The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Factorization Machinesを用いたCoxハザードモデル

The Cox hazard model using Factorization Machines

佐竹 哉太 *1 Kanata Satake　山田 誠 *1*2 Makoto Yamada　松井 孝太 *2 Kota Matsui　松井 茂之 *2*3 Shigeyuki Matsui　鹿島 久嗣 *1*2 Hisashi Kashima

*1京都大学 Kyoto University　*2理化学研究所 RIKEN Center　*3名古屋大学 Nagoya University

Survival analysis is an approach to analyze the time until the occurrence of an event of interest. In many cases, the Cox proportional hazard model is the most popular regression method without specifying the underlying time-to-event distribution. However, it can only use linear information of features. In this research, we introduce the factorization machines (FM) to the Cox hazard model. In contrast to the previous models, The proposed model can use the interaction between covariates and retain interpretability. The experiments using a medical dataset show that our model yields better performance than baseline models.

1. はじめに

ある出来事が起こるまでの時間を分析する問題は、多くの分野で共通する重要な問題である。このイベントが起こるまでの時間である生存時間を推定する手法として、生存時間解析がある。一般に生存時間解析では、ある瞬間にイベントが生じる可能性をハザード関数 $h(t)$ として定義し、それをもとに時刻 t 以上生存する確率である生存関数 $S(t)$、時刻 t までに死亡する確率である死亡関数 $F(t)$ を求める。イベントの振る舞いが詳細に分かっていない場合には、モデルの一部のみを定義するセミパラメトリックな手法が用いられることが多い。例えば医療分野では個別の病気に関するデータが十分ではない場合が多いため、セミパラメトリックな手法がよく用いられる。

生存時間解析のセミパラメトリックなモデルのうち、最もよく用いられるものに Cox 比例ハザードモデル [Cox 72] がある。Cox 比例ハザードモデルは、2 つのサンプルのハザード関数の比であるハザード比が時間に依存しないという重要な性質をもつ。またモデルに対して対数をとると線形になるという対数線形性を仮定したモデルで、時間 t による変化は定義しない。Cox 比例ハザードモデルは対数線形性を持つため、各特徴量のモデルに対する貢献を解釈することができる。しかし一方で非線形の情報を利用することはできない。後述するカーネル法を用いたモデルでは特徴間の非線形な情報を利用できるよう工夫されている [Li 02] が、それらのモデルでは解釈性が失われるという問題がある。そこで本論文では、非線形性を利用でき解釈も可能なモデルについて考察した。

本研究の貢献は以下の通りである。

- Cox 比例ハザードモデルに FM を導入し、解釈性を保ったまま特徴間の相互作用を利用できるモデルを提案した。
- FM による提案手法が実際の医療データに対し既存手法よりも良い精度を示すことを確認した。

2. 関連研究

Cox 比例ハザードモデルにおいて非線形な情報を利用した例としては、Li と Luan が Cox モデルにカーネル法を適用した研究がある [Li 02]。この論文では高次元データから p 値をもとに特徴を選択し、モデルを構築している。このモデルではカーネル法により変数間の非線形な関係性を利用できるが、解釈ができないという問題がある。

また最近の研究として、ニューラルネットワークを用いる DeepSurv と呼ばれる手法を Katzman らが提案している [Katzman 16]。DeepSurv では生存期間をただニューラルネットワークで推定するのではなく、ハザード関数自体はニューラルネットワークを用いて出力し、観測打ち切りデータに対応するためにハザード関数の推定法を用いることで精度の高い推定に成功している。この手法では特徴間の非線形な相関を利用できる。ただし解釈性は高くなく、またこの論文中の実験では人力で数十程度の特徴を選択して実験するなど高次元小サンプルの問題への適応は難しい。

連絡先: 佐竹 哉太, k-satake@ml.ist.i.kyoto-u.ac.jp

3. 問題設定

本研究では Cox モデルをセミパラメトリックなまま評価するため、遺伝子情報、イベント・打ち切りの有無、生存期間のデータから、ハザード関数の大きさを求め、それによりイベントの順序を予測する問題を扱う。ただしそのまま生存期間を予測すると打ち切りデータを評価に用いることができないため、ある時点で死亡しているかどうかをラベルとして評価を行う。

今回の実験設定では観測済みデータとして特徴量 $\boldsymbol{x}_i \in \mathbb{R}^d$、生存期間 $t_i \in \mathbb{R}$、そして生存期間の終わりに死亡したかどうか $c_i \in \{0, 1\}$ を扱う。ここで $c_i = 1$ であれば死亡、$c_i = 0$ であれば観測打ち切りを意味する。本論文では、訓練データ $\{(\boldsymbol{x}_i, c_i, t_i)\}_{i=1}^N$ を用いてモデルを学習し、未知の入力データ $\boldsymbol{x}$ に対する死亡リスクの大きさ $f(\boldsymbol{x})$ を出力する問題を扱う。

4. 提案手法

まず基本となる Cox ハザードモデルについて述べ、次にそれを改善した提案手法について述べる。

4.1 Cox 比例ハザードモデル

Cox 比例ハザードモデルはセミパラメトリックなハザードモデルの代表的なもので、以下の式で定義される [Cox 72]。

$$h(\boldsymbol{x}, t) = h_0(t) \exp(\boldsymbol{\beta}^T \boldsymbol{x})$$

ここで $h_0(t)$ はベースラインハザードと呼ばれ、各 $\boldsymbol{x}$ に依存しない形で定義する。一般にベースラインハザードは陽に定義

しないことが多い。

Cox 比例ハザードモデルは、ハザード比が時間 t に依存しないという重要な性質がある。これは以下のように確かめられる。

$$\frac{h(\boldsymbol{x}_i,t)}{h(\boldsymbol{x}_j,t)} = \exp(\boldsymbol{\beta}^T(\boldsymbol{x}_i - \boldsymbol{x}_j))$$

この性質により、ベースラインハザードが分からない場合でもリスク比 $\frac{h(\boldsymbol{x}_i,t)}{h(\boldsymbol{x}_j,t)}$ を求めることができる。

Cox 比例ハザードモデルではベースラインハザードを定めないため、尤度を直接最適化することができない。そこで部分尤度関数 $L(\boldsymbol{\beta})$ を定義し、パラメータの推定を行う。

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\ L(\boldsymbol{\beta})$$

$L(\boldsymbol{\beta})$ は以下のようになる。

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n}\left[\frac{h(\boldsymbol{x}_i,t_i)}{\sum_{j\in R_i} h(\boldsymbol{x}_j,t_i)}\right]^{c_i} \\ &= \prod_{i=1}^{n}\left[\frac{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}{\sum_{j\in R_i}\exp(\boldsymbol{\beta}^T\boldsymbol{x}_j)}\right]^{c_i} \end{aligned}$$

ここでデータは $\boldsymbol{t}$ の昇順に並び替えたものとする。また R_i はリスク集合で、その時点でイベントが生じる可能性がある集合を表す。これには $c_i = 0$ のサンプルも含まれる。部分尤度の最適化は、$\boldsymbol{x}_i$ によるハザード関数の大きさを次にそのデータのイベントが選択される確率とみなしたときの条件付き確率に相当する。

この対数尤度 $l(\boldsymbol{\beta})$ は以下のようになる。

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(c_i\boldsymbol{\beta}^T\boldsymbol{x}_i - c_i\log\sum_{j\in R_i}\exp(\boldsymbol{\beta}^T\boldsymbol{x}_j)\right)$$

Cox 比例ハザードモデルのパラメータ推定では、一般に上の対数尤度関数 $l(\boldsymbol{\beta})$ が用いられる。

4.2 FM の導入

4.2.1 Factorization Machines (FM)

Factorization Machines (FM) は各特徴間の相互作用をモデル化するための手法の一つである [Rendle 10]。FM では各特徴量間の相互作用を直接パラメータとして保持するのではなく、特徴ごとに持たせた k 次元のパラメータの内積で表現する。2 次元までの特徴の相互作用を考える場合、FM を用いた回帰モデルは以下のように表される。

$$y(\boldsymbol{x}) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d}\sum_{j=i+1}^{d}\langle\boldsymbol{v}_i,\boldsymbol{v}_j\rangle x_i x_j$$

ここで $w_0 \in \mathbb{R}$、$\boldsymbol{w} = (w_1,\ldots,w_d)^T \in \mathbb{R}^d$、$\boldsymbol{V} = [\boldsymbol{v}_1,\ldots,\boldsymbol{v}_d] \in \mathbb{R}^{d\times k}$ であり、

$$\langle\boldsymbol{v}_i,\boldsymbol{v}_j\rangle = \sum_{l=1}^{k} v_{i,l}v_{j,l}$$

である。ただし $v_{i,l}$ は $\boldsymbol{v}_i$ の l 番目のパラメータを表す。ハイパーパラメータ k は自由度である。k が十分大きければ任意の交互作用をモデリングできるが、一方でスパースなデータの場合オーバーフィッティングにつながる。k が大きすぎなければ、相互作用をパラメータ化する場合に比べてパラメータ量を削減できる

4.2.2 FM による Cox モデル

Cox 比例ハザードモデルの線形和の部分に FM を導入する形で拡張する。ここでは 2 次の交互作用を考慮したモデルを用いる。

$$h(\boldsymbol{x},t) = h_0(t)\exp(\sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d}\sum_{j=i+1}^{d}\langle\boldsymbol{v}_i,\boldsymbol{v}_j\rangle x_i x_j)$$

このモデルは特徴間の相互作用の情報をパラメータ k の自由度の分だけ利用できる。また比例ハザード性を満たすため、特徴の重要度を解釈することができる。ただし他のモデルに比べモデル自体の計算量は大きくなる。

4.3 カーネル法による線形なモデル

FM は特徴間の相互作用を利用できるが、提案したモデルでは 2 次の関係までしか利用できない。そこで特徴間の高次の関係性を利用するために、ガウスカーネルによる情報を用いる拡張も考える。ここでは解釈性を残すため、線形な拡張を考える。モデルは以下のようになる。

$$h(\boldsymbol{x},t) = h_0(t)\exp((\boldsymbol{D}\boldsymbol{\alpha})^T\boldsymbol{x})$$

ただし $\boldsymbol{D}$ はグラム行列で、以下で求められる。

$$\boldsymbol{D}_{ij} = \exp(-\gamma|\boldsymbol{f}_i - \boldsymbol{f}_j|^2)$$

γ はパラメータであり、$\boldsymbol{f}_i$ は訓練データの i 番目の特徴量を並べたベクトルである。このモデルは各特徴間の非線形な情報を利用することができ、各特徴の係数により解釈性を持つ。

5. 評価実験

5.1 データセット

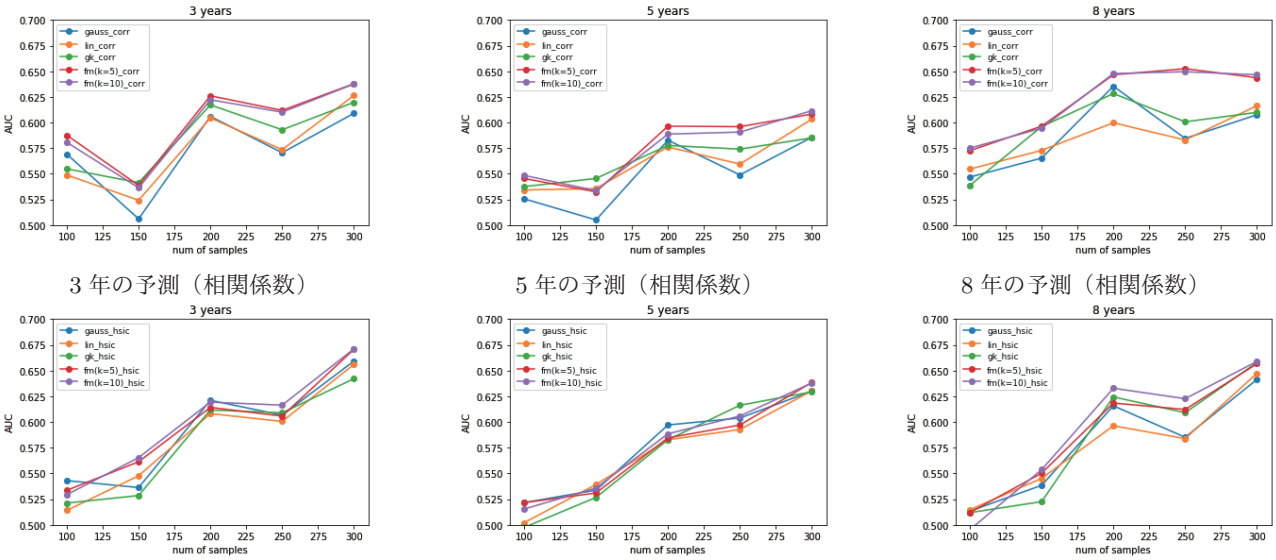
データセットは病気と生存期間に関する実データを用いる。用いたデータセットについて、サンプル数が 351、特徴の次元数が 54675 であり、打ち切りデータは 196 である。

5.2 評価手順

本研究では特徴数やサンプル数を変化させ精度を検証する。なお本研究では Cox モデルの最適化が煩雑になることを防ぐため、t_i が同一であるタイデータは用いない。今回扱うデータセットではタイデータが存在するので、まずランダムに一方を選択したのち実験を行う。

実験では、10 分割のクロスバリデーションを異なるサンプリングで 3 回行い、精度を検証する。サンプリングを行いデータを分割したのち、訓練データによって決められた特徴数だけ特徴選択を行う。この特徴選択は生存期間 t と特徴 x によって計算される。特徴選択については打ち切りデータも用いた方が良い精度を示したので、全ての訓練データで行う。特徴が選択されると訓練データからモデルを学習し、テストデータに対して予測を行い、精度を得る。なお全てのモデルで L2 正則化を用いる。

精度の計算では、まず 3 年、5 年、8 年で生存しているかどうかでテストデータをラベル付けする。このとき指定の期間未満の打ち切りデータはラベルが分からないため利用しない。次にラベル付けされた各データに関して、ベースラインハザードを除くハザード関数の大きさを求める。第 4.1 項よりハザード比は時間 t に依存せず一定であるため、このハザード関数の大きさが大きいほど死亡する確率が高いとみなすことができる。そこでハザード関数の大きさを死亡する確率と考え、ラベルに対して AUC を求めてこれを精度とする。



3 年の予測（相関係数）　5 年の予測（相関係数）　8 年の予測（相関係数）

3 年の予測（HSIC Lasso）　5 年の予測（HSIC Lasso）　8 年の予測（HSIC Lasso）

図 1: サンプル数に対する実験

5.3　比較手法

5.3.1　Cox 比例ハザードモデル

本研究で提案しているモデルの原型である Cox 比例ハザードモデルをベースラインとして用いる。定義等は第 4.1 項で述べた通りで以下のようになる。

$$h(\boldsymbol{x}, t) = h_0(t)\exp(\boldsymbol{\beta}^T\boldsymbol{x})$$

5.3.2　カーネル Cox モデル

Cox モデルの非線形な拡張として、ガウスカーネル法による手法 [Li 02] をベースラインとする。ハザードモデルは以下のようになる。

$$h(\boldsymbol{x}, t) = h_0(t)\exp(\boldsymbol{K}\boldsymbol{\alpha})$$

$\boldsymbol{K}$ はグラム行列で、以下で求められる。

$$\boldsymbol{K}_{ij} = \exp(-\gamma|\boldsymbol{x}_i - \boldsymbol{x}_j|^2)$$

γ はパラメータであり、$\boldsymbol{x}$ は各サンプルのベクトルである。ここでモデルの訓練時には $\boldsymbol{x}_i$、$\boldsymbol{x}_j$ がともに訓練データのサンプルのベクトルで、テストデータの予測時には $\boldsymbol{x}_i$ がテストデータ、$\boldsymbol{x}_j$ が訓練データのサンプルのベクトルとなる。このモデルは非線形の特徴間の情報をカーネルによって利用できるが、一方各特徴の貢献を解釈することは難しい。

5.4　特徴選択

5.4.1　相関係数による特徴選択

Cox 比例ハザードモデルでは、2. で述べたように線形の特徴選択が用いられることが多い。そこで線形な特徴選択として、目的である生存時間 t と各特徴との相関係数による特徴選択を用いる。具体的には生存時間 $\boldsymbol{t}$ と全ての特徴ベクトル $\boldsymbol{f}$ でそれぞれ相関係数を求め、その絶対値をスコアとする。その後スコアの高いものから特徴を抽出する。

5.4.2　HSIC Lasso による特徴選択

特徴間の非線形な相関を利用するため、非線形性を利用した特徴選択が可能な HSIC Lasso による特徴選択も用いる [Yamada 14]。具体的には生存時間 $\boldsymbol{t}$ と全ての特徴ベクトル $\boldsymbol{f}$ を用いて HSIC Lasso によるスコアを求め、スコアの高いものから特徴を抽出する。

6.　実験結果

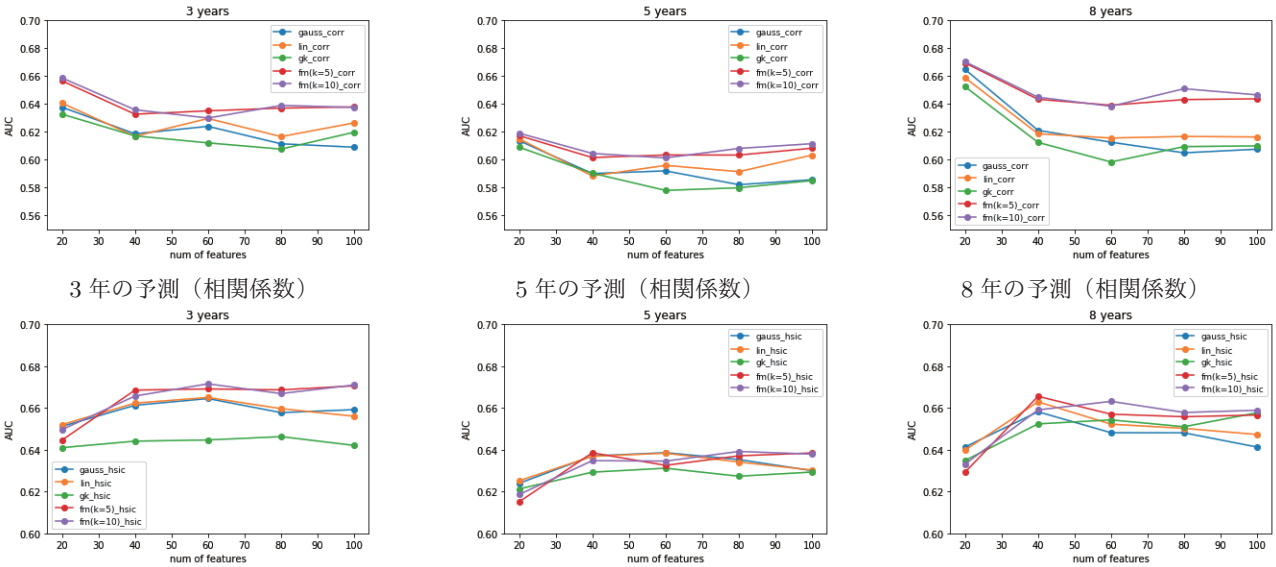
各ラベルは lin が通常の Cox モデル、gk がカーネル Cox モデルによる比較手法を表し、gauss が 4.3 で述べたカーネル法による線形な提案手法、fm が 4.2 で述べた FM による提案手法を表す。

6.1　サンプル数に対する実験

まずはじめに選択する特徴の数を 100 で固定し、サンプル数を 100 から 300 まで変化させた実験について述べる。相関係数を用いて特徴選択を行った場合の結果を図 1 の上側に示す。この結果を見ると、ほとんど全ての場合で FM による提案手法が最も良い精度を示している。これは特にサンプル数が多い場合と予測する期間が長い場合に顕著である。一方カーネル法による提案手法では精度が良くないが、これはガウスカーネルに有用な特徴を選択できていないからだと考えられる。

次に HSIC Lasso による特徴選択で同様の実験を行った結果を図 1 の下側に示す。この結果を見ると、こちらも多くの場合で FM による提案手法が最も良い精度を示している。ただし特に 5 年の予測などでは比較手法の方が良い精度を示しており、その他の場合でも相関係数による特徴選択の場合ほど比較手法と差があまりない。比較手法で精度が良いものを見ると、ガウスカーネルを用いた線形手法（gauss）や非線形手法（gk）である。HSIC Lasso による特徴選択は出力と各特徴の非線形な相関で行われるので、その情報を利用できるグラム行列を用いるこれらの手法の精度が良いと考えられる。

最後に両方の特徴選択手法で良い精度を示した FM による提案手法について比較を行った。1 による結果から、どの期間の予測でもサンプル数が少ない場合には相関係数による特徴選択が良い結果を示し、サンプル数が多い場合には HSIC Lasso による特徴選択が良い結果を示している。特にサンプル数が少ない場合にはその差が大きく、線形な特徴選択が非線形な特徴選択に比べて有用だと言える。これはサンプル数が少ない場合に複雑な特徴選択は過学習を起こしているからだと考えられる。また予測する期間が短いと HSIC Lasso が、予測する期間が長いと相関係数がそれぞれ優位な結果となっている。今回の特徴選択では出力に生存期間 t をそのまま利用しているので、



3 年の予測（相関係数）　5 年の予測（相関係数）　8 年の予測（相関係数）

3 年の予測（HSIC Lasso）　5 年の予測（HSIC Lasso）　8 年の予測（HSIC Lasso）

図 2: 特徴数に対する実験

相関係数では長期の生存に関わる特徴を選ぶことができていると考えられ、また HSIC Lasso では短い期間でも頑健である特徴を選ぶことができていると考えることができる。

6.2 特徴数に対する実験

次にサンプル数を 300 に固定し、特徴数を 10 から 100 まで変化させた実験について述べる。相関係数による特徴選択の結果を図 2 の上側に示す。この結果をから、どの年数・どのモデルの予測でも特徴数が最も少ない 10 のときに特に精度が高く、次の特徴量数 20 で精度が大幅に下がっている。これは予測に十分に寄与する特徴量が少数であるためと考えられる。また特徴量が少ない場合ではどのモデルでも差異は小さいが、特徴数が大きくなるとどの年数でも FM による提案手法と比較手法の差が大きくなっている。このことから、FM による提案手法では特に重要でない特徴も比較的うまく扱えると考えられる。

HSIC Lasso による特徴選択で同様の実験を行った結果を図 2 の下側に示す。こちらの場合は相関係数による特徴選択の場合と異なり、特徴数が少ない場合では精度が低く、その数が増えるに従って精度が向上している。このことから、HSIC Lasso による特徴選択では重要な特徴を少数で確保することはできないが、一方で満遍なく意味のある特徴を選択できていると考えられる。モデルの精度を見ると FM による提案手法がほとんどの場合で最も良い精度を示しているが、サンプル数が少ない場合には比較手法に比べて良いとは言えない。これは特徴数が少ない場合にはパラメータ数が大きすぎて過学習が起こっていると考えられる。

7. 終わりに

本研究では Cox ハザードモデルに FM による拡張を取り入れたモデル、さらに HSIC Lasso による特徴選択を提案した。実験の結果、FM によるモデルの有効性を示すことができた。また HSIC Lasso による特徴選択を Cox モデルに適用し、従来の線形な相関係数に対して不利な条件・有利な条件を示した。さらに生存期間に対する特徴選択の性質と、提案手法が特徴数の大きい場合にとくに良い精度を示すことを確認した。

今後の研究としては、まず FM で 3 次以上の交互作用を用いたモデルを適用することが考えられる。その場合パラメータを抑えても計算量は大きくなるが、特徴の数を増やすとさらに精度が良くなる可能性がある。また今回の実験は提案したモデルの精度を確かめることが目的であったため、議論が複雑になるタイデータの処理は考えていない。そのため実運用するためには実際の条件に近いタイデータを含むデータセットでも精度を確かめる必要がある。さらに実運用を考える上では特徴の数を増やしたり、別のデータセットでも提案手法の優位性が保たれるかを確認することが必要である。

参考文献

[Cox 72] Cox, D. R.: Regression models and life-tables, Journal of the Royal Sta- tistical Society: Series B (Methodological), Vol. 34, No. 2, pp. 187202 (1972).

[Li 02] Li, H. and Luan, Y.: Kernel Cox regression models for linking gene expres- sion profiles to censored survival data, Biocomputing 2003, World Scien- tific, pp. 6576 (2002).

[Katzman 16] Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T. and Kluger, Y.: Deep survival: A deep cox proportional hazards network, stat, Vol. 1050, p. 2 (2016).

[Rendle 10] Rendle, S.: Factorization machines, Data Mining (ICDM), 2010 IEEE 10th International Conference on, IEEE, pp. 9951000 (2010).

[Yamada 14] Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P. and Sugiyama, M.: High-dimensional feature selection by feature-wise kernelized lasso, Neural computation, Vol. 26, No. 1, pp. 185207 (2014).