２のヒストグラム。右側の上は各ブロックの（1 位、2 位）の乱数の組み 100 個の散布図、下は相対頻度で描いた閾値超過 TE のヒストグラムである。

（3）極値統計の手法

極値統計には、**ブロック最大法**（BM 法: block maximum method）と**閾値超過法**（POT 法: peaks over threshold method）という二つの代表的な手法がある。BM 法では、各ブロックの最大データが従う分布を、POT 法では、閾値超過データが従う分布を求める。最大データに加え、上位 5 位のデータを用いる場合は、それが従う r(r=5)次元の分布を求める。

（4）データ選択

BM 法と POT 法では、用いるデータが同じではない。閾値 u を超えるデータでブロック最大データではないデータもあれば、ブロック最大データで閾値 u を超えないものもある。BM 法で用いるデータを上位 r 個にすることで、閾値 u を下げることで、それぞれ利用できるデータを増やすことができるが、極値統計では、必ずしも多くのデータを使う方がよいということにはならず、使うデータの選び方、上位 r 個の r 、閾値 u の適切な設定が重要であり、これは難しい問題でもある。また、シミュレーション実験とは異なり、実際の極値解析では、入手できるデータが限られる場合がある。たとえば、各年の最高気温（年最大値（AMS: annual maximum series）だけといった場合では、必然的に 1 年の気温を 1 ブロックとみなした BM 法を使うことになるし、反対に、３０度を超えた日のデータ（POT 或いは PDS: partial duration series）に対しては、POT 法を用いることになる。このように適切な手法の選択は、データなど状況に大きく左右される。

（５）極値解析の数学モデル

BM 法と POT 法の数学モデルを説明する。独立で同分布に従う（independent and identically distributed, i.i.d.）確率変数列 $X_1, X_2, \dots, X_n, \dots$で，$F(x) = P(X_i \leqq x)$ とする。ここで、最初から n 個までの $X_1, X_2, \dots, X_n$ を大きさの順に並べた順序統計量（order statistics）を $X_{(1:n)} \leqq X_{(2:n)} \leqq \dots \leqq X_{(n:n)}$ 、極値統計量（extreme statistics）を $Z_n := X_{(n:n)} = \max\{X_1, X_2, \dots, X_n\} = \max_{1 \leqq k \leqq n} X_i$ とする。これがブロック最大データのモデルである。図表 2.1.1 の緑の丸は Z_{100} からの 20 個の乱数になる。上位 r 個の順序統計量（$X_{(n:n)}, X_{(n-1:n)}, \dots, X_{(n-r+1:n)}$）の同時分布が上位 r 個のデータのモデルとなる。これに対し、十分大きな閾値 u にたいして、u よりの大の条件の下での超過 $X_i - u \mid X_i > u$ が閾値超過データのモデルである。図 1 の乱数で u より大きいものの u を超えた部分に相当する。

これらのモデルで、数学的な極限操作を行うと、モデルごとに確率分布（極限分布）が現れ、この極限分布を漸近分布として、モデルの分布を近似する。これをみていこう。

分布 F の上限点を $\omega_F := \sup\{x: F(x)<1\} \leqq \infty$ と書く。$n \to \infty$ のとき $Z_n \to \omega_F$ となるのはいうまでもないが、分布 F が'最大値吸引領域なるものに属していれば、定数 $a_n > 0$ と $b_n \in \mathbf{R}=(-\infty,\infty)$ により、大きさと位置を基準化した $(Z_n - b_n)/a_n$ の分布が 1 点に縮退しない分布に収束し、このときの極限分布を極値分布という。極値分布は、フレシェ分布、グンベル分布、（極値）ワイブル分布の３種類とされることが多いが、それらをひとつのパラメータにまとめて表現したものが一般極値分布である（$\xi > 0$ がフレシェ分布、$\xi = 0$ がグンベル分布、$\xi < 0$ が（極値）ワイブル分布に対応する）。

【定義】

次の分布を**一般極値分布**（generalized extreme value distribution）といい、

$GEV(\mu, \sigma, \xi)$ ($\mu \in \mathbf{R}, \sigma > 0, \xi \in \mathbf{R}$)で表す。

$$G(x) = \exp\{-[1+\xi((x-\mu)/\sigma)]_+^{-1/\xi}\} = G_\xi((x-\mu)/\sigma). \quad \text{ここで、} (a)_+ = \max\{a, 0\}.$$

３つのパラメータ μ, σ, ξ をそれぞれ位置、尺度、形状パラメータと呼ぶ。ξ は極値指数とも呼ばれる。ここで、G_ξ は標準一般極値分布 GEV(0,1,ξ)であり，$\xi=0$ のときは、極限 $G_0((x-\mu)/\sigma) = \exp\{-\exp[-(x-\mu)/\sigma]\}$ (グンベル分布)となる。

BM 法では、各ブロックの最大データに一般極値分布をあてはめ、(μ,σ,ξ) を推定する。

BM 法を一般化した上位 r 個の場合も同じ定数で基準化することができ、漸近分布である r 次元分布を $rGEV(\mu, \sigma, \xi)$ とかく。

（6）最大値吸引領域

極値統計では､分布Fが最大値吸引領域に属していることを仮定することが多い｡分布Fにたいして、$(Z_n - b_n)/a_n$ の分布が極値分布 G へ収束するとき、分布Fは、極値分布 G に吸引されるという。極値分布 G にたいして、G に吸引される分布全体を G の最大値吸引領域(Maximum domain of attraction)といい、D(G)とかく。多くの確率分布、特にほとんどの連続分布がいずれかの極値分布の最大値吸引領域に属するが、裾が軽い、すなわち極端に大きい値を取る確率が大きくない離散分布であるポアソン分布、幾何分布は属さない。

次に閾値超過データの極限分布である一般パレート分布を説明する｡X を $D(G_\xi)$に属する分布Fをもつ確率変数とするとき、閾値を超えたときの超過の分布 $P(X-u \leqq x \mid X>u)$ $(x \geqq 0)$ の分布は、$u \rightarrow \omega F$ のとき、$H_\xi(x/a(u))$ で近似できる。ここで、a(u) は正の関数で、H_ξは次の標準一般パレート分布である。

【定義】

次の分布を**一般パレート分布** (generalized Pareto distribution) といい、GP(σ,ξ) (σ >0, ξ∈**R**)で表す。

$$H(y) = 1-(1+\xi x/\sigma)_+^{-1/\xi} = H_\xi(x/\sigma).$$

H_ξは標準一般パレート分布 $GP(1,\xi)$である。２つのパラメータ σ 、 ξ をそれぞれ尺度、形状パラメータと呼ぶ。一般パレート分布は、ξ >0 のときパレート分布、ξ =0 のとき指数分布、ξ <0 のときベータ分布になる。

POT 法では、a(u) = σ (>0) とおき、十分大きな閾値 u を選び、u の超過データに一般パレート分布 $H_\xi(y/\sigma)$にあてはめて、(σ,ξ)を推定する。

BM 法では、ブロックサイズの選び方（ブロックへの分け方)、上位 r 個では r の選び方、POT 法では､閾値の選び方が問題になる｡いずれの選択もそれぞれにトレードオフの状態にあり､容易ではない。たとえば、適切な r を選べばブロック最大法(r=1)より精度を上げることができるが、r は大きければよいわけではなく、適切な r の選択は、Q-Q plot (データとモデルの適合度を視覚的にみる手法のひとつ)などでなされる。

以上、極値統計について主に代表的なふたつの手法について簡単に説明をした。詳しいことを知りたい方のために､極値統計を扱った和書と関連した訳書を挙げる｡本文中の２つの図は､極値統計学 (2016) から引用した。

【参考文献】

[1] 極値統計学 (2016) 高橋 倫也・志村 隆彰 近代科学社

[2] R による極値統計学(2020) 西郷 達彦・有本 彰雄 オーム社

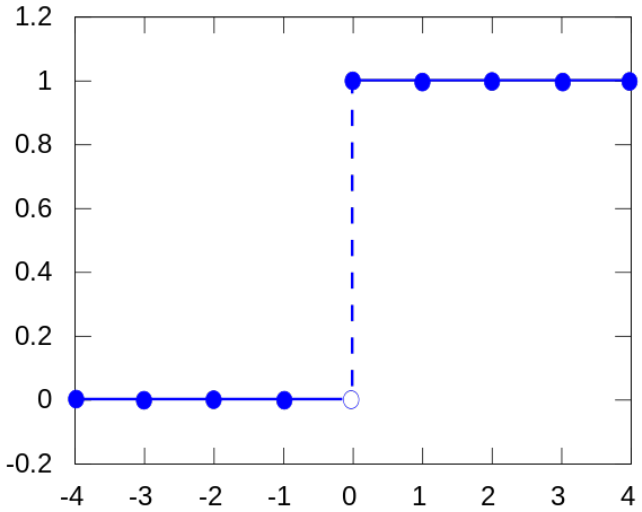
[3] 極値現象の統計分析-裾の重い分布のモデリング- (2021)S.I. レズニック(著)／国友 直人・栗栖 大輔(訳) 朝倉書店

2.2. 生体認証適用のポイント

導入する極値理論にもとづく推定手法を適用するために満足すべき基本仮定および実務上留意しておくべき点について説明する。

（1）スコア分布が非退化分布であること

対象となる分布が非退化分布であることは、極値理論適用の際の基本仮定のひとつである。図 2.2.1 のように退化した分布とはある値しかとらないような確率分布である。生体認証システムでは、どのような入力に対しても一定のスコア値しか返さないアルゴリズムがそれにあたる。そのようなアルゴリズムは言うまでも無く生体認証システムとして役に立たず、この要件を満たさないシステムは対象から外してもなんら問題はない。



An example of univariate degenerate distribution

By IkamusumeFan - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=36245162

図表 2.2.1. 非退化分布の例

（2）スコア分布が極値分布の吸引領域に属すること

本条件も極値理論の基本仮定のひとつである。FMR の推定においては、照合アルゴリズムの出力するスコアのヒストグラムの右裾のトレンドがゼロに漸近していくような分布であれば、多くの場合適用できると考えられる。

一方、スコアがある値を超えたら一定値を出力する(いわゆる打ち切り)アルゴリズムの場合はそのままでは本手法を利用することはできない。そのようなアルゴリズムの場合は、打ち切り処理をやめて生のスコア値を出力するように変更しておく必要がある。ただし、このような打ち切り処理は、通常あまりに登録データとかけ離れた照合データが入力されたときにタイムアウトを待たず終了するために実施されるケースが多く、他人受入する可能性のある高い類似度スコアではあまり実例は無いと思われる。

（3）各試行が独立同一分布に従うこと

独立同一分布に従うとは、各試行が前の試行の結果に影響を受けず(独立)、同じ確率密度分布に従って発生する(同一分布)ということであり、極値理論の基本仮定である。生体認証においては、各試行とは照合データの取得にあたり、生体部位を提示する際には毎回同じ条件で無作為に繰り返す必要がある。

独立試行が条件であるということは、生体部位を提示した後にフィードバックを受けるようなシステムではそのままでは本手法を適用できない。例えば、顔画像を撮影する際に、上下左右位置や角度を修正するよう指示がでるシステムで取得したスコアは、前のスコアが後のスコアに影響を及ぼすと考えられるので独立試行とはいえない。その場合は、フィードバックを無くして照合データを収集する必要がある。

また、入力された画像ストリームの各フレームについて判定を実施し、複数回連続で照合に成功した場合に認証成功となるようなアルゴリズムも、各フレームを 1 試行と捉えると前後に相関があるため独立試行とはいえない。その場合は、1 ストリームを以て 1 試行とすれば、独立性は担保されると考えられる。

2.3. 既存精度評価方法との違い

生体認証技術の評価方法は、主に ISO/IEC 19795 シリーズにて規定されている。現在、発行済の規格を図表 2.3.1.に示す。汎用的な用途での生体認証アルゴリズムの精度評価方法は、主に ISO/IEC 19795-1 と 19795-2 に規定されている。

図表 2.3.1. 発行済の ISO/IEC 19795 シリーズ

規格番号	規格名称
19795-1:2021	情報技術 - バイオメトリック性能試験及び報告 - 第 1 部：原則および枠組み
19795-2:2007	第 2 部：技術及びシナリオ評価のための試験方法論
19795-2/Amd:2015	修正票 1－マルチモーダルバイオメトリック導入の試験
TR 19795-3:2007	第 3 部：モダリティ固有の試験
19795-4:2008	第 4 部：相互運用性の性能試験
19795-5:2011	第 5 部：アクセスコントロールシナリオ及び格付けスキーム
19795-6:2012	第 6 部：運用評価の試験方法
19795-7:2011	第 7 部：カード内バイオメトリック比較アルゴリズムの試験
TS 19795-9:2019	第 9 部：モバイルデバイスでの試験

枠組み(フレームワーク)を規定する ISO/IEC19795-1 においては、評価で使われる様々な概念や評価尺度が定義されている。具体的な試験仕様というよりも、評価を実施する設計者・責任者が考慮すべき事柄が記載されている。図表 2.3.2.に ISO/IEC 19795-1:2021 の内容を示す。本報告に関わる内容としては、7.5.3 に評価に必要なサンプル数の仕様が記載されている。そこでは、認証精度および信頼水準に基づく数を超えるように、十分なサンプル数が収集される必要があり、Rule of 3 や Rule of 30 のルールが使用されることが記載されている。なお、2006 年発行の初版では、Rule of 3 と Rule of 30 のいずれかが使用されることが必須要件となっていた。2021 年発行版では、これらのルールは使用可能なルールの一例として示してあり、新たなルールも使用可能となっている。Rule of 3 の具体的な内容は、Annex B に記載されているため、参照されたい。[1]

図表 2.3.2. ISO/IEC 19795-1:2021 の内容

章	タイトル	内容
1 章	Scope	スコープ
2 章	Normative reference	引用規格
3 章	Terms and definitions	用語、定義
4 章	Abbreviated terms	略語
5 章	Conformance	適合性
6 章	General biometric system	一般的なバイオメトリックシステム
7 章	Planning the evaluation	評価計画
8 章	Data collection	データ収集

9 章	Analyses	分析
10 章	Graphical presentation of results	結果の可視化方法
11 章	Record keeping	記録管理
12 章	Reporting performance results	評価結果の報告

誤照合率(false match rate ： FMR)を評価するために必要なサンプル数は統計的に決められる。Rule of 3 では、誤り率 p 以下であることを有意水準 95%で示すには、$3/p$ 回規模の試験において誤りが 1 回も起きなければよい、というものである。ここで、各試行(照合行為)は互いに独立で同一の分布に従い(independent identically distributed ： i.i.d.)、誤りの生起回数はパラメータ n(サンプル数)および p(誤り率)の二項分布に従うものとする。誤照合率を評価するために必要なサンプル数を図表 2.3.3.に示す。言い換えると、集めたサンプル数が決まれば、評価可能な誤照合率が決定される。なお、ISO/IEC 19795-1:2021 の 7 章では、i.i.d.に従うようにデータ収集で注意すべき事項がまとめられている。

図表 2.3.3. 誤照合率の評価に必要なサンプル数

誤照合率	必要な照合件数	必要なサンプル数
0.001% (10 万分の 1)	30 万以上	775 以上
0.0001% (100 万分の 1)	300 万以上	2450 以上
0.00001% (1000 万分の 1)	3000 万以上	7746 以上

本報告で提案する極値統計を用いた精度推定方法は、ISO/IEC WD5152 として規格開発が進んでいる。本報告執筆時(2022 年 2 月)では、Committee draft(CD)に進むことが決まったところである。ISO/IEC WD 5152 は、ISO/IEC 19795-1、19795-2、19795-6 を引用し、特に ISO/IEC 19795-1 を意識しており、精度推定方法で差分となる内容を中心に記載している。評価指標としては、精度推定方法を適用した誤照合率を Extrapolated FMR として新たに定義しているが、それ以外の指標は ISO/IEC 19795-1 の指標をそのまま使用している。ISO/IEC WD5152 の内容を表 4 に示す。極値統計を用いた精度推定方法では、既存評価方法と異なり、照合スコアの分布の裾をモデル化して外挿することで、誤照合率を推定する。既存の精度評価方法(ISO/IEC 19795-1)では、集めたサンプル数で評価可能な誤照合率が決まっていたが、精度推定方法(ISO/IEC WD5152)では、サンプル数ではこれまで評価できなかった誤照合率を推定することが期待される。

また、精度評価方法への適用においては、既存の精度評価方法と比べて、難しさが増している。極値統計モデル(rGEV、GP)による近似でのパラメータ選定手順、近似結果の妥当性検証、どこまで誤照合率を外挿してよいかの判断などが難しい手順として挙げられる(3 章、4 章で報告)。図表 2.3.5.に、既存の精度評価方法と精度推定方法との主な違いをまとめる。

図表 2.3.4.　ISO/IEC WD5152 の内容

章	タイトル	内容
1 章	Scope	スコープ
2 章	Conformance	適合性
3 章	Normative reference	引用規格
4 章	Terms and definitions	用語、定義
5 章	Details of estimation	精度推定方法の詳細
6 章	Performance metrics	評価指標
7 章	Record keeping	記録管理
8 章	Reporting estimation results	推定結果の報告

図表 2.3.5. 既存の精度評価方法との主な違い

	既存の精度評価方法 ISO/IEC 19795-1:2021	精度推定方法 ISO/IEC WD 5152(開発中)
評価指標	バイオメトリック技術全般	1:1 照合の誤照合率 (Extrapolated FMR)のみ
評価に必要な サンプル数	Rule of 3、Rule of 30 で決定 (ただしそれ以外の方法も許容)	Rule of 3 と比べて少数で評価可能 (具体的な数値指針はなし)
サンプル数の見積りで 使用する統計モデル	二項分布	極値統計(rGEV、GP)
誤照合率算出	サンプル数(試行数)に従い、ほぼ一意に決まる	サンプル数(試行数)だけでなく、統計モデル近似の適合性や推定結果(信頼区間)によって算出される結果が変動
		照合スコア分布が i.i.d.を満たさない場合には統計モデルで近似できないなど適用できない条件がある

【参考文献】

[1] ISO/IEC 19795-1:2021

第3章 新精度評価方法

本章では極値統計モデルを用いた精度評価方法（評価手順と報告方法）のポイントを紹介する。

3.1. 極値統計を用いた精度評価方法の概要

今回導入する Extrapolated FMR は統計的外挿による FMR の推定方法である。スコアを外挿するということは、実際には得られていない瓜二つの生体特徴をもつ別人のデータを想定することに相当する。生体認証に使われる身体的特徴(顔や指紋、静脈など)や行動的特徴(歩容やキーストロークなど)は、そもそも個人差が大きいことを理由に採用されているため、類似度スコアが高い他人対を見つけるためには大量の生体サンプルを集めなければならない。

従来は、この貴重な他人対を持っていることが一つの財産的価値があったため、外挿による推定の導入には慎重な意見もあった。その一方で、行動的特徴にもとづく生体認証システムの場合、他人の動作・所作を模倣することによって意識的に類似度スコアの高い他人対を作り出すことが可能であり、或る意味外挿データを作り出すことが比較的簡単に出来るという現状がある。そこで、今回はモダリティに関わらず、単に極値理論の基本仮定を満たすスコア分布の統計的特性のみにもとづいて得らえる近似直線の延長部分として、外挿を実施することにした。

適用評価に使用したデータベースは、指紋[1]、顔・音声[2]、歩容［3］の大規模データベースを用いた。音声や歩容は模倣した他人データが含まれている可能性は排除できず、統計的には想定を超えた高類似度を示すいわゆる外れ値となっている可能性がある。また、顔については同一被験者が異なる感情の表情を記録したデータベースが含まれていたが、十分なサンプル数がなかったため検証には利用しなかった。

Extrapolated FMR の算出には一般化パレート分布(GP)モデルおよび上位 r 位を用いた一般極値分布(rGEV)モデルの二つを用いた。まず、大規模データベースを用いて、すべてのスコアから算出された FMR を仮の真値とし、推定用サンプルをどの程度まで低減できるかの検討をおこなった。また、同一のデータベースに対して、GP と rGEV の２つのモデルで Extrapolated FMR を算出し、二つの手法の比較検討を行った。Extrapolated FMR 算出の手順の概要を下記に示す。

（1）診断図等を参照しながらハイパーパラメータを決定する

各モデルで用いるサンプルスコアを選択するためのハイパーパラメータを、Q-Q プロットや ξ - θ プロットを見ながら決定する。rGEV ではブロックサイズ n および各ブロック上位から取り出すサンプル個数 r、GP では極値データと見做すスコア閾値 θ がハイパーパラメータとなる。

（2）最尤推定法でパラメータ母数の最尤値を得る

rGEV、GP の両モデルのパラメータ母数を、1 で得られたサンプルスコアを元に最尤推定法で求める。rGEV、GP のパラメータ母数はそれぞれ$\{\mu, \sigma, \xi\}$,$\{\sigma, \xi\}$である。

（3）パラメータ母数の最尤値を使って近似曲線を描く

得られたパラメータ母数の最尤値を使って、実測値と近似曲線を同一平面にプロットする。求めたい値が Extrapolated FMR であるため、CDF(Cumulative Distribution Function、累積分布関数)値を 1 から減じた値を、横軸にスコア値、縦軸に Extrapolated FMR 値としてプロットする。見やすさのために、縦軸は対数軸とする。

（4）近似曲線の信頼区間を描く

3 で描いた Extrapolated FMR のグラフ上に、各推定点における信頼区間をプロットする。2 で得られたパラメータ母数の最尤推定値を RU 法で生成した疑似乱数によって変化させ、100 本以上の近似曲線を得る。上下 2.5%部分を除外した範囲を 95%信頼区間としてプロットする。

それぞれのモデルにおける詳細な手順と解析結果、考察については 3.2 節以降に示す。
なお、実データ[1，2，3]を用いた解析は JAISA バイオメトリクス部会精度評価技術グループメンバーが実施した。その解析結果を、検討委員会メンバーに報告してレビュー、極値統計技術の適用について考察した。

【参考文献】

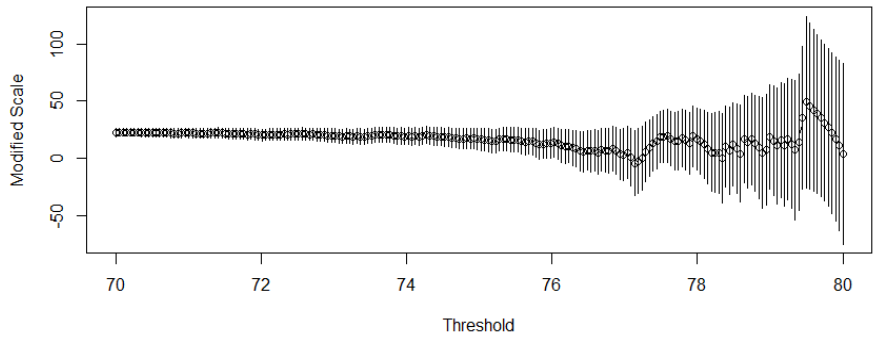
[1] National Institute of Standards and Technology, “NIST Biometric Scores Set （BSSR1)”,
[2] Idiap bioscote database, Laurent El Shafey, “Scalable Probabilistic Models for Face and Speaker Recognition”, PhD thesis,
[3] Y. Makihara, H. Mannami, A. Tsuji, M.A. Hossain, K. Sugiura, A. Mori, and Y. Yagi, ``The OU-ISIR Gait Database Comprising the Treadmill Dataset,'' IPSJ Trans. on Computer Vision and Applications, Vol. 4, pp. 53-62, Apr., 2012.

3.2. GPによる評価手順と報告形式

【評価手順】

（1）閾値を選択

スコアの閾値を変化させた場合に、閾値とスケールパラメータの関係をプロットすることで、パラメータが安定しており、かつなるべく値が大きくなるような閾値を選択する．閾値とパラメータの関係は ismev パッケージの gpd.fitrange 関数により得られ、例えば、図表 3.2.1.の結果を得られたとする。図 1 によると、閾値が 78 を超えると水平にならず急激な変化が生じている。この範囲では推定結果の分散が大きくなり、良好な推定結果が得られないため、閾値として 78 を選択する。



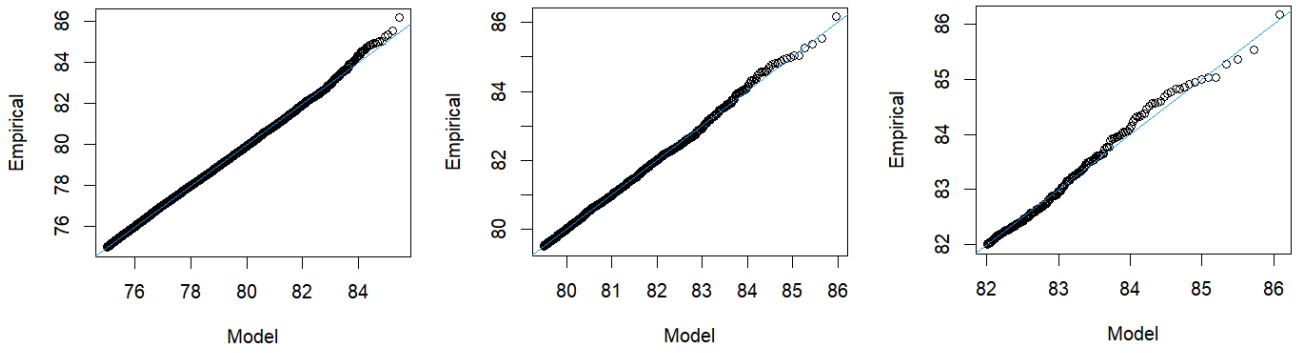
図表 3.2.1. 閾値とスケールパラメータの関係

（2）閾値超過データに GP を当てはめ、最尤法でパラメータ推定

ismev パッケージの gpd.fit 関数で行う。その結果として出力される診断図を基に閾値を調整する。

（3）診断図を用いて閾値を調整し、最適なパラメータを決定

診断図の例を図 2 に示す。このパラメータの良し悪しは左上の Q-Q プロットを参照するのが良い。Q-Q プロットは、横軸が GP の推定結果、縦軸が実測値を表している。推定結果が実測値を良く表している場合は、プロットが y=x の直線にフィットするため、Q-Q プロットが直線に最もフィットするパラメータを探すことが目標になる。図表 3.2.2.には、閾値を 75、78、82 とした場合の Q-Q プロットを示している。初めに 78 を選択したが、最も 78 の Q-Q プロットが直線にフィットしている。閾値は 78 とするのが最適と考えられる。他方、閾値が 75 や 82 では Q-Q プロットが直線からずれており、閾値が大きすぎるのも小さすぎるのも良くないことがわかる。

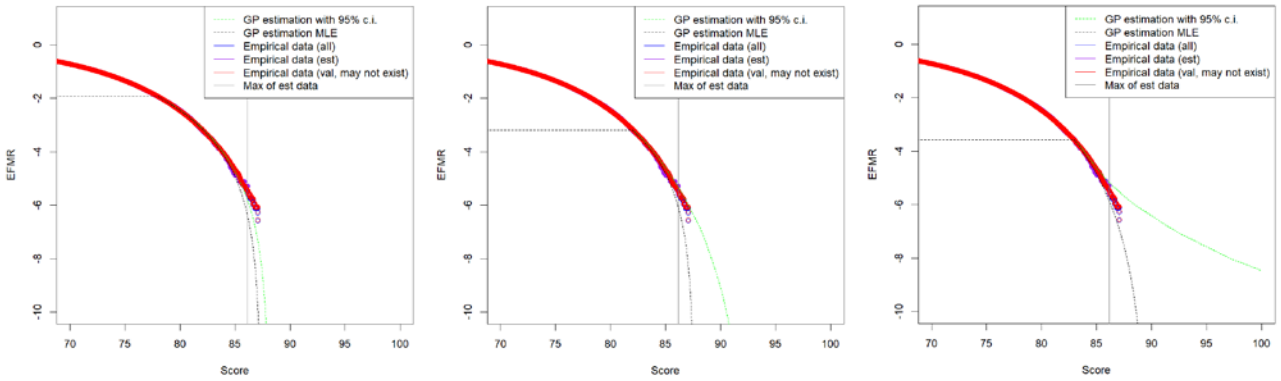


図表 3.2.2. Q-Q プロット（右から閾値が 75, 78, 82）

（4）確率密度関数から誤照合率を推定

gpd.fit 関数で得られた結果は、確率密度関数（PDF）である。誤照合率は、累積分布関数（CDF）を用いて 1-CDF を算出する。CDF は extRemes パッケージの pevd 関数を用いて算出できる。その後、95%信頼区間を以下のように算出する。極値理論において、分布の裾野は測分布の推定に関して、分布の裾野部分は中心極限定理が適用できるほどデータ量が十分でないため、従来用いられてきた中心極限定理を仮定する最尤法よりも、その仮定が必要ないベイズ推定を用いる方が優位であると考察されている。これを生体認証の照合スコアにも適用できると考え、ベイズ推定を用いる例を述べる。まず、ratio-of-uniforms 法によるベイズ推定によりスケールパラメータと形状パラメータの乱数を N 組発生させ、各組において 1-CDF を算出する。続いて、N 本の 1-CDF をプロットし、上端 5%を取り除いた N x 0.95 本に収まる範囲を片側 95%信頼区間として算出する。誤照合率の推定においては信頼区間の上端が意味を持つため、上側の片側信頼区間として算出した。

閾値を 75、78、82 の場合の 1-CDF の算出結果および、N=512 として算出した信頼区間を図表 3.2.3. に示す。閾値が小さすぎる場合は信頼区間が小さいが、極値でない部分で推定を行った結果、極値部分での推定精度が悪い。一方で閾値が大きすぎる場合は信頼区間が大きく広がってしまい安定した推定ができていないことになる。



図表 3.2.3. 誤照合率と信頼区間の推定結果（右から閾値が 75、78、82）

【報告形式】

GP により誤照合率を推定した場合、以下の項目が報告すべき事項である。

・誤照合率推定に統計モデルとして GP を用いたこと

・使用したパラメータとして、スコア閾値、形状パラメータ、スケールパラメータ

・Q-Q プロットなどの診断図（任意）

例えば R の ismev パッケージにおいては、実測値とモデル推定値のパーセンタイルを x、y 軸に取った Probability Plot （通常はこれを Q-Q プロットと呼ぶ）、実測値とモデル推定値のパーセンタイルを x、y 軸に取った Quantile Plot、サンプル数と期待される最大スコア値をそれぞれ x、y 軸に取った Return Level plot、スコアの実測値のヒストグラムとモデルにより算出された確率密度分布を重ねた Density Plot が診断図として得られる。

・モデルによる推定値と 95%信頼区間がそれぞれ明確になるように記述

95%信頼区間の算出方法も記載する。上述したモデルのパラメータをベイズ推定で発生させる方法の他に、プロファイル尤度を用いる方法などがある。

3.3. rGEVによる評価手順と報告形式

本節では、性能評価に使われるデータセットを複数のブロックに分け、その上位 r 位まで極値として取り扱う一般化極値分布（Generalized Extreme Value distribution）を用いる評価手順とその評価結果の報告形式を説明する。以下、この評価方法を r GEV と呼ぶ。また、本節の用語定義は以下の通りである。

(1) データサイズ(N)　：精度評価に使用する母集団（サンプル）のスコア値の数。
(2) ブロックサイズ(n)　：各ブロックに含まれるスコア値の数。
(3) ブロック数(m)　：推定に用いる母集団のスコア値の数 N を n で除した数。
m が小数値の場合は切り上げ。
(4) 順位(r)　：各ブロックから取り出される極値データの個数。
スコア値が類似度の場合は、各ブロックに含まれるスコア値を降順に、距離値の場合は昇順にソートしたときのそれぞれ上位 r 位までのデータが極値データとして用いられる。

【評価手順】

rGEV による評価手順は精度推定と検証の手順から構成される。その手順を図 3.2.1.に示す。
それぞれのステップにおける処理手順と注意事項を以下の通りである。

(1)ブロックサイズ n の選定

ブロックサイズｎは、各々のブロックがある程度の極値データを含むように選定する必要がある。ｎを大きな値にすれば､各ブロックに多くの極値データが含まれる一方､ブロック数 m が少なくなるため、推定に用いるための極値データの数が減るというトレードオフの関係にあることに注意が必要である。

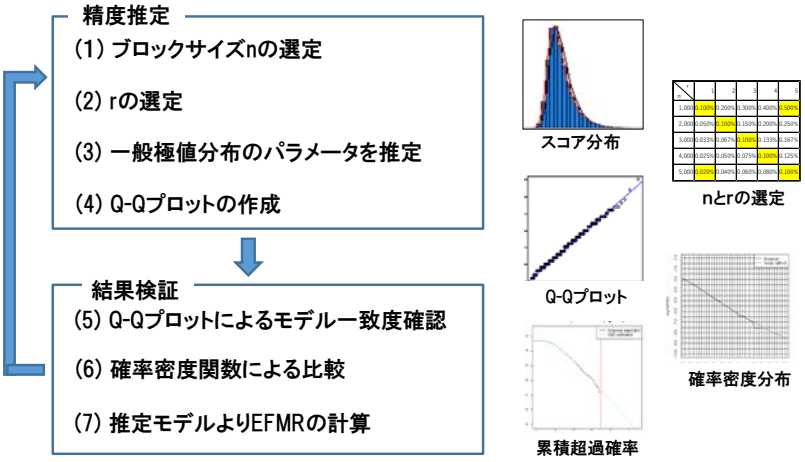


図 3.2.1. rGEV による精度精度手順※の概要

※本手順はＪＡＩＳＡ精度評価技術グループの研修会での実証実験的な例。N5152WD1 の手順とは若干異なり、精度推定（前段）でｎとｒの組み合わせによりモデル推定評価を行い、Q-Q プロットによるモデル適合性の確認を行い、サンプル（実測）データと推定モデルの比較（検定）を行い、推定モデルを積算して、累積確率密度関数を求め、閾値 θ のときの Extrapolated FMR を求める。

(2) r の選定

r は各ブロックからサンプリングされる極値データの個数を決定する値であり、推定に用いられる全データ数は m × r 個となる。一般的に r を大きな値にすれば、より多くの極値データを用いて推定することができるため、精度の高い推定が期待できる一方、あまりに r を大きな値にとると極値データと見做すことができないデータまで推定に使われることになり、逆に推定精度が劣化する恐れがある。

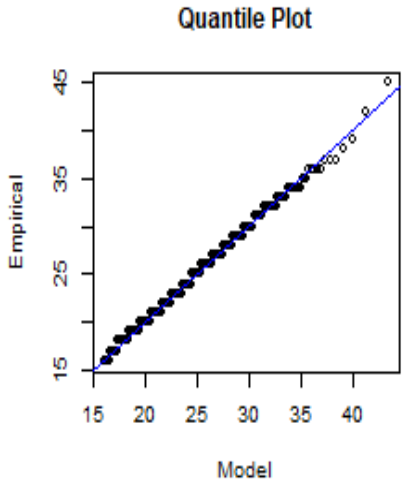
なお、自然現象を対象とする極値推定では一般的に r は 1～5 程度が使われることが多い。

(3) 一般極値分布のパラメータを推定

手順 1、2 を経て得られたデータを推定用極値データとする。推定用極値データの分布が一般極値分布に従うと仮定して、一般極値分布関数の位置(μ)、尺度(σ)および形状(ξ)の各パラメータを最尤推定により求める。最尤推定で求めた各パラメータは ^ を付して($\hat{\mu}$, $\hat{\sigma}$, $\hat{\xi}$)とする。

(4) Q-Q プロットの作成

手順 3 で得られた分布モデルと、推定用極値データの双方のパーセンタイル値をそれぞれ x 軸、y 軸にとって Q-Q プロット（**図表 3.3.2.**）を作成する。



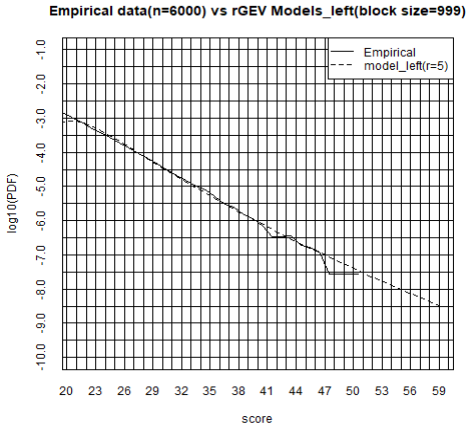
図表 3.3.2.　Q-Q プロットの例

(5) Q-Q プロットを用いてモデルの適合性を確認

手順４で作成した Q-Q プロットを観察し、手順３で得られたモデルが実測値(推定用極値データ)にどの程度適合しているかを確認する。特に Q-Q プロットの右上部分(パーセンタイル値が高い領域)適合の程度、すなわち直線 y=x との乖離の程度に着目してモデルの適合性を確認する。具体的には、直線 y=x 上に多くの Q-Q プロット点が集まっていれば適合性が高く、直線 y=x から離れた値が多ければ適合性が低いと考えられる。

(6) 確率密度関数で比較

推定用極値データのヒストグラムの各頻度を全データ数で正規化して確率密度関数(Probability Density Function, PDF)とし、手順 3 で得たモデルの確率密度関数と同一平面にプロットすることにより、目視で一致の程度を観察する。この時、外挿データ領域での観察をしやすくするために、x 軸にはスコア値を、y 軸には確率をとり、y 軸は対数目盛でプロットする。また、モデルの両側 95%信頼区間も併せてプロットする。



図表 3.3.3. 確率密度関数による比較

手順(5)、(6)でモデルが実測値に十分に適合していると考えられる場合は手順 7 に進む。適合の度合いが不十分な場合は、手順 2 に戻り、異なる r 値を選択して手順 3～6 を繰り返す。もし、適合の良い r が得られなかった場合は、手順 1 に戻り、異なる n 値を設定して手順 2～6 を繰り返す。

(7) 推定モデルより Extrapolated FMR を求める

図 3.3.3.のモデルは、あるスコア値が発生する確率を示す確率密度関数である。ここからある閾値を設定したときに推定される Extrapolated FMR を求めるためには、その閾値までの累積確率を 1 から引けば得られる。例えばモデルの累積分布関数(Cumulative Distribution Function, CDF)を C(x)とすれば、閾値 θ のときの Extrapolated FMR は

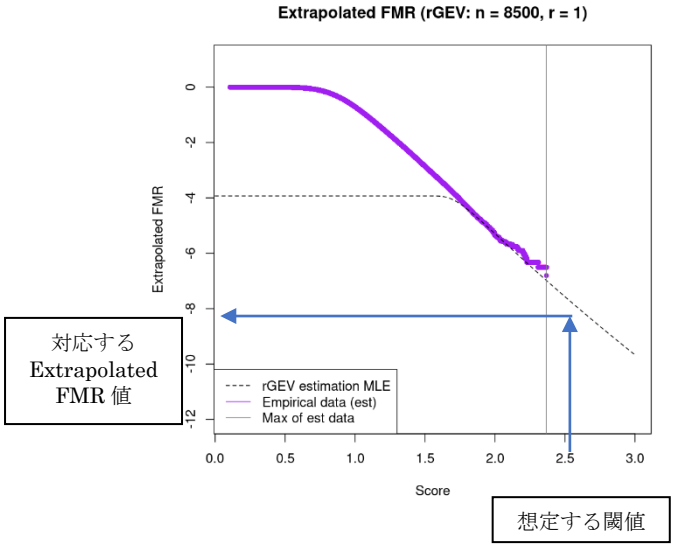
$$\text{Extrapolated FMR} = 1\text{-}C(\theta)$$

で求められる。

図表 3.3.4.にスコア値に対して rGEV 法で推定した Extrapolated FMR をプロットしたグラフの一例を示す。視認性向上のために y 軸は常用対数軸でプロットしている。

黒の実線は実測値（推定用極値データ）、黒色の破線は rGEV による推定値を示す。また、灰色の縦線は実測値の最大値を示しており、それより大きなスコア値は外挿値であることを示している。

このグラフから任意の閾値における Extrapolated FMR を求めることができる。例えば、閾値を 2.5 に選んだ場合、rGEV の推定値は 10e-7.568388 = 2.701545×10^{-08} となる。



図表 3.3.4. Extrapolated FMR の決定

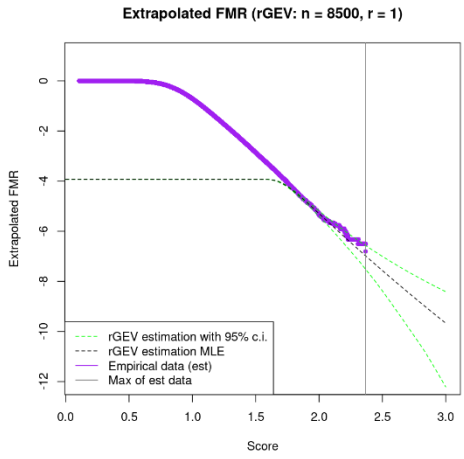
得られた Extrapolated FMR は n, m, r, θ および C(θ)の信頼区間と共に報告する。また、閾値 θ のときの FRR が分かっていれば、併せて報告する。

(8) 信頼区間を求める

報告に必要な θ および C(θ)の信頼区間算出には、様々な方法が考えられる。ここでは、(疑似) 乱数生成に、ベイズ推定を利用した、一様乱数の比による受理・棄却法 (RU 法) を使う場合について示す。

RU 法を用いるにあたり、ベイズ推定の事前分布を正規分布とする。この時、無情報となるように、その平均値を$(\hat{\mu}, \hat{\sigma}, \hat{\xi})$、その分散を非常に大きくとる。さらに、この事前分布と rGEV の対数尤度関数を合わせたものをベイズ推定の事後分布 (母数のベイズ分布) とする。

この事後分布から M (≧100) 個の乱数組 $\theta_i=(\mu_i, \sigma_i, \xi_i)$を生成する。$\theta_i$の各パラメータを 1 次元プロットし、左右 2.5%分を除外すれば θ の 95%信頼区間を得られる。また、Extrapolated FMR (=C(θ)) の信頼区間は、θ_iによる推定曲線を描き、想定する定義域の各点で上下 2.5%を除外したものを集めれば得られる。図表 3.3.5.に Extrapolated FMR の信頼区間を併せて描画した例を示す。



図表 3.3.5. Extrapolated FMR と 95%信頼区間

【報告形式】 (前年報告書より)

r GEV の報告形式は GP の報告形式とほぼ同じである。大きな違いは r GEV で用いたブロックサイズ n と r を評価条件として記載しておくことと、m 個のグループに分けた場合、グループ分割の再現性がなくなるため、グループ分割してデータセットを残しておくことである。また、r と n をどのような範囲で変化させた結果であるかも記録しておくと、本方式による探索的データ解析をどの範囲で実施したかを示すことができる。

なお、以下では R の極値統計パッケージ ismev による診断図を例に、モデルの適合性検証方法および報告内容のポイントを示す。

(1) Probability Plot

実測値およびモデルによる推定値をそれぞれ x、y 軸に取り、モデルの適合性を視覚化したグラフ。x,y 軸の値はともに[0,1]の値を取る。通常はこのプロットを Q-Q プロットと呼ぶ。直線 y=x 状に多くのプロットが載っているほど、モデルの適合性が高いと考えられる。既知の実測値との比較であるが、サンプル数が少なくなるグラフの部分の適合の程度がよければ、それだけ外挿値の高い妥当性が期待できる。

（2）Quantile Plot

実測値およびモデルによる推定値のスコア値をそれぞれ x、y 軸に取り、モデルの適合性を視覚化したグラフ。x,y 軸の値はともに推定に用いた極値領域のスコア値を取る。Probability plot 同様、直線 y=x 上に多くのプロットが載ることがモデルの適合性を示す。スコア値で表現しているため、特定の(生体認証スコアの)閾値を設定したときにモデルが実測とどの程度一致しているかを判断しやすい。

（3）再現レベルプロット（Return Level plot）

サンプル(スコア)数、期待される最大スコア値をそれぞれ x、y 軸に取ったグラフ。より多くのスコアを収集したときに平均 1 回の観測が期待できる最大スコア値を示す。実際に収集できるスコア数は、通例、時間・経済的制約等により限られているが、再現レベルプロットでは理論上無限のサンプル数での最大スコアを推定することができる。

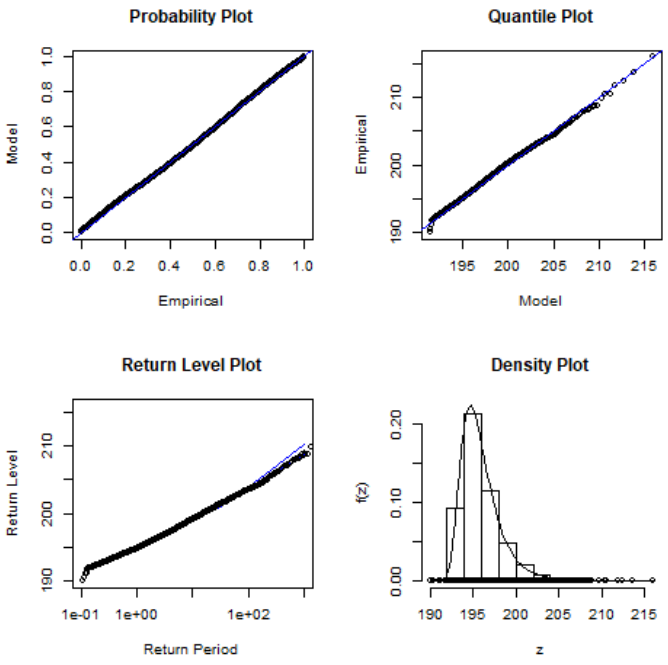
（4）確率密度プロット（Density Plot）

推定に用いたスコア値の領域における実測値のヒストグラムと、モデルによる推定分布を重ねてプロットしたもの。一般極値分布関数が極値領域においてどの程度適合しているかを視覚的に確認できる。外挿の領域においてはリニアスケールでは差異が小さく目視での適合性の判断が困難であるため、縦軸を対数軸にプロットした方が分かりやすい。

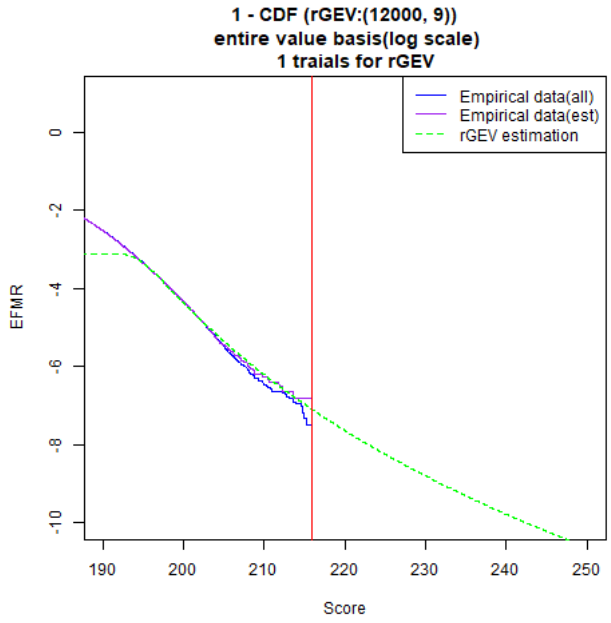
図 3.3.5.および図 3.3.6. に適合性の高いモデルと低いモデルの一例を示す。両図共に、同じデータベースに対して、同じブロックサイズ(12,000)、異なる r 値(9 および 2)で抽出された評価用データに対して一般極値モデルを適用し、パラメータ母数を最尤推定して得られたものである。

適合性の高いモデル(図表 3.3.6.(a))の Quantile Plot を見ると、グラフの右上部分までプロットが直線 y=x 上に載っているのに対し、低いもでるの Quantile Plot はかなり上方にずれていることが分かる。これはモデルの方が実測値よりも精度的に楽観的な数値を示すことを表しており、実施、Extrapolated FMR 値を示す図表 3.3.6. (b)図で比較すると、適合性の低いモデルはかなり小さな Extrapolated FMR、すなわち他人受入率を小さく見積もっていることが分かる。

なお、この例では、データベース全体のスコア出現順序はランダマイズしていないため、順序を変えると同じブロックサイズ、同じ r 値であっても異なるモデルが得られる可能性が高い。実際にはスコア出現順序をランダマイズして rGEV 法を繰り返し実施して、モデルの安定性、妥当性を検証することが必要である。



(a)



(b)

図表 3.3.6. モデル適合性の高い事例 （n=12,000, r=9）

(a)診断図 (b) 1-CDR(Extrapolated FMR)プロット

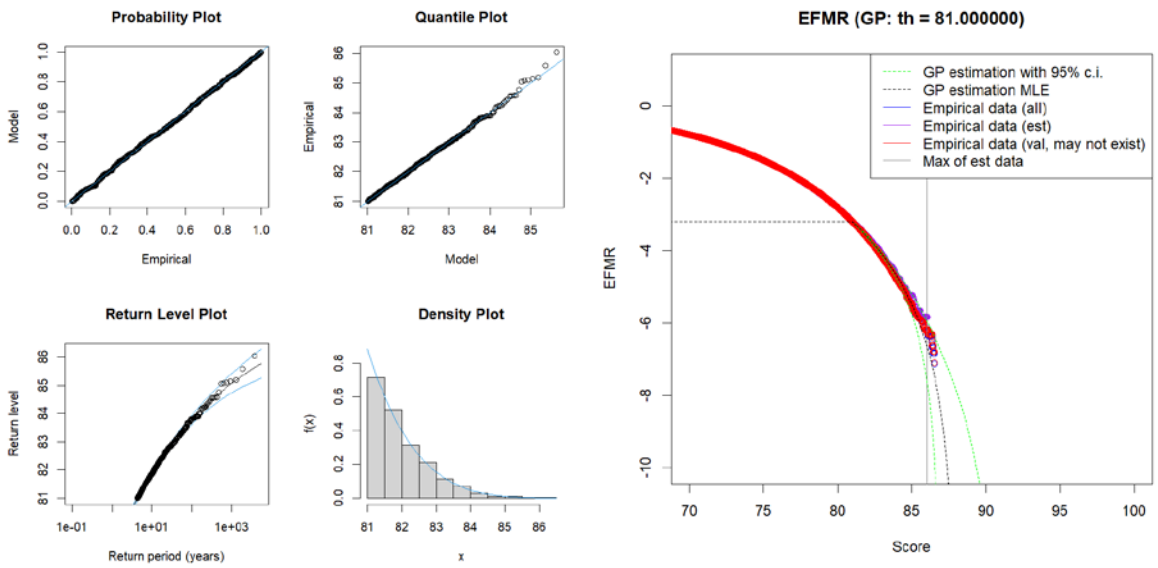
第4章 適用評価例

4. 1. 適用評価の良い例と悪い例

本節で極値統計モデルにフィットする良い例とフィットしない悪い例について説明する。

4.1.1. GPの例

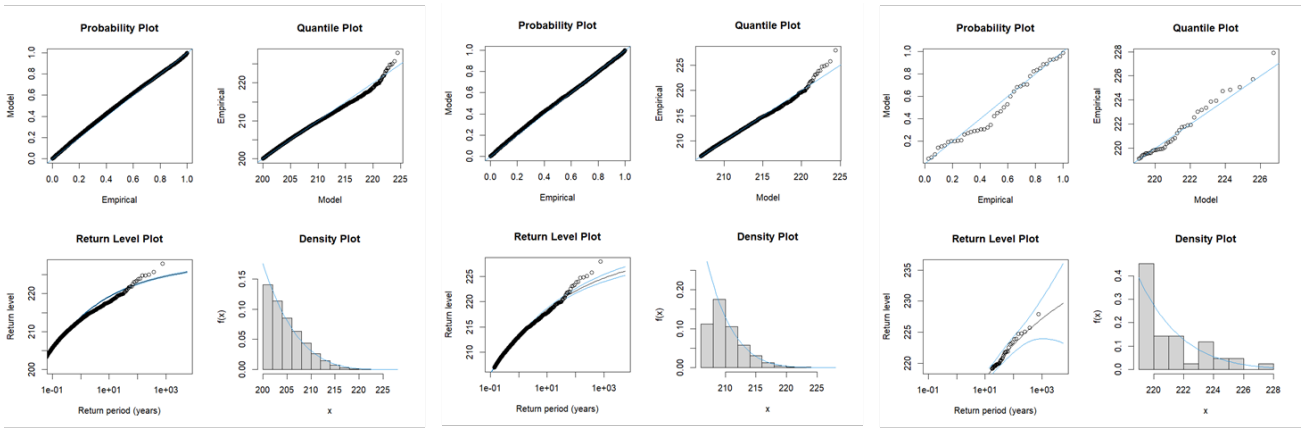
ここでは、GP を用いて誤照合率を推定した 2 例を示す。第一の例は、歩容の他人対約 1,370 万件のスコアデータを 10%ランダムサンプリングした約 130 万件のデータを用いて誤照合率を推定した結果である。Q-Q プロットが直線にフィットするパラメータを選択することで、良好な推定結果を得られた。図表 4.1.1.1.において、赤が元の 1,370 万件のデータにおける誤照合率を示す正解値で、黒点線が GP による推測値、緑点線が 95%信頼区間の両端になる。今回の評価では、推定すべき正解データが判明している状態で実施しているため、正解データが 95%信頼区間内に収まっていることで、良好な結果を得たと判断している。実運用の際は正解データが分からないため、診断図によって判断することになり、特に Q-Q プロットが y=x の直線にフィットするパラメータを選択すればよい。



図表 4.1.1.1.歩容データの診断図と誤照合率推定結果

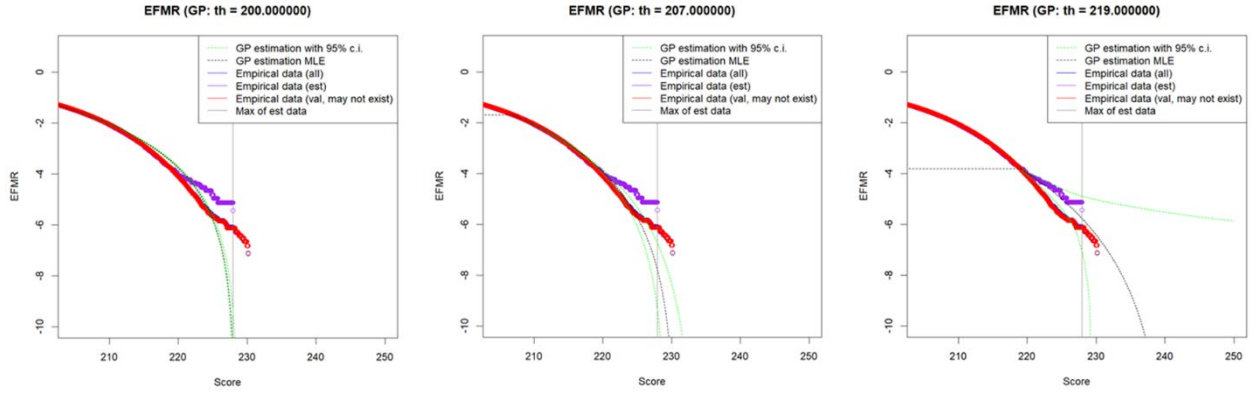
第二の例は音声の他人対約 1,300 万件のスコアデータを 20%ランダムサンプリングした約 1,300 万件のデータを用いて誤照合率を推定した結果である。第二の例では、図表 4.1.1.2.のように閾値を変えても Q-Q プロットが直線状にならないことがわかった。このような場合はいずれの閾値でも推定結果は大きな誤差を含み（図表 4.1.1.3.）、うまく推定できない例だと言える。この結果は、スコア分布の裾野部分に外れ値が生じていたためである。図表 4.1.1.4.に歩容データと音声データの対数軸のスコアヒストグラムを示したが、歩容データは極値の部分でも比較的きれいな分布をしているのに対し、音声データは外れ値が生じており、これがうまく指定できなかった主要因だと考えられる。なお、この音声スコアデータに関しては、rGEV でも同様にうまく推定できなかったため、使用した極値分布モデルが原因ではないと考えられる。今回の評価では推定すべき正解データが分かっていたため、推定できない理由が明瞭であるが、実適用では、同じ ID で別の特徴量で試す、同じ特徴量で別な ID で試す、などの試行

により，原因を類推する必要がある。以上のように、Q-Q プロットが直線にフィットするという観点が推定の良し悪しを決める重要なファクターである。



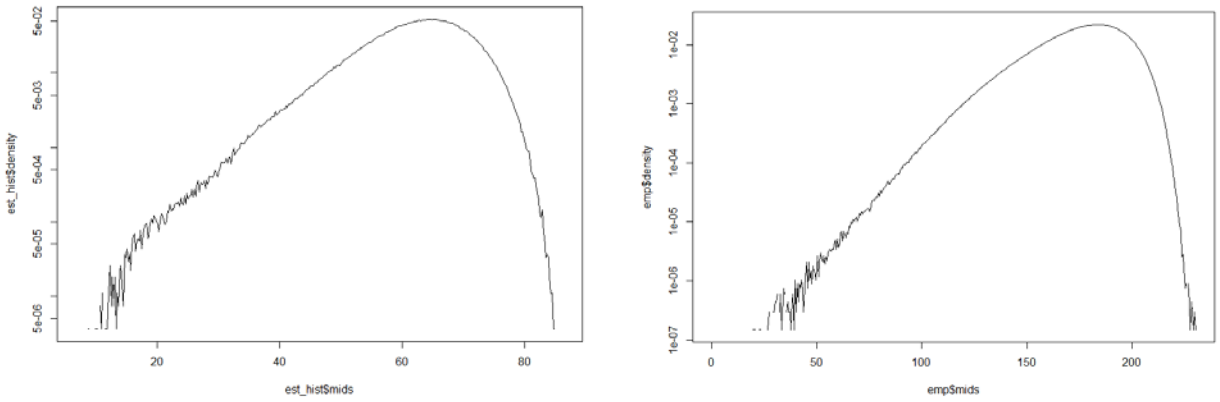
閾値＝200の診断図　　閾値＝207の診断図　　閾値＝219の診断図

図表 4.1.1.2. 音声データの診断図



閾値＝200 の推定結果　　閾値＝207 の推定結果　　閾値＝219 の推定結果

図表 4.1.1.3. 音声データの誤照合率推定結果



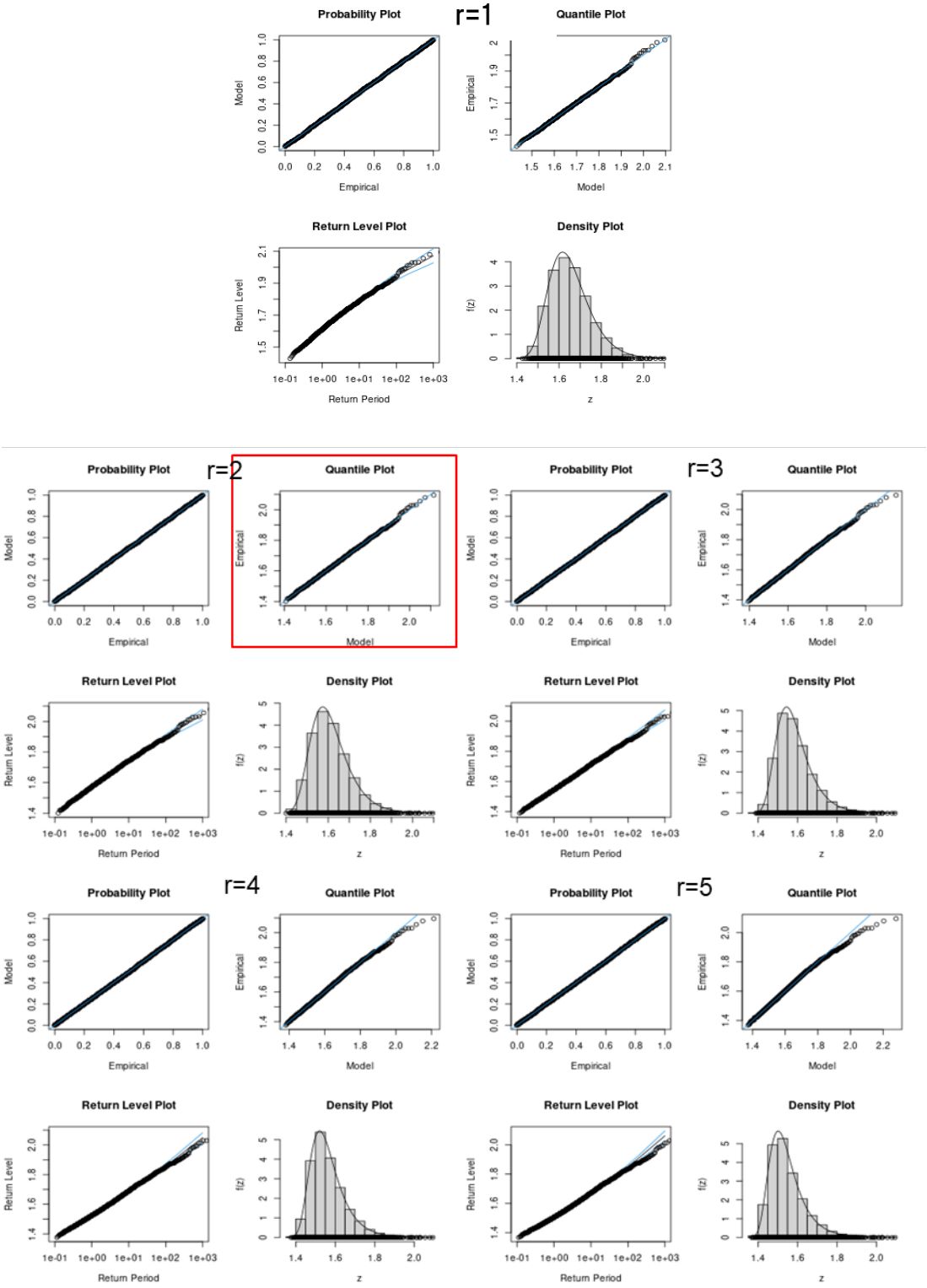
歩容データ　　音声データ

図表 4.1.1.4. スコア分布の違いによる影響

4.1.2. rGEVの例

本項では、rGEV 評価を同一のデータセットに適用し評価する過程を通じ、推定結果の適合性を判断する例（良い例、悪い例）を示す。なお、使用データセット（顔の静止画）ついては次節で詳しく紹介する。

この例(図表 4.1.2.1) は、r=1 と固定し、QQ プロットの乖離が小さい、最大のブロック数 n(4,000) を見つけ、r＝5 まで変化させた診断図である。



図表 4.1.2.1　rGEV 評価例（r＝1～5, n=4000）

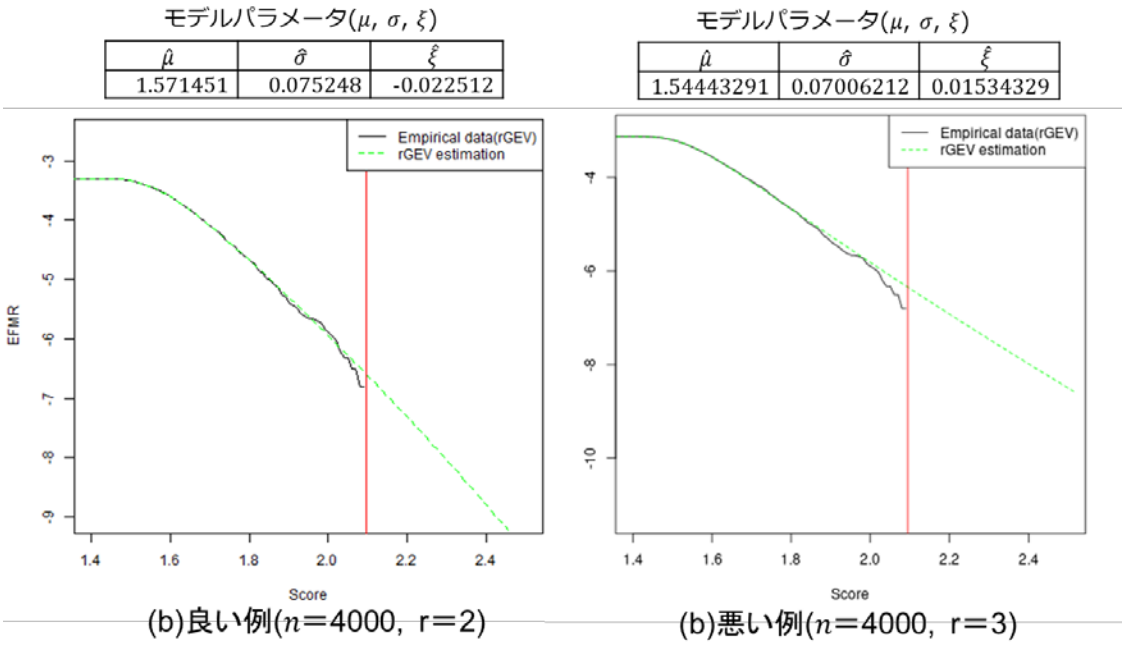
図表 4.1.2.1（n＝4000）の QQ プロット(診断図の右上図)において、r=1 は実績値が上振れしていて、r=2（赤枠部分）もっともフィットしている。r=3、r=４、r=5 と r が増えるについて、実績値が下振れしてしている。そこで、ブロックサイズｎ=4,000、r=２と r=3 の CDF プロットを図表 4.1.2.2.に示す。

【良い例】 ｎ=4,000、r=２（左図）

良い例では、分析結果(緑の点線)がスコア 1.9 あたりから実データ(黒の実線)との上下に乖離があるものの、概ね良好（乖離が少ない）である。

【悪い例】 ｎ=4,000、r=3（右図）

一方、悪い例では、r スコア 1.9 あたりから実データ(黒の実線)との分析結果(緑の点線)の差が広がり、分析結果(緑の点線) と実データ(黒の実線)の乖離が大きくなってきている。



図表 4.1.2.2. CDF プロット図の比較

これは極値域のデータ数が増え、裾野ではないデータが多くなることで、極値モデルに適さなくなってきているものと考えられる。サンプリングされたデータの中で、どこまでを極値データとするかが重要な観点となっている。

次節では極値統計モデルを用いて得られた評価結果、特に信頼区間を中心に、評価結果の信頼性について r GEV と GP の比較を紹介する。

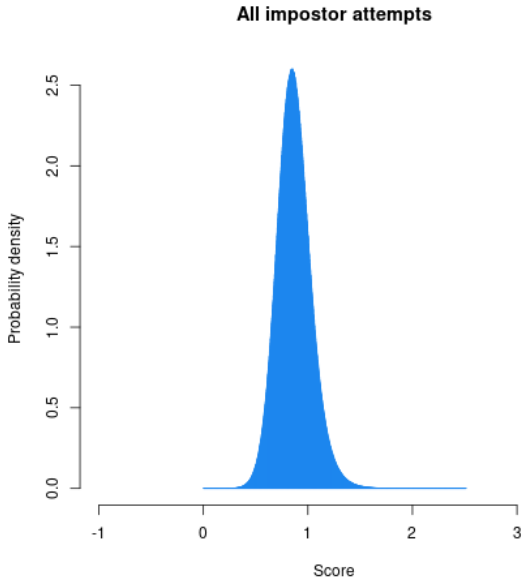
4.2. 極値統計を用いた評価手順と導入効果

4.2.1.　２つの極値統計モデルを用いた精度評価　－rGEVとGPの比較－

本節では、rGEV と GP による評価を同一のデータセットに適用し評価する過程を通じ、推定結果の適合性を判断する例を与える。本節で対象とするデータセット(以下、対象データセット)の概要を図表 4.2.1.1.に、対象データセットに含まれる他人照合スコアの頻度分布を図表 4.2.1.2.に示す。なお、今回はスコアの最小値が 0 となるように offset を適用して評価に用いた。

図表 4.2.1.1. 適用対象データセットの概要

モダリティ	出典 [1]	スコア値属性	データ件数
顔(静止画像)	bioscote/frgc/gmm/scores/2.0.4/nonorm/scores-dev	類似度	31,927,840 件



図表 4.2.1.2. 他人照合スコアの頻度分布(offset 適用後)

次に、本節で利用する評価環境を述べる。本節では、データセットの評価は R (バージョン 4.0.3)上で実施する。また、極値統計に関わる処理(フィッティングや評価グラフの出力など)は、R のライブラリである ismev （バージョン 1.42）および extRemes （バージョン 2.0-12）を、信頼区間算出には revdbayes（バージョン 1.3.9）をそれぞれ利用する。

以下では、まず、rGEV および GP を適用した適合性評価の過程を示す。続いて、両者の結果を比較し考察を述べる。その後、評価手順を効率良く進めるにあたって幾つかのノウハウを示して結びとする。

---

1 命名規則は https://www.idiap.ch/dataset/bioscote を参照

（1）rGEV を用いた Extrapolated FMR 推定

rGEV を用いた評価手順を以下に示す。

a) ランダマイズした対象データセットの先頭から 20%を推定用データセットD_{est}、残りを検証用データセットD_{val}とする。

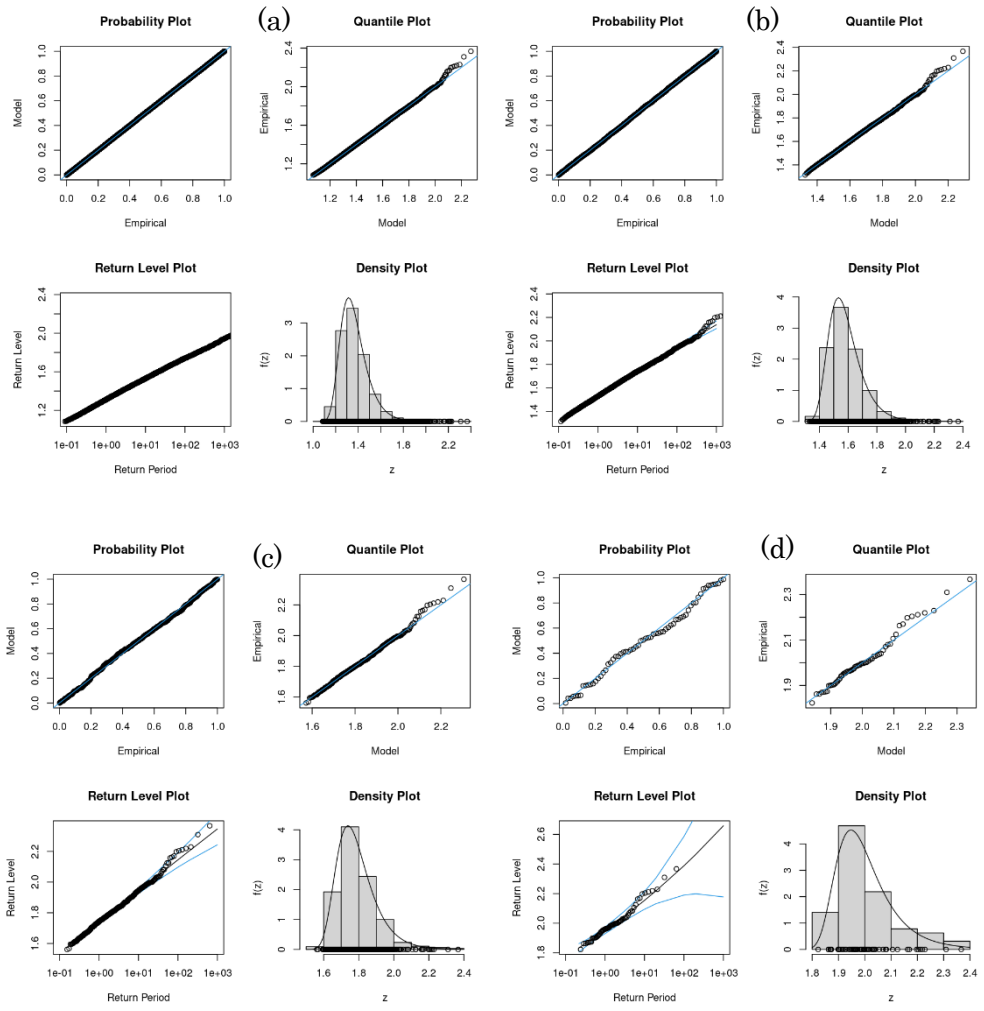
b) D_{est}のスコアの出現順序をランダマイズする。

c) 3.2 節で示した手順により、D_{est}と最も適合性が高いと考えられるモデルのハイパーパラメータ(n, r)を得る（フィッティング及び適合性評価には rlarg.fit、rlarg.diag を利用する）。

d) 得られたモデルの CDF と、D_{est}の累積確率密度分布を求め、それぞれを 1 から減じた値（Extrapolated FMR）をプロットする。

e) Extrapolated FMR の 95%信頼区間を求めてプロットし、外挿の妥当性を検証する。

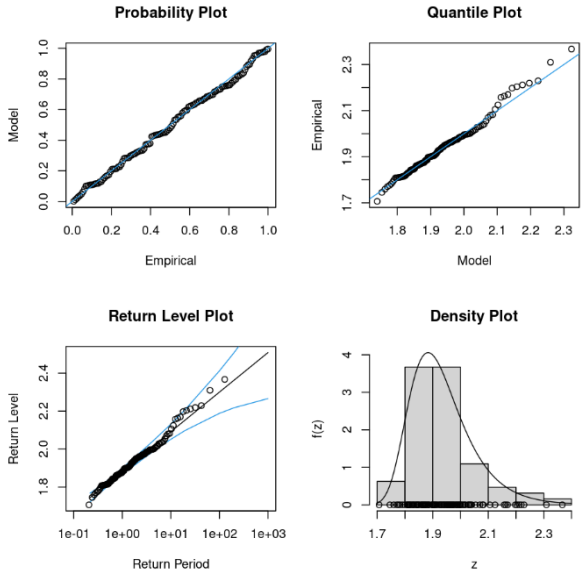
モデルのハイパーパラメータは、3.2 節に沿ってブロックサイズから探索（r = 1 固定）する。探索範囲を絞るため、n = 100, 1000, 10000, 100000 とした診断図（rlarg.diag 結果）を図表 4.2.1.3.に示す。



図表 4.2.1.3. 様々な n における rlarg.diag の出力

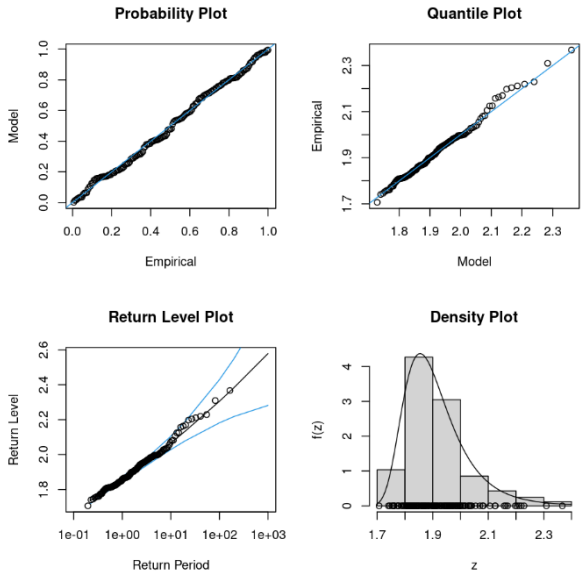
(a) n = 100、(b) n = 1000、(c) n = 10000、(d) n = 100000

図表 4.2.1.3.より、10000 ≦ n ≦ 100000 を詳しく探索する。ここで、n = 100000 の場合は Return Level Plot の信頼区間が大きく開いていることから、10000 ≦ n ≦ 50000 に適当な n があると予想し、n = 50000 とした診断図を確認する。



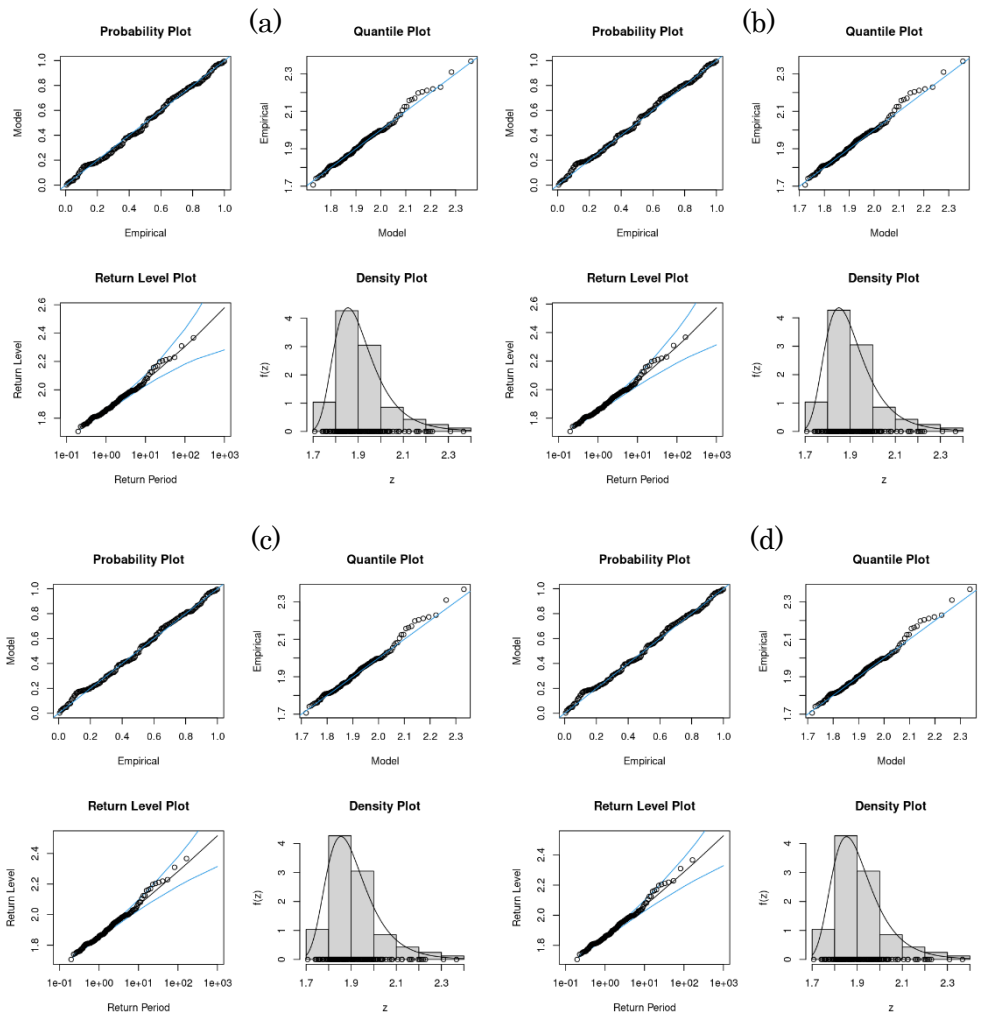
図表 4.2.1.4. $(n,r) = (50000,1)$における rlarg.diag の出力

図表 4.2.1.4.より、n = 50000 の時も、Return Level Plot の信頼区間の広がりがみられることから、10000 ≦ n ≦ 50000 に絞って探索する。今回の実験では、1000 区切りで探索した結果、n = 39000 が最も Q-Q Plot の適合性が高いブロックサイズであると判断した。



図表 4.2.1.5. $(n,r) = (39000,1)$における rlarg.diag の出力

次に、n = 39000 とし、r を変化させる。ここで、r の定義域は自然現象の推定などでよく使われる 1 ≦ r ≦ 5 とした。

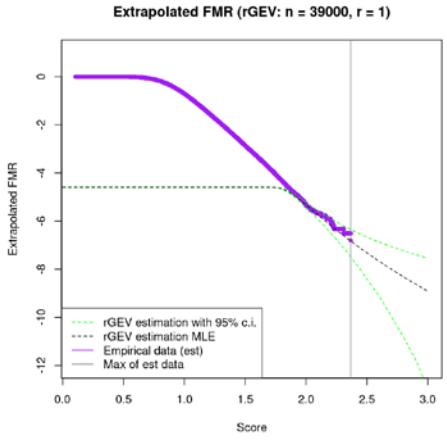


図表 4.2.1.6. 様々な r における rlarg.diag の出力

(a) r = 1、(b) r = 2、(c) r = 3、(d) r = 4

r≧3 では y = x からの乖離が見られるため、r = 1、2 が候補に挙がる。微小な差ではあるが、r = 2 は r = 1 と比較して右端の点がやや y = x から離れている。従って、最も適合性が高い r=1 を採用する。

(n, r)= (39000, 1)とした場合の、Extrapolated FMR を図表 4.2.1.7.に示す。



図表 4.2.1.7. $(n, r) = (39000, 1)$における Extrapolated FMR

データ数が少なくなるスコア値 1.9 近辺からモデルと実測値の間に若干の乖離が見られるが、全体としては 95%信頼区間の中に納まっており、概ね良好な推定結果が得られていることが分かる。

(2)GP を用いた Extrapolated FMR 推定

GP を用いた評価手順を以下に示す。

(a) 対象データセットの先頭から 20%を推定用データセットD_{est}、残りを検証用データセットD_{val}とする。

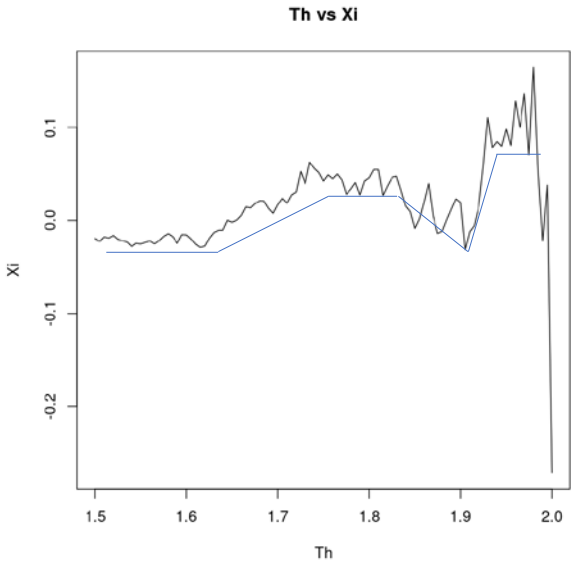
(b) D_{est}から$\theta - \xi$プロファイルを作成し、θを絞り込む（ξ算出には gpd.fit を利用）

(c) 適合性評価により最適なθを決定する（評価には gpd.diag を利用）。

(d) 得られたモデルの CDF とD_{est}の累積確率密度分布を求め、それぞれ 1 から減じた値（Extrapolated FMR）をプロットする。

(e) Extrapolated FMR の 95%信頼区間を求めてプロットし、外挿の妥当性を検証する。

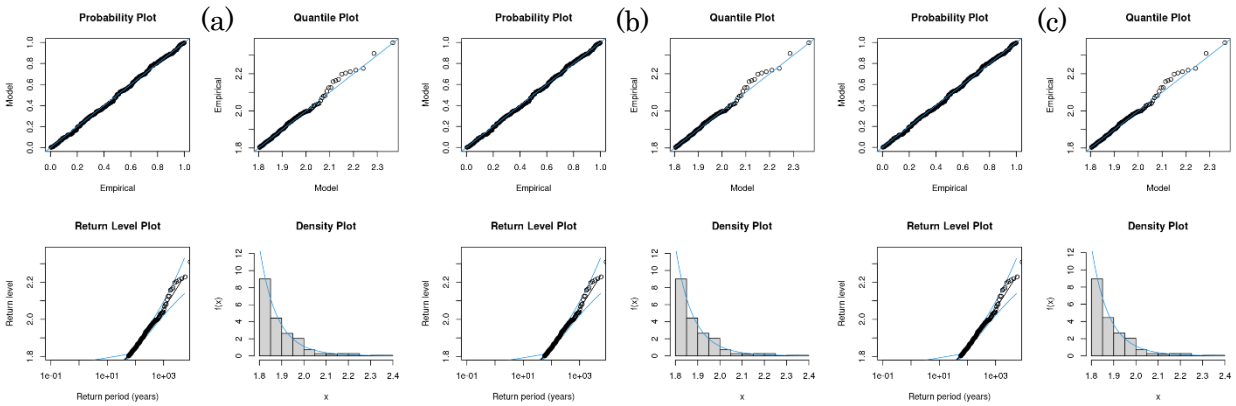
図表 4.2.1.8.は対象データセットの θ - ξ プロファイルとその概形である。



図表 4.2.1.8. $\theta - \xi$プロファイル

曲線の傾きが 0 に近い部分を閾値 θ の候補として絞る。概形で見ると、$\theta \leqq 1.64$、$1.75 \leqq \theta \leqq 1.85$、$\theta \geqq 1.9$ の三つが候補として挙がる。しかし、$\theta \geqq 1.9$ ではサンプルの数が少なく（～100 個程度）、プロファイル（ξ の値）が乱高下している。従ってこれを除き、$1.75 \leqq \theta \leqq 1.85$ に絞り込む。

$1.75 < \theta \leqq 1.85$ に含まれるサンプルは約 350 個であるため、これらの値を θ として診断図を確認する。

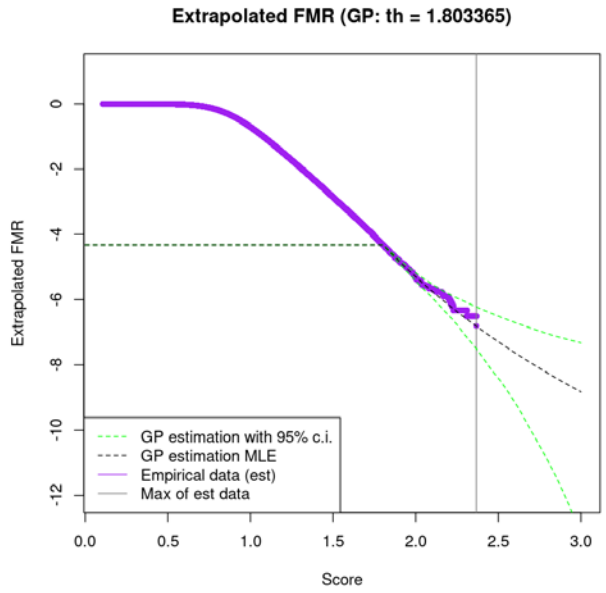


図表 4.2.1.9. 様々なθにおける gpd.diag の出力

(a) θ = 1.803159、(b) θ = 1.803365、(c) θ = 1.803413

図表 4.2.1.9.を見ると、(a)、(b)はほとんど変わらず、(c)は(a)、(b)と比較して右端の点が y = x から乖離している。従って、より大きいθ = 1.803365 を採る。

θ = 1.803365 とした場合の、Extrapolated FMR を図表 4.2.1.10.に示す。

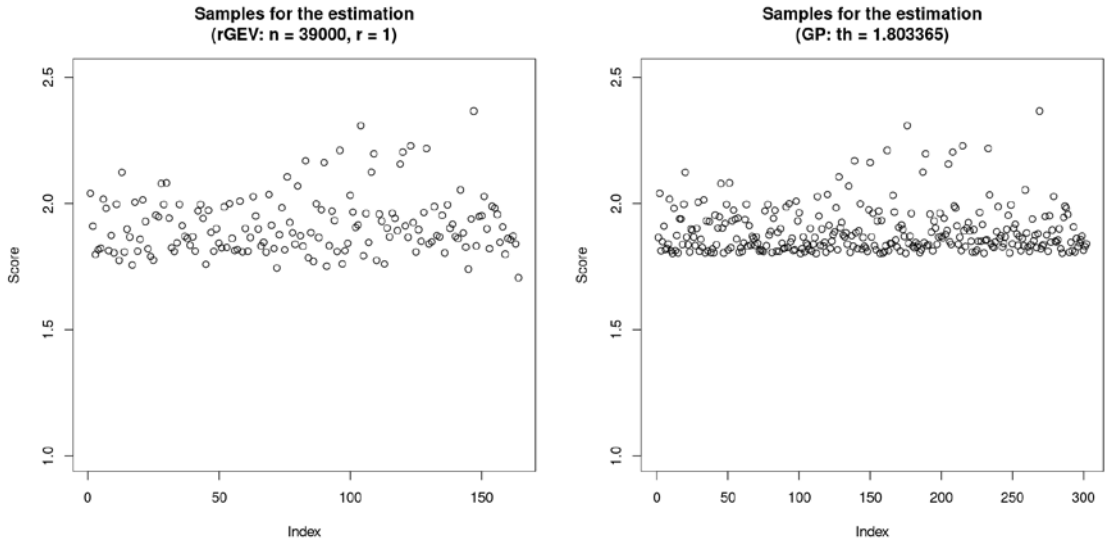


図表 4.2.1.10. θ = 1.803365 における Extrapolated FMR

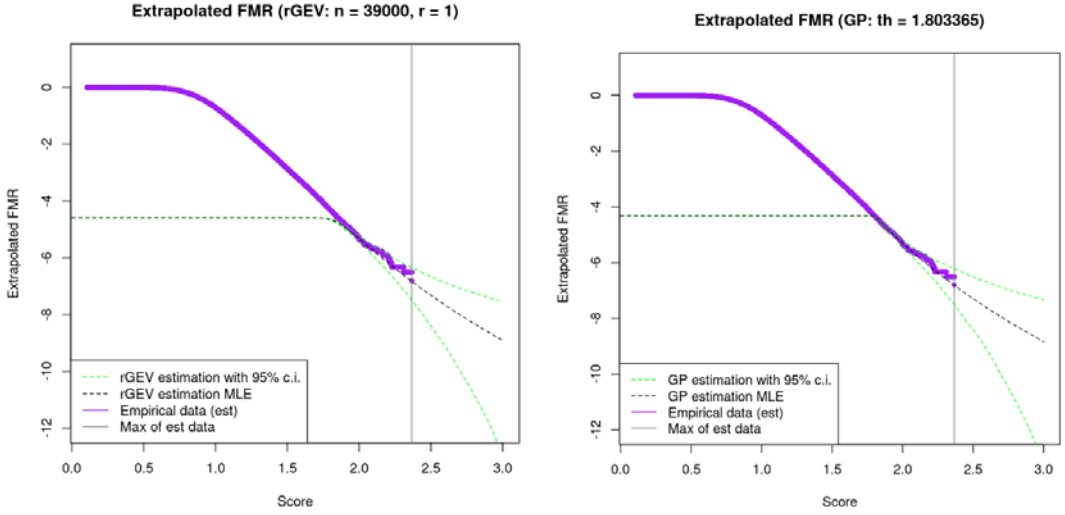
データ数が少なくなるスコア値 1.9 近辺からモデルと実測値の間に若干の乖離が見られるが、全体としては 95%信頼区間の中に納まっており、概ね良好な推定結果が得られていることが分かる。

(3)比較と考察

図表 4.2.1.11.は rGEV および GP の推定時に極値と見做したスコア値、すなわち rGEV では各ブロックの上位 r 位、GP ではスコア値θ以上のデータの散布図である。それぞれ独立して極値領域を設定したにもかかわらず、ほぼ同じ範囲を推定に利用していることがわかる。



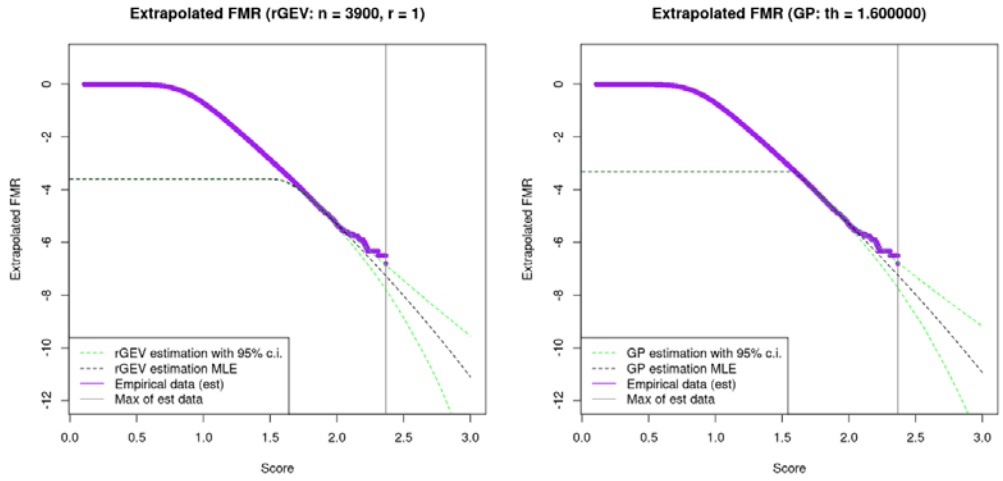
図表 4.2.1.11. モデル推定時極値とみなされたサンプル



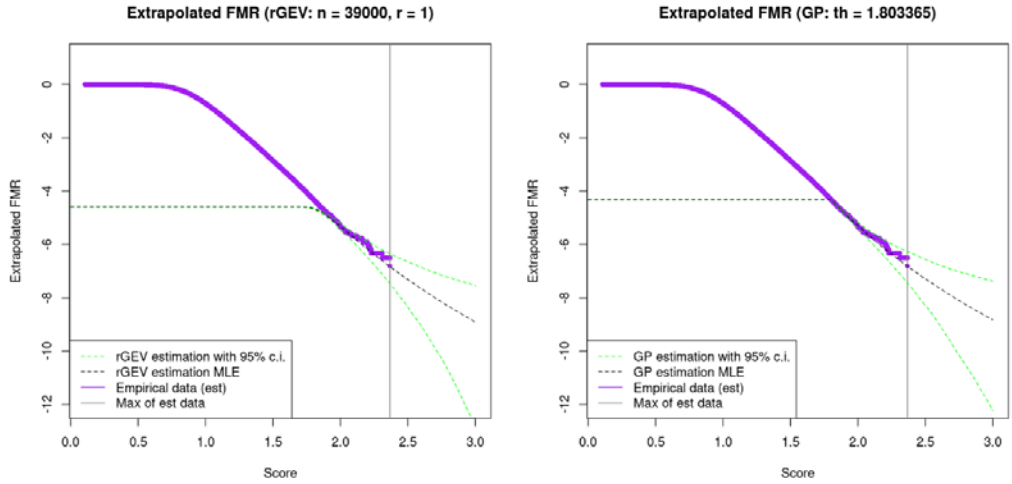
図表 4.2.1.11. Extrapolated FMR の比較

図表 4.2.1.11.の両者を比較すると、GP の方が下側信頼区間にやや広がりがあるものの、ほぼ両者同等の結果を示している。信頼区間の幅については、推定に用いたサンプルの数や、revdbayes で発生させたパラメータ組（疑似乱数）の数による影響が考えられる。実際に、より多くのサンプル（図 4-1-13 描画時の約 10 倍）を用いた推定の結果を図表 4.2.1.12.に、より多くのパラメータ組（図 4-1-13 描画時の 2 倍）を用いた推定の結果を図表 4.2.1.13.に示す。

図表 4.2.1.13.では、図表 4.2.1.12.と比較して信頼区間の幅は狭くなっているものの、区間からはみ出しているサンプルがあり、推定結果として良いものとは言えない（ただし、より良いパラメータを探索して、推定精度を向上させられる可能性はある）。また、図表 4.2.1.13.では図表 4.2.1.11.と比較して信頼区間の幅がより広くなっている。これは、疑似乱数の生起確率によるものと考えられる。



図表 4.2.1.12. より多くのサンプルを用いた Extrapolated FMR 信頼区間の比較



図表 4.2.1.13. より多くのパラメータ組を用いた Extrapolated FMR 信頼区間の比較

(4)評価を効率的に進めるために

極値モデルのハイパーパラメータ探索は、闇雲に進めると膨大な数の実験を行い、大量の診断図を読み、微細な差を比較することになる。そこで本項では、極値モデルのハイパーパラメータ探索をより効率よく進めるためのノウハウをいくつか示す。

(a) rGEV

rGEV は、データをサイズ n のブロックに分けブロック内の上位 r 番目までを極値とみなす。n は自然現象であれば月や年といった時間軸であることも多い。しかし、生体認証では、分割するブロックサイズに意味付けができるケースは少ない。従って、効率化のためには、はじめに二分探索的な手法によってある程度範囲の見当をつけ、その中で最も適当なものを総当たりで探索する。

二分探索的な手法においてポイントとなるのは次の二点である。

① 推定に使われると目されるデータ数

② Return Level Plot の信頼区間幅

推定に使われると目されるデータ数は、単純に探索範囲の上限を決めるために用いる。今回の実験では、用いたRライブラリでは、少なくとも100~1000個程度のデータがあると推定結果が安定しやすい。そこで、探索範囲の上限として、100個程度のサンプルが得られる $n \leqq 63855$ を目安としている（$50000 \leqq n \leqq 100000$ を探索範囲から外した判断要因の一つ）。今回の実験では念のため、二分探索開始時の上限を目安より一桁大きい、$n = 100000$（データ数約63個）とした。

Return Level Plotは、データ数、期待される最大（スコア）値をそれぞれx、y軸に取ったグラフであり、より多くのスコアを収集したときに平均1回の観測が期待できる最大（スコア）値を示す。そのため、Extrapolated FMR（の右側先端部分）とみなすことができ、信頼区間（Extrapolated FMRの信頼区間は計算方法が異なるため、厳密には一致しない）も併せて推定結果を素早く確認することができる。診断図を確認した時点で、信頼区間が極端に広がっている場合は、そのnに対する探索を打ち切ることができる。ただし、信頼区間幅は狭ければ良いというものでもなく、少なくとも推定に用いたデータがその中に納まるのが適当と言える。

一方rは、先に述べたように、自然現象であれば経験的に $1 \leqq r \leqq 5$ を用いることが多いようである。しかしこちらも、生体認証に関しては目安となるrの範囲の指針はない。従って、nを決めた後にrを増やしていき、診断図で適合性を判断することになる。ここでも、Q-Q Plotの他Return Level Plotも有用である。

(b) GP

GPは、閾値 θ を超えるデータすべてを極値とみなす。

ハイパーパラメータである閾値の見当には、大局的には$\theta - \xi$プロファイルが有用である。$\theta - \xi$プロファイルは、計算時にとる θ の個数（θ の分解能）によって、同じ定義域を持つ場合も形が変わる。従って、移動平均を取って、その概形を基に見当をつけるのが良い。その後、見当をつけた範囲について、適合性の高い θ を細かく調べることになる。

$\theta - \xi$プロファイルの眺め方のポイントは次の二点である。

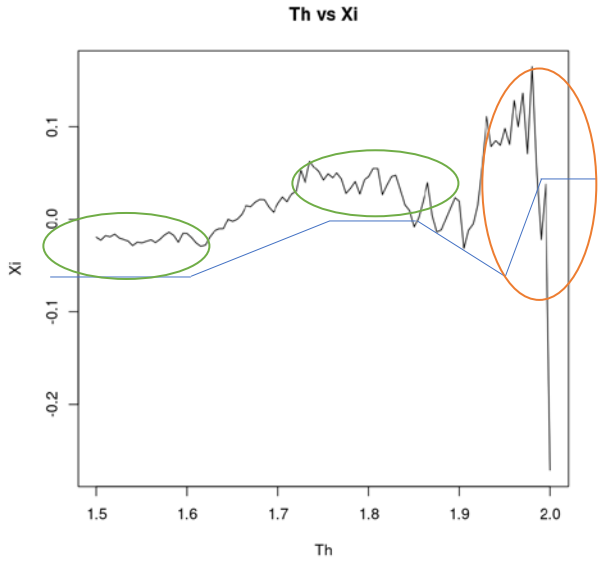
① 概形の傾きが0に近いところ

② 曲線（概形でない）の乱れが大きい部分は除外

$\theta - \xi$プロファイルの傾きが0に近いところが閾値候補、というのは先に述べた通りだが、実際に描いた曲線の傾きが0になることは、我々が実験で扱ったデータでは少なかった。そこで本節でも、その概形に着目し、$\theta \leqq 1.64$、$1.75 \leqq \theta \leqq 1.85$、$\theta \geqq 1.9$ の三つを候補として挙げている。曲線の乱れが大きい部分を除外する理由は、データ数が少ないため結果が安定していないことを示しているからである。実際、本節で扱ったデータでは、1.9を超えるデータは99個あり、1.95を超えるデータは59個ある。このようにごく少ないデータを推定に使っている場合、データの出入りが$\theta - \xi$プロファイルに大きな影響を及ぼしていることがわかる。これはすなわち、推定結果の良し悪しに大きな影響を及ぼすことを示す。なお、本節ではgpd.fitの結果を基に$\theta - \xi$プロファイルを出力したが、ismevであれば、gpd.fitrangeという関数で$\theta - \xi$プロファイルを得ることができる。

$\theta - \xi$プロファイルによってあたりを付けた後、診断図を見ていくことになるが、多くの場合、この時点で当該閾値の範囲に含まれるデータの数は1000個未満になっている。推定結果がデータの出入りに大きな影響を受けることを踏まえると、闇雲に閾値を区切るのではなく、当該範囲に含まれている値（スコア値）をそのまま閾値として用いることで、より効率の良い探索が可能である。

GPではもう一つ、やや強引だが効率の良い探索方法を採れることがある。それは、ヒストグラム（対数軸）を用いて見当をつける方法である。

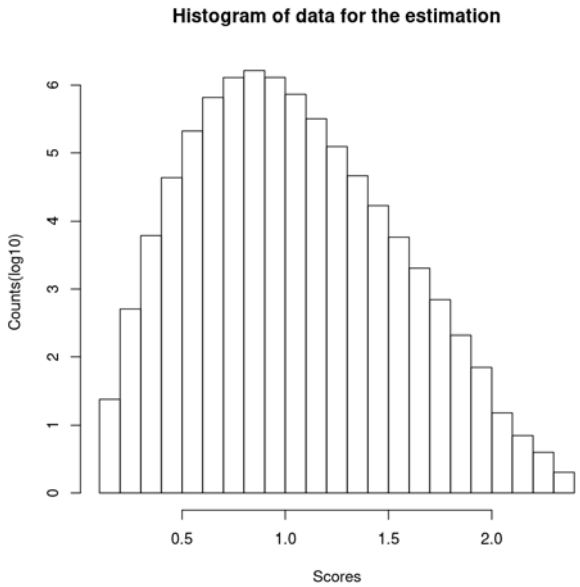


図表 4.2.1.14. $\theta - \xi$プロファイルとその概形（青線）

（緑囲いは閾値の候補とした範囲、オレンジ囲いは曲線が乱れている部分）

図表 4.2.1.15.は本項で扱った推定用データのヒストグラム（対数軸）である。中には、Score ≧ 2.0 が極値と見当をつける人もあるだろう。しかし、実際には Score ≧ 2.0 となるデータ数は 30 個程度と極端に少なく、今回の実験環境では安定した結果を得られなかった。100~1000 個程度のデータがあると推定結果が安定しやすいことを踏まえて少し範囲を広げ、例えば Score ≧ 1.8 とすると、Score ≧ 2.0 の場合よりも安定した結果が得られている。

このように範囲を絞った後は、先に述べたように、1.8 ≦ Score ≦ 2.0 に含まれる実際の値を閾値として診断図を出力し、適応性を判断することで効率の良い探索が可能である。ただし、$\theta - \xi$プロファイルを併用して俯瞰し、部分最適になっていないか（今回の例で言えば、1.75 ≦ Score ≦ 1.8 により良い閾値がないか）を確認しておくべきである。



図表 4.2.1.15. 推定用データのヒストグラム（対数軸）

4. 2. 2.　生体認証精度評価における極値統計導入の効果

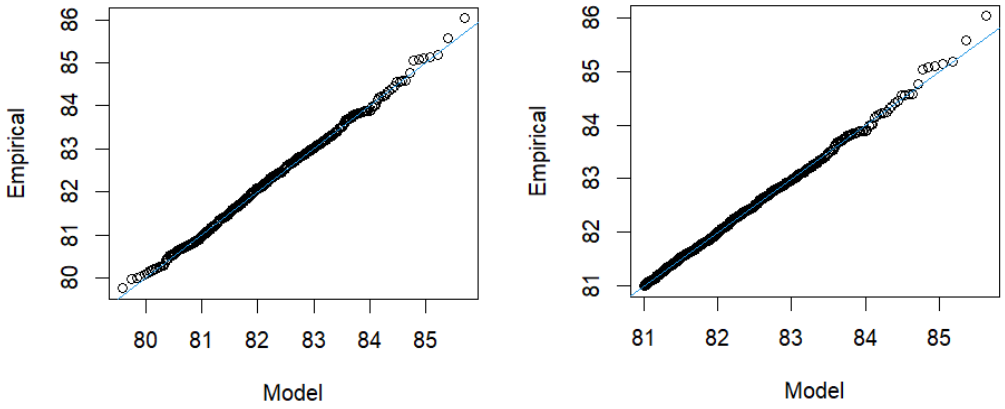
（1）既存手法（rule of three）との比較

本章では、生体認証精度評価における極値統計導入の効果について述べる。従来の精度評価では、国際標準 ISO/IEC 19795-1 の規定に基づき rule of three によって必要なサンプル数を求めており、高精度な認証精度を検証しようとすると膨大なサンプル数が必要だった（図表 4.2.2.1.）。そこで、極値統計モデルを取り入れることで、裾野の部分に焦点を当て、少ないサンプル数から小さい誤照合率が推定できることが期待できる。

前項でうまく推定できた例である、歩容の他人対約 1,370 万件のスコアデータを用いて考える。図 rGEV と GP でそれぞれ適切なパラメータを選択し結果、Q-Q プロット(図表 4.2.2.2.)は直線によくフィットしており、誤照合率と 95%信頼区間の推定結果は図表 4.2.2.3.の通りになった。

図表 4.2.2.1. rule of three における誤照合率と評価に必要なサンプル数

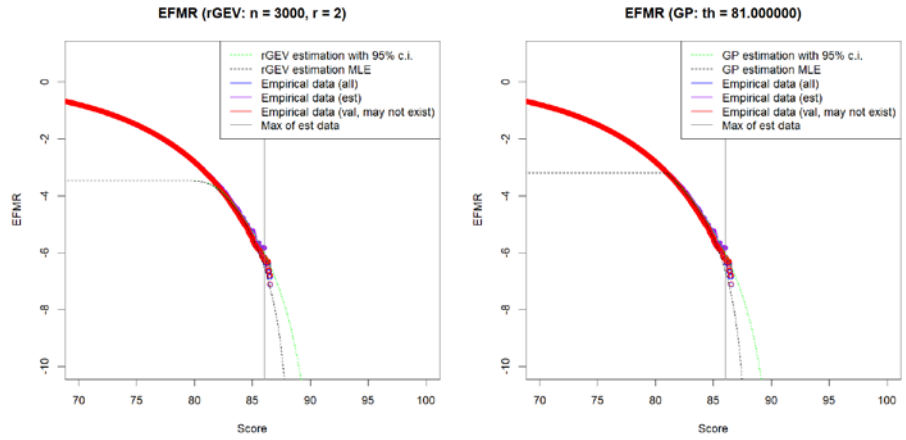
誤照合率	必要な照合件数	必要なID数
1/10万	30万以上	775以上
1/100万	300万以上	2,450以上
1/1000万	3,000万以上	7,746以上



図表 4.2.2.2.Q-Q プロット（左：rGEV、右：GP）

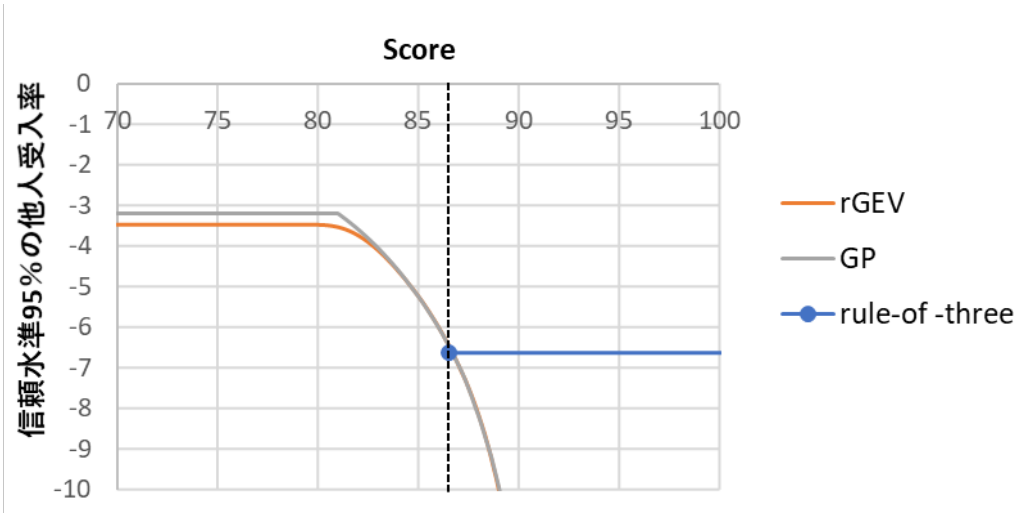
rGEV と GP の手法によらずにほぼ同等の推定結果を得られている。図表 4.2.2.3.の緑点線が 95%信頼区間の上下端であるが、誤照合率の場合では小さい分には問題ないため、上限を考慮する必要がある。

一方で、従来方式である rule of three では、誤照合率が p 以下であることを信頼水準 95%で示すには 3/p 回の試験において誤照合が 1 回も起きなければよいとされる。すなわち、例えば誤照合率が 100 万分の 1 となる閾値を信頼水準 95%で求めたい場合、300 万回の照合試験を行って、1 回も誤照合が発生しない閾値を設定すればよい。今回の歩容スコアデータセットは約 1,370 万件で、最大のスコアは 86.52 であり、この値を超えると誤照合は発生しない。したがって、この値より大きい閾値とすれば、信頼水準 95%で誤照合率が 457 万分の 1 以下だといえる。このようにして算出した、rule of three による信頼水準 95%の誤照合率と、rGEV および GP の信頼水準 95%の誤照合率を同時にプロットした結果が図表 4.2.2.4.である。



図表 4.2.2.3. 誤照合率と 95%信頼区間の推定結果（左：rGEV、右：GP）

縦の黒点線がスコア最大値の 86.52 を示しており、rGEV と GP 共に閾値 86.52 とした時の誤照合率推定結果は rule of three の結果とほぼ同等の結果を得ることができている。すなわち、rGEV や GP を用いることで、rule of three を用いる場合に比べて 10%のスコアデータセット数で、誤照合率を推定できたことを示している。さらに、rule of three の考え方では 86.52 を超えた閾値の場合でも信頼水準 95%の誤照合率は同じ値と計算されるが、rGEV や GP の結果では大きいスコアの誤照合率を外挿できているため、極値統計モデルを用いることで、従来よりも小さい誤照合率まで推定可能になっている。

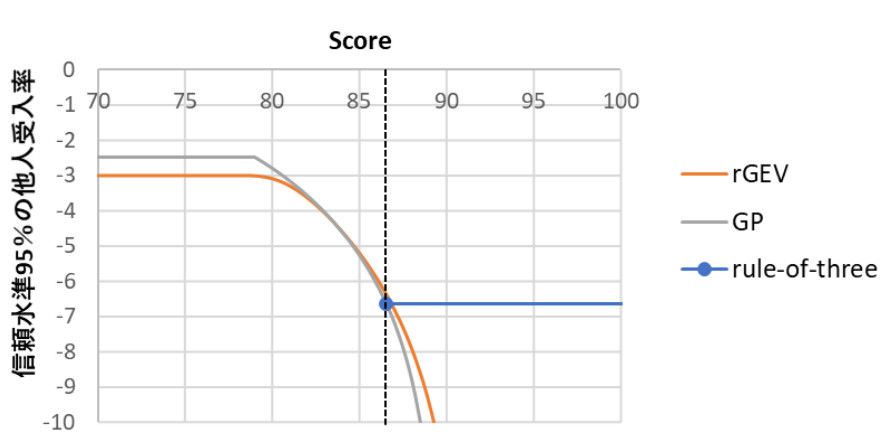


図表 4.2.2.4. Rule of three，rGEV，GP の信頼水準 95%の誤照合率の推定結果

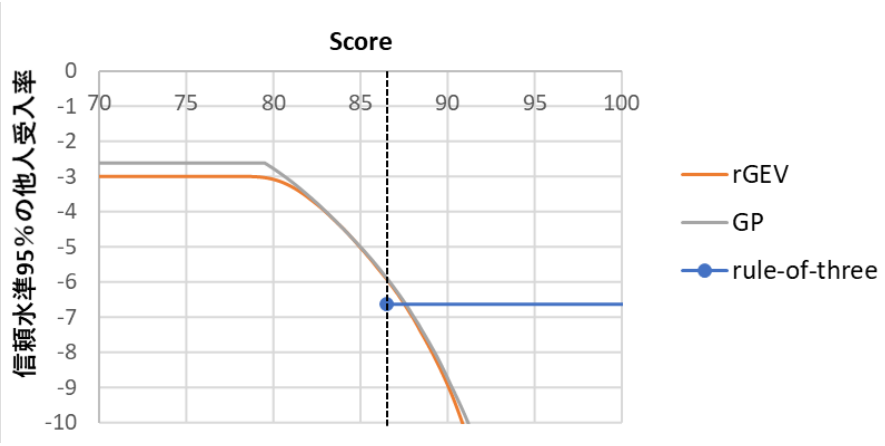
（2）スコアデータ数と推定精度の関係

前項の結果より、極値分布を用いた推定により、従来方式の 10%のスコアデータ数から同等の精度で誤照合率を推定可能なことが分かったが、続いて、更にスコアデータ数を減らした場合でも同様な推定ができるかに関して考察した。約 1,370 万件の歩容スコアデータのサンプリング数を 5%、2%、1%とデータ数を減らしていったときに、誤照合率の推定結果が同様に変化するかについて評価した結果を図表 4.2.2.5.に示す。また、元データの 1,370 万件において、rule of three で推定可能な信頼水準 95%で誤照合率が 457 万分の閾値（スコア 85.2）において GP および rGEV の推定結果との比較を図表 4.2.2.6.に示す。結果を参照すると、5%では rGEV も GP も rule of three と誤差 2%以内で同等の結果を得られているが、2%以下になると誤差が大きくなり、うまく推定できているとは言えなくなる。これは図

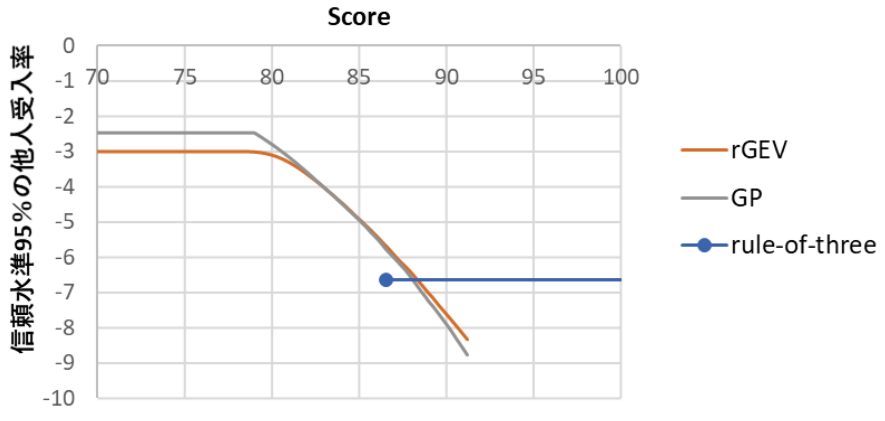
表 4.2.2.7.に示した通り、元データに対するサンプリング割合が小さくなるほど、極値分布のパラメータが不安定になり、信頼区間が広がっているためである。結果としては、今回評価した歩容スコアデータに関しては、rule of three に比べて 5%のサンプル数を収集すれば十分だと言える。ここで、他人対 1,370 万件は、ID 数に換算すると 1,660 人になる。収集すべきサンプル数が 5%の 69 万件になった場合、ID 数では 375 人となり、従来よりも 1/4 以下の ID 数で同等の誤照合率を推定可能になり、少ないサンプル数からの誤照合率の推定を実現できた。



(a) 5% (69 万件)からの推定結果



(b) 2% (27 万件)からの推定結果



(c) 1% (14 万件)からの推定結果

図表 4.2.2.5. 各手法による信頼水準 95%の誤照合率推定結果の比較

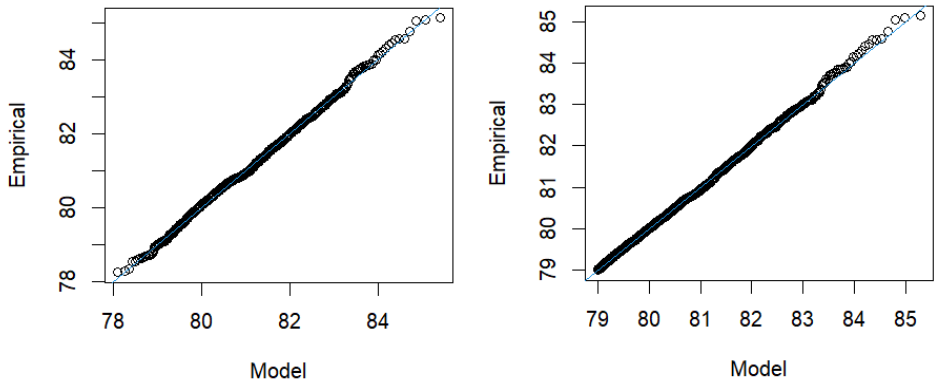
図表 4.2.2.6. 閾値 85.2 における誤照合率推定結果の実数値での比較
（各数字は 10 の累乗の指数部分を表す）

	GP	rGEV	rule-of-three
10%(137万件)から推定	-6.460134221	-6.450432347	-6.477375
5%(69万件)から推定	-6.611732439	-6.348844006	
2%(27万件)から推定	-5.902935321	-5.931377659	
1%(14万件)から推定	-5.762551544	-5.665910318	

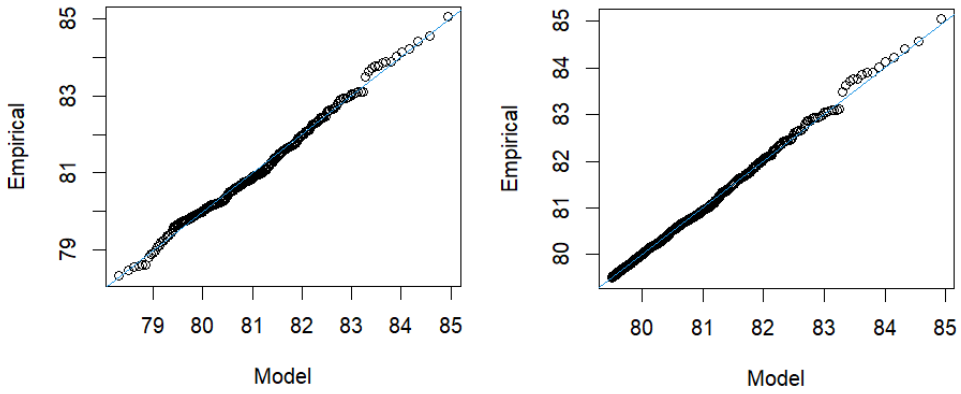
図表 4.2.2.7. サンプリング割合と GP のパラメータの関係

サンプリング割合	形状パラメータk	スケールパラメータσ	閾値θ
10%（137万件）	-0.1159	1.0941	80.5
5%（69万件）	-0.1073	1.2103	78.5
3%（46万件）	-0.1256	1.0688	80.5
2%（27万件）	-0.1281	1.2083	79.5
1%（14万件）	-0.1457	1.9733	75.5
0.5%（7万件）	-0.2524	3.3581	71.5
0.1%（1.4万件）	-0.2149	3.4846	70.5

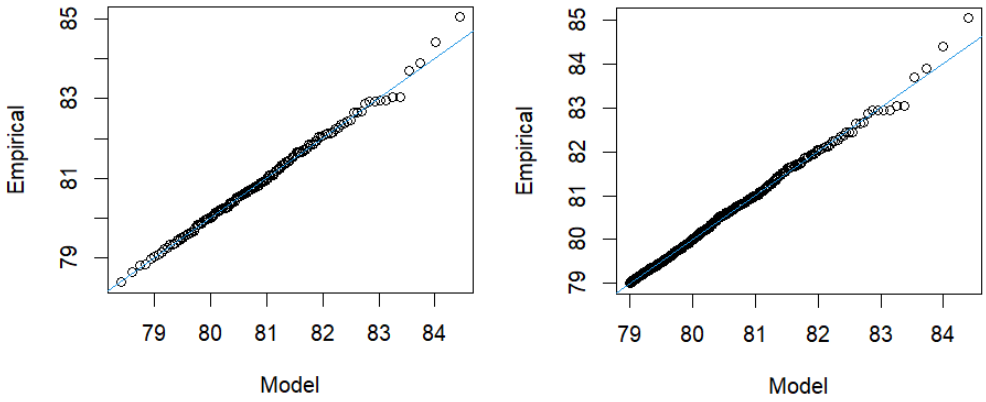
歩容スコアデータでは5%のデータセット数でrule of threeと同等の精度で信頼水準95%の誤照合率を推定できるという結果が得られたが、この5%というのはデータ依存であるため、実運用時はQ-Qプロットにより判断すべきある。サンプリング数が5%、2%、1%の場合のQ-Qプロットを図表4.2.2.8.に、誤照合率推定結果と95%信頼区間を図表4.2.2.9.に示す。Q-Qプロットに関しては、これまでの評価と同様に推定精度が高い場合はy=xの直線にフィットし、誤差が大きい場合は直線からのずれが大きくなっている。すなわち、実際の評価においても、Q-Qプロットが直線状になっているスコアデータであれば、十分な精度で高いスコアにおける誤照合率を推定できると言える。



(a) サンプリング割合 5%（69 万件）の Q-Q プロット

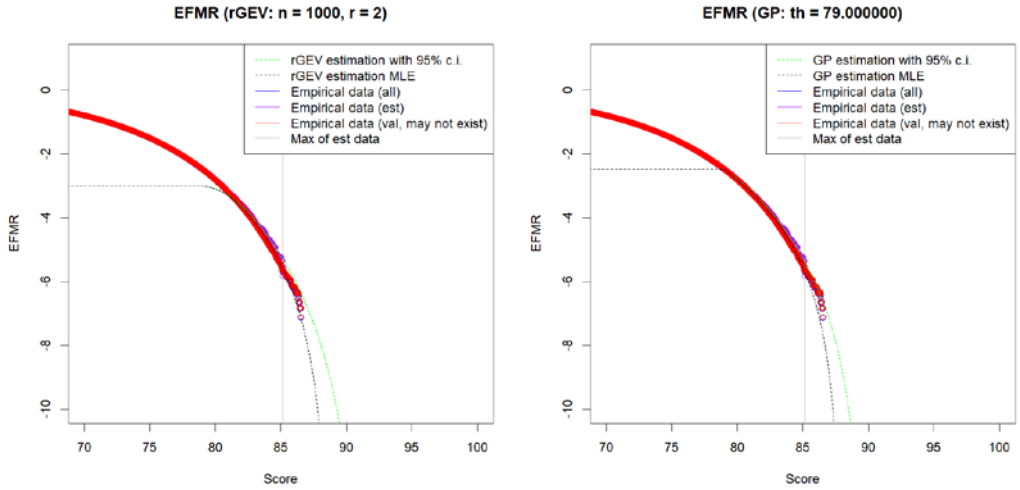


(b) サンプリング割合 2 %（27 万件）の Q-Q プロット

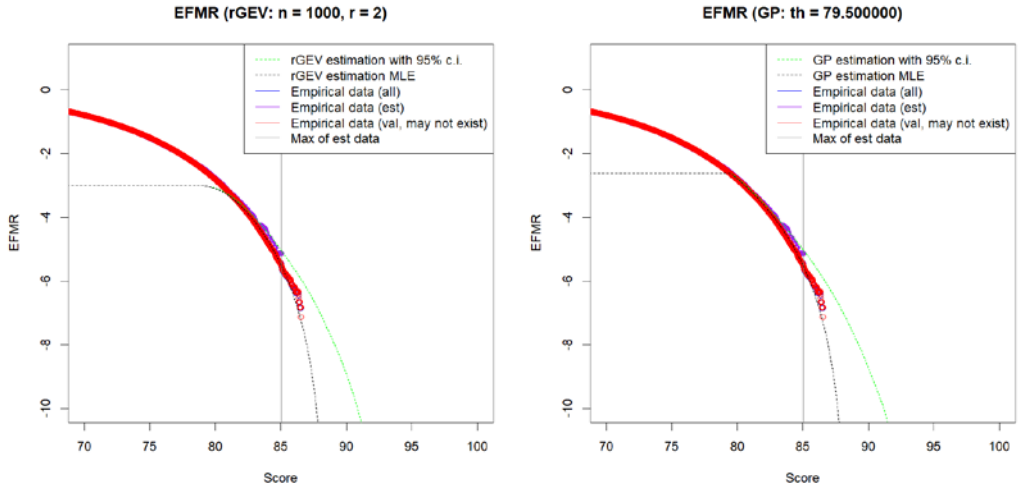


(c) サンプリング割合 1%（14 万件）の Q-Q プロット

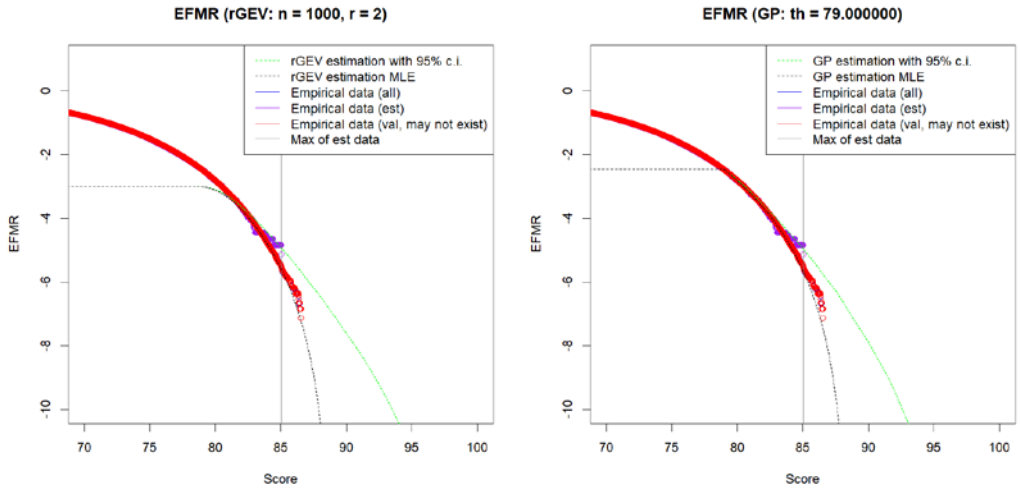
図表 4.2.2.8. データ数の減少と Q-Q プロットの関係（左：rGEV 右：GP）



(a) サンプリング割合 5%（69 万件）の Q-Q プロットと誤照合率推定結果



(b) サンプリング割合 2 %（27 万件）の誤照合率推定結果

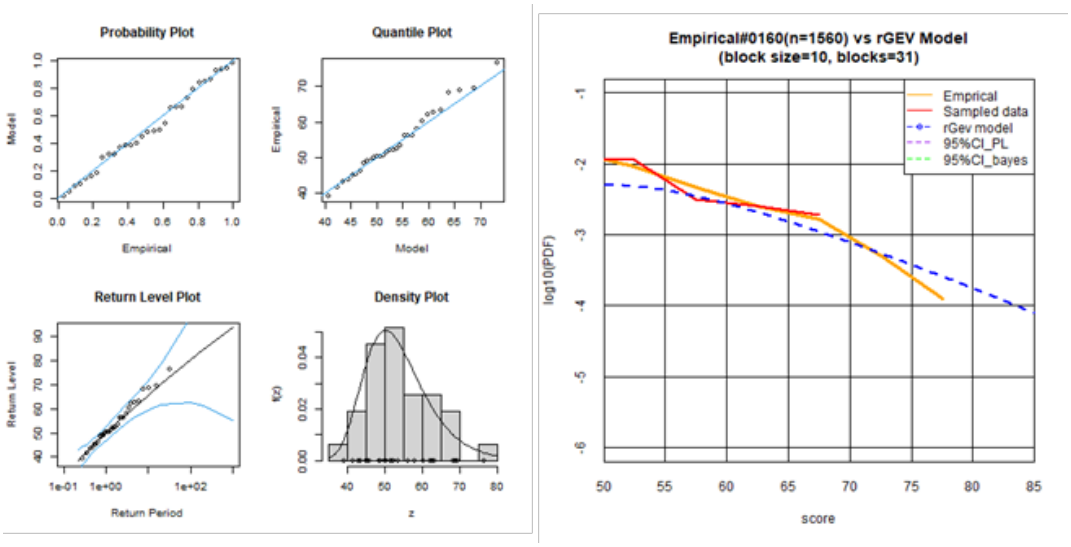


(c) サンプリング割合 1%（14 万件）の誤照合率推定結果

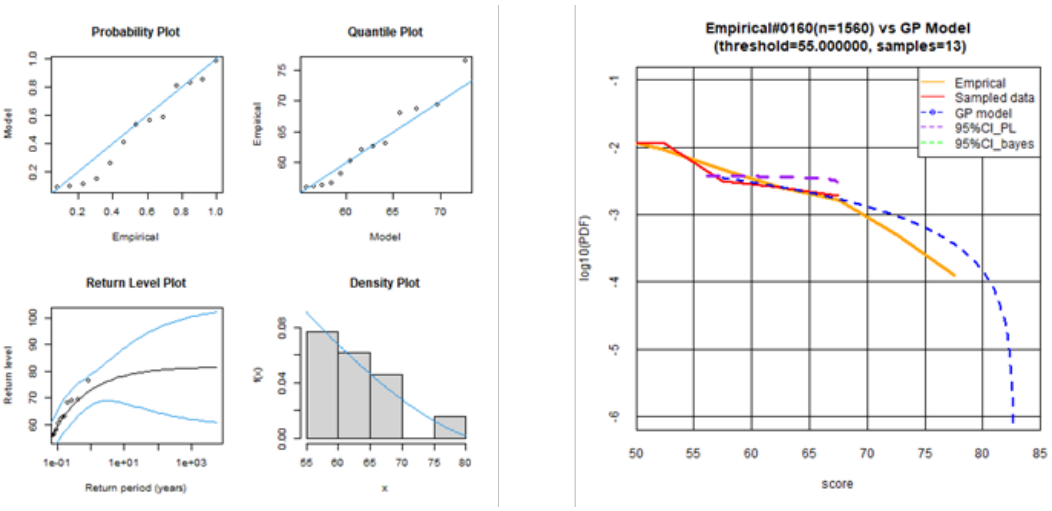
図表 4.2.2.9. データ数の減少と 95%信頼区間の広がりの関係（左：rGEV 右：GP）

（3）小規模なデータセットへの適用

前項では、1,370 万件の大規模なデータセットで評価を行ったが、より小規模なデータセットでの評価も実施したので報告する。顔＋声の GMM スコア 1,560 件を用いて rGEV および GP で推定した結果を図表 4.2.2.10 に示す。結果としては、Q-Q プロットはパラメータを変えても直線にはフィットせず、R の関数において 95%信頼区間が算出できなかった。これは、極値に当たるスコアの数が少なすぎたためである。経験則として、少なくとも極値分布モデルを用いる場合は少なくとも 50 個程度の極値が必要だが、今回の例では全体のデータ数が少なかったため、極値の必要数を満たすことはできなかった。すなわち、数千件の規模のデータセットから極値分布モデルを用いて誤照合率を外挿することは、困難であるといえる。



(a) rGEV の推定結果



(b) GP の推定結果

図表 4.2.2.10. 小規模なデータセットでの推定結果

（4）導入効果のまとめ

本節では、既存手法である rule of three と比較することで、生体認証精度評価における極値統計導入の効果について考察した。前節でうまく推定できた例であった 1,370 万件の歩容スコアデータで評価した結果、rule of three と比較して 5%のサンプル数で同等の精度の誤照合率 95%信頼区間を算出できることを確かめられた。このことにより、評価の際に収集すべき被験者 ID 数を従来の 1/3 以下にすることが可能となり、評価者の負担を軽減することができる。しかし、データ数が少なくなるに従い推定誤差が大きくなるという結果によって、本手法がいかなるスコアデータセットにも適用できるわけではないことも示された。

今後は、異なるモダリティや特徴量のスコアデータセットにも適用し、スコアヒストグラムの特性やデータ数と推定精度の関係を定量的に評価していき、生体認証の評価方法 として確立していく。

検討委員会メンバー

本検討委員会にご参画いただきましたメンバーの方々に深謝いたします。

No.	役割	氏名	所属
1	委員長	鷲見 和彦	青山学院大学
2	委員	志村 隆彰	統計数理研究所
3	委員	北野 利一	名古屋工業大学
4	委員	溝口 正典	東京理科大学
5	委員	山田 茂史	富士通 株式会社
6	委員	森原 隆	富士通 株式会社
7	委員	松濤 智明	富士通 株式会社
8	委員	日間賀充寿	株式会社日立製作所
9	委員	鈴木彦太郎	株式会社日立製作所
10	委員	坂本 静生	日本電気株式会社
11	委員	杉澤 正俊	日本電気株式会社
12	委員	岩田英三郎	株式会社ノルミー
13	委員	川路 雅博	バイオニクス株式会社
14	委員	甲斐 成樹	独立行政法人 情報処理推進機構
15	委員	金子 浩之	みずほリサーチ&テクノロジーズ株式会社
16	ｵﾌﾞｻﾞｰﾊﾞ	木村 英和	経済産業省 産業技術環境局
17	ｵﾌﾞｻﾞｰﾊﾞ	吉永恭平	株式会社三菱総合研究所
18	事務局	川嶋 一宏	一般社団法人 日本自動認識システム協会

おわりに

少ないサンプル数で実現する新しい生体認証精度評価方法によって、高性能な生体認証の装置やシステムの研究開発や改良がより迅速に行えるようになることが、本委員会活動の目標の一つである。

今回の適用評価では昨年度国際標準化に提案した評価手順で、大規模(1 千万件以上)および中規模なデータ（400 万件程度）により、顔、音声、歩容などによる生体認証の精度評価ができることが分かった。

一方、極値領域に外れ値がある場合や 2 つの分布がある場合では 1 つのモデルで精度評価をすることが困難である。外れ値や 2 つの分布を分けて精度評価をすることも必要である。

また、大規模なデータの 10%（100 万件程度）、1%（10 万件程度）のサンプルデータによる精度評価も実験した。データが少なくなるについて、閾値 θ の設定が困難になることも分かってきた。今後、GP や rGEV の特徴を生かした精度評価ノウハウを蓄積していくことも重要である。

本報告書の執筆にご尽力いただいた、統計数理研究所 志村 隆彰准教授、富士通 山田茂史氏、松濤智明氏、日立製作所 日間賀充寿氏、鈴木彦太郎氏に感謝します。また、スコアデータをご提供いただいた大阪大学 八木康文教授に感謝します。

本報告が生体認証装置やシステムの研究開発および事業化に貢献できることを祈念します。

2022 年 2 月 10 日
精度評価方法に関する国際標準化検討委員会
委員長　鷲見 和彦

この報告書は、経済産業省からの委託事業を株式会社三菱総合研究所からの再委託として実施したものの成果である。

本件についてのお問合せ先

（内容等）
〒101-0032
東京都千代田区岩本町 1-9-5 FK ビル 7 階
TEL　03-5825-6651
一般社団法人日本自動認識システム協会
研究開発センター