ECサイトの商品特性を考慮した2次元確率表による購買予測

西村　直樹，鮭川　矩義，高野　祐一，岩永　二郎，水野　眞治

1. はじめに

インターネットの普及に伴い，現在では商品販売やサービスをウェブサイト上で提供するEC（Electronic Commerce：電子商取引）サイトを多くの企業が運営するようになった．消費者にとっては，店舗に足を運ばずとも多数の商品やサービスを比較し利用できることがECサイトの最大の魅力であろう．一方で，運営する企業にとっても来訪者のアクセスログを収集・解析し，ウェブサイト上でさまざまな施策を容易に実行できるというマーケティング上の利点がある．アクセスログ解析を成果につなげるためのさまざまな方法も提案されている [1].

本論文では，ECサイトに来訪する顧客の購買商品を予測することを分析課題とする．顧客が購買する商品を予測することで，効果的なマーケティング施策を実施することができ，商品需要を見積もることで在庫管理の効率化にも役立てることができる．経営科学系研究部会連合協議会主催のデータ解析コンペティションの課題設定部門では，平成24年度は不動産賃貸サイトにおける商品閲覧と資料請求の予測が，そして平成25年度はファッションECサイトにおける購買予測が課題として設定されており，ECサイトにおける顧客の閲覧や購買の予測は実務上の重要な課題であると言える．

購買商品を予測するための手法としては，ロジットモデルやプロビットモデルに代表されるブランド選択モデル [2] や，決定木やニューラルネットワークなどの予測モデル [3] が考えられる．特に，ECサイトにおける購買予測では，来訪した顧客がどのようにサイト内を遷移して離脱したかの履歴を集めたアクセスログのデータを利用することができる．予測モデルの性能を向上させるためには，このアクセスログ情報を有効に活用する必要がある．

岩永ら [4] は，アクセスログのデータから閲覧商品に対する「関心度」と「忘却度」を数量的に定義し，これら2種類の特徴量から購買につながる商品を予測する手法を提案した[1]．この方法では，閲覧商品への関心度と忘却度に応じた購買確率を表す，2次元の確率表を作成する．さらに，関心度と忘却度に対する購買確率の単調性を満たすように確率表を補正することで，予測精度の改善につなげている．上述のコンペティションでは，この手法を採用したチームが平成24・25年度の2年連続で課題設定部門の最優秀賞を受賞しており，このことは手法の有用性を実証するものと言える．しかし，既存手法 [4] では全顧客・全商品に対して単一の確率表を参照して購買予測をしており，ECサイトの商品特性が十分に考慮されていない．

そこで本論文では，多種多様な商品を扱い，幅広い年齢層の顧客を抱える企業のECサイトを想定して，その商品特性を考慮した2次元確率表の作成方法を提案する．具体的には，顧客や商品の多様性を考慮して顧客と商品を類型化し，各類型に対して確率表を作成する．さらに，類型数の増加に起因する過剰適合を軽減するために，確率表間の乖離を抑制する制約条件を提案する．また，購買確率の関心度に対する増加率と忘却度に対する減少率は逓減するという性質を制約条件として加えたモデルも提案する．

本論文の構成は以下のとおりである．次節では，本論文で着目する既存手法 [4] と関連研究を紹介する．3節では提案手法を説明し，4節では数値実験を通してその有用性を検証する．5節では本論文のまとめと今後の課題を述べる．

にしむら　なおき，すけがわ　のりよし，みずの　しんじ
東京工業大学 大学院社会理工学研究科 経営工学専攻
〒152–8552 東京都目黒区大岡山 2–12–1
たかの　ゆういち
専修大学 ネットワーク情報学部
〒214–8580 神奈川県川崎市多摩区東三田 2–1–1
いわなが　じろう
（株）NTTデータ 数理システム
〒160–0016 東京都新宿区信濃町 35 信濃町煉瓦館 1 階

[1] 先行研究 [4] では再閲覧確率表を作成し，閲覧および資料請求が行われる商品を予測しているが，本論文では購買予測という目的に合わせて手法を説明する．

2. 既存手法

本節では，先行研究 [4] で提案された，EC サイトにおける購買予測手法を説明し，関連研究についても述べる．

先行研究 [4] では，閲覧商品に対する関心度と忘却度をアクセスログから数量化する．関心度を表す特徴量としては，当該商品に対する閲覧回数，閲覧時間，閲覧セッション数などがあり，忘却度に関しては当該商品に対する最終閲覧以降の経過日数，（他の商品に対する）閲覧回数，セッション数などが考えられる．これらの特徴量によって関心度と忘却度を整数値として定義すれば，関心度が $i \in I$ で忘却度が $j \in J$ の商品が購買される確率（実績購買確率）p_{ij} は過去データから計算することができる．本論文では関心度と忘却度の組 $(i,j) \in I \times J$ をセルと呼ぶことにする．

購買確率は商品に対する関心度が高いほど増加し，商品に対する忘却度が高いほど減少することが期待される．しかし，データ数が少ないセルでは実績購買確率が真の購買確率から乖離する可能性が高く，実績購買確率の単調性が満たされない場合がある．そこで，先行研究 [4] では関心度と忘却度に対する単調性制約の下で，各セルのデータ数によって重み付けられた残差 2 乗和が最小となるように購買確率 x_{ij} を推定する問題を，以下の凸 2 次最適化問題[2]として定式化した：

$$
\begin{aligned}
&\text{最小化}: \min_{x_{ij}} \sum_{i\in I}\sum_{j\in J} c_{ij}^2 (x_{ij} - p_{ij})^2 \\
&\text{制約条件}: x_{i_1 j} \le x_{i_2 j} \quad (i_1 < i_2\ (\in I),\ j \in J), \\
&\qquad\qquad x_{ij_1} \ge x_{ij_2} \quad (i \in I,\ j_1 < j_2\ (\in J)), \\
&\qquad\qquad 0 \le x_{ij} \le 1 \quad (i \in I,\ j \in J).
\end{aligned}
$$

ここで，c_{ij} はセル (i,j) の実績購買確率 p_{ij} の計算に用いたデータ数とする．そして，顧客が閲覧した商品の中から，確率表を参照して購買確率が上位の商品を購買商品として予測する．

2 次元数値相関ルール [6〜8] は 2 種類の数値属性からなる領域と事象の生起を関連付けるルール[3]であり，既存手法 [4] と関連が深い．特に，数値属性に対する単峰性を仮定してデータを近似する場合は最適ピラミッド問題 [11, 12] と呼ばれる．ただし，数値相関ルールに対しては組合せ論的な解法が提案されており [13]，連続最適化問題として定式化した先行研究 [4] とは異なる．

[2] この問題は単調回帰 [5] とみなすこともできる．
[3] 詳細については，文献 [9, 10] などを参照されたい．

3. 提案手法

前節の定式化からもわかるように，既存手法 [4] では単一の確率表によって購買商品を予測している．しかし，男女を問わず幅広い年代の顧客を抱え，価格帯や用途が異なる多種多様な商品を扱う EC サイトに対しては，単一の確率表では十分な予測精度を達成できない可能性がある．本節では EC サイトの商品特性を考慮した確率表の作成方法を説明する．

3.1 顧客と商品の類型化

顧客や商品の多様性を考慮した方法として，顧客や商品の類型 $k \in K$ に対応させて複数の確率表を作成することを考える．例えば，男性と女性とで購買傾向が異なるとすれば類型を $K = \{$ 男性, 女性 $\}$ と設定し，商品分類によって購買傾向が異なるとすれば $K = \{$T シャツ, 時計, . . . , 財布 $\}$ のように設定する．顧客と商品の類型の組に対して確率表を作成することも可能である．

複数の確率表を作成することは，顧客や商品の多様性を表現できるという利点がある．しかし，類型の個数 $|K|$ が多くなると各類型に割り当てられるデータ数が減少し，過剰適合が生じて逆に予測精度が悪化する可能性がある．過剰適合を軽減するためには，モデルに罰則項や制約条件を加えて推定する正則化と呼ばれる方法が有効であることが知られており [14]，本論文では確率表間の乖離を抑制する制約条件を提案する．具体的には，類型 k の確率表のセル (i,j) の購買確率を変数 x_{ijk} とし，補助変数 $\hat{x}_{ij}$ を導入する．そして，x_{ijk} $(k \in K)$ が一定範囲内に収まるように以下の制約条件を追加する：

$$
\frac{1}{1+\lambda}\hat{x}_{ij} \le x_{ijk} \le (1+\lambda)\hat{x}_{ij} \quad (i \in I,\ j \in J,\ k \in K).
$$

ここで，λ は確率表間の乖離の度合いを調節するパラメータである．この乖離度パラメータの値を小さくすると，各類型の確率表が同一の確率表に近づいていく．このパラメータの値を適切に設定することで，各類型の多様な購買傾向と全体の購買傾向をバランスよく捉えた確率表を作成できると期待される．

各類型の確率表を求める問題は，以下の凸 2 次最適化問題として定式化できる：

$$
\text{最小化}: \min_{x_{ijk},\,\hat{x}_{ij}} \sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c_{ijk}^2 (x_{ijk} - p_{ijk})^2,
$$

制約条件: $\dfrac{1}{1+\lambda}\hat{x}_{ij} \le x_{ijk} \le (1+\lambda)\hat{x}_{ij}$
$(i \in I,\ j \in J,\ k \in K)$,

$x_{i_1jk} \le x_{i_2jk}$
$(i_1 < i_2\ (\in I),\ j \in J,\ k \in K)$,

$x_{ij_1k} \ge x_{ij_2k}$
$(i \in I,\ j_1 < j_2\ (\in J),\ k \in K)$,

$0 \le x_{ijk} \le 1 \quad (i \in I,\ j \in J,\ k \in K)$.

ここで，p_{ijk} は類型 k の確率表のセル (i,j) の実績購買確率とし，c_{ijk} は p_{ijk} の計算に用いたデータ数とする．

3.2 凹／凸性制約

商品の購買確率は関心度に対して単調に増加し，忘却度に対して単調に減少するという仮定は，我々の直感とも合致する．実際に先行研究 [4] では，単調性制約を満たすように確率表を補正することで，購買予測の精度が改善することを示している．一方で，購買確率の性質を表す制約条件は単調性以外にも考えられる．

例えば，購買確率は関心度に対して単調に増加していくが，関心度が増えるにつれてその効果が薄れ，購買確率の増加率は徐々に 0 へと近づいていくことが予想される．同様に，忘却度に対する購買確率の減少率も徐々に 0 へ近づいていくと予想される．

そこで本論文では，購買確率を表す関数の傾きが関心度に対して単調に減少し，忘却度に対して単調に増加することを表す以下の線形制約を提案する：

$x_{i-1,j} - x_{i-2,j} \ge x_{ij} - x_{i-1,j}$
$(i \in I\backslash\{1,2\},\ j \in J)$,

$x_{i,j-1} - x_{i,j-2} \le x_{ij} - x_{i,j-1}$
$(i \in I,\ j \in J\backslash\{1,2\})$.

上記の制約条件は凹関数/凸関数の特徴付け [15] とも一致するため，以降では凹/凸性制約と呼ぶこととする．凹/凸性制約の有効性については次節で検証する．

4. 数値実験

本論文では，経営科学系研究部会連合協議会主催，平成 25 年度データ解析コンペティションで提供された 2011 年 9 月から 2013 年 4 月までのファッション EC サイトにおける顧客データ，商品データ，注文履歴，閲覧履歴を用いた．レコード数は顧客データが約 10 万件，商品データが約 450 万件，注文履歴が約 86 万件，閲覧履歴が約 6,400 万件であった．関心度を表す特徴量は「当該商品に対する閲覧回数」とし，$I = \{1, 2, \ldots, 20\}$ とした．忘却度を表す特徴量は「当該商品の最終閲覧日からの経過日数」とし，$J = \{1, 2, \ldots, 28\}$ とした．

表 1 数値実験で扱う類型

類型	説明	類型数
性別	男性，女性	2
年代別	～19, 20～34, 35～49, 50～	4
商品大分類別	トップス，パンツなど	25
商品小分類別	ポロシャツ，タンクトップなど	215
ショップ別	EC サイト上の店舗	537
ゾーン別	似た系統のショップ	35

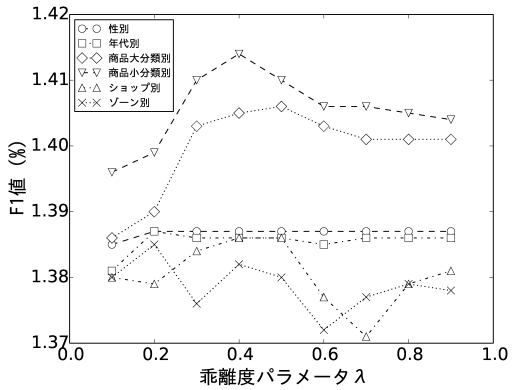

図 1 類型化と予測精度の関係

数値実験では，2013 年 3 月に 1 回以上商品を閲覧した顧客 27,590 人を対象として，2013 年 4 月の購買商品を予測した．また，2013 年 3 月までの連続した 2ヶ月間のデータにおいて，前半 1ヶ月で閲覧された商品が後半 1ヶ月で購買された回数を集計することで実績購買確率 p_{ij}，p_{ijk} を計算した．顧客が閲覧した商品に対する関心度と忘却度から，確率表を参照して購買確率を求め，1 顧客に対して購買確率が上位の 6 商品を購買商品として予測した．予測精度の評価指標としては適合率と再現率の調和平均である F1 値[4]を用いた．なお，F1 値が大きいほど予測精度が高いことを表す．

4.1 類型化による改善効果の検証

本節では，顧客と商品の類型を考慮した方法（3.1 節の定式化）の予測性能を検証する．表 1 の 6 種類の類型を用いた提案手法の予測精度を図 1 に示す．横軸は乖離度パラメータ λ の値であり，この値が 0 に近づくほど各類型の確率表は同一の確率表に近づく．逆に λ の値を大きくすると，各類型の特性が反映された確率表を作成することができる．

図 1 から，商品大分類別・商品小分類別といった商品の類型化によって予測精度が改善することがわかる．

[4] 詳しい定義については，文献 [10] などを参照されたい．

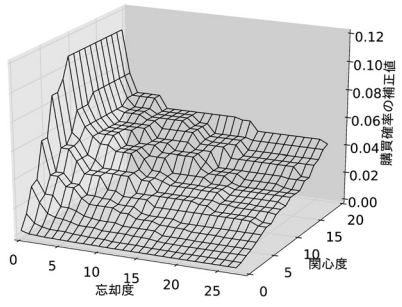

図 2　トップスの購買確率

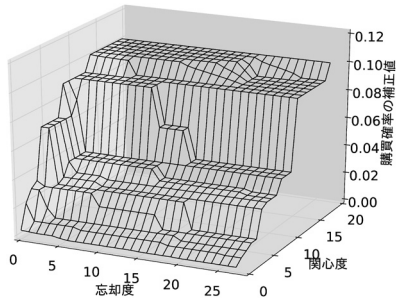

図 3　時計の購買確率

一方で，性別や年代別といった顧客の類型化は予測精度の改善効果が小さく，ショップ別やゾーン別といった類型化は予測精度を逆に悪化させる結果となった．また，乖離度パラメータを適切な値に設定することで予測精度が向上していることもわかる．例えば，商品小分類による類型化の場合は，$\lambda = 0.4$ のときに F1 値が最も高くなった．確率表間の乖離を抑制することで，類型化に起因する過剰適合が軽減され，予測精度の向上につながったと考えられる．

4.2　商品大分類別の購買確率表の比較

図 2，3 はそれぞれ商品大分類のトップスと時計の購買確率である．トップスは忘却度が高くなるにつれて購買確率が大きく減少している．一方で，時計は忘却度の増加に対する購買確率の減少が比較的小さい．時計はトップスと比べて価格帯が高く購買頻度が少ないために，長い時間をかけて検討された後に購買に至るという傾向を反映していると考えられる．また，スーツやバッグなどの商品大分類の確率表が時計と同様の傾向を示す．このように商品ごとの差異を考慮した確率表を作成することによって，図 1 に示したように予測精度が改善されたと考えられる．

4.3　予測モデルの性能の比較

本節では，表 2 の 5 種類の予測モデルの性能を比較する．単調性モデルが既存手法 [4] に対応する．また，類型化は商品小分類別とし，凹/凸性制約を課した場合は単調性制約も同時に課している．前述のように 2013 年 4 月の購買商品の予測精度を比較するが，各モデルの乖離度パラメータ λ は 2013 年 3 月の購買商品を予測して最も精度が高かった値を採用した．

表 2　予測モデルの概要

予測モデル	説明
集計	実績購買確率による 単一の確率表を参照
単調性	単調性制約により補正した 単一の確率表を参照
単調性+類型化	単調性制約により補正した 商品小分類別の確率表を参照
凹/凸性	単調性制約と凹/凸性制約により 補正した単一の確率表を参照
凹/凸性+類型化	単調性制約と凹/凸性制約により 補正した商品小分類別の確率表を参照

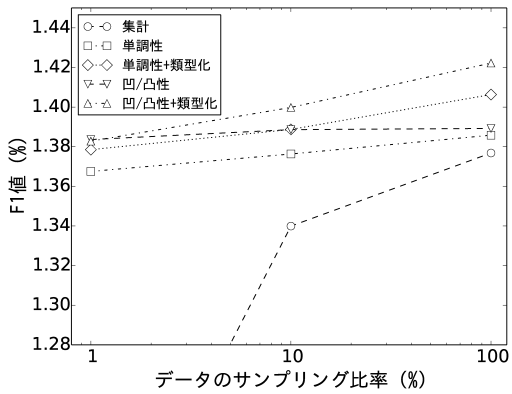

図 4　各予測モデルの予測精度の比較

確率表の作成に用いたデータ数と予測精度との関係を調べるために，顧客と閲覧商品のすべての組合せ約 4,000 万件からランダムに 1%，10%，100%をサンプリングし，1%と 10%の場合は 10 回のサンプリングによる F1 値の平均を計算した．データのサンプリング比率に対する各モデルの予測精度は図 4 のようになった．

集計モデルはサンプリング比率が減少することで予測精度が急激に悪化するが，他のモデルは単調性制約や凹/凸性制約の補正により予測精度の悪化が抑えられている．また，単調性モデルと凹/凸性モデルを比較すると，すべてのサンプリング比率で凹/凸性モデルのほうが予測精度が高い．したがって予測精度の改善のためには，単調性制約に凹/凸性制約を加えて補正することが有効であると言える．

サンプリング比率が 1%の場合は凹/凸性モデルの予測精度が最も高いが，サンプリング比率が 100%の場合

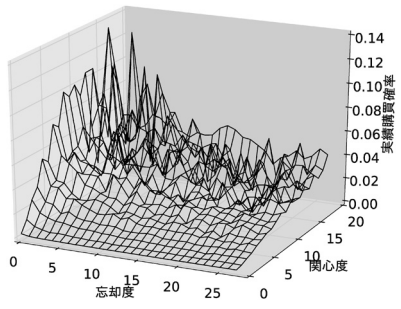

図 5　実績購買確率

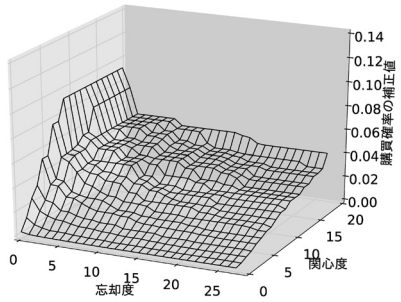

図 6　単調性制約により補正した購買確率

は凹/凸性+類型化モデルの予測精度が最も高い．データ数が十分にある場合は各類型に割り当てられるデータ数も多くなり，類型化に起因する過剰適合が軽減される．それゆえ，類型化はその性能を十分に発揮することができ，予測精度が向上していると考えられる．

4.4　補正された購買確率表の比較

最後に，データのサンプリング比率を 100%とした場合の集計モデル，単調性モデル，凹/凸性モデルの購買確率をそれぞれ図 5，6，7 に示す．図 5 ではいたるところで単調性制約が満たされておらず，このことが予測精度を悪化させていると考えられる．一方で，図 6 では単調性制約によって，関心度に対して単調に増加し，忘却度に対して単調に減少する確率表が作成されている．図 7 では凹/凸性制約によって，関心度に対しては逓増し，忘却度に対しては逓減する確率表が作成されている．図 5，6 と比較すると，図 7 では購買確率が滑らかに変化していることもわかる．

5.　おわりに

本論文は，EC サイトに来訪する顧客の購買商品を予測することを目的として，岩永ら [4] の既存手法の改良に取り組んだ．既存手法 [4] は，アクセスログのデータから関心度と忘却度という 2 種類の有効な特徴量を抽出し，単一の 2 次元確率表を作成して購買商品を予測する．一方で本論文では，類型ごとに複数の確率表を作成する方法を提案し，特に十分なデータ数が確保できる場合に予測精度を向上させることができた．また，提案手法では複数の確率表を作成することで，類型間の購買傾向の差異を分析することができるという，分析モデルとしての利点もある．さらに本論文では，関心度/忘却度に対する購買確率の凹/凸性制約を提案し，数値実験によって有効性を確認した．

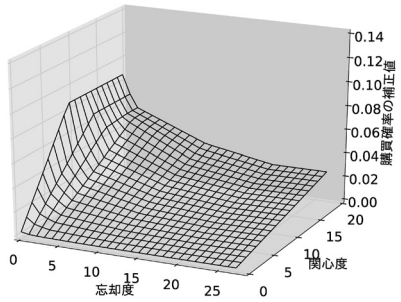

図 7　単調性制約と凹/凸性制約により補正した購買確率

本論文で用いた手法は，顧客が閲覧した商品の中から購買確率が高い商品を選択するものである．ゆえに，（例えば新発売の商品などの）顧客が閲覧していない商品に対しては購買を予測することができず，このことは商品推薦などの用途を考えると大きな欠点である．今後は，相関ルールや協調フィルタリングなどの手法と組み合わせることで，この欠点を解消したいと考えている．また，今回は関心度と忘却度の特徴量として「閲覧回数」と「最終閲覧日からの経過日数」を採用したが，より有効な特徴量を構築することも考えられる．また，他の予測モデルとの比較も行いたいと考えている．

謝辞　本論文を執筆する機会をくださりました中央大学の生田目崇先生に，この場を借りて御礼申し上げます．また，貴重なデータを提供していただいたデータ解析コンペティション関係者の皆様に，心より感謝申し上げます．

参考文献

[1] 小川卓，『入門ウェブ分析論（増補改訂版）』，ソフトバンククリエイティブ，2012.
[2] 古川一郎，守口剛，阿部誠，『マーケティング・サイエンス入門（新版）』，有斐閣，2011.
[3] ゴードン S. リノフ，マイケル J. A. ベリー，『データマイニング手法 予測・スコアリング編』，海文堂出版，2014.
[4] 岩永二郎，鍋谷昴一，梶原悠，五十嵐健太，“関心度と忘却度に基づくレコメンド手法—単調性制約付きレコメンドモデルの構築—，”オペレーションズ・リサーチ，**59**，72–80, 2014.

[5] R. L. Dykstra and T. Robertson, “An algorithm for isotonic regression for two or more independent variables,” *The Annals of Statistics*, **10**, 708–716, 1982.
[6] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, “Data mining using two-dimensional optimized association rules: scheme, algorithms, and visualization,” In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 13–23, 1996.
[7] T. Fukuda, Y. Morimoto, S. Morishita and T. Tokuyama, “Data mining with optimized two-dimensional association rules,” *ACM Transactions on Database Systems*, **26**, 179–213, 2001.
[8] N. Katoh, “Finding an optimal region in one-and two-dimensional arrays,” *IEICE Transactions on Information and Systems*, **83**, 438–446, 2000.
[9] 福田剛志，森本康彦，徳山豪，『データマイニング』，共立出版，2001.
[10] 加藤直樹，羽室行信，矢田勝俊，『データマイニングとその応用』，朝倉書店，2009.
[11] D. Z. Chen, J. Chun, N. Katoh and T. Tokuyama, “Efficient algorithms for approximating a multi-dimensional voxel terrain by a unimodal terrain,” In *Proceedings of the Computing and Combinatorics: 10th Annual International Conference*, pp. 238–248, 2004.
[12] 全眞嬉，D. Z. Chen，加藤直樹，徳山豪，“高次元ピラミッドを用いた数値属性ルールの生成とデータマイニングへの応用，”日本データベース学会 Letters，**2**, 83–86, 2003.
[13] 徳山豪，“関数近似における幾何学アルゴリズムの最近の進展—データ解析への応用に向けて—，”電子情報通信学会論文誌 A, **J89-A**, 419–429, 2006.
[14] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, 2009.
[15] D. L. Hanson and G. Pledger, “Consistency in concave regression,” *The Annals of Statistics*, **4**, 1038–1050, 1976.

