Discovering Subgroups with Exceptional Survival Characteristics

Anonymous Authors[1]

Abstract

In many applications, it is important to identify subpopulations that survive longer or shorter than the rest of the population. In medicine, for example, it allows determining which patients benefit from treatment, and in predictive maintenance, which components are more likely to fail. Existing methods for discovering subgroups with exceptional survival characteristics require restrictive assumptions about the survival model (e.g. proportional hazards), pre-discretized features, and, as they compare average statistics, tend to overlook individual deviations. In this paper, we propose SYSURV, a fully differentiable, non-parametric method that leverages random survival forests to learn individual survival curves, automatically learns conditions and how to combine these into inherently interpretable rules, so as to select subgroups with exceptional survival characteristics. Empirical evaluation on a wide range of datasets and settings, including a case study on cancer data, shows that SYSURV reveals insightful and actionable survival subgroups.

1. Introduction

Survival analysis traditionally focuses on estimating whether a *given* group of individuals has different survival characteristics than a reference population. This has applications in many fields, but most obviously in precision medicine, as it allows characterizing patients who benefit from a treatment. What if the subgroups are not yet known? Can we *learn* easily interpretable rules that select subgroups with exceptional survival characteristics? Can we learn these in a flexible, differentiable manner, without restrictive assumptions or pre-discretizing features, allowing even heavy censoring, while keeping individual deviations in mind? That is exactly the topic of this paper.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

We consider time-to-event data, in which for each sample we have a vector $\mathbf{x}$ of descriptive features (covariates), a time t since we started observing the subject, and an indicator δ of whether the event of interest (e.g. death) has occurred since then. We are interested in learning conjunctive rules that define conditions on the covariates (e.g. $x_1 \in [0.2, 1.0]$ **and** $x_2 \in [0.8, 0.9]$) and so select a subgroup for which the corresponding survival curves are exceptional with regard to the overall population.

To find these subgroups, we use a *non-parametric* measure of exceptionality which is based on *individual* survival curves. That is, unlike the state of the art, we do *not* rely on group averages, as this obscures individual deviations. We show how to learn rules that single out subgroups with exceptional survival characteristics in a differentiable manner, automatically finding relevant features and intervals in a way that allows easy optimization, yet crisp logical interpretation after learning.

Through an extensive set of experiments, we show that our method works well in practice. It outperforms state-of-the-art methods by a wide margin. Through a case study on cancer treatment, we confirm that SYSURV finds biomarkers that are associated with good, respectively, poor response to treatment, while also identifying novel subgroups that warrant further investigation.

In a nutshell, our main contributions are as follows

1. We propose a differentiable, non-parametric measure of survival subgroup exceptionality based on individual survival estimates.
2. We show how to differentiably learn rules that select subgroups with exceptional survival characteristics.
3. We empirically evaluate our method on a wide range of experiments, comparing to 3 state-of-the-art methods.

2. Related Work

Subgroup discovery First introduced by Klösgen (1996), is the task of finding and describing subpopulations that are exceptional in terms of a target property. Typically, subgroup discovery approaches leverage different exceptionality measures to accommodate the data type of a single target variable (Song et al., 2016; Kalofolias & Vreeken, 2022).

Whereas most methods for subgroup discovery rely on combinatorial search, Xu et al. (2024) recently showed that gradient-based optimization can efficiently learn subgroups in large datasets and removes the need for pre-discretization.

Survival subgroup discovery That is, subgroup discovery applied to survival data. Existing approaches to survival subgroup discovery are typically based on the logrank statistic (Mantel, 1966). RULEKIT (Gudyś et al., 2020) relies on heuristic search, while FIBERS (Urbanowicz et al., 2023) is evolutionary-based, and, more recently, ESMAMDS (Vimieiro et al., 2025) was based on Ant Colony Optimization. In contrast, we propose a gradient-descent-optimizable objective that uses flexible individual survival functions.

3. Preliminaries

We consider time-to-event (survival) data, where we have a dataset $D = \{(\mathbf{x}^{(i)}, t^{(i)}, \delta^{(i)})\}_{i=1}^{n}$ consisting of n i.i.d. realizations, called *subjects*, from a joint distribution $\mathbb{P}(\mathbf{X}, T, \Delta)$. Here, $\mathbf{x} \in \mathbb{R}^p$ is a vector of p covariates that describes the subject, $t \in \mathbb{R}_{\geq 0}$ measures time since the start of its observation, and δ indicates whether the event of interest (e.g. relapse) has occurred yet. If the event did occur ($\delta = 1$), t is the time at which it was observed. If it did not occur, t is the latest time we observed the subject to still be fine, and so the outcome is said to be *censored*. A survival function $S(t)$ gives the probability of a subject surviving beyond time t, i.e. it is defined as the cumulative distribution function $\mathbb{P}(T > t)$. To ease notation, we denote the domains of $\mathbf{x}$ and t by $\mathcal{X}$ and $\mathcal{T}$, respectively.

A subgroup Q is a set of subjects selected from D using a rule $\sigma_Q \colon \mathcal{X} \to \{0, 1\}$ that indicates whether a subject belongs to Q or not, i.e. $Q = \{\mathbf{x} \in D \mid \sigma_Q(\mathbf{x}) = 1\}$. The survival function of subgroup Q is denoted by $\hat{S}_Q(t)$.

4. Learning Survival Subgroups

We are interested in learning rules that select subgroups with survival trends that are exceptional compared to those of the general population. For that, we need three ingredients: a way to model subgroup survival, a measure for subgroup exceptionality, and a way to learn these rules.

4.1. Subgroup Survival Model

Precision medicine overcomes the traditional "one-size-fits-all" model by incorporating individual characteristics into treatment outcome predictions (Feuerriegel et al., 2024). This is particularly important in survival analysis where subject heterogeneity can significantly impact outcomes. For us, this translates to estimating individual survival.

Individual survival functions $\hat{S}(t \mid \mathbf{X} = \mathbf{x})$, or $\hat{S}(t \mid \mathbf{x})$ for short, can leverage the covariates $\mathbf{x}$ of a subject and so provide more accurate and more robust survival estimates (Cox, 1972) than those estimated marginally (Kaplan & Meier, 1958). Given individual estimates, the survival function for a subgroup Q selected by rule σ_Q is defined as

$$\hat{S}_Q(t) \coloneqq \mathbb{E}_{\mathbf{x} \sim \mathbf{X}}[\hat{S}(t \mid \mathbf{x}) \mid \sigma_Q(\mathbf{x}) = 1] \,.$$

To obtain individual estimates, we fit a random survival forest (RSF) (Ishwaran et al., 2014) over the entire dataset. This is a fully-non-parametric model that extends random forests (Breiman, 2001) to time-to-event data. Unlike well-known parametric models, it does not restrict us to specific survival distributions (e.g. Weibull) or structures (e.g. Cox).

4.2. Subgroup Exceptionality Measure

In traditional survival analysis, the subgroup of interest is assumed to be *given* and exceptionality versus the reference population is measured on average, at the group level. This may lack in sensitivity to per-subject deviations.

Individual estimates can *reveal* differences from the reference survival that group-level estimates *obscure*. An average curve obscures the curves of individual subjects when they are very different, effectively underestimating the true exceptionality. In general, individual-level exceptionality measures offer increased sensitivity to deviations of individual survival functions making them particularly suited for optimization of subgroup membership rules. Regardless whether we measure exceptionality at the group or individual level, should survival functions cross, it is important to also consider the absolute difference to the reference.

Given two groups A and B, selectable by rule σ_A and σ_B, resp., for which the expected group-level survival at any time t are $\hat{S}_A(t)$ and $\hat{S}_B(t)$, and for which individual-level survival is denoted by $\hat{S}(t \mid \mathbf{x})$. For our purposes, A would typically represent a group of interest, e.g. a subgroup, and B a reference group, e.g. the entire dataset. To measure the exceptionality between two survival functions $S_A, S_B \geq 0$, we consider the L^1 distance $\ell^1_{\mathcal{T}}(S_A, S_B) \coloneqq \int_{t \in \mathcal{T}} |S_A(t) - S_B(t)| \, dt$. Finally, we define our exceptionality measure as the expected difference in survival of the subjects of a subgroup selectable by σ from the estimated survival in the population $\hat{S}_D(t)$ throughout $\mathcal{T}$ as

$$\phi(\sigma, \sigma_D) \coloneqq \mathbb{E}_{\mathbf{x} \sim \mathbf{X}}[\ell^1_{\mathcal{T}}(\hat{S}(t|\mathbf{x}), \hat{S}_D(t)) \mid \sigma(\mathbf{x}) = 1] \,. \quad (1)$$

Next, we will show how to learn rules that select subgroups with high exceptionality according to this measure.

4.3. Rule Learner

Armed with our exceptionality measure, we now present SYSURV for learning subgroups with exceptional survival characteristics using gradient-based optimization.

Learnable rules Subgroups are selected via rules, so we first formalize the language of these rules. Traditionally, a subgroup is defined by a hard rule $\sigma\colon \mathbf{x} \mapsto \bigwedge_{j=1}^{p} \pi(x_j; \alpha_j, \beta_j)$, where each π is a Boolean condition evaluating to true (1) if a covariate x falls within the interval defined by lower and upper bounds $\alpha, \beta \in \mathbb{R}$, e.g. "$18 < \mathit{age} < 32$".

To enable differentiable subgroup discovery, we employ a continuous relaxation of these logical expressions. Specifically, we use soft rules $\hat{\sigma}\colon \mathcal{X} \to [0,1]$ consisting of soft conditions $\hat{\pi}\colon \mathbb{R} \to [0,1]$. These soft conditions model the probability of a covariate being inside the specified interval via a composition of two opposing sigmoids located at the learnable bounds α and β. A temperature hyperparameter $\tau > 0$ controls the strictness of the bounds; as $\tau \to 0$, the soft condition $\hat{\pi}$ converges to a Boolean interval.

Following Xu et al. (2024), we define the soft condition as

$$\hat{\pi}(x_j; \alpha_j, \beta_j, \tau) \coloneqq \frac{e^{\frac{1}{\tau}(2x_j - \alpha_j)}}{e^{\frac{1}{\tau}x_j} + e^{\frac{1}{\tau}(2x_j - \alpha_j)} + e^{\frac{1}{\tau}(3x_j - \alpha_j - \beta_j)}} ,$$

and the differentiable rule learner using a weighted harmonic mean of soft conditions as

$$\hat{\sigma}(\mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \tau) = \frac{\sum_{j=1}^{p} w_j}{\sum_{j=1}^{p} w_j \; \hat{\pi}(x_j; \alpha_j, \beta_j, \tau)^{-1}} ,$$

where $\mathbf{w} \in \mathbb{R}^p$ represents a vector of learnable weights. These weights allow the model to perform feature selection, i.e. a covariate x is actively included in the conjunction when $w > 0$ and effectively ignored as $w \to 0$. The harmonic mean structure is chosen because it serves as a smooth approximation of the logical-and operator. Henceforth, the output $\hat{\sigma}(\mathbf{x})$ can be interpreted as the probability that a subject belongs to the subgroup $\mathbb{P}(\hat{\sigma}(\mathbf{X}) = 1 \mid \mathbf{X} = \mathbf{x})$.

Objective Our goal is to learn those rules to select the subgroups that maximize our objective. For rule updates, we need to take the gradient of our objective w.r.t the parameters of the soft rule $\hat{\sigma}$. Hence, we rewrite the expectation in Eq. (1) to derive our objective where $\hat{\sigma}$ explicitly appears as $\phi(\hat{\sigma}, \hat{\sigma}_D) =$

$$\int_{\mathbf{x} \in \mathcal{X}} \ell^1_{\mathcal{T}}(\hat{S}(t|\mathbf{x}), \hat{S}_D(t)) \; \frac{\mathbb{P}(\mathbf{X} = \mathbf{x})}{\mathbb{P}(\hat{\sigma}(\mathbf{X}) = 1)} \; \hat{\sigma}(\mathbf{x}) \; d\mathbf{x} , \quad (2)$$

by using the expected value definition, marginalization rule and Bayes' theorem. We estimate the integral over $\mathcal{X}$ and $\mathcal{T}$ in Eq. (2) using the standard Monte Carlo and trapezoidal methods, respectively. Our objective is defined as

$$\hat{\phi}(\hat{\sigma}, \hat{\sigma}_D) \coloneqq \frac{1}{|\hat{\sigma}|} \sum_{i=1}^{n} \ell^1_{\mathcal{T}}(\hat{S}(t \mid \mathbf{x}^{(i)}), \hat{S}_D(t)) \; \hat{\sigma}(\mathbf{x}^{(i)}; \boldsymbol{\theta}) ,$$

where the subgroup size $|\hat{\sigma}|$ is estimated as $\frac{1}{n}\sum_{i=1}^{n} \hat{\sigma}(\mathbf{x}; \boldsymbol{\theta})$, and $\boldsymbol{\theta}$ stands for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\mathbf{w}$, collectively.

Rule generality To avoid learning overly specific subgroups, as is standard in subgroup discovery, we introduce a size penalty $|\hat{\sigma}|^\gamma$ to our objective. We give the size-adjusted objective as

$$\arg\max_{\boldsymbol{\theta}} |\hat{\sigma}|^\gamma \; \hat{\phi}(\hat{\sigma}, \hat{\sigma}_D) .$$

We control the trade-off between exceptionality and subgroup size via the hyperparameter $\gamma \in [0,1]$.

Optimization Gradient-based optimization allows us to efficiently learn the parameters of our rules, and therewith scale SYSURV to large datasets. For every subject, each feature is subjected to its corresponding soft condition according to the learned bounds, which are then weighted and combined into a soft rule that predicts a membership probability. We iteratively learn the weights and bounds using standard first-order gradient-based optimization techniques (Kingma & Ba, 2017) for a number of epochs, while annealing the temperature to arrive at crisp discretizations.

5. Experiments

Next, we empirically evaluate SYSURV on synthetic and real-world data. We compare it to state-of-the-art methods RULEKIT (Gudyś et al., 2020), FIBERS (Urbanowicz et al., 2023), and ESMAMDS (Vimieiro et al., 2025), all of which optimize the logrank for exceptionality.

5.1. Synthetic Data

We start by evaluating on data with known ground truth. To this end, we generate data of $n = 10\,000$ subjects subject to 10% censoring. We split the subjects into in-subgroup (20%) and out-of-subgroup (80%). We sample the population survival time from a $\mathit{Weibull}(1.5, 5)$. We make the in-subgroup subjects experience the event $5\times$ earlier on expectation by sampling from a $\mathit{Weibull}(1.5, 1)$. The in-subgroup subjects are defined by a rule that selects on their features, which are sampled from a standard normal distribution, where some features are noise.

Scalability We assess the performance on data where we vary the number of features. We report the average F1-scores over 10 runs in Fig. 1 (left). We see that all methods are stable as the dimensionality increases, and that RULEKIT does not finish within 48 hours for more than 500 features. SYSURV outperforms the closest competitor, ESMAMDS, by a wide margin.

Censoring We assess robustness to censoring when increasing the percentage of censored subjects. We give the results in Fig. 1 (left); we see that SYSURV is very stable with average F1-scores of approx. 0.8, up to very high rates of censoring ($> 80\%$) outperforming the baselines.

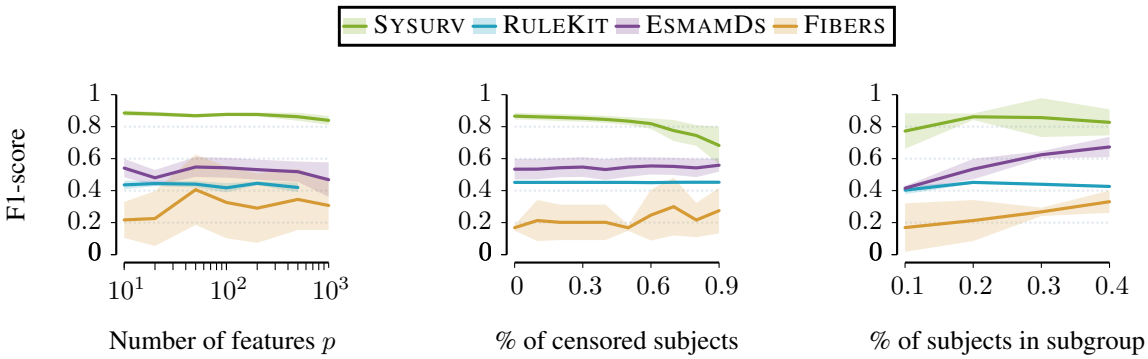

Figure 1. Synthetic setting. Comparison of SYSURV and baselines in terms of F1-scores recovering planted subgroups with increasing dimensionality **(Left)**, censoring **(Center)** and subgroup sizes **(Right)**. Higher is better.

Subgroup Size We assess retrieval quality as the percentage of subjects that belong to the planted subgroup increases. We see in Fig. 1 (right) that all methods improve as the planted subgroup size increases. We attribute this behavior in the baselines to them preferring larger subgroups, as observed. Nevertheless, SYSURV remains the best performing method across all subgroup sizes.

SYSURV performs very well and is very stable with average F1-scores of approx. 0.8, up to very high rates of censoring ($> 80\%$). FIBERS is highly unstable, with very large standard errors. ESMAMDS is the closest competitor, hovering around 0.55, closely followed by RULEKIT consistently at 0.45, while FIBERS lags behind in all experiments by a wide margin, with average F1-scores of approximately 0.35.

5.2. Case Study: Neck Cancer

We qualitatively evaluate SYSURV via a case study. We consider a high-dimensional dataset of patients with locally advanced head and neck squamous cell carcinoma (HNSCC) undergoing radiochemotherapy (Schmidt et al., 2020). The data consists of patients who received radiochemotherapy after a tumor removal operation ($n = 190$). The covariates are tumor gene expressions ($p = 158$) and the outcome is the time until tumor recurrence at 85% censoring. From a clinical perspective, we are both interested in subgroups of subjects that respond better, as well as those that respond worse to treatment, as this allows more targeted treatment.

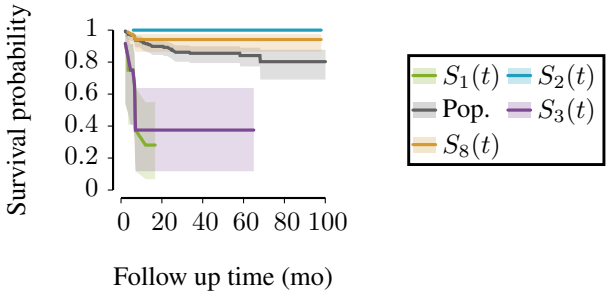

Figure 2. Case study. SYSURV discovers two subgroups of poor responders of HNSCC data.

We run SYSURV on the dataset and show the survival curves in Fig. 2. SYSURV finds two subgroups that respond better, and two that respond worse, to treatment than the overall populations. Almost 80% of the patients survive for more than 80 months (Fig. 2). We are specifically interested in subgroups S_1 (green, n=12) and S_3 (purple, n=12) as these identify subjects that respond very poorly. The corresponding rules make biological sense: they both select on genes that relate to cellular communication, metabolism, and DNA repair, which are related to highly aggressive tumor subtypes (Lendahl et al., 2009; Toustrup et al., 2011) for which conventional treatment not only fails to improve patient condition but even promotes recurrence (Barker et al., 2015). The learned rules allow clinicians to pre-screen these subjects and offer them alternate treatment instead.

6. Conclusion

We propose SYSURV, a method for survival subgroup discovery. SYSURV leverages non-parametric survival regression to learn individual survival functions. SYSURV exploits individual-level deviations in survival with respect to the population in the way it quantifies subgroup exceptionality. SYSURV employs a rule learner whose parameters can be learned to automatically select features and cutoffs along their domains to select members of a subgroup. By doing so, SYSURV effectively results in human-interpretable rules that describe subgroups with exceptional survival behavior as it maximizes subgroup exceptionality. Extensive experiments on synthetic datasets demonstrate that SYSURV consistently outperforms existing baselines. In a case study on neck cancer patients, SYSURV discovers biomarkers that are associated with good, respectively, poor response to treatment while also identifying novel subgroups that may warrant further investigation.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Barker, H. E., Paget, J. T., Khan, A. A., and Harrington, K. J. The tumour microenvironment after radiotherapy: mechanisms of resistance and recurrence. *Nature Reviews Cancer*, 15(7):409–425, 2015.

Breiman, L. Random Forests. *Machine learning*, 45(1): 5–32, 2001.

Cox, D. R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(2):187–202, 1972.

Feuerriegel, S., Frauen, D., Melnychuk, V., Schweisthal, J., Hess, K., Curth, A., Bauer, S., Kilbertus, N., Kohane, I. S., and van der Schaar, M. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30(4): 958–968, 2024.

Gudyś, A., Sikora, M., and Wróbel, Ł. RuleKit: A comprehensive suite for rule-based learning. *Knowledge-Based Systems*, 194:105480, 2020.

Ishwaran, H., Gerds, T. A., Kogalur, U. B., Moore, R. D., Gange, S. J., and Lau, B. M. Random survival forests for competing risks. *Biostatistics*, 15(4):757–773, 2014.

Kalofolias, J. and Vreeken, J. Naming the Most Anomalous Cluster in Hilbert Space for Structures with Attribute Information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4057–4064, 2022.

Kaplan, E. L. and Meier, P. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization, 2017. URL https://arxiv.org/abs/1412.6980.

Klösgen, W. *Explora: a multipattern and multistrategy discovery assistant*, pp. 249–271. American Association for Artificial Intelligence, 1996. ISBN 0262560976.

Lendahl, U., Lee, K. L., Yang, H., and Poellinger, L. Generating specificity and diversity in the transcriptional response to hypoxia. *Nature Reviews Genetics*, 10(12): 821–832, 2009.

Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170, 1966.

Schmidt, S., Linge, A., Grosser, M., Lohaus, F., Gudziol, V., Nowak, A., Tinhofer, I., Budach, V., Sak, A., Stuschke, M., Balermpas, P., Rödel, C., Schäfer, H., Grosu, A.-L., Abdollahi, A., Debus, J., Ganswindt, U., Belka, C., Pigorsch, S., Combs, S. E., Mönnich, D., Zips, D., Baretton, G. B., Buchholz, F., Baumann, M., Krause, M., and Löck, S. Comparison of GeneChip, nCounter, and Real-Time PCR–Based Gene Expressions Predicting Locoregional Tumor Control after Primary and Postoperative Radiochemotherapy in Head and Neck Squamous Cell Carcinoma. *The Journal of Molecular Diagnostics*, 22 (6):801–810, 2020.

Song, H., Kull, M., Flach, P., and Kalogridis, G. Subgroup Discovery with Proper Scoring Rules. In *Machine Learning and Knowledge Discovery in Databases*, pp. 492–510. Springer, 2016.

Toustrup, K., Sørensen, B. S., Nordsmark, M., Busk, M., Wiuf, C., Alsner, J., and Overgaard, J. Development of a Hypoxia Gene Expression Classifier with Predictive Impact for Hypoxic Modification of Radiotherapy in Head and Neck Cancer. *Cancer Research*, 71(17):5923–5931, 2011.

Urbanowicz, R., Bandhey, H., Kamoun, M., Fogarty, N., and Hsieh, Y.-A. Scikit-FIBERS: An'OR'-Rule Discovery Evolutionary Algorithm for Risk Stratification in Right-Censored Survival Analyses. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pp. 1846–1854, 2023.

Vimieiro, R., Mattos, J. B., and de Mattos Neto, P. S. EsmamDS: A more diverse exceptional survival model mining approach. *Information Sciences*, 690:121549, 2025.

Xu, S., Walter, N. P., Kalofolias, J., and Vreeken, J. Learning Exceptional Subgroups by End-to-End Maximizing KL-Divergence. In *International Conference on Machine Learning*, volume 235, pp. 55267–55285. PMLR, 2024.