Forecasting Demand Distributions for New Products: Combining Subjective Rankings with Sales Data

Marat Salikhov
Yale School of Management, marat.salikhov@yale.edu

Nils Rudi
Yale School of Management, nils.rudi@yale.edu

A major obstacle to wider adoption of the newsvendor model is the difficulty of obtaining its key input—the demand distribution forecast, specifically when the products are new and no historical data are available. In such cases, judgmental forecasting methods are a commonly suggested solution, in particular, the Sport Obermeyer approach which collects point forecasts of demand quantity from a panel of experts and uses the degree of disagreement between experts as a proxy for demand uncertainty. However, our attempt to implement this approach at fashion retailer Moods of Norway was a failure. We were not able to recruit a sufficiently large and diverse crowd because many potential experts found it difficult to provide quantity inputs. In response to this issue, we started asking the experts to rank the products within their respective categories. While this new type of input boosted participation, its conversion to quantities requires additional data and new methodology. To that end, we propose to use category-wise historical data, and we constructed a framework for this conversion based on a tripartite decomposition of the demand vector into total demand, ordered proportions, and ranking. We also propose several new evaluation metrics and test our framework on a dataset from Moods of Norway.

Key words: demand forecasting; probabilistic forecasting; wisdom of crowds; newsvendor model; subjective rankings

1. Introduction

To avoid stagnation in the race for market share, many companies undertake a cycle of high-risk bets on new product launches. In the absence of data on historical demand, prediction of future demand is particularly difficult; the result is a high level of demand uncertainty. Hence companies face major risks when deciding on the quantities of their new products. The key aspect of that gamble—namely, the potential mismatch between supply and demand—is captured by the newsvendor model.

There cannot, of course, exist any past sales data for a new product. It follows that time-series forecasting methods, which are the natural choice for established products, cannot be employed directly. Yet new products in some cases share objectively defined attributes with existing ones, which enables the use of historical data from similar, "like-to-like" products. In contrast, this paper focuses on new products whose attractiveness is largely determined by subjective attributes that are difficult to capture and/or not clearly related to past products. For such products, a natural alternative approach is so-called judgmental demand forecasting—an approach based on subjective inputs.

Judgmental forecasting and estimation are most accurate when the responses (a.k.a. inputs) from many individuals are combined, a phenomenon known as "the wisdom of crowds" (Surowiecki 2004, Galton 1907). Such aggregate judgments often outperform *the* best of the experts, picked with the benefit of hindsight. For example, a commonly used demonstration of wisdom of crowds in the MBA classroom asks the students to estimate the quantity of jelly beans in a jar (Treynor 1987); in most instances, none or almost none of the students are able to give an estimate that is better than the crowd average.

Lyon and Pacuit (2013) structure the wisdom of crowds into six core aspects, illustrated in Figure 1 (in bold italics) and defined as follows (and specified for the jelly beans example in parentheses):

Output. The target result of the wisdom of crowds (estimated quantity of jelly beans).

Inputs. The relevant pieces of information collected from the crowd members (each student's estimated quantity).[1]

Aggregation. Mapping the set of inputs to the output (average of all provided quantities).

Recruiting. Selecting the crowd members, also known as *experts*, from a pool of candidates (ask the students that show up for the class).

Elicitation. Collecting inputs from the crowd (hand out and collect back the pieces of paper on which each student writes down the quantity of jelly beans).

Evaluation. Measuring the performance of the output (compute percentage deviation of the output from the actual quantity of jelly beans).

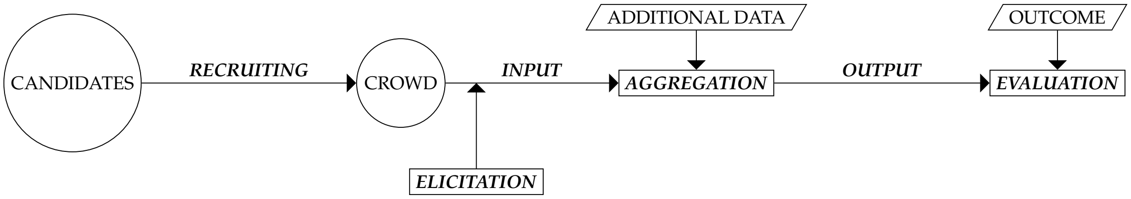

Figure 1 Forecasting process flowchart.

The most prominent application of wisdom of crowds to demand distribution forecasting was done at Sport Obermeyer as described by Fisher and Raman (1994). It can be characterized in terms of Lyon and Pacuit's six aspects as follows. The target *output* is a probability distribution

[1] A *form* of an input is the set of all possible values it might take (similar to *scale* in psychometrics).

forecast of the demand quantity. The *inputs* are subjective point forecasts of demand provided by the experts. The *aggregation* produces a Gaussian distribution with the mean equal to the average of experts' inputs and the standard deviation proportional to the standard deviation of experts' inputs. The seven-person buying committee was *recruited* to be the experts, and *elicitation* was conducted by asking each expert to independently provide their input for each product. The resulting distribution forecast was *evaluated* graphically, using predicted vs. actual and quantile-quantile plots. The implementation was successful, and the approach is standard advice in MBA classes for demand distribution forecasting for new products.

We collaborated with the three-person supply chain department of fashion retailer Moods of Norway to implement demand distribution forecasting for new products.[2] Our initial attempt aimed to collect quantity inputs in line with the Obermeyer approach, but failed due to recruiting issues. Only three experts outside the supply chain department ended up participating, bringing the total up to six, which made the supply chain department concerned about (i) the lack of diversity in the crowd, and (ii) their own ability to provide valuable market foresight (because they were among the employees most distant from actual product and market knowledge). Both diversity and expertise are crucial for crowds to be wise (Mannes et al. 2012). The lack of participation was in spite of our efforts to recruit broadly within the company.[3] Most of those invited simply refused to participate. The major complaint from those who declined to participate was that giving precise quantity estimates was too complicated and unfamiliar a task. A minor complaint from those who participated was that the process of giving inputs was time-consuming and inconvenient.

In response to the major complaint, we proposed an alternative form of input: the *ranking* of products by their expected future demand. Some of the experts who declined to provide quantity inputs stated that rankings would be much easier for them to provide. However, this new form of input requires (i) the products to be comparable to each other and (ii) the number of compared products to be sufficiently small. Consequently, we relied on the company's product categories, each of which consisted of products that are functional substitutes but have different subjective attributes (such as style). The experts we recruited were then asked to rank the products within each category. In the category of "female shirts," for example, an expert might rank the solid white shirt first, the striped green one second, and the dotted red shirt third.

In response to the minor complaint of a cumbersome user experience, the old process of entering quantities on their laptop placed on the table in the showroom was replaced by a mobile app that an expert could operate while assessing the product samples on the rack.

[2] As a part of a larger project on supply chain optimization.

[3] Due to confidentiality concerns, no experts external to the company were invited.

Following the introduction of ranking as input form, participation from outside the supply chain department increased sixfold, resulting in a diverse crowd from six different roles. The boost in participation suggests that the new form of input was successful. Moreover, several participants made comments about the process being fun and appreciated that they were asked for their opinion, suggesting that the new user experience contributed positively in the elicitation.

While the new form of input created excitement at Moods of Norway, we were short of a way to turn it into the target output. Therefore, we needed to develop a new methodology which would, among other things, address the following key challenge. Since the new form of subjective inputs (ranking) is on a different scale from the form of the target output (the probability distribution forecast of demand quantity) a conversion step between the two scales is required. The conversion is complicated by the fact that rankings do not contain enough information to produce the target output on their own. First, a ranking does not represent the relative difference between the ranks. Second, it is not in the same units as the output, namely, quantity.

To address this issue, we propose a tripartite decomposition of the *demand vector* for a category. First, we sum the entries of the demand vector to produce the first component, the *total demand*. Dividing the demand vector by the total demand yields the demand proportions vector. This vector can be further decomposed into the second component, the *ordered proportions vector* (obtained by sorting its entries in the descending order) and the third component, the *ranking* of products. These three components can also be recombined back into the demand vector: multiply the ordered proportion vector by the total demand and reorder according to the ranking.

This decomposition serves as a basis for the aggregation process, represented formally as a *forecasting rule*—a function that maps the input data to a distributional forecast. We construct the forecasting rule for our target output, the demand vector, in two steps. First, we establish the forecasting rules for each of the three components of the decomposition. A forecasting rule for the ranking component is based on subjective ranking inputs, while forecasting rules for the total and the ordered proportions are based on category-wise historical data. Second, we put the resulting component forecasts together via a *recombination rule*—a function that combines the distributional forecasts for ranking, total, and ordered proportions into a distributional forecast for the demand vector.

The quality of the *final output*, which is produced in the second step of the aggregation, is what we are ultimately concerned about. However, the *improvement* of the output will predominantly be driven by the choice of the forecasting rules for each component, employed in the first step of the aggregation. The central component in our method is the ranking component, which is also the only one that relies on subjective inputs. Metrics for evaluation of distribution forecasts over rankings appear to be understudied. Therefore, we propose several new such metrics, all of which are demonstrated to be proper scoring rules. These metrics are also normalized by the number of products in the category making them comparable across different studies.

We evaluate the performance of our method using field data. Because of our particular interest in ranking, we compare the out-of-sample performance of several alternative forecasting rules for that component, holding fixed a) the recombination rule and b) the forecasting rules for the total and the ordered proportions. We propose the following candidates: A natural one is the *empirical distribution* that assigns equal weight to each of the collected inputs. The *Plackett-Luce model* is a smoothed version of the empirical distribution which assigns non-zero probabilities to rankings not provided by the crowd. The *Borda rule* produces a single-point distribution assigning probability 1 to a single ranking obtained by combining subjective inputs.

For evaluation, we posit a worst-case (least-informed) *baseline* and a best-case (best-informed) *benchmark* ranking forecasting rules. As the baseline, we use a completely non-informative forecasting rule that assigns an equal probability to each possible ranking of products. As the benchmark, we consider a single-point distribution that assigns probability 1 to the actually observed ranking. It is reasonable to expect that the benchmark outperforms the baseline if the recombination rule and the component forecasts for total and ordered proportions are specified appropriately. For the three ranking forecasting rules introduced above, it is not obvious *a priori* what their relative performance would be or even to what extent they would outperform the least-informed baseline.

Since our interest in forecasting was motivated by its application to the newsvendor problem, we also use the optimal newsvendor profit induced by the forecast as an economic metric. To facilitate comparisons, we normalize the newsvendor profit by its theoretical upper bound: the *known-demand* profit which would result from ordering the ex-post demand quantity. This way, the profit is always bounded from above by 1.

We performed the comparison on the two seasons of historical data from Moods of Norway. Using the season[4] in which our forecasting methodology performed worse as an illustration, we obtain the following results. The profit of the baseline is equal to 0.59 of the theoretical upper bound. The ranking benchmark results in a profit of 0.65, implying that a better forecasting rule for the ranking can improve the newsvendor profit considerably. Out of three subjective input-based forecasting rules we consider, the empirical one is the best, giving a profit of 0.61; Plackett-Luce does not improve over the baseline, and the Borda rule performs worse than the baseline at 0.55. A casual interpretation of these results is that the Borda rule does not sufficiently account for the uncertainty of the aggregate forecast; Plackett-Luce does not adequately reflect the information contained in subjective inputs; and the empirical one, balancing the two, performs the best.

[4] For the other season, the relative performance of the ranking forecasting rules is similar.

2. Literature review

A survey of 100 firms by Kahn and Chase (2018) shows that new product forecasting relies more on subjective judgment than on statistical modeling. These authors provide few details about the elicitation and aggregation methods used in the papers surveyed. According to Lawrence et al. (2006), field studies on judgmental demand forecasting that are published in the operations management and marketing literature focus on quantitative inputs. The following papers are related to our application. Blattberg and Hoch (1990) find that an equal-weighted average of a linear regression forecast and the judgmental point forecast is more accurate than either of those alone. Seifert et al. (2015) study the information that should be given to new-product forecasters in the music industry; these authors find that forecasters should have access not to historical data but rather to contextual data, which should direct their attention toward aspects for which their judgment is most relevant. They also find that historical data should be used as input for a separate statistical model, whose forecasts are combined with judgmental forecasts. Bassamboo et al. (2015) propose prediction markets for demand distribution forecasting. They partition the set of all possible demand values into several ranges, represent each range by a virtual stock, and instruct participants to trade those stocks via a specified trading mechanism. The probabilities of each range are then calculated by normalizing the final prices of the corresponding stocks. Bassamboo et al. use probability integral transform plots to evaluate the resulting predictions and find that forecasts resulting from prediction markets are, from a probabilistic standpoint, well-calibrated. A practical application in the bicycle industry, which is based on recalibrating the forecasts with reference to past actual/forecast ratios, is described in the case study presented by Diermann and Huchzermeier (2017).

Judgmental demand forecasting has been investigated not only in field studies but also in controlled laboratory environments. For instance, Kremer et al. (2015) analyze hierarchical demand forecasting; they identify relevant behavioral biases and classify the conditions under which top-down is preferable to bottom-up forecasting. Another example is Lee and Siemsen (2016) who study the decomposition of the newsvendor ordering decision into point forecasting, range forecasting, and the service-level decision. These authors show experimentally that—because of the behavioral biases that might arise at different decision-making steps—the choice between directly setting the order quantity and breaking down the order quantity decision into the three parts should depend on the product's underlying demand uncertainty and profit margins. Scheele et al. (2017) use the data from a pharmaceutical company to demonstrate that sales departments might have incentives to inflate their demand forecasts. The authors build on this finding and construct a game-theoretic model to analyze the incentive misalignment, address it by introducing a compensation scheme, and validate their theoretical findings by way of a controlled laboratory experiment.

An important theoretical study is that of Gaba et al. (2019) who construct a model for combining point forecasts given by multiple experts into a distributional forecast. The combination proceeds by Bayesian updating of a prior distribution based on the generating model for expert inputs. Their model implies a positive relationship between the dispersion of expert inputs and forecast uncertainty. This result is also corroborated by the theoretical model of Tong and Feiler (2016), under which experts provide their input based on small mental samples from the demand distribution.

Ranking could be thought of as a form of qualitative (as opposed to quantitative) elicitation. On this subject, Windschitl and Wells (1996) compare verbal assessments of uncertainty to numerical ones. They suggest that verbal measures reflect psychological uncertainty better and are more predictive of preferences and behavioral intent. Similarly, Zimmer (1984) considers the choice of input form to the forecasting process. Zimmer compares two groups of bank clerks experimentally. Both groups were asked to forecast the USD/DM exchange rate; the first group was asked to provide verbal forecasts, which were then converted to numbers algorithmically, while the second group was asked to provide numerical ones. Zimmer states, “The comparison of the verbal predictions with the estimates of the numerical forecasting group revealed that the first group was more correct and more internally consistent.” Considering comparative vs. absolute judgment, Nunnally and Bernstein (1994) assert, in their classic *Psychometric Theory*, that “One of psychology’s truisms is that people are almost invariably better (more consistent and/or accurate) at making comparative responses than absolute responses.” Along the same lines is the study of Por and Budescu (2016), who show that eliciting joint probability distributions of binary events by asking for pairwise comparisons (“How much more likely is event A compared to event B?”) dominates direct elicitation (“What is the joint probability of events A and B?”) in terms of estimation accuracy. MacGregor (2001) suggests that it may be beneficial to decompose the inputs into separable components, giving the following example: rather than directly estimating the number of pieces of mail handled by the U.S. Postal Service per year, performance can be improved by recombining the estimates for the number of post offices for each state and the average number of pieces of mail per day handled by each post office. MacGregor (2001) stresses that decomposition is particularly helpful for problems with high uncertainty. Focusing on the fit of skill level to the input form, Önkal and Muradoglu (1996) find that for lower-skilled forecasters, asking for binary (up-or-down) inputs instead of more detailed interval predictions tends to improve their performance; for more experienced forecasters, the effect was the opposite.

Rankings as an input have been studied in a variety of contexts. Chevalier and Goolsbee (2003) address a problem similar to ours: determining how to convert a book’s sales rank (the only information listed by Amazon) into an estimate of its sales quantity. Their method is based on the observations that (i) a book's sales rank—when normalized by the total number of books—can be well fitted to a Pareto distribution and (ii) the parameters of this distribution can be estimated using best-seller data from the *Wall Street Journal*. This approach is not directly applicable to our case because the validity of the continuous Pareto approximation relies on a large number of books being sold; recall that we typically have only a few products in each category. Also, we focus not only on point forecasting but also on distributional forecasting, for which Chevalier and Goolsbee do not account. One of the few studies of the wisdom of crowds with ranking inputs is offered by Lee et al. (2014) who apply a Thurstonian cognitive model to aggregate rankings for several different recollection and prediction problems. However, these authors do not consider the specific case of demand forecasting; neither do they convert rankings into quantities or address distributional forecasting. Finally, consumer preference rankings have been used as an input to the multiproduct pricing problem by Rusmevichientong et al. (2006). Our paper differs in that it considers demand forecasts instead of consumer preference rankings; also, we focus not on pricing but rather on the forecasting of demand distribution.

An alternative to judgmental forecasting is forecasting based on objective product attributes. The operations management literature features several works in this vein. Ferreira et al. (2015) consider the problem of pricing new products for an online fashion retailer; they use regression trees with bagging to predict demand for a product based on its attributes such as style and color. Baardman et al. (2018) introduce a "cluster while estimate" model that simultaneously (i) clusters products based on their attributes and (ii) estimates regularized predictive models for each cluster. The authors then apply this model to two data sets, one from a consumer goods company and another from a fashion retailer. Hu et al. (2019) focus on forecasting product life-cycle curves by clustering the products, estimating a life-cycle curve for each cluster (while adjusting for seasonality and other effects), and assigning a new product to one of those clusters; the corresponding life-cycle curve then generates the forecast. We remark that all of the studies cited here focus on *point* forecasting, whereas we are interested in *distributional* forecasting. Note also that except for the Hu et al. paper, these studies require that objectively defined attributes be specified for all products.

3. Model

In this section we formalize the forecasting method described in the Introduction. The method is fairly elaborate, so we present an overview of the notation employed in Table 1. Fundamental to our forecasting method is the decomposition of a category's demand vector D into the following components: total demand A, which is the sum of the elements of the demand vector; ordered proportions B—the demand vector divided by total demand and then sorted in descending order;

and ranking C, whereby the highest-demand product is ranked 1, the second-highest is ranked 2, and so forth. The *input space* Ξ is linked to the *outcome space* Ω through the *forecasting rule* $\Xi \to \Delta(\Omega)$ where $\Delta(\Omega)$ is the set of all probability distributions on Ω. The subscripts A, B, C, and D (as applied to Ξ or Ω) refer to the subspace that is relevant to the decomposition's respective component.

Concept	Total demand	Ordered proportion	Ranking	Demand vector
Outcome space Ω				
Example of an element	100 units	[.5, .3, .2]	$D_1 > D_3 > D_2$	$[D_1 = 50, D_2 = 20, D_3 = 30]$
Algebraic representation	$\begin{pmatrix} 100 & 0 & 0 \\ 0 & 100 & 0 \\ 0 & 0 & 100 \end{pmatrix}$	(.5 .3 .2)	$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$	(50 20 30)
Notation for element	A	B	C	D
Definition of a set	$\Omega_A = \mathbb{R}_+$	$\Omega_B^m = \{B \mid B \in \mathbb{R}_+^m, \sum_i B_i = 1, (B_i - B_{i'})(i - i') \geq 0 \; \forall i, i' : 1 \leq i, i' \leq m\}$	$\Omega_C^m = \{C \mid C \in \mathbb{R}_+^{m \times m}, C_{i,i'} \in \{0,1\}, \sum_{i=1}^m C_{i,i'} = 1, \sum_{i'=1}^m C_{i,i'} = 1\}$	$\Omega_D^m = \mathbb{R}_+^m$
Input space Ξ				
Example of an element	k t A_{kt} m_{kt} 1 14 140 4 1 15 60 3 1 16 120 5 2 16 150 2	t B_t 14 [.4, .3, .2, .1] 15 [.6, .3, .1] 16 [.3, .25, .2, .15, .1]	j $r[j, \cdot]$ Ali $D_1 \succ D_2 \succ D_3$ Olga $D_3 \succ D_2 \succ D_1$ Mary $D_1 \succ D_3 \succ D_2$	Product of all elements to the left.
Notation for element	ξ_A	ξ_B	ξ_C	ξ_D
Definition of an element	$\{(A_{kt}, m_{kt}) \mid 1 \leq t < T, 1 \leq k \leq K\}$	$\{B_t \mid 1 \leq t < T\}$	$\{r[j, \cdot] \mid 1 \leq j \leq n\}$	Product of all sets to the left.
Forecasting rule $\Xi \to \Delta(\Omega)$				
Example	Estimate $\hat{\beta}$ and $\hat{\sigma}$ from a regression: $\log A_{kt} \sim N(\beta_k + \beta_m m_{kt}, \sigma)$ using data in ξ_A. Return a distribution: $\log A \sim N(\hat{\beta}_K + \hat{\beta}_m m, \hat{\sigma})$.	Estimate $\hat{\lambda}$ for the following model: $B_t \sim \text{OrderedDirichlet}(\lambda \cdot \mathbf{1}_{m_t})$ using data in ξ_B. Return a distribution: $B_T \sim \text{OrderedDirichlet}(\hat{\lambda} \cdot \mathbf{1}_m)$.	Empirical distribution over rankings in ξ_C.	Recombination: $\mathcal{D}(\xi_D) = \mathcal{C}(\xi_C) \times \mathcal{A}(\xi_A) \times \mathcal{B}(\xi_B)$ assuming that $\mathcal{A}(\cdot), \mathcal{B}(\cdot), \mathcal{C}(\cdot)$ are independent for any value of the inputs ξ_A, ξ_B and ξ_C.
Notation	$\mathcal{A}(\cdot)$	$\mathcal{B}(\cdot)$	$\mathcal{C}(\cdot)$	$\mathcal{D}(\cdot)$

Table 1 Summary of notation.

For time period T, we forecast demand for the m_k products in categories $k = 1, \ldots, K$; recall that all products within a given category have similar characteristics. To ease exposition we focus on category K, and omit the category and time-period subscripts whenever it is implied by the context. The ranking forecast is based on the subjective ranking inputs $r[j,i]$—from $j = 1, \ldots, n$ experts—that are given for $i = 1, \ldots, m$ products. Total demand and ordered proportions rely on historical data from periods $t = 1, \ldots, T-1$. Because there may not be many periods of relevant historical data, we supplement the data for category K by also using data from categories $1, \ldots, K-1$. Realizations for the m_{kt} products in category k in period t are denoted by D_{kt} for the demand vector, A_{kt} for total demand, B_{kt} for ordered proportions, and C_{kt} for ranking. Forecasting rules are likewise denoted by $\mathcal{D}(\cdot)$ for the demand vector, $\mathcal{A}(\cdot)$ for total demand, $\mathcal{B}(\cdot)$ for ordered proportions, and $\mathcal{C}(\cdot)$ for ranking.

3.1. Decomposition

Our main objective is to design a method by which the demand for new products can be forecast using rankings as (simpler) subjective inputs. As we have seen, these inputs need to be supplemented—with forecasts of total and ordered proportions—so they can be mapped to the demand space.

Forecasting total demand may initially seem like a task that is suitable for standard time-series models. Yet we have already discussed how a category's limited number of representative time periods, each of which could include a different number of products, may impose additional challenges. That said, we seek a pragmatic approach to this task because our objective is to find predictive patterns and not to isolate causal mechanisms. Such reasoning leads us to propose forecasting total demand via the log-linear regression model $A_{kt} = \exp(\beta_k) m_{kt}^{\gamma} \varepsilon_{kt}$. In this expression, β_k is the category-specific scaling factor; γ is the power exponent (common across categories), which captures the effect of a category's *number* of products; and the logarithms of the error terms ε_{kt} are assumed to be independent and identically distributed (i.i.d.) normal with zero mean.

In contrast to the forecasting of total demand, there are few methodologies for the probabilistic forecasting of ordered proportions. Moreover, none of these seem to apply directly to our setting. The varying number of products further complicates this task.

To design such a rule, we introduce a family of log-likelihoods $\mathcal{L}_{\tilde{m}}(\cdot \mid \theta^B)$ which would, for any fixed number of products $\tilde{m}$ and a fixed value of the parameter θ^B, assign a log-probability to any ordered proportions vector of length $\tilde{m}$. We proceed in two steps. First, we estimate θ^B by maximizing the log-likelihood of historical data $\sum_{t=1}^{T-1} \mathcal{L}_{m_t}(B_t \mid \theta^B)$ with respect to θ^B and denote the resulting estimate by $\hat{\theta}^B$. Second, we forecast B_T by a distribution with log-density $\mathcal{L}_{m_T}(\cdot \mid \hat{\theta}^B)$.

The only missing piece is the specification of the likelihood family $\mathcal{L}_{\tilde{m}}(\cdot \mid \theta^B)$. In this paragraph, we provide a general construction for such likelihood families. Assign utility u_i to each of $\tilde{m}$ products and map the resulting vector of utilities to the $\tilde{m}$-simplex of choice probabilities $\Delta([\tilde{m}])$; we write it as $p_{\tilde{m}}(\cdot, \ldots, \cdot) \colon \mathbb{R}^{\tilde{m}} \to \Delta([\tilde{m}])$. We call any family of such mappings indexed by $\tilde{m} \in \mathbb{N}$ a *choice model*. For example, the expression for choice probabilities of the *multinomial logit model* is given by

$$p_{\tilde{m}}(u_1, \ldots, u_{\tilde{m}}) = \left(\frac{\exp(u_1)}{\sum_i \exp(u_i)}, \ldots, \frac{\exp(u_{\tilde{m}})}{\sum_i \exp(u_i)} \right),$$

which means that each product's expected share is proportional to the exponent of its utility. To model the forecast uncertainty of ordered proportions, we consider product-level utilities drawn independently from a distribution of the *utility-generating process* $F_U(\cdot \mid \theta^B)$ parameterized by θ^B. The choice model maps the random vector of utilities to the random vector of unordered proportions, which in turn is subjected to the *sorting map* $\Delta([\tilde{m}]) \to \Omega_B^{\tilde{m}}$ and thereby yields the final random vector of ordered proportions, the log-density of which serves as the log-likelihood $\mathcal{L}_{\tilde{m}}(\cdot \mid \theta^B)$. For the multinomial logit model and the utility-generating process $F_U(\cdot \mid \theta^B)$ specified by the cumulative distribution function (CDF) of the logarithm of $\text{Gamma}(1, \theta^B)$ random variable with scalar θ^B, the result is the ordered $\text{Dirichlet}(\theta^B \cdot \mathbf{1}_{\tilde{m}})$ model. For this model, larger values of θ^B imply a smaller variance in the randomly drawn utilities; as a consequence, the expected ordered proportions approach a more evenly spread vector. The ordered Dirichlet model is especially convenient because the maximum likelihood estimate can be computed directly without the need for simulation.

Ranking is the last component of our decomposition, and it is the only one for which the forecast depends on subjective inputs. Following the common approach of judgmental forecasting literature (Armstrong 2001), we focus on simple forecasting rules that aggregate multiple judgments; these rules are compared in terms of empirical out-of-sample performance. One complication is that the outcome space differs from those typically studied—namely, binary events and continuous numbers—because it has a combinatorial structure. We must therefore consider forecasting rule candidates that are appropriate for this particular space.

The *empirical* forecasting rule assumes that the possible ranking outcomes are limited to the set of subjective rankings, where each ranking has an equal likelihood of $1/n$. For the example in Table 1, the rankings of Ali, Olga, and Mary are each assigned a probability of $1/3$. The *Borda* rule (Levin and Nalebuff 1995) results in the equivalent of a ranking point forecast; it assigns the probability 1 to the ranking that orders products by their Borda scores, where each score is calculated as the sum of a product's subjective ranks. So in the same example: product 1 receives a score of 5; product 2 a 7; and product 3 a 6. Thus we may write $D_1 \succ D_3 \succ D_2$; to avoid confusion with the ordering of the actual outcome values, we use $\succ$ to denote the subjective rankings. The *Plackett-Luce* rule (Plackett 1975, Luce 1959) works in the opposite direction: rather than combining all rankings into a single one, it expands the set of rankings with non-zero probability beyond the elicited subjective ones. This rule assigns a positive "strength parameter" θ_i^C to each product and represents the process by which the probability of each ranking is determined as follows: The top-ranked product is selected with probability proportional to its strength, and this procedure is repeated for the remaining products until they are depleted. At first glance, one might suppose that a maximum likelihood estimate for $\boldsymbol{\theta}^C$ would be appropriate. Yet this approach might be infeasible—as when, for example, a particular product is always ranked first—and so we add a regularization term that penalizes large values of $\boldsymbol{\theta}^C$. The estimates of this method approach the estimates of the maximum likelihood as the regularization penalty approaches zero.

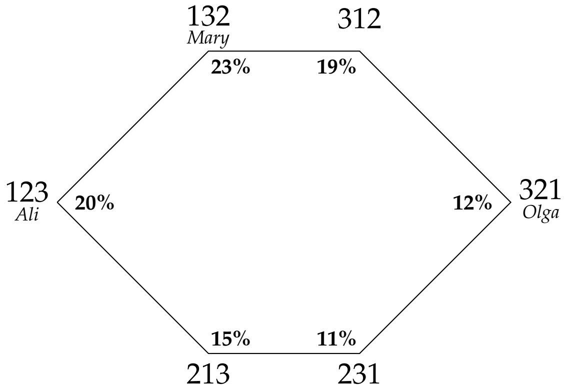

Figure 2 **Plackett-Luce probability distribution over rankings for the example under discussion; to simplify the presentation, we show only the rankings' product *subscripts* (so, for instance, "132" corresponds to the ranking $D_1 \succ D_3 \succ D_2$).**

In the context of our example, consider the Figure 2 hexagon that depicts all six possible rankings; here the number of arcs between two rankings is equal to the number of their pairwise disagreements, a reflection of their dissimilarity. The products' Plackett-Luce estimates—which are interpreted as their likelihood of being ranked first—are $\theta_1^C = 0.43$, $\theta_2^C = 0.26$, and $\theta_3^C = 0.31$. We observe that Mary's ranking $D_1 \succ D_3 \succ D_2$ has the highest assigned probability $\left(\frac{0.43}{0.43+0.26+0.31} \times \frac{0.31}{0.26+0.31} \times 100\% = 23\%\right)$ because it is only either one or two steps away from the other inputs. Because it is closer to Mary's ranking, the probability (20%) assigned to Ali's ranking $D_1 \succ D_2 \succ D_3$ is higher than that (12%) assigned to Olga's ranking $D_3 \succ D_2 \succ D_1$. For the rankings that were not in the set of inputs, $D_3 \succ D_1 \succ D_2$ is given the highest probability (19%) since it is next to both Mary and Olga and is only one step away from being next to Ali; the ranking $D_2 \succ D_1 \succ D_3$ is assigned 15% and $D_2 \succ D_3 \succ D_1$ is given 11%. It is interesting that, because of its distance from the other two inputs, Olga's rank is assigned only the fifth-highest probability.

Finally, as a worst-case benchmark we include the *random* forecasting rule; this rule assigns an equal probability $1/m!$ to each possible ranking. Here we generalize the notation, and state that the forecasting rule $\mathcal{C}(\cdot)$ maps a list of subjective inputs ξ_C to a probability distribution over $m!$ possible rankings. Those rankings are represented as $m \times m$ permutation matrices whose ith row and i'th column have value 1 if product i has rank i' (and have value 0 otherwise). Permutation matrices are used here because they minimize the notation needed during the recombination stage. An alternative representation, useful for evaluation purposes, is in the form of an m-vector whose ith component is equal to product i's rank. When we use this representation, the forecasting rule is written as $\tilde{\mathcal{C}}(\cdot)$ and the outcome is denoted $\tilde{C}$.

3.2. Recombination

Once each component of the decomposition is estimated, the final demand forecast is produced through recombination. This recombination is expressed conveniently by the formula

$$\mathcal{D} = \mathcal{C} \times \mathcal{A} \times \mathcal{B},$$

which highlights the modular nature of our approach: the quality of the forecast could be improved by selecting a better candidate for any of the components. Another advantage of this modularity is that it offers a clear framework on which future research can build.

Whereas recombining point forecasts is typically straightforward, recombining probabilistic forecasts is much more complicated. This difference mainly reflects the potential dependency among components—and the dependence is needed to complete our specification of the recombination rule. Toward that end, we first employ the most convenient choice: assuming that $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ are independent. In the next section, we test our demand forecasting method's performance while assuming such independence when recombining the three components into $\mathcal{D}$.

Given this problem's generality, the recombined demand forecast $\mathcal{D}$'s probability distribution function is unlikely to have an analytically tractable form. We therefore propose a bootstrap-based algorithm to approximate $\mathcal{D}$. Thus we run L iterations as follows: first draw a sample (A, B, C)—which includes total demand, the ordered proportion vector and ranking—from the distribution $(\mathcal{A}, \mathcal{B}, \mathcal{C})$; next, multiply the ordered proportion vector by the total demand and then reorder according to the ranking. The empirical distribution of the resulting L demand vectors is used as the final demand forecast. Note that this algorithm also accounts for the possibility of intercomponent dependence.

3.3. Evaluation methodology

The two primary objectives of the evaluation described here are (i) testing our method and (ii) providing companies with a way to measure their forecasting performance that will guide them in improving it. In essence, testing our method boils down to evaluating the final result—the demand distribution forecast $\mathcal{D}$. Better forecast accuracy will be driven by improving the forecasting rules for the components $\mathcal{A}$, $\mathcal{B}$, and/or $\mathcal{C}$—or the recombination rule. It is therefore important to evaluate the components not only in combination but also individually. With regard to forecasting rules, one complication is that each component has a different outcome space: whereas both A and D are real-valued vectors, B is an ordered proportion vector (that must sum to 1) and C is a discrete ranking. It follows that each outcome space will be best served by its own set of evaluation methods. For instance, evaluation methods that work well for total demand or ordered proportions are likely to be unsuitable for rankings. The performance criteria for a recombination rule should reflect that rule's ability to use the inputs for each component appropriately.

Action space	Metric name	Action function	Metric definition	Comment	Best value
Outcome space: Quantity $(\Omega = \mathbb{R}_+)$					
Point forecasts/decisions $(S = \mathbb{R}_+)$	Newsvendor profit	$Q_i = \arg\max_{q \in \mathbb{R}} \mathbb{E}_{\mathcal{D}_i}(r_i \min(\mathcal{D}_i, q) - w_i q))$	$\sum_{i=1}^{m} r_i \min(D_i, Q_i) - w_i Q_i$	Economic profit; for product i, r_i is revenue and w_i is purchase cost.	$\sum_i (r_i - w_i) D_i$
	Root-mean-squared error (RMSE)	$\hat{D}_i = \arg\min_{d \in \mathbb{R}} \mathbb{E}_{\mathcal{D}_i}(\mathcal{D}_i - d)^2$	$\sqrt{\frac{1}{m} \sum_{i=1}^{m} (D_i - \hat{D}_i)^2}$	Most commonly used statistical metric.	0
	Mean absolute percentage error (MAPE)	$\hat{D}_i = \arg\min_{d \in \mathbb{R}} \mathbb{E}_{\mathcal{D}_i} \left\lvert \frac{\mathcal{D}_i - d}{\mathcal{D}_i} \right\rvert$	$\frac{1}{m} \sum_{i=1}^{m} \left\lvert \frac{D_i - \hat{D}_i}{D_i} \right\rvert$	Comparable across studies, commonly used for demand.	0
Interval forecasts $(S = \mathbb{R}_+ \times \mathbb{R}_+)$	Coverage	$[\underaccent{\bar}{D}_i, \bar{D}_i] = \left[F_{\mathcal{D}_i}^{-1}\left(\frac{\alpha}{2}\right), F_{\mathcal{D}_i}^{-1}\left(1 - \frac{\alpha}{2}\right)\right]$	$\frac{1}{m} \sum_{i=1}^{m} \mathbb{I}(\underaccent{\bar}{D}_i < D_i < \bar{D}_i)$	When averaged over multiple categories, should be equal to α: nominal coverage.	α
Full distribution forecasts $S = \Delta(\mathbb{R}_+)$	Continuous ranked probability score (CRPS)	Identity action function	$\frac{1}{m} \sum_{i=1}^{m} \int_0^{+\infty} (\mathbb{P}(\mathcal{D}_i > t) - \mathbb{I}(D_i > t))^2 \, dt$	Proper scoring rule for distributions.	0
Outcome space: Ranking $(\Omega = \Omega_C^m)$					
Point forecasts	Spearman's ρ correlation	Borda aggregation: $\hat{\tilde{C}}_i$-rank of product i when ordered by $\mathbb{E}\tilde{\mathcal{C}}_i$	$1 - \frac{6}{m^3 - m} \sum_{i=1}^{m} (\tilde{C}_i - \hat{\tilde{C}}_i)^2$	Correlation between the ranks.	1
Distribution forecasts	Spearman–Brier score	Identity action function	$\frac{1}{2}\left(1 - \mathbb{E}\rho(\tilde{\mathcal{C}}, \tilde{C}) + \frac{1}{2}\mathbb{E}\rho(\tilde{\mathcal{C}}, \tilde{\mathcal{C}}')\right)$	Proper scoring rule for rankings; $\rho(\cdot, \cdot)$ is the Spearman's ρ correlation.	0

Table 2 Summary of metrics.

The numerical metrics that we consider can be broadly divided into those that are *statistical* and those that are *economic*. In our context, the relevant economic metric is the realized newsvendor profit resulting from the optimal order quantity prescribed by our demand distribution forecast. This economic metric is useful because it assigns weights according to economic importance; yet it is also limited by being applicable only to D, which means that statistical metrics are essential for evaluating the other components.

Formally, our sequence of evaluation first maps the forecast to an *action space* S via an *action function* $\Delta(\Omega) \to S$. Once the outcome is observed, a *metric* $\Omega \times S \to \mathbb{R}$ maps the observed outcome and the action from S to a number. Table 2 lists the different outcome spaces, action spaces, action functions, and metrics used.

When the distributions are evaluated as a whole, which corresponds to the case $S = \Delta(\Omega)$, it is desirable to use *proper scoring rules*—defined as metrics that incentivize experts to report their inputs honestly. The proper scoring rules are not only well-suited for *ex ante* elicitation, but also excel at *ex post* evaluation (see Winkler and Jose 2011). In their construction of proper scoring rules for general spaces, Gneiting and Raftery (2007) propose an approach based on the notion of a *negatively semidefinite kernel* that represents dissimilarity between outcomes. Formally, a kernel is a function which satisfies the following definition (note that hereafter we omit "negatively semidefinite").

			Score		
Outcome space	**Kernel**	**Scoring rule**	**Best**	**Worst**	**Uniform**
Boolean	$d(x,y) = \frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2$	Mean Brier score	0	1	0.25
Quantity	$d(x,y) = \frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2$	Mean squared error	0	$+\infty$	
	$d(x,y) = \frac{1}{m}\sum_{i=1}^{m}\lvert x_i - y_i\rvert$	Mean CRPS	0	$+\infty$	
Ranking	$d(x,y) = \frac{6}{m^3-m}\sum_{i=1}^{m}(x_i - y_i)^2$	Spearman–Brier score	0	1	0.25
	$d(x,y) = \frac{2}{m(m-1)}\sum_{1\le i<i'\le m}(\mathbb{I}(x_i > x_{i'}) - \mathbb{I}(y_i > y_{i'}))^2$	Kendall–Brier score	0	1	0.25
	$d(x,y) = \frac{1}{2}\sum_{1\le i\le m}(\mathbb{I}(x_i = l) - \mathbb{I}(y_i = l))^2$	Top-*l* Brier score	0	1	$\frac{m-1}{2m}$

Table 3 Kernel scoring rules (rankings are represented in vector form).

DEFINITION 1. A *kernel* is a function $d(\cdot,\cdot)$ that maps $\Omega \times \Omega$ to $\mathbb{R}$ and that has the following properties.

(i) *Symmetry.* For any x_1, x_2 in Ω, the equality $d(x_1,x_2) = d(x_2,x_1)$ holds.

(ii) *Negative semidefiniteness.* For any $x_i \in \Omega$ and any $\lambda_i \in \mathbb{R}$ such that $\sum_i \lambda_i = 0$, we have that $\sum d(x_i,x_j)\lambda_i\lambda_j \le 0$.

Now if $d(\cdot,\cdot)$ is a kernel, then there always exists a corresponding *kernel proper scoring rule* given by $\mathbb{E}d(\mathcal{Y},Y) - \frac{1}{2}\mathbb{E}d(\mathcal{Y},\mathcal{Y}')$ where $\mathcal{Y}$, $\mathcal{Y}'$ are two independent copies of the forecast and where Y is the outcome. There are widely used kernel proper scoring rules for quantitative outcome spaces, but we are not aware of any for distributions over rankings. Hence we define three such rules—which we call the Spearman–Brier, Kendall–Brier, and top-*l* Brier scores (where *l* refers to a position in the ranking)—by using different kernels defined for the ranking space. Table 3 describes the extant kernel scoring rules for a quantitative outcome space as well as our new scoring rules for a ranking outcome space. This table reports the best and the worst possible values for each scoring rule, and, where appropriate, each rule's value for a distribution that is uniform over the outcome space. Whereas the Spearman–Brier and Kendall–Brier scores evaluate the ranking forecast's overall accuracy, top-*l* Brier scores (namely, top-1 and top-*m*) focus on a particular position. The following proposition establishes the essential properties of our scoring rules for rankings.

PROPOSITION 1. *With regard to the kernel scoring rules for rankings as defined in Table 3, the following statements hold.*

(a) *All functions $d(\cdot,\cdot)$ listed in the second column of Table 3 are kernels.*

(b) *The best-case, worst-case, and uniform distribution-case values of the scoring rules are given by the values listed in (respectively) the three last columns of Table 3.*

(c) *For the simulated ranking forecast with L samples, all kernel scores listed in Table 3 can be computed in $O(L)$ iterations.*

Proof. See Appendix.

We supplement these numerical metrics with two graph-based evaluation methods. The first such method is a *predicted versus actual* plot, which is a scatter plot of the observed outcomes against the medians of the forecast distributions (see Figure 3 in the next section). For a well-calibrated forecasting process, those points would concentrate near the 45° line; about half of the points should be above the line and the rest below. The second graphical method is the *probability integral transform* plot (Diebold et al. 1998), which is a histogram of actual outcomes plugged into the corresponding forecast's CDF (see Figure 4). For a well-calibrated forecast, the resulting histogram should appear to be uniform in shape.

4. Empirical study

In this section, we illustrate our proposed methodology using two seasons of data from Moods of Norway: Spring-Summer 2015 (SS15) and Autumn-Winter 2015 (AW15). Table 4 summarizes the data. Not all products were ranked, and not all products that were ranked were eventually launched. After we removed the unlaunched products, the ranking inputs were then re-ranked. Because few stockouts were observed in the data, uncensoring was not required; hence we use the raw sales quantities to represent demand.

Variable	**SS15**	**AW15**
Number of products sold	467	507
Number of products ranked	311	347
Number of new products ranked and launched	235	253
Number of categories	37	52
Historical seasons used	SS13, SS14	AW13, AW14
Number of historical seasons	2	2
Number of historical products	1,116	1,112
Number of experts	21	21
Number of inputs for launched products	1,936	2,460

Table 4 Summary statistics.

Inputs were elicited from employees of diverse ages, genders, and educational backgrounds. The positions these individuals held in the company are summarized in Table 5. As mentioned in the Introduction, when we initially asked for quantity inputs, only three experts outside the supply chain department agreed to participate. Table 5 shows that after a switch to rankings, participation outside the supply chain department increased sixfold.

Role	**SS15**	**AW15**
Design/sourcing	9	9
Marketing	4	3
Merchandising	3	3
Retail management	1	2
Store manager	1	1
Supply chain	3	3
Total	21	21

Table 5 Experts by professional role.

We use the metrics defined in Table 2 to evaluate each of the three component forecasts' out-of-sample-performance. For this purpose, we ran simulations with $L = 1{,}000$ samples. Table 6 reports the metrics for total demand. The log-linear model performs well in terms of coverage and leads to reasonably small values of mean absolute percentage error (MAPE); it appears to be slightly overdispersed.

Season	50% cov.	95% cov.	MAPE
SS15	0.68	1.00	54%
AW15	0.62	0.97	89%

Table 6 Total demand performance.

Table 7 evaluates the ordered proportions forecasts from the ordered Dirichlet model. The values reported in the MAPE column show that the forecast's expected values are close to the realized outcomes. In addition, the nearness of each interval coverage (cov.) to the respective nominal coverage values suggests that this model captures the forecast uncertainty well.

Season	50% cov.	95% cov.	MAPE
SS15	0.39	0.88	40%
AW15	0.51	0.91	25%

Table 7 Ordered proportion performance.

Table 8 presents the accuracy of the forecasting rules for ranking. As the best-case benchmark, we consider an additional forecasting rule that has perfect foresight and forecasts the distribution that assigns probability 1 to the realized ranking. As the worst-case baseline, we use a non-informative forecasting rule which assigns equal probability to each possible ranking. The metrics are interpreted as follows: for Spearman's ρ, the closer it is to 1, the better; for the other three metrics, the closer each is to 0, the better. We do not report the Kendall–Brier score because the results are similar to those for the Spearman–Brier score. The empirical forecasting rule performs the best overall, followed by the Plackett-Luce rule. The Borda rule, despite its high Spearman's ρ metric (which reflects the quality of a rule's point forecasts), underperforms in terms of the metrics which account for forecast uncertainty: Spearman-Brier score, top-1 and top-m Brier scores.

Season	Forecasting rule for ranking	Spearman's ρ	Top-m Brier score	Spearman–Brier score	Top-1 Brier score
SS15	Baseline	0.01	0.38	0.25	0.38
	Plackett–Luce	0.27	0.37	0.22	0.35
	Borda	0.50	0.51	0.25	0.49
	Empirical	0.50	0.33	0.17	0.34
	Benchmark	1.00	0.00	0.00	0.00
AW15	Baseline	–0.02	0.36	0.25	0.36
	Plackett–Luce	0.33	0.34	0.22	0.34
	Borda	0.40	0.46	0.30	0.57
	Empirical	0.37	0.34	0.20	0.32
	Benchmark	1.00	0.00	0.00	0.00

Table 8 Ranking performance.

Next, Table 9 evaluates the recombined demand forecasts for each of the three alternative forecasting rules as well as the baseline and the benchmark. The sum of newsvendor profits for all products ("NV profit") is normalized by the total known-demand profit—that is, by $\sum_{k=1}^{K} \sum_i (r_i - w_i) D_i$. This metric achieves its maximum value of 100% if, for each product, the company predicted the demand perfectly and set its order quantity equal to that demand. In terms of NV profit, the best performance results when we use the empirical forecasting rule for the ranking. This rule also tends to perform well on the MAPE and continuous ranked probability score (CRPS) metrics. We observe that all of the ranking forecasting rules consistently overshoot the nominal coverage (defined in Table 2 on page 14). Possible drivers of this phenomenon are (i) that the forecast uncertainty is overestimated for the individual components and/or (ii) our assumption of independence among components in the recombination rule.

Season	Forecasting rule	50% cov.	95% cov.	CRPS	MAPE	NV profit
SS15	Baseline	0.73	0.98	67.80	143%	70%
	Plackett–Luce	0.71	0.98	65.52	129%	71%
	Borda	0.55	0.97	66.69	109%	70%
	Empirical	0.69	0.99	60.89	107%	73%
	Benchmark	0.65	0.99	53.38	66%	77%
AW15	Baseline	0.62	0.99	61.88	103%	59%
	Plackett–Luce	0.63	0.98	61.13	98%	59%
	Borda	0.54	0.92	65.15	92%	55%
	Empirical	0.65	0.99	58.32	82%	61%
	Benchmark	0.62	0.97	51.50	64%	65%

Table 9 Recombined demand forecast performance.

To shed additional light on the performance of the component recombination, Table 10 presents various combinations of replacing the estimation of a component with its realized outcome. It suggests that a better forecast of total demand is the best way to improve newsvendor profit. The improvement resulting from using the realized ordered proportion vector over its forecast appears to be negligible, while a perfect knowledge of ranking could improve the profit by as much as 6 percentage points. The empirical forecasting rule yields an improvement of 2–3 percentage points.

	Total	Ordered		Plackett–			
Season	demand	proportion	Baseline	Luce	Empirical	Borda	Benchmark
SS15	Known	Known	83%	85%	89%	82%	100%
		Estimated	80%	82%	86%	80%	92%
	Estimated	Known	70%	72%	74%	74%	79%
		Estimated	71%	71%	73%	70%	77%
AW15	Known	Known	79%	80%	86%	79%	100%
		Estimated	78%	78%	82%	76%	91%
	Estimated	Known	59%	60%	62%	60%	67%
		Estimated	59%	59%	61%	55%	65%

Table 10 Sensitivity analysis of newsvendor profits: Effect of each component's accuracy.

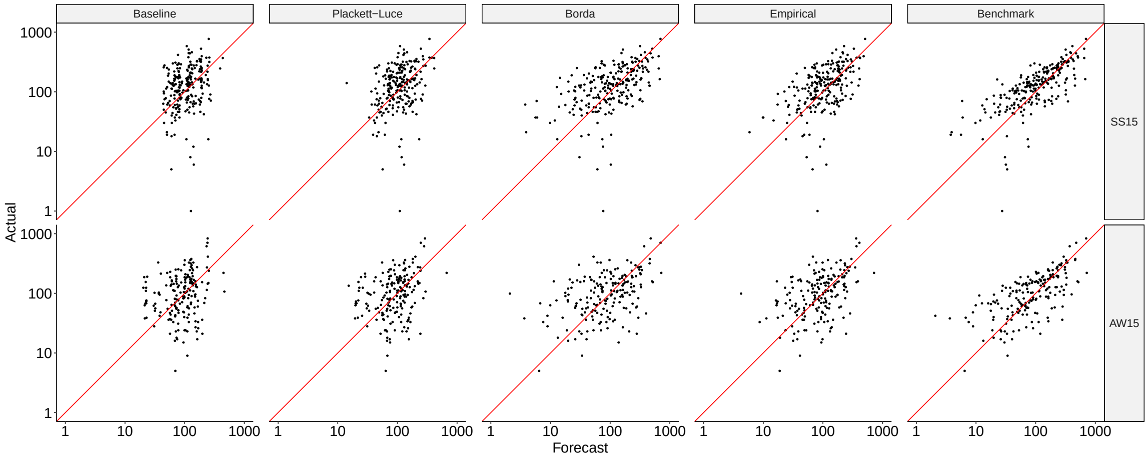

Figure 3 **Predicted versus actual demand (plotted on a log-log scale).**

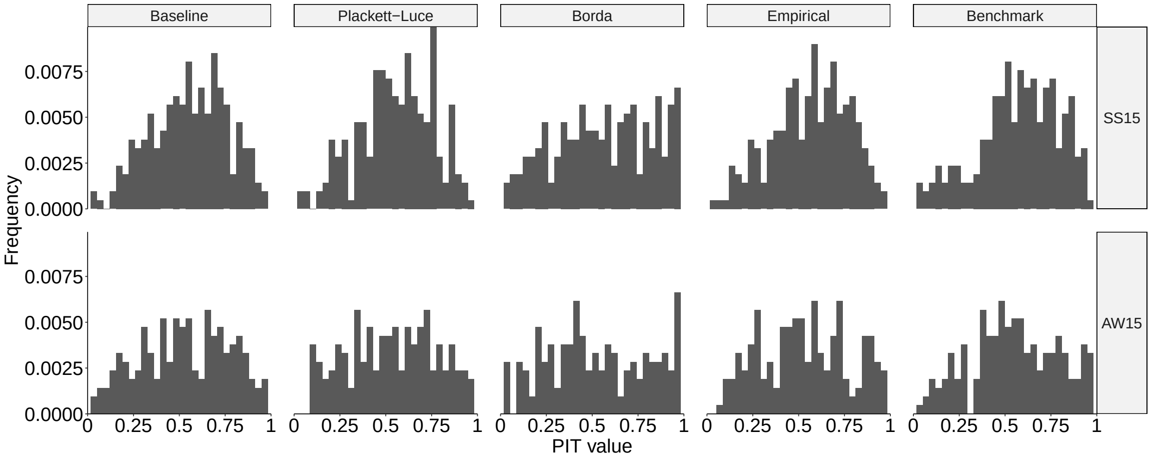

Figure 4 **Probability integral transform plots for demand distribution forecasts across seasons.**

Figure 3 presents a grid of “predicted versus actual” plots (as described in Section 3.3), with different forecasting rules on the horizontal axis and different seasons on the vertical one. We can see that the forecasts are well-calibrated in that the points are almost evenly split by the 45° line. Figure 4 consists of probability integral transform plots (also described in Section 3.3), where the rules and seasons are arranged just as in Figure 3. Most of the histograms are reasonably close to the uniform (with a slight tendency for overdispersion), which indicates that also the recombined forecast is well-calibrated probabilistically.

5. Discussion and concluding remarks

To highlight how our contribution relates to the field of demand distribution forecasting, it can be helpful to consider the following four scenarios. The go-to method for demand distribution forecasting is time-series analysis, but it requires that historical data are available for the same product. If the product is new, but has objectively defined attributes shared with products that were sold in the past, and if these attributes are linked to demand in predictable ways (which is often the case for utilitarian goods), a "like-to-like" approach can be applied instead. If the attributes are subjective and/or not predictably linked to demand (which is commonly the case with products for which the demand is driven by fashion), but it is possible to recruit experts who are able to provide quantity inputs, the Obermeyer approach is feasible. If it is not possible to recruit a large and diverse crowd which is capable of providing quantity inputs, we suggest to collect ranking inputs instead.

The first three of these scenarios are addressed by the existing literature on forecasting. We introduce the fourth scenario and address it by demonstrating that: (i) asking experts for rankings can improve participation and diversity; (ii) well-calibrated demand distribution forecasts can be produced from ranking inputs combined with historical data for total demand and ordered proportions in the same category; and (iii) experts who are asked to provide ranking inputs perform better than a baseline forecast which assigns every possible ranking the same probability.

The proposed methodology can enable companies that are facing our fourth scenario to produce demand distribution forecasts, helping them to adopt stochastic inventory optimization algorithms that often require a demand distribution as an input.

Our paper opens up multiple avenues for future research. Collecting more data, preferably from a varied set of companies, would allow testing the robustness of the methods at scale. The forecasting method itself can be further developed, particularly in the aspects of aggregation, input forms, and elicitation.

The current version of the framework recombines distribution forecasts for the three components (total demand, ordered proportions, and ranking) under the assumption that the components are independent. The modular structure of the framework invites further studies of the forecasting rules for each component as well as of the recombination rule. A natural development of the recombination rule would account for potential dependency across the components. For the forecasting of the ranking component, inputs could be weighted non-equally. The most obvious approach would be to give higher weight to experts who were more accurate in the past. A more sophisticated approach would also adjust the weights to reflect an expert's contribution to the diversity of inputs.

Other forms of inputs, such as star ratings or adaptive pairwise comparisons, can serve as an alternative to ranking inputs. Further studies could assess their relative strengths and weaknesses. Finally, it would be interesting to study further the processes and methods of elicitation. One direction is to consider the value of information, for example, comparing the quality of inputs between experts who have access to physical product samples versus those who only have access to photos. Another is to develop, analyze, and implement incentive schemes suitable for rankings and other comparison-based subjective inputs.

Acknowledgments

The authors are grateful to Moods of Norway—in particular, to Jørn Peder Sætre and Hanna Rytter Hasle—for helpful support and to Hans Petter Hübert for his assistance at various stages of the project. The authors also thank Stephen Chick, Santiago Kraiselburd, Jürgen Mihm, and Ville Satopää for many useful comments and suggestions.

Appendix

Proof of Proposition 1. In this proof, we use the subscripts 1, 2, and 3 with reference to (respectively) the Spearman–Brier score, the Kendall–Brier score, and the top-l Brier score. Also, we use $J(\mathcal{C}, C)$ to denote the kernel scoring rule $\mathbb{E}d(\mathcal{C}, C) - \frac{1}{2}\mathbb{E}d(\mathcal{C}, \mathcal{C}')$. Observe that the functions $d(\cdot, \cdot)$ defined in Table 3 can all be represented in the form

$$d(x,y) = (\Phi(x) - \Phi(y))^T(\Phi(x) - \Phi(y)), \tag{1}$$

where $\Phi(x)$ is a function that maps a ranking vector to a real vector space. For the Spearman–Brier score, $\Phi_1(x)$ normalizes the ranking vector x by $\sqrt{(m^3 - m)/3}$. For the Kendall–Brier score, $\Phi_2(x)$ maps x to a $((m^2 - m)/2)$-vector whose indices correspond to ordered pairs (i, i') of products such that the (i, i')th component of vector $\Phi_2(x)$ is equal to $\sqrt{2/(m^2 - m)}$ if $i > i'$ in x and is equal to 0 otherwise. For the top-l Brier score, $\Phi_3(x)$ maps the ranking x to an m-vector whose ith element is equal to $1/\sqrt{2}$ if product i is ranked lth (and to 0 otherwise). We now prove each part of the proposition in turn.

(a) We must show that Definition 1 applies to the functions $d(\cdot, \cdot)$ that satisfy equation (1). Symmetry follows immediately from the equation. To establish negative semidefiniteness, choose any sequence c_π of real numbers indexed by rankings π such that $\sum_\pi c_\pi = 0$. Then, rewriting the quadratic form given in the definition as $\sum_{\pi,\pi'}(\Phi(\pi) - \Phi(\pi'))^{\mathrm{T}}(\Phi(\pi) - \Phi(\pi'))c_\pi c_{\pi'}$, we obtain $-2 \times \| \sum_\pi c_\pi \Phi(\pi) \| + \sum_{\pi,\pi'}(\|\Phi(\pi)\| + \|\Phi(\pi')\|)c_\pi c_{\pi'}$. The first term is negative and the second term is zero; hence $d(\cdot, \cdot)$ is a kernel.

(b) Because $J(\cdot, \cdot)$ is a strictly proper scoring rule, evaluating this rule at a single-point distribution which assigns probability 1 to the true outcome C yields (i) $J(C, C) \leq J(\mathcal{C}, C)$ for any distribution $\mathcal{C}$ but (ii) $J(C, C) = 0$ for all the kernels listed in Table 3. Therefore, the minimum possible value of $J(\cdot, \cdot)$ is 0.

Now we can use straightforward combinatorial calculation to show that, for all three metrics, the maximum possible value attained by $d(\cdot, \cdot)$ is 1. This value is also the highest possible for the score function because, in $\mathbb{E}d(\mathcal{C}, C) - \frac{1}{2}\mathbb{E}d(\mathcal{C}, \mathcal{C}')$, the first term cannot exceed unity and the second term is nonnegative.

Finally, we consider the case of a uniform distribution $\mathcal{C}_U$ over all rankings. We start by showing that for none of the three metrics does $\mathbb{E}d(\mathcal{C}_U, C)$ depend on the actual outcome C. Define the ranking *opposite* to x, or $-x$, as the one that ranks product i as $m + 1 - x_i$. For the Spearman–Brier kernel, we can easily demonstrate that $d_1(x, C) + d_1(-x, C) = 1$ for any rankings x and C. Since x is never equal to $-x$ and since $-(-x) = x$, it follows that the set of all rankings can be partitioned into $m!/2$ pairs of opposite rankings. Averaging over all such pairs now yields the equality $\mathbb{E}d_1(\mathcal{C}, C) = 0.5$. The exact same reasoning applies to the Kendall–Brier score. For the top-l Brier score, $1/m$ is the probability that a uniformly drawn ranking has the same top-l product as the actual outcome. If so, then the kernel value is 0; otherwise, it is 1. As a result, $\mathbb{E}d_3(\mathcal{C}_U, C) = (m - 1)/m$ irrespective of C.

So if $\mathbb{E}d(\mathcal{C}_U, C)$ does not depend on C, then we can use iterated conditioning to deduce that $\mathbb{E}d(\mathcal{C}_U, C) = \mathbb{E}d(\mathcal{C}_U, \mathcal{C}_U')$. This equality implies that $J(\mathcal{C}_U, C) = \frac{1}{2}\mathbb{E}d(\mathcal{C}_U, C)$ for all the scoring rules we consider; hence the Spearman–Brier and Kendall–Brier scores for the uniform distribution are equal to 0.25 and the top-l Brier score is equal to $(m-1)/2m$.

(c) First we prove the following auxiliary result. For an arbitrary embedding function $\Phi(\cdot)$, if $\mathcal{C}$ and $\mathcal{C}'$ are arbitrary i.i.d. random variables over the rankings then

$$\mathbb{E}(\Phi(\mathcal{C}) - \Phi(\mathcal{C}'))^{\mathrm{T}}(\Phi(\mathcal{C}) - \Phi(\mathcal{C}')) = 2\mathbb{E}\|\Phi(\mathcal{C}) - \bar{\Phi}\|^2; \tag{2}$$

here $\bar{\Phi} = \mathbb{E}\Phi(\mathcal{C})$. Let $\tilde{\Phi}(x) = \Phi(x) - \bar{\Phi}$, and note that $\mathbb{E}\tilde{\Phi}(\mathcal{C})$ is a zero vector. We can therefore derive the equality

$$\mathbb{E}(\Phi(\mathcal{C}) - \Phi(\mathcal{C}'))^{\mathrm{T}}(\Phi(\mathcal{C}) - \Phi(\mathcal{C}')) = \mathbb{E}(\tilde{\Phi}(\mathcal{C}) - \tilde{\Phi}(\mathcal{C}'))^{\mathrm{T}}(\tilde{\Phi}(\mathcal{C}) - \tilde{\Phi}(\mathcal{C}')).$$

After simplifying the expression, we may write

$$\mathbb{E}\big(\|\tilde{\Phi}(\mathcal{C})\|^2 + \|\tilde{\Phi}(\mathcal{C}')\|^2\big) - 2\mathbb{E}\tilde{\Phi}(\mathcal{C}')^{\mathrm{T}}\tilde{\Phi}(\mathcal{C}).$$

Here the first term is equal to $2\mathbb{E}\|\Phi(\mathcal{C}) - \bar{\Phi}\|^2$, and iterated conditioning reveals that the second term is zero.

Now our desired result follows immediately from the representation in equation (2). We can calculate $\mathbb{E}d(\mathcal{C}, C)$ in L evaluations of the kernel function, after which the average vector embedding $\bar{\Phi}$ can be calculated in L iterations. Next we compute, again in L iterations, the value of the dispersion term $\mathbb{E}d(\mathcal{C}', \mathcal{C}) = 2\mathbb{E}\|\Phi(\mathcal{C}) - \bar{\Phi}\|^2$. The final result is $\mathbb{E}d(\mathcal{C}, C) - \frac{1}{2}\mathbb{E}d(\mathcal{C}', \mathcal{C})$. □

References

Armstrong JS (2001) Combining forecasts. Armstrong JS, ed., *Principles of Forecasting: A Handbook for Researchers and Practitioners*, 417–439 (Boston, MA: Springer US).

Baardman L, Levin I, Perakis G, Singhvi D (2018) Leveraging comparables for new product sales forecasting. *Production and Operations Management* 27(12):2340–2343.

Bassamboo A, Cui R, Moreno A (2015) The wisdom of crowds in operations: Forecasting using prediction markets. Mimeo, Available at SSRN: http://ssrn.com/abstract=2679663.

Blattberg RC, Hoch SJ (1990) Database models and managerial intuition: 50% model + 50% manager. *Management Sci.* 36(8):887–899.

Chevalier J, Goolsbee A (2003) Measuring prices and price competition online: Amazon.com and BarnesandNoble.com. *Quantitative Marketing and Economics* 1(2):203–222.

Diebold F, Gunther TA, Tay AS (1998) Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39(4):863–83.

Diermann C, Huchzermeier A (2017) Case—canyon bicycles: Judgmental demand forecasting in direct sales. *INFORMS Transactions on Education* 17(2):63–74.

Ferreira KJ, Lee BHA, Simchi-Levi D (2015) Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing Service Oper. Management* 18(1):69–88.

Gaba A, Popescu DG, Chen Z (2019) Assessing uncertainty from point forecasts. *Management Sci.* 65(1):90–106.

Galton F (1907) Vox populi (the wisdom of crowds). *Nature* 75(7):450–451.

Gneiting T, Raftery A (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 102(477):359–378.

Hu K, Acimovic J, Erize F, Thomas DJ, Van Mieghem JA (2019) Forecasting new product life cycle curves: Practical approach and empirical analysis. *Manufacturing Service Oper. Management* 21(1):66–85.

Kahn KB, Chase CW (2018) The state of new-product forecasting. *Foresight: The International Journal of Applied Forecasting* 51.

Kremer M, Siemsen E, Thomas DJ (2015) The sum and its parts: Judgmental hierarchical forecasting. *Management Sci.* 62(9):2745–2764.

Lawrence M, Goodwin P, O'Connor M, Önkal D (2006) Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting* 22(3):493–518.

Lee MD, Steyvers M, Miller B (2014) A cognitive model for aggregating people's rankings. *PLOS ONE* 9(5):1–9.

Lee YS, Siemsen E (2016) Task decomposition and newsvendor decision making. *Management Sci.* 63(10):3226–3245.

Levin J, Nalebuff B (1995) An introduction to vote-counting schemes. *Journal of Economic Perspectives* 9(1):3–26.

Luce RD (1959) *Individual Choice Behavior: A Theoretical analysis* (New York, NY, USA: John Wiley & Sons).

Lyon A, Pacuit E (2013) The wisdom of crowds: Methods of human judgement aggregation. Michelucci P, ed., *Handbook of Human Computation*, 599–614 (New York, NY: Springer New York).

MacGregor DG (2001) Decomposition for judgmental forecasting and estimation. Armstrong JS, ed., *Principles of Forecasting: A Handbook for Researchers and Practitioners*, 107–123 (Boston, MA: Springer US).

Mannes AE, Larrick RP, Soll JB (2012) The social psychology of the wisdom of crowds. Kruger JI, ed., *Frontiers of Social Psychology: Social Psychology and Decision Making* (Philadelphia, PA: Psychology Press).

Nunnally JC, Bernstein IH (1994) *Psychometric theory* (New York, NY: McGraw-Hill).

Önkal D, Muradoglu G (1996) Effects of task format on probabilistic forecasting of stock prices. *International Journal of Forecasting* 12(1):9–24.

Plackett RL (1975) The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 24(2):193–202, URL `http://www.jstor.org/stable/2346567`.

Por HH, Budescu DV (2016) Eliciting subjective probabilities through pair-wise comparisons. *Journal of Behavioral Decision Making* 30(2):181–196.

Rusmevichientong P, Van Roy B, Glynn PW (2006) A nonparametric approach to multiproduct pricing. *Oper. Res.* 54(1):82–98.

Scheele LM, Thonemann UW, Slikker M (2017) Designing incentive systems for truthful forecast information sharing within a firm. *Management Sci.* 64(8):3690–3713.

Seifert M, Siemsen E, Hadida AL, Eisingerich AB (2015) Effective judgmental forecasting in the context of fashion products. *Journal of Operations Management* 36:33–45.

Surowiecki J (2004) *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations* (New York, NY: Doubleday).

Tong J, Feiler D (2016) A behavioral model of forecasting: Naïve statistics on mental samples. *Management Sci.* 63(11):3609–3627.

Treynor J (1987) Market efficiency and the bean jar experiment. *Financial Analysts Journal* 43(3):50–53.

Windschitl PD, Wells GL (1996) Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied* 2(4):343.

Winkler RL, Jose VRR (2011) Scoring rules. Cochran J, ed., *Wiley Encyclopedia of Operations Research and Management Science*, volume 7, 4733–4744 (New York, NY: John Wiley & Sons).

Zimmer AC (1984) A model for the interpretation of verbal predictions. *International Journal of Man-Machine Studies* 20(1):121–134.