MultiFlags: A Probabilistic Framework for Article-Based Size Advice in Fashion E-Commerce

Matthias Späth[1,*], Andrea Nestler[1,*], Henry Böddeker[1,*], Leonidas Lefakis[1], Yevgeniy Puzikov[1], Rodrigo Weffer[1], Nour Karessli[1], Nadja Klein[2] and Reza Shirvany[1]

[1]*Zalando SE, Berlin, Germany*

[2]*Technische Universität Dortmund, Germany*

Keywords: Fashion, Article-Based Size Advice, Recommendation Systems, Size and Fit, Size-Related Returns, E-Commerce, Probabilistic Framework.

Abstract: Accurately modeling the size behavior of fashion articles at scale is a critical task for fashion e-commerce. However, it has proven to be highly challenging due to inconsistent sizing systems across countries, inconsistent garment design processes, and brand-specific sizing specifications. Widespread methods in the field focus primarily on giving customers rudimentary size recommendations (e.g., we recommend you size S) based on the customers' purchase behavior and/or their size and fit preferences. These approaches fail to take into account the size and fit behavior of the article, for example their design cut, shape, material, etc. (or at best treat it with simplistic *ad hoc* assumptions), and in turn, not effectively reducing the high volume of online article returns due to size and fit. In this work, we propose a theoretically-motivated probabilistic framework, *MultiFlags*, which can significantly reduce size-related returns in fashion e-commerce thanks to modeling multiple aspects of article's size and fit behavior. We also highlight how this framework enables a principled approach to article-based size advice, while leveraging data from multiple modalities. The results validate the competitiveness of the proposed framework in the state-of-the-art in several size advice scenarios that are critical for fashion e-commerce. The framework is deployed in production in a large e-commerce site, serving millions of customers and driving significant results.

1 INTRODUCTION

The rise and steady growth of fashion e-commerce has introduced customers to a novel shopping experience. On the one hand, customers are enabled to browse and search clothing items from a multitude of brands and different trends simply from their phones. On the other hand, they can no longer physically interact with the clothing items. In order to aid customers in navigating this new shopping environment, a host of fashion-focused algorithmic products have been developed in recent years (see Jaradat et al. (Jaradat et al., 2022) and references therein).

One of the main challenges in choosing the right clothing item to wear is its size, which has been shown to be a major factor in returning clothes and shoes (C, 2014). In the brick-and-mortar experience, a customer typically has access to a fitting room to try clothes and shoes on before purchasing them, and to shop attendants or family and friends who can give feedback on the right size and fit of the item. In online fashion, purchasing clothes and shoes in the wrong size not only will result in that article being returned, it most importantly will cause frustration to the customer, increase CO_2 emissions, reduce profitability for the online shop, and increase logistical costs and delivery trucks on the roads. Considering the CO_2 emissions, for example, a recent e-commerce carbon report indicates that reducing one return could cut a whooping ~ 0.99 kg of CO_2 equivalent emissions (Aso, 2020). Considering that millions of clothes and shoes are returned each year in online fashion due to size and fit issues, the CO_2 impact of reducing the size related returns is significant. In fact, these effects are pivotal today and in coming years where returns are seeing a steady increase (Choi, 2016). We consider the size-related returns, in theory, to be an unnecessary hassle, espe-

*These authors contributed equally to this research.

Späth, M., Nestler, A., Böddeker, H., Lefakis, L., Puzikov, Y., Weffer, R., Karessli, N., Klein, N. and Shirvany, R.
MultiFlags: A Probabilistic Framework for Article-Based Size Advice in Fashion E-Commerce.
DOI: 10.5220/0013709900004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 313-322

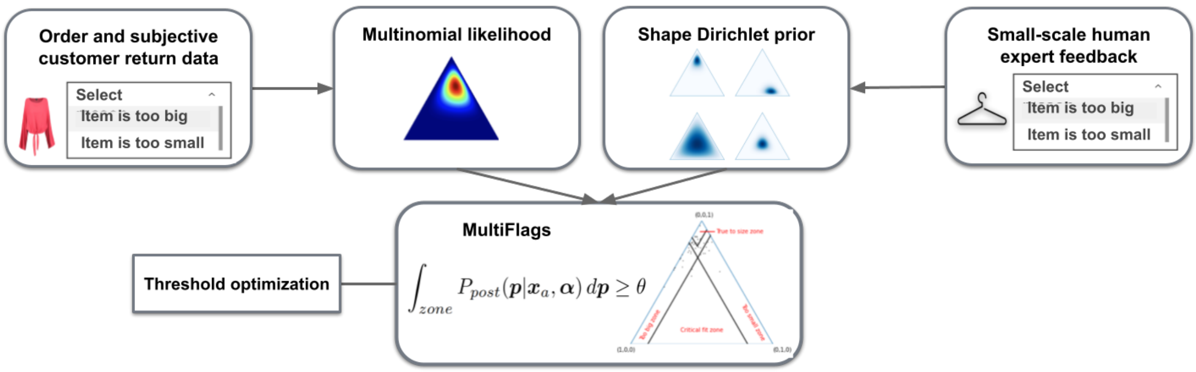

Figure 1: High level overview of the new article-based size advice framework *MultiFlags*.

cially considering that the customers are often satisfied with all other attributes of their order (e.g., item's color, price, material, delivery, etc.) and only due to the fact that the ordered size of the clothes or shoes does not fit them, they have no choice but to return the item. This is an unsolved problem today in fashion e-commerce due to: the lack of standardized sizing systems (e.g., discrepancies in a medium size across different brands), incompatible sizing variations (e.g., the difference between EU and UK sizes), the prevalence of vanity sizing (intentional mislabeling to make customers feel smaller, e.g., make customers believe they are a size 6 when in reality they are a size 8), inconsistent size charts (e.g., conflicting measurements for the same size across brands), limited size range (e.g., excluding plus-size or petite options), variations in body shapes (e.g., inadequate consideration of different proportions), inconsistent fit (e.g., the same size fitting differently in different styles), and insufficient size information in product descriptions.

Addressing these challenges requires the implementation of creative solutions to enhance the sizing experience and improve overall customer satisfaction in the realm of online fashion. Prior work on developing algorithmic products to assist customers often focuses on providing personalized size recommendations, relying rather on customers' personal data (Lefakis et al., 2021; Yuan et al., 2021), or their prior purchases (Hajjar et al., 2020; Sheikh et al., 2019; Lasserre et al., 2020), and often without considering the articles' sizing issues (Guigourès et al., 2018; Sembium et al., 2017; Sembium et al., 2018; Dogani et al., 2019; Abdulla and Borar, 2017; Abdulla et al., 2019). This approach to size advice excludes customers without a purchase history or customers who buy for others within a single account.

A notable exception is *SizeFlags* (Nestler et al., 2021) where articles are "flagged" as too-big and too-small based on historic customer returns for this article, expert assessments, and the fit issue probability assessment of an image based convolutional neural network, *SizeNet* (Karessli et al., 2019). Such advice allows customers to decide whether to go "one size down" or "one size up" compared to the usual size they would be purchasing. In this work, we present a novel joint probabilistic framework, called *MultiFlags*, which similarly provides article-based size advice. *MultiFlags* leverages a multinomial-Dirchlet model to build a recommender system that eliminates the need for separate treatment of the article size issues while relaxing the limiting assumptions made in previous works. Figure 1 shows a high level overview of the framework and is detailed in Section 3.

2 RELATED WORK

2.1 Size Recommendations

Previous work on the issue is mostly focused on suggesting a specific size of an article for the customer's preferences, i.e. size recommendation. Abdulla and Borar (Abdulla and Borar, 2017) formulate the recommendation task as binary classification, circumventing data sparsity by learning article latent representations from categorical features including brand, size, occasion, etc. Customer representations are built by aggregating article vectors from their order history. Sembium et al. (Sembium et al., 2017) propose a latent factor model for customers and products, corresponding to their physical true size. Sheikh et al. (Sheikh et al., 2019) exploit the correlations between different categories in a scalable deep learning system that can ingest broader article and customer features, while Dogani et al. (Dogani et al., 2019) leverage transfer learning from brand to product level. Hajjar et al. (Hajjar et al., 2020) further extend this ability and explicitly consider the sequential and temporal nature of the problem by applying an attention-based model. Inspired by advancements in the computer vision field, Lasserre et al. (Lasserre et al., 2020) utilize a meta-learning approach that efficiently learns new customer sizes with little order data. Other approaches further deal with extreme cases of cold-start customers with little or no purchase history using physical body-related features (Lefakis et al., 2021; Yuan et al., 2021). From another angle on the problem, computer vision techniques and article measurement data are used to learn article fit and body shape compatibility (Hidayati et al., 2018; Hsiao and Grauman, 2020). With the aim of simulating the in-store customer fitting experience, recent efforts developed virtual try-on solutions to visualize garments on a target body. While 3D methods (Bhatnagar et al., 2019; Patel et al., 2020) accurately render garments on virtual avatars, they often require expensive real body scans. Overcoming this requirement, generative adversarial models are used to warp clothing items on a target person in 2D images (Han et al., 2018; Dong et al., 2019; Issenhuth et al., 2019; Pecenakova et al., 2022; Pang et al., 2024), however, the physical feasibility of such methods in real-life applications is not guaranteed. There have been earlier Bayesian models in the size and fit space focused on personal size recommendation. Sembium et al. (Sembium et al., 2018) used Bayesian inference to find the best article size for a given customer, by modelling the joint posterior of a customer's true size and article sizes by using the purchase histories. Guigourès et al. (Guigourès et al., 2018) use a similar strategy by modelling the joint probability distribution of return and article size in a hierarchical model. Size recommendation poses several challenges. Some approaches require sensitive customer information, such as height and weight, or scans of the customer's physical body to be able to generate recommendations or simulate the fitting experience (Yuan et al., 2021; Lefakis et al., 2021; Bhatnagar et al., 2019; Patel et al., 2020).

2.2 Article-Based Size Advice

The status quo on article-based size advice is that the existing state-of-the-art is notably and sadly very sparse, where only a handful of publications focus on the critical article-level size advice capability unconditioned on a target customer's orders or body properties (Nestler et al., 2021; Karessli et al., 2019; Baier, 2019; Misra et al., 2018), which is key in alleviating the issues with data sparsity, sizing systems and the customer-specific cold start problem. This has been approached by analyzing customer reviews (Baier, 2019; Misra et al., 2018), and by evaluating cold-start articles with a computer vision-based teacher-student framework (Karessli et al., 2019). Building on that work, (Nestler et al., 2021) propose *SizeFlags*, a method capable of incorporating different priors and evaluated on a large-scale A/B test scenario showing SizeFlags reduce the size-related return rate across different fashion categories of textile and shoes. Due to space limitations, we do not provide a detailed description of *SizeFlags* (Nestler et al., 2021) and refer the reader to the original paper. In short, *SizeFlags* (Nestler et al., 2021) is reducing size-related returns by providing article-based size advise with the recommendation of selecting a larger/ smaller size when the article runs small/ large. It models the probability of an article being too-small or too-big separately, using two separate Beta-Binomial distributions. This implies the assumption of independence between the probability of an article being too-big or too-small, which is unlikely to hold in practice and constitutes one of the main limitations of *SizeFlags* (Nestler et al., 2021) . Furthermore, *SizeFlags* (Nestler et al., 2021) uses point estimates for predictive posterior calculation and does not estimate the confidence of the model in its predictions.

2.3 Baselines for Size-Related Return Reduction

To our knowledge, *SizeFlags* (Nestler et al., 2021) is the only state-of-the-art approach that has demonstrated a positive impact on reducing size-related returns—the key objective of this line of work in fashion e-commerce—and will therefore serve as our baseline for article-based size advice. Other article-based size advice approaches have neither claimed nor have been evaluated by their authors in terms of reducing size-related returns, making them not directly applicable as baselines for the *MultiFlags* approach. Previous approaches using personalized recommenders as described in Section 2.1 are often evaluated based on acceptance rate (usually defined as the share of sales where a customer follows the recommendation and keeps it). This in turn is not possible for article-based size advice like the presented *MultiFlags*, because the true size of the customer is unknown, so it remains unclear whether the customer followed the advice or not.

To provide a broader context beyond article-based advice, we also consider a personalized size-recommendation paradigm outlined in Section 2.2 for comparison, which utilizes the meta-learning methodology proposed by Lasserre et al. (Lasserre et al., 2020). In this approach, the individual customer gets a direct size recommendation (e.g., ”We recommend size S”) that is dependent on both the customer and the article. This allows us to frame the new algorithm described in this paper against both article-based and personalized recommendation approaches (Section 4.3).

The scarcity of prior research on data science solutions for reducing size-related returns underscores the urgent need for advancements in this area. This paper addresses this gap by presenting a novel, comprehensive solution that not only overcomes existing limitations but also makes a significant contribution to the field through our probabilistic framework and its demonstrably impactful results.

3 *MULTIFLAGS:* A MULTI-CLASS SIZE ADVICE ALGORITHM

The aim of the proposed *MultiFlags* approach is to lower size-related returns in fashion e-commerce by providing article-based size advice to the customer. It leverages the article's sales and return reason information, such as "too small" and "too big", which is typically collected in fashion e-commerce. The output is a flag that provides the customer with additional sizing information on the article, such as "This article runs large, we recommend going one size down". Optional additional input for the model is used as prior information and can be based on expert knowledge on the article or visual sizing cues based on models such as *SizeNet* (Karessli et al., 2019). The prior information can help raising flags faster and earlier in the product life cycle, so that customers can benefit early on. The approach is illustrated in Figure 1. *MultiFlags* addresses the limitations described in Section 2.2, leading to novelty in five points: 1) Instead of using separate models, the estimation of all flags is unified into a single model, acknowledging the dependence between an article being too-big (tb; an article fits bigger than expected) or too-small (ts; an article fits smaller than expected). 2) The approach naturally gives rise to two new flag types, namely true-to-size (tts; an article has good fit in their indicated size.) and critical-fit (cf; an article is ts and tb at the same time) when the probability vertex is split into different zones. 3) The framework provides a principled approach to include prior information. 4) The model has a flagging logic based on confidence and 5) the model is extendable to prior mixtures and hyperpriors.

3.1 A Unified Probabilistic Framework

The framework leverages a conjugate Bayesian Model with a Dirichlet prior $p|\alpha \sim Dir(\alpha)$.

The return realizations $x = (x_{ts}, x_{tb}, x_{tts})$ (with $x_k =$ number return reasons $k \in \{tb, ts\}$; $x_{tts} =$ number sales $-x_{tb} - x_{ts}$) given $p = (p_{ts}, p_{tb}, p_{tts})$ (class probability vector with $\sum_{k\in\Omega} p_k = 1$) are modelled using a multinomial likelihood $x|p \sim Mult(p)$. Thus, the probability density functions of the likelihood $P(x|p)$ and prior $P_{prior}(p|\alpha)$ read as

$$P(x|p) = \frac{n!}{x_{ts}!\,x_{tb}!\,x_{tts}!} p_{ts}^{x_{ts}}\, p_{tb}^{x_{tb}}\, p_{tts}^{x_{tts}}, \quad (1)$$

$$P_{prior}(p|\alpha) = \frac{1}{B(\alpha)} \prod_{i\in\Omega} p_i^{\alpha_i - 1}, \quad (2)$$

where $B(\alpha)$ is the multivariate Beta function, $\Omega = \{ts, tb, tts\}$ is the set of indices. The posterior probability density function, by conjugacy, is defined as

$$P_{post}(p|x,\alpha) = \frac{P(x|p)P_{prior}(p|\alpha)}{\|\cdot\|} \quad (3)$$

$$= P_{prior}(p|x+\alpha) \quad (4)$$

$$= \frac{1}{B(x+\alpha)} \prod_{i\in\Omega} p_i^{x_i+\alpha_i-1}. \quad (5)$$

Note that, while *SizeFlags* (Nestler et al., 2021) computes a score value for each flag directly and then compares to a threshold for raising a flag, *MultiFlags* estimates the posterior probability, determining flags at a later step.

3.2 Estimation of the Posterior Probability

We define the observed size-related return rate $srr_k(a)$ of a given article $a \in C$ that belongs to a fashion category $C \subset C_{all}$ (*e.g.* jeans, shirts, shoes) as

$$srr_k(a) = \frac{x_k}{n}, k \in \{tb, ts\}. \quad (6)$$

with n the number of sales. The true size-related return rate $srr_k^\star(a)$ of an article is unknown, but we know that the higher the number of article orders n, the more confident we are that $srr_k(a)$ is close to $srr_k^\star(a)$:

$$srr_k^\star(a) = \lim_{n\to\infty} srr_k(a). \quad (7)$$

Each fashion category C poses different size and fit challenges. Therefore, we evaluate each article against its category.

The mean and the standard deviation

$$\mu_{C,k} := \text{mean}(\{srr_k(a)\}_{a\in C}), \quad (8)$$

$$\sigma_{C,k} := \text{std}(\{srr_k(a)\}_{a\in C}), \quad (9)$$

for $k \in \{ts, tb\}$ differ for each category. For example, $\mu_{C,k}$ of the category $C = C_{t-shirt}$ is in general much smaller than $\mu_{C,k}$ with $C = C_{dresses}$ as t-shirts are less complex in terms of fit than dresses, which results in a lower size-related return rate.

Based on the category averages $\mu_{C,k}$ and standard deviations $\sigma_{C,k}$, we define the zones that are interesting for the different flag types. The tb/ts zone is defined by the $srr^{\star}_{k}(a)$ of an article a being more than one standard deviation $\sigma_{C,k}$ above the category mean $\mu_{C,k}$ for $k \in \{ts, tb\}$ (red and green zone in Figure 2). The cf zone is defined by having both the $srr^{\star}_{tb}(a)$ and $srr^{\star}_{ts}(a)$ above the category specific mean (blue zone in Figure 2). The tts zone is defined exactly opposite by having both the $srr^{\star}_{tb}(a)$ and $srr^{\star}_{ts}(a)$ below the category specific mean (yellow zone in Figure 2). The four zones are the most obvious approach to zoning the probability vertex and hence naturally give rise to the two new flag types. The four zones result in the following logic for $k \in \{tb, ts\}$:

$$\left.\begin{array}{lcl} a \text{ is k} & \Leftrightarrow & srr^{\star}_{k}(a) \geq \mu_{C,ts} + \sigma_{C,k}, \\ a \text{ is cf} & \Leftrightarrow & srr^{\star}_{ts}(a) \geq \mu_{C,ts} \;\; \& \\ & & srr^{\star}_{tb}(a) \geq \mu_{C,tb}, \\ a \text{ is tts} & \Leftrightarrow & srr^{\star}_{ts}(a) \leq \mu_{C,ts} \;\; \& \\ & & srr^{\star}_{tb}(a) \leq \mu_{C,tb}. \end{array}\right\} \quad (10)$$

Example 1. *Let us illustrate the zoning rules (10) with one example for a real world category C_{rw}, which has an average size-related return rate $\mu_{C,k} = 10\%$ for $k \in \{tb, ts\}$, meaning 10% of sold items are returned as too-big or too-small each. The standard deviations are $\sigma_{C,tb} = 0.05$ and $\sigma_{C,ts} = 0.06$. C_{rw} contains more than 100 000 different articles.*

In practice $srr^{\star}_{k}$ for $k \in \{tb, ts\}$ is not known, but let's suppose we pick one article $a \in C_{rw}$ with a high $srr^{\star}_{ts}(a) = 33\%$ and low $srr^{\star}_{tb}(a) = 8\%$. The article a has an exceeding amount of too-small returns in the category C_{rw} with $srr^{\star}_{ts}(a) = 33\% \geq \mu_{C,ts} + \sigma_{C,ts} = 10\% + 6\% = 16\%$. At the same time this article shows a usual too-big returns since $srr^{\star}_{tb}(a) = 8\% \leq \mu_{C,tb} + \sigma_{C,tb} = 10\% + 5\% = 15\%$. This article would therefore be a suitable too-small candidate. But we do not know the true return rate. Therefore we integrate over the probability distribution in the zone defined by $\mu_{C,k}$ and $\sigma_{C,k}$ for $k \in \{tb, ts\}$ to get a confidence measurement. Based on this, we decide whether the item is actually marked as too-small. This number helps us to make a confident decision as to whether an item will be finally be labeled as too-small. This approach is outlined in the next paragraph.

With definition (10), for a given article $a \in C$ with return realization x and fixed parameter α, the probabilities are defined as

$$\begin{aligned} \mathcal{P}_{ts} &= \mathcal{P}(p_{ts} \geq \mu_{C,ts} + \sigma_{C,ts} = l^{(1)}_{ts}), & (11) \\ \mathcal{P}_{tb} &= \mathcal{P}(p_{tb} \geq \mu_{C,tb} + \sigma_{C,tb} = l^{(2)}_{tb}), \\ \mathcal{P}_{cf} &= \mathcal{P}(p_{ts} \geq \mu_{C,ts} = l^{(1)}_{cf}, p_{tb} \geq \mu_{C,tb} = l^{(2)}_{cf}), \\ \mathcal{P}_{tts} &= \mathcal{P}(p_{ts} \leq \mu_{C,ts} = u^{(1)}_{tts}, p_{tb} \leq \mu_{C,tb} = u^{(2)}_{tts}). \end{aligned}$$

With the zone boundaries $u^{(i)}_{k}$ and $l^{(i)}_{k}$ from (11) and additionally $u^{(i)}_{k} = 1$ for $k \neq tts$ and $l^{(2)}_{ts} = l^{(1)}_{tb} = l^{(i)}_{tts} = 0$ we get $\forall k \in \Omega^{+}$ and $i \in \{1, 2\}$

$$\mathcal{P}_{k} = \int_{u^{(1)}_{k}}^{l^{(1)}_{k}} \int_{u^{(2)}_{k}}^{l^{(2)}_{k}} \int_{0}^{1} P_{post}(p|x, \alpha)\, dp \quad (12)$$

For each $k \in \Omega^{+}$ ($=\Omega \cup \{cf\}$) this results in a posterior probability $\mathcal{P}_{k} \in [0, 1]$, which can be interpreted as the probability that $srr^{\star}_{k}(a)$ of an article a is within the specified zone. For example, $\mathcal{P}_{ts} = 97\%$ implies that, based on the current return realizations of the article, it's 97% likely that the $srr^{\star}_{k}(a)$ is within the too-small zone. With this approach the posterior probability can be interpreted as a confidence measure in the flag.

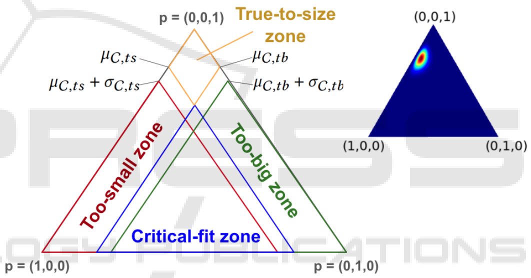

Figure 2: (Bottom Left) The different zones are defined based on their respective $\mu_{C,k}$ and $\sigma_{C,k}$.
(Top Right) Probability density distribution for a too-small flagged article $a \in C_{rw}$.

In Figure 2 we illustrate the different zones on the probability vertex and re-visit the Example 1 with category $C = C_{rw}$ and article $a \in C_{rw}$ with $srr^{\star}_{ts}(a) = 33\%$ and $srr^{\star}_{tb}(a) = 8\%$. Assuming return realizations $x + \alpha = (8.4, 2.9, 28.5)$ (e.g. based on return observations $x = (6, 0, 6)$ and prior $\alpha = (2.4, 2.9, 22.5)$), the posterior zone value integrates to $\mathcal{P}_{ts} = 97\%$, $\mathcal{P}_{tb} = 0.01\%$, $\mathcal{P}_{cf} = 1\%$ and $\mathcal{P}_{tts} = 1\%$. Here we used the prior concentration parameter $\alpha = \alpha^{\star}_{C}$ which will be described in the next Section 3.3.

The approach *MultiFlags* raises a flag when $\mathcal{P}_{k} \geq \theta \in [0, 1]$. In Section 4.1 we will show how an optimal threshold $\theta^{\star} = \theta$ for raising a flag can be determined.

3.2.1 Computational Complexity and Efficient Integration Methods

The biggest bottleneck is the numerical integration to get the posterior probability $\mathcal{P}_{k}(x(a_i), \alpha_k)$ in equation

(12). To increase efficiency, the double integration of the two free parameters is performed only in zones with overlap, reducing the runtime by a factor of 10. Zones that depend on one parameter are integrated using the marginal distribution of the Dirichlet distribution. To derive the marginal distribution of a Dirichlet distribution with three dimensions, we use the definition of conditional probability after integrating the third dimension, $f(p_1, p_2) = f(p_1) \times f(p_2|p_1)$. Rewriting the equation shows that the marginal is a Beta distribution, $\frac{1}{1-p_1} P_2|P_1 \sim Beta(\alpha_2, \alpha_3)$, which has a closed-form cumulative density function, so the numerical integration is not necessary. When the too-small and too-big zones overlap, we calculate the individual zones using the marginal Beta distributions, and handle the overlap separately. To reduce the time, the full computation is parallelized on a distributed computing system, i.e. Spark, on a cluster with 8 workers and 8 cores each, taking 3 hours for several million articles.

3.3 Estimation of the Parameter α

3.3.1 Estimation Based on Category Sales and Return Data

All articles within a category C have similar size-related return properties, which can be incorporated as a prior in the Bayesian approach, as $\mu_{C,k}$ and $\sigma_{C,k}$ for $k \in \{tb, ts\}$ are known at the time of calculation.

To estimate the concentration parameter α for the prior, we use the Dirichlet-Multinomial distribution. It is a closed-form compound distribution of the Dirichlet and multinomial distribution, where the variable p of the Dirichlet distribution is integrated out. With $A = \sum_{k \in \Omega} \alpha_k$ and the Gamma function $\Gamma()$ the likelihood function reads

$$\begin{aligned} f(x|\alpha, n) &= \int P(x|p) P_{prior}(p|\alpha) dp \\ &= \frac{\Gamma(A)\Gamma(n+1)}{\Gamma(n+A)} \prod_{k \in \Omega} \frac{\Gamma(x_k + \alpha_k)}{\Gamma(\alpha_k)\Gamma(x_k+1)}. \end{aligned} \tag{13}$$

The parameter $\alpha = \alpha^\star_C$ for each category C can be estimated using the maximum likelihood method

$$\alpha^\star_C = \arg\max_\alpha \prod_{a \in C} f(x|\alpha, n). \tag{14}$$

Applying method (14) for the category C_{rw} in Example 1, we obtain the optimal solution $\alpha^\star_C = (2.4, 2.9, 22.5)$. This problem is solved by a combined Gradient-Newton method.

3.3.2 Estimation Based on Prior Information

We aim to have articles flagged as early as possible to ensure both customers and the business benefit respectively from the supportive advice and the lower return rates thanks to those early flags. Using the prior information as outlined here helps to tackle the cold-start problem for articles where not enough return information is available and aims at raising flags with less processed returns available. For *MultiFlags* we introduce prior information from human expert feedback from fashion models. The feedback is based on fashion models who try on articles before they are activated on the platform to provide first hand feedback on a subset of articles. They provide size and fit feedback similar to the customer return information, indicating whether an article runs ”too small”, ”too big” or ”true to size”. The averaged feedback of multiple human models for the same article is used to estimate the article-specific prior parameters.

Let $C_{exp}(k) \subset C$ be the subset of all articles a for which the feedback of the fashion models is exactly k (i.e. $feedback(a) \equiv k$ for $k \in \{tb, ts, tts\}$). Then, similar to equation (14), the *Expert Prior* $\alpha^\star_{C_{exp}(k)} = \alpha$ can be determined via the maximum likelihood method. Revisiting Example 1 with category $C = C_{rw}$ and focusing on all articles with human expert feedback $C_{exp}(k)$, the maximum likelihood approach for articles with too-small human expert feedback ($k \equiv ts$) results in the concentration parameters $\alpha_{ts} = 6.3$, $\alpha_{tb} = 2.2$, and $\alpha_{tts} = 15.0$.

4 *MULTIFLAGS:* DATA AND RESULTS

MultiFlags is compared to the state-of-the-art benchmark for article-based size advise, *SizeFlags* (Nestler et al., 2021) . The comparison creates an additional complexity that goes beyond the complexity of benchmarking against no size advise at all, as the comparison of the two models, gives rise to four distinct cases, namely when a flag is added by *MultiFlags* (*Added*), dropped by *MultiFlags* (*Dropped*) or when both approaches raise the same flag (*Same*) or a different flag (*Different*) for the same article. A graphical overview of these four sets can be found in Figure 3. In the coming sections we will focus on the evaluation of the *Same*, *Dropped* and *Added* flags as the number of *Different* flags is close to zero and negligible in our experiments.

The following evaluation of the models is separated into two sections: Section 4.1 addresses the *Same* and *Dropped* flags, by optimizing the θ threshold, which was introduced in Section 3.2, on a training set and evaluating the results on a holdout set via Difference-in-Differences approach. Section 4.2 extends the evaluation of the two models to the *Added*

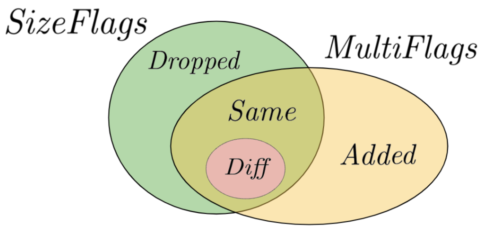

Figure 3: The comparison of the two approaches gives rise to the *Same*, *Different*(*Diff*), *Dropped* and *Added* flags.

flags via A/B test. Finally, Section 4.3 discusses the complementary effects between our article-based approach and personalized size recommendations.

4.1 Threshold Optimization and Holdout Set Performance Evaluation

4.1.1 Datasets

The data acquisition is based on the return process of the e-commerce platform. Customers can return articles at no charge with or without providing a return reason from a predefined list of reasons. With respect to size or fit, customers have the possibility to state ”The item is too small” or ”The item is too big” among other non-size-related options (such as ”The item isn't as described”, ”The item is too expensive” etc.). The dataset used is a random subset of anonymized data coming from two years [June 2021 - May 2023] of article purchases and returns of a major fashion e-commerce platform for multiple European countries. It contains the sales and return information for 3.6 million textile articles. The additional prior information was available for ~ 5000 articles for the human expert feedback. The top five categories are dresses, knitwear, jerseys, blouses, trousers and these together make up 55% of articles in the dataset. We highlight that this method is applicable for most fashion e-commerce platforms as it requires a dataset that contains only the sales as well as ”too small” and ”too big” returns on an article level, which fashion e-commerce platforms typically collect.

To our knowledge, from publicly available fashion datasets (fas, 2022), only two are relevant for the size topic in fashion, ‘ModCloth’ and ‘RentTheRunWay’ (Misra et al., 2018). These contain fashion products with fit labels, however, both datasets lack the sales and return information required to test the proposed approach and for drawing any conclusion on reducing size-related returns.

We perform a training-test split and use the first year of data [June 2021 - May 2022] for parameter optimization as outlined in Section 4.1.3 and the second year [June 2022 - May 2023] for the results presented in Section 4.1.4. *MultiFlags* is tested experimentally by running the algorithm on the second year of historical data (holdout set), where the results of the reference model *SizeFlags* (Nestler et al., 2021) are also known.

4.1.2 Evaluation Method DiD

Given the lack of a ground truth, the causal effect of showing size advice to customers can only be captured through the relative reduction of size-related returns (srr_{red}) of the two approaches. The nearest neighbor Difference-in-Differences (DiD) approach of Nestler et al. (Nestler et al., 2021) is used as a quasi-experimental signal of the srr_{red} by a flag. This method, originally presented in Heckmann et al. (Heckman et al., 2019), compares the srr_{red} for articles which received a flag against similar articles which did not receive a flag over a time frame of 6 weeks before and after raising the flag. We aggregate the srr_{red} to calculate the number of saved returns per article. It is important to note, that srr_{red} cannot be calculated with DiD for all flags $\mathcal{M}(\theta)$ that are raised by *MultiFlags* with fixed threshold θ, as $\mathcal{M}(\theta)$ includes new flags for which no historical results exist. It is, however, possible to calculate the srr_{red} for all *SizeFlags* (Nestler et al., 2021) $\mathcal{S}$ that have already been shown to the customers. It follows that srr_{red} can be calculated for the two subsets *Same* $= \mathcal{S} \cap \mathcal{M}(\theta)$ and *Dropped* $= \mathcal{S} \setminus \mathcal{M}(\theta)$. The additional flags for which we cannot calculate the srr_{red} via the DiD method are denoted by *Added* and are A/B tested in Section 4.2. The sets are illustrated in Figure 3.

4.1.3 Threshold Optimization on Training Set

The practical approach outlined here focuses on determining θ for comparing the proposed approach with *SizeFlags* (Nestler et al., 2021) . We use the data over a period of one year [June 2021 - May 2022] and find the best value which satisfies two constraints:

(i) $\theta \in [0.5, 1]$, which corresponds to the probability that the true size-related return rate $srr^{\star}_{k}(a)$ is within the respective zone is above the threshold of 0.5, i.e. $\mathcal{P}_k \geq 0.5, k \in \{ts, tb, cf, tts\}$, with $\mathcal{P}_k$ from equation (11).

(ii) Find θ as such that the overall srr rate $srr_{red}(\mathcal{M}(\theta))$ is reduced as much as possible.

Choosing a high value of θ implies that *MultiFlags* $\mathcal{M}(\theta)$ raises flags only when it is very confident about the prediction. With $\theta^{\star} = 1$, $\mathcal{M}(\theta)$ raises only $\approx 5K$ *tb* and *ts* flags, which is ≈ 30 times less than the number of flags raised by *SizeFlags* (Nestler et al., 2021)

$\mathcal{S}$ in the same timeframe. Therefore, to find the optimal value of θ, the effectiveness per single flag based on (i) needs to be balanced vs. the total number of flags.

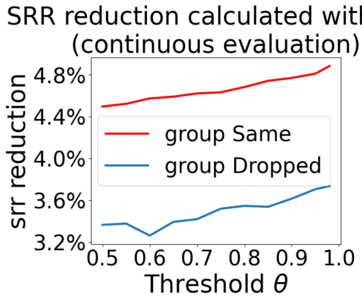

Figure 4: The highest srr reduction is achieved for $\theta \to 1$.

Figure 4 shows that for $\theta \in [0.5, 1]$ the srr reduction varies only slightly, providing a relatively free choice of θ value in that range. Given the free choice of θ in the specified range and in order to be able to compare $\mathcal{S}$ vs. $\mathcal{M}(\theta)$ on equal footing, we select θ as such that

1. we achieve a balance between the groups *Dropped* and *Added* ($|\mathcal{M}(\theta)| \approx |\mathcal{S}|$) and
2. the flag recall $R(\theta) = \frac{|Same(\theta)|}{|\mathcal{S}|}$ should be as large as possible.

By using grid search, we found $\theta^\star = 0.695$ to be the optimal value, resulting in $|\mathcal{M}(\theta)| \approx |\mathcal{S}|$ with the numbers of *Dropped* and *Added* balanced (Figure 5), and the flag recall $R(0.695) = 78.2\%$ sufficiently high.

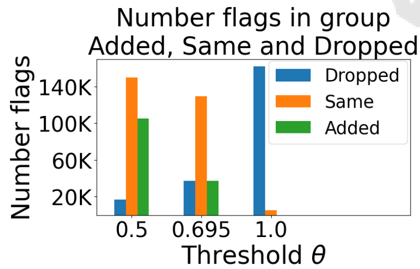

Figure 5: For $\theta = 0.695$ there are as many flags in *Added* as in *Dropped*.

With the resulting $\theta^\star$, we also see a first indication that $\mathcal{M}(\theta^\star)$ performs well on the cold-start problem, as the necessary number of returns needed to raise a flag are smaller in the case of $\mathcal{S}$ ($\hat{r}_{MF} = 161$ vs. $\hat{r}_{SF} = 165$).

4.1.4 Results on Holdout Set

We evaluate *MultiFlags* $\mathcal{M} := \mathcal{M}(\theta)$ with fixed $\theta = 0.695$ vs. *SizeFlags* (Nestler et al., 2021) $\mathcal{S}$ on the holdout set. Despite both sets covering all four seasons by choosing one year for the training and the holdout set each, both sets are slightly different. The holdout set contains 2.6 million articles, which is 9.3% more than in the training set and the number of *SizeFlags* (Nestler et al., 2021) $|\mathcal{S}| = 176\,386$ is 5.5% higher than in the training set. With $|\mathcal{M}| = 193\,103$, $\mathcal{M}$ contains 9.5% more flags than $\mathcal{S}$. $\mathcal{M}$ raises 144 378 *Same* flags and achieves a recall of 81.9% compared to $\mathcal{S}$, showing that it performs equally well as on the training set. The number of *Dropped* is 32 008 and *Added* is 48 725. The same DiD approach is used to calculate the srr_{red}. Similar to the training set, the size-related return reduction of the group *Same* with $srr_{red} = 3.8\%$ is greater than the reduction of the group *Dropped* with $srr_{red} = 2.8\%$. Lastly, $\mathcal{M}$ requires $\hat{r}_M = 151$ returns to raise a flag versus $\hat{r}_S = 166$ for $\mathcal{S}$, indicating a more than 10% improvement in terms of speed for $\mathcal{M}$, potentially increasing the average number of saved returns per item over an extended period, by offering a flag earlier in the article's life cycle.

4.1.5 New Flags Critical-Fit and True-to-Size

In addition to the results in comparison with the *SizeFlags* (Nestler et al., 2021) $\mathcal{S}$ baseline, we explore the results obtained for the new flags offered by *MultiFlags* $\mathcal{M}$, critical-fit and true-to-size. $\mathcal{M}$ raises 62 408 critical-fit flags on the holdout set. These enable new actionable information for customers about articles with difficult fit and provide new opportunities for businesses. Note that critical-fit denotes ill fit of an article, because it does not fit the majority of the customers, which is a problem that can be addressed in several ways. Depending on the criticality of the fit issue, businesses can, for instance, down-sort the article in it's catalog views, exclude the articles from discovery, communicate the article to brands for their improvement, or ultimately remove the article from the assortment. $\mathcal{M}$ raised 159 411 true-to-size flags, which enable high confidence for customers when selecting a size; contrary to the articles with a critical-fit flag, businesses are able to up-sort such articles in the article ranking shown to the customers.

4.2 A/B Test Evaluation Against Article-Based Size Advice

The flags added by the $\mathcal{M}$ approach have been A/B tested against $\mathcal{S}$. As those flags are added, it's an A/B test where a flag from $\mathcal{M}$ shown to the treatment group is compared to the same article without a flag from $\mathcal{S}$ shown to the control group. The A/B test was performed November 2023 to January 2024 in the whole textile category with more than 132 000 customers and 150 000 orders per group from 12 different countries. The A/B test demonstrated that the size-related return rate was significantly reduced by $\mathcal{M}$ (-4.5% relative change, p-value < 0.001). The biggest impact was observed in the category of full body garments (e.g. dresses and tracksuits; -7.3%), followed by upper garments (e.g. shirts and pullovers; -3.6%) and lower garments (e.g. pants and skirts; -2.1%). The A/B test also showed an increase in selection orders (in which a customer orders multiple sizes of the same article; $+4.4\%$) for the treatment group and a reduction of cost-intensive re-orders (in which a customer returns an article and re-orders in a different size; -6.3%).

Table 1: *Added* flags by *MultiFlags* $\mathcal{M}$ show the strongest srr reduction.

Group	*Dropped*	*Same*	*Added*
Evaluation	DiD	DiD	A/B test
Dataset	Holdout	Holdout	Production
Srr reduction	2.8%	3.8%	4.5%

The results of Section 4.1 and 4.2 are summarized in Table 1. The DiD results show that the flags dropped by $\mathcal{M}$ have a clearly lower srr reduction than *Added* and *Same*. We highlight that for the sake of this study we chose a single global θ such that the number of flags each method raises is roughly equal. Within this context it is important to note that though *Dropped* have a positive srr reduction (as estimated by DiD), *Added* have an even larger srr reduction (as estimated by the A/B test), and consequently for the same number of flags we expect a higher overall srr reduction by $\mathcal{M}$. Furthermore, if we wish to reduce the number of *Dropped* flags, this can be achieved by lowering the θ parameter or even having different θ_k for each flag k. The choice here depends, amongst other considerations, on the aggressiveness of the strategy.

4.3 Complementary Effects with Personalized Size Recommendations

MultiFlags $\mathcal{M}$ requires only article-level information and as such it is available for *all* customers shopping that flagged article without limitations or needs with respect to customer data. In contrast, size recommenders leverage the combined data from both customers and articles and as such are only available to customers from whom the size matching optimization given an article is confidently achieved thanks to a pre-requisite depth level in the customer data. In other words, $\mathcal{M}$ marks only those articles with a sizing flag that exhibit a systemic sizing behavior (e.g., articles that run too large) and enables all customers to improve their sizing choices on those articles and reduce sizing returns. However, most articles display regular sizing behavior. This is precisely where size recommenders play a crucial role: providing customers with personalized size advice when no size flag is available, ensuring comprehensive size guidance throughout the product catalog. This indicates that both approaches are not interchangeable but rather complementary solutions, which will be the subject of future work.

5 CONCLUSION

A probabilistic framework was presented to build theoretically motivated article-based size advice solutions, with a confidence measure for the raised flags. This framework enables the true-to-size and critical-fit flags, which can be leveraged in practical approaches for businesses. With empirical results, the strengths of the approach were shown: the proposed approach reduces size-related returns significantly more than *SizeFlags* (Nestler et al., 2021) baseline. Fewer returns save the customer time as the return process can be time consuming and also improve sustainability and environmental footprint. Future work involves experimenting with different prior setups for including more signals (e.g., leveraging sentiment analysis of reviews, leveraging LLMs to gather richer return information) and developing the approach into hierarchical recommender system.

REFERENCES

(2019/2020). Asos operational carbon report.

(2022). Fourth workshop on recommender systems in fashion.

Abdulla, G. M. and Borar, S. (2017). Size recommendation system for fashion e-commerce. In *Machine Learning Meets Fashion Workshop at ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Abdulla, G. M., Singh, S., and Borar, S. (2019). Shop your right size: A system for recommending sizes for fashion products. In *Companion Proceedings of The 2019 World Wide Web Conference*.

Baier, S. (2019). Analyzing customer feedback for product fit prediction. *arXiv*.

Bhatnagar, B. L., Tiwari, G., Theobalt, C., and Pons-Moll, G. (2019). Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

C, R. (2014). *How Fashion Ecommerce Retailers Can Reduce Online Returns*. Econsultancy.

Choi, T.-M. (2016). *Analytical Modeling Research in Fashion Business*. Springer.

Dogani, K., Tomassetti, M., Vargas, S., Chamberlain, B. P., and Cnudde, S. D. (2019). Learning embeddings for product size recommendations. In *eCOM@SIGIR*.

Dong, H., Liang, X., Shen, X., Wang, B., Lai, H., Zhu, J., Hu, Z., and Yin, J. (2019). Towards multi-pose guided virtual try-on network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9025–9034.

Guigourès, R., Ho, Y. K., Koriagin, E., Sheikh, A.-S., Bergmann, U., and Shirvany, R. (2018). A hierarchical Bayesian model for size recommendation in fashion. In *Proceedings of the 12th ACM Conference on Recommender Systems*, pages 392–396. ACM.

Hajjar, K., Lasserre, J., Zhao, A., and Shirvany, R. (2020). Attention gets you the right size and fit in fashion. In *14th Conference on Recommender Systems*. ACM.

Han, X., Wu, Z., Wu, Z., Yu, R., and Davis, L. S. (2018). Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552.

Heckman, J., Ichimura, H., and Todd, P. (2019). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

Hidayati, S. C., Hsu, C.-C., Chang, Y.-T., Hua, K.-L., Fu, J., and Cheng, W.-H. (2018). What dress fits me best? fashion recommendation on the clothing style for personal body shape. In *26th ACM International Conference on Multimedia*.

Hsiao, W.-L. and Grauman, K. (2020). Vibe: Dressing for diverse body shapes. In *Computer Vision and Pattern Recognition*.

Issenhuth, T., Mary, J., and Calauzènes, C. (2019). End-to-end learning of geometric deformations of feature maps for virtual try-on. *arXiv preprint arXiv:1906.01347*.

Jaradat, S., Dokoohaki, N., Pampín, H. J. C., and Shirvany, R. (2022). *Fashion Recommender Systems*, pages 1015–1055. Springer US, New York, NY.

Karessli, N., Guigoures, R., and Shirvany, R. (2019). SizeNet: Weakly supervised learning of visual size and fit in fashion images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*.

Lasserre, J., Sheikh, A.-S., Koriagin, E., Bergmann, U., Vollgraf, R., and Shirvany, R. (2020). Meta-learning for size and fit recommendation in fashion. In *SIAM International Conference on Data Mining (SDM20)*.

Lefakis, L., Koriagin, E., Lasserre, J., and Shirvany, R. (2021). Towards user-in-the-loop online fashion size recommendation with low cognitive load. In Dokoohaki, N., Jaradat, S., Corona Pampín, H. J., and Shirvany, R., editors, *Recommender Systems in Fashion and Retail*, pages 59–76, Cham. Springer International Publishing.

Misra, R., Wan, M., and McAuley, J. (2018). Decomposing fit semantics for product size recommendation in metric spaces. In *Proceedings of the 12th ACM Conference on Recommender Systems*, RecSys '18, New York, NY, USA.

Nestler, A., Karessli, N., Hajjar, K., Weffer, R., and Shirvany, R. (2021). SizeFlags: Reducing size and fit related returns in fashion e-commerce. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Pang, K., Zou, X., and Wong, W. (2024). Learning visual body-shape-aware embeddings for fashion compatibility. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8056–8065.

Patel, C., Liao, Z., and Pons-Moll, G. (2020). Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Pecenakova, S., Karessli, N., and Shirvany, R. (2022). Fitgan: Fit- and shape-realistic generative adversarial networks for fashion. In *26th International Conference on Pattern Recognition (ICPR) 2022*. arXiv.

Sembium, V., Rastogi, R., Saroop, A., and Merugu, S. (2017). Recommending product sizes to customers. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, New York, NY, USA.

Sembium, V., Rastogi, R., Tekumalla, L., and Saroop, A. (2018). Bayesian models for product size recommendations. In *Proceedings of the 2018 Web Conference*. ACM.

Sheikh, A.-S., Guigourès, R., Koriagin, E., Ho, Y. K., Shirvany, R., and Bergmann, U. (2019). A deep learning system for predicting size and fit in fashion e-commerce. In *Proceedings of the 13th ACM Conference on Recommender Systems*. ACM.

Yuan, Y., Park, M.-J., and Huh, J.-H. (2021). A proposal for clothing size recommendation system using chinese online shopping malls: The new era of data. *Applied Sciences*, 11(23).