A Scalable Algorithm for Subset Selection and Rank Probabilities in Contests and Latent Variable Choice Models

Abstract

A k-subset of items will be chosen from n according to values taken by n auxiliary variables $X_1, \ldots, X_n$ interpreted as performances in a contest. Item i is chosen if $X_i \leq X^{(k)}$ where $X^{(k)}$ is the k'th order statistic. A numerical algorithm is presented for computing many k-combination choice probabilities quickly, for small k but potentially large $n \gg 1,000,000$. No assumption is made on the 1-margin distributions of the X_i, and the analytical convenience survives the introduction of dependence via a factor model also. The computation of rank probabilities for k items is a corollary. The algorithm is provided in the winning package, on PyPI.

Keywords: choice models, order statistics, ranking models

1. Motivation

Since the 1930's bettors have taken an interest in the probability of choosing two specific items from many. In a quinella wager one picks two contestants to finish first and second, in either order. It is somewhat surprising that despite the popularity of this format no fast, compelling, deterministic means of estimating quinella probabilities has been disseminated that is consistent with an underlying performance model. Models that are computationally convenient have tended to have weakly motivated generative models, such as exponential performance, and not surprisingly are at odds with empirical evidence.

Contest theory considers all-pay auctions, where only k participants receive an item but everybody pays regardless (in work output). Explicit "races" such as this, with rewards accruing to the top k performers, are not uncommon. For instance, a keyword search for documents might be considered a race and the importance of finishing on page one rather obvious. Likewise, manufacturers compete for distribution and most channels will limit product range to a manageable number $k \ll n$.

Securities trading provides another broad category. A sealed bid auction with k rewards is a reasonable approximation for the trading of fungible securities subject to market impact. That is true when this activity is facilitated by a dealer community (over-the-counter), or when there are multiple ways to route orders via alternative exchanges, dark pools and so forth.

As an example, a client might solicit several bids and offers, and then split flow amongst the k most competitive respondents. On some electronic trading venues, additional information will be disseminated only to those whose bids or offers are deemed competitive (sometimes $k = 2$). This information reward can be as important as the winning of the trade itself.

Students applying to colleges face a similar contest, even if the yardstick and performance variable may vary greatly in its transparency from one country to another. (In Australia a single quantitative score holds sway whereas in the United States bespoke evaluations are performed).

Government and private tender processes can resemble contests with more than one rank-induced reward, assuming that the task is not likely to be completed by a single respondent. Similar calculus can apply to athletes wishing to make their way through qualifying rounds, to job applicants, or to anyone, or anything hoping to make a shortlist of any variety.

Hundreds of billions of contests are performed daily when web pages contained space for advertising are opened. The browser initiates a contest between automated bidding engines looking to supply content. In other settings the performance variable may be price, quality or some combination of the two. The variables $X_1, \ldots, X_n$ might represent any measure of selection desirability.

A fast method of determining whether a contestant is likely to lie in the top k out of n (or bottom k) can assist strategy, especially if there is an ability to shift the performance distribution at some cost. A fast, noiseless calculation makes what-if scenarios, and the computation of marginal rewards, far more convenient than the alternative, Monte Carlo.

These explicit contests are not, however, the sole motivation for this work. Indeed, the task of choosing k items from a basket of n possibilities is so commonplace, and sometimes unconscious, that we do not stop to consider it. Even when this choice is not ostensibly an exercise in drawing random variables and ranking them, the positing of a latent continuous variable for each item representing its fitness for selection can be a strongly motivated approach. For example it offers coherent probabilities when we compare subset selection probabilities in overlapping groups of items.

It is not surprising, therefore, that latent variable models for preference are popular in the literature. They might hope to explain decisions, or provide hypothetical probabilities for item group selections based on limited data.

For instance if only a single choice of candidate is made on a ballot, presumably for a first-past-the-post voting procedure, how might we infer the hypothetical outcome if preferential voting were used instead - or if a candidate were to drop out? A simple renormalization of probability might be no more convincing in this context than it is at the racetrack. After all, this assumption, known as Luce's Axiom of Choice, was criticised by Luce himself.

Ranking models might also benefit. When contestants, who might be sporting teams or students participating in exams, are assessed based on longitudinal performance it is possible to assess likelihood of underlying ability models quickly using the approach presented herein. That is especially helpful when rank performance or grade data is available but not individual question responses (the subject of item response theory).

That algorithm presented here is scalable in n. After an initial computation involving all participants is performed, the marginal cost of computing $p(s)$, the probability of subgroup $s \in S(n,k)$ of cardinality k is chosen, is shown to be small and independent of n.

Because rank probabilities are differences of sums of symmetric group-choice probabilities, the method provided also yields rank probabilities for moderate k more accurately than previously disseminated methods.

2. Approach

Figure 1 depicts some hypothetical performance distributions for five contestants. It is assumed that variables can be approximated by random variables $\tilde{X}_1, \tilde{X}_2, \dots$ where the $\tilde{X}_i$ are supported on an evenly spaced lattice. Expected prices of rank-determined contingent claims (paying 1 in the event of no tie, but proportionately less otherwise) will serve as an approximation to rank-probabilities for the continuous variables.

The key insight borrowed from Cotton (2021) is the characterization of a group of contestants by two quantities: the density of the first order statistic, but also the multiplicity (expected score-contingent number of winning ties).

2.1. Distributions supported on a lattice, approximate translation, and state prices

Dropping tildes we use F to denote the (cumulative) distribution function and f the density of variables taking values on a lattice. Both are considered vectors. This is not a serious loss of generality, provided the algorithm supports reasonably sized lattices, which it does. Nor is there loss of generality in assuming the distributions are supported on integers.

Following Cotton (2021) we define, for any $f(\cdot) : \mathcal{Z} \to \mathcal{R}$ and any $a \in \mathcal{R}$ a shifted distribution $f^{\to a}(\cdot)$ also supported on the integers $\mathcal{Z}$:

$$f^{\to a}(j) := (1-r) f^{\to \lfloor a \rfloor}(j) + r f^{\to \lfloor a \rfloor + 1}(j) \tag{1}$$

where $r = a - \lfloor a \rfloor$ is the fractional part of the shift a. This extends the obvious right shift operator applicable when a is an integer. Formerly $f^{\to \lfloor a \rfloor}(j) := f(j - \lfloor a \rfloor)$. If $f(\cdot)$ is the distribution for X then $f^{\to a}$ approximates the distribution $X + a$.

The quantity

$$p_i = E\left[\frac{\iota_{X_i = X^{(1)}}}{\sum_{j=1}^{n} \iota_{X_j = X^{(1)}}}\right] \tag{2}$$

is, ignoring ties, a winning probability for the i'th item (probability it is selected first). It is the winning state price. Here ι is the indicator function and the denominator counts ties.

2.2. Review of multiplicity calculus and removal of one item

Consider a subgroup A minimally characterized by $\Upsilon_A = (S_A(), m_A())$ where $S_A(j) = Prob(X^{(1)} > j) = 1 - F_A(j)$ is the survival function for the first order statistic, and here F_A is the cumulative distribution for the same. Obviously we can quickly determine f_A, the density of the first order statistic, from F_A by prepending zero and differencing the vectors.

As was justified in Cotton (2021) the combination of two item groups A and B into a single group relies on the following estimate for combined multiplicity:

$$m_{A \cup B}(j) = \frac{m_A(j) f_A(j) S_B(j) + (m_A(j) + m_B(j)) f_A(j) f_B(j) + m_B(j) f_B(j) S_A(j)}{f_A(j) S_B(j) + f_A(j) f_B(j) + f_B(j) S_A(j)} \tag{3}$$

The corresponding operation for $S's$ is mere multiplication, whereupon we recover winning distribution F_A also and density $f_{A \cup B}$. It is clear, at noted in that paper, that the density of the first order statistic and also the multiplicity can be determined for the entire race by repeated application.

Furthermore, the multiplicity relations can be inverted to enable us to remove the i'th contestant. The multiplicity with item i left out is related to the multiplicity with i left in as follows:

$$m_{\hat{i}}(j) = \frac{m(j)f_i(j)S_{\hat{i}}(j) + m(j)f_i(j)f_{\hat{i}}(j) + m(j)f_{\hat{i}}(j)S_i(j) - m_i(j)f_i(j)S_{\hat{i}}(j)}{f_{\hat{i}}(f_i + S_i)} \tag{4}$$

as also discussed in Cotton (2021). The corresponding operation for survival functions is mere division, and thus we can characterize a race with one contestant removed.

2.3. Approximate state price for lowest score

Losing probability for asset i by removing it from the race, using formula 4. Then an approximate win state price can be estimated by considering all possible values j taken by X_i and also the lowest value j' taken by the other item performances. The density of the latter is $f_{\hat{i}}$ and

$$\begin{aligned} p_i &= \sum_{j,j'} f_i(j) f_{\hat{i}}(j') E\left[\frac{W}{M} | X_i = j, X_{\hat{i}}^{(1)} = j'\right] \\ &\approx \sum_j f_i(j) \left\{ \frac{f_{\hat{i}}}{1 + m_{\hat{i}}(j)} + S_{\hat{i}}(j) \right\} \end{aligned} \tag{5}$$

as shown in more detail in Cotton (2021).

2.4. The illusive quinella calculator, and generalization.

Although the calculus above ostensibly applies when comparing one horse against the rest, it is also possible to compare the worst performance amongst a subgroup A to the best performance of the complement $B = \{1, \dots, n\}\ A$. The accounting for multiplicity in the discrete variables brings their state prices into close alignment with the probabilities for their continuous counterparts.

So as an example, a fast numerical method for estimating quinella probabilities under arbitrary performance distribution assumptions, for independent performances, can proceed as follows:

1. Remove two horses i and j from the race using repeated application of Equation 4.
2. Compute the distribution of the worst performance of horses i and j, and the multiplicity.
3. Compare the worst performance of the two horses to the rest of the runners, represented by the density and multiplicity achieved in the first step.

Clearly the quinella example generalizes to any subgroup of cardinality $k > 2$ by repeated application of Equation 4. Contests with hundreds of thousands of participants are amenable to this procedure.

2.5. Rank probabilities

Let $S(k,n)$ denote the set of all of k-combinations and $p(s)$ the probability, computed as above, that a particular k-combination corresponds to the first k items chosen. Also let $p_{i,j}$ denote the probability that item i is chosen j'th (i.e. that X_i is the j'th smallest value. Then we have

$$\sum_{j=1}^{k} p_{i,j} = \sum_{s \in S(j,k) | i \in s} p(s) \tag{6}$$

and both equal the probability that item i is selected when all items compete for the top k spots. We can use the right hand side to compute the left, and since this holds for every k we can then compute $p_{i,j}$ by subtraction.

To place this approach in horseracing terms for $k=2$: the probability that a horse finishes second is equal to the probability it finishes in the top two minus the probability it wins, and the former is a sum of quinella probabilities.

By symmetry, the probability of a k-combination of items being drawn last can also be computed, merely by reversing the performance distributions. Working from the front and the rear, we can compute all group selection probabilities subject only to the combinatorial increase in the number of terms on the right hand side of Equation 6 - a burden that represents the limitation of this approach for large k.

This computational caveat aside, it is clear that for moderate n *all* rank probabilities can be computed if the sum on the right hand side extends to all sets of cardinality $\lfloor \frac{n-1}{2} \rfloor$, since in the case of odd n it is only necessary to compute up to $k = (n-1)/2$. (Again speaking loosely if we know the probabilities that a horse finishes first, second, though seventh and also last through seventh-last, we clearly know the probability it finishes eighth in a 15 horse race. The same rationale applies with ties and state prices.)

2.6. Dependence

Having established a route to the computation of group and rank probabilities in the case of independent performances, we now consider the case where performances are dependent. Depending on the computational budget and structure assumed, quadrature can be used. The example of a Gaussian copula with common off-diagonal correlation $0 \le \rho < 1$ is given.

The performance X_i of the i'th contestant can be assumed to satisfy

$$X_i = F_i^{-1}\left(\Phi(Z_i)\right)$$

where F_i is the cumulative distribution on the lattice (the 1-margin), Φ is the standard cumulative normal distribution, and Z_i is an auxiliary random variable with standard normal distribution. (As written this is merely a restatement of the distributional transform, or the definition of F_i.)

In the independent case all assets are assumed to satisfy this generative model, with corresponding Z_i's that are independent and identically distributed. A simplistic dependence structure now introduces a connection between the Z_i's, such as

$$Z_i = \rho Z + \sqrt{1-\rho^2}\epsilon_i$$

where $Z \sim N(0,1)$ is a common factor, ρ is a correlation parameter and the ϵ_i's are independent N(0,1). All state prices can be viewed as iterated expectations, first conditioning on a choice of Z. If we let $F_i(\cdot; z)$ denote the cumulative distribution of the i'th asset return knowing $Z = z$ then rearranging we have

$$F_i(x; z) = \Phi\left(\frac{\Phi^{-1}(F_i(x)) + \rho z}{\sqrt{1-\rho^2}}\right) \tag{7}$$

Thus given any vector valued linear functional G acting on collections of densities, we can define $g(z)$ as the action of G on the set of densities transformed according to Equation 7 with parameter z. Then, the integral

$$\int_{z=-\infty}^{\infty} g(z)\phi(z)dz$$

can be estimated using standard methods such as Gaussian quadrature.

2.7. Example

Five-way rank probabilities are shown in Table 2 the performances take on a financial interpretation - the realized return of a stock (the market, personified, chooses companies). This table uses a correlation parameter $\rho = 0.1$, whereas when $\rho = 0.25$ is considered instead slightly different rank probabilities are computed, reported in Table 3.

The results are quite similar. Because the common factor influences conditional probabilities in monotone fashion, the results are not overly sensitive to choice of correlation. The reader will observe that the rounded answers are almost identical to two significant digits.

The dependence structure can be generalized by allowing more than one factor to vary. The accuracy and computational convenience is then subject to the efficacy of two, three or higher dimensional quadrature schemes.

	Asset 1	Asset 2	Asset 3	Asset 4	Asset 5
location	-0.50	-0.25	0.00	1.00	1.50
scale	1.00	1.50	1.20	1.30	2.00

Table 1: Location and scale parameters for skew-normal return distributions used to illustrate the rank probability approach.

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Asset 1	0.37	0.33	0.20	0.08	0.02
Asset 2	0.32	0.24	0.21	0.15	0.08
Asset 3	0.20	0.26	0.28	0.19	0.07
Asset 4	0.04	0.10	0.19	0.37	0.31
Asset 5	0.07	0.08	0.12	0.21	0.52

Table 2: Example five-way rank probabilities, $\rho = 0.1$.

3. Application to crowd-sourced forecasting

Prediction markets, financial markets and betting markets represent a crucial piece of information that can and should assist forecasting - but often this information represents only a slice, or a projection, or a margin, of some more important “full” picture worthy of forecasting.

3.1. Relative location parameter inference

While group selection and rank probabilities may be desirable, they can be awkward to elicit from survey respondents. Elsewhere, rank probabilities might be only partially revealed by market prices.

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Asset 1	**0.38**	0.33	0.20	0.08	0.02
Asset 2	0.32	0.24	0.21	0.15	**0.07**
Asset 3	0.20	0.26	0.29	**0.18**	0.07
Asset 4	0.04	**0.09**	0.19	0.37	0.31
Asset 5	**0.06**	**0.08**	0.12	0.21	0.53

Table 3: Example rank probability computation, $\rho = 0.25$.

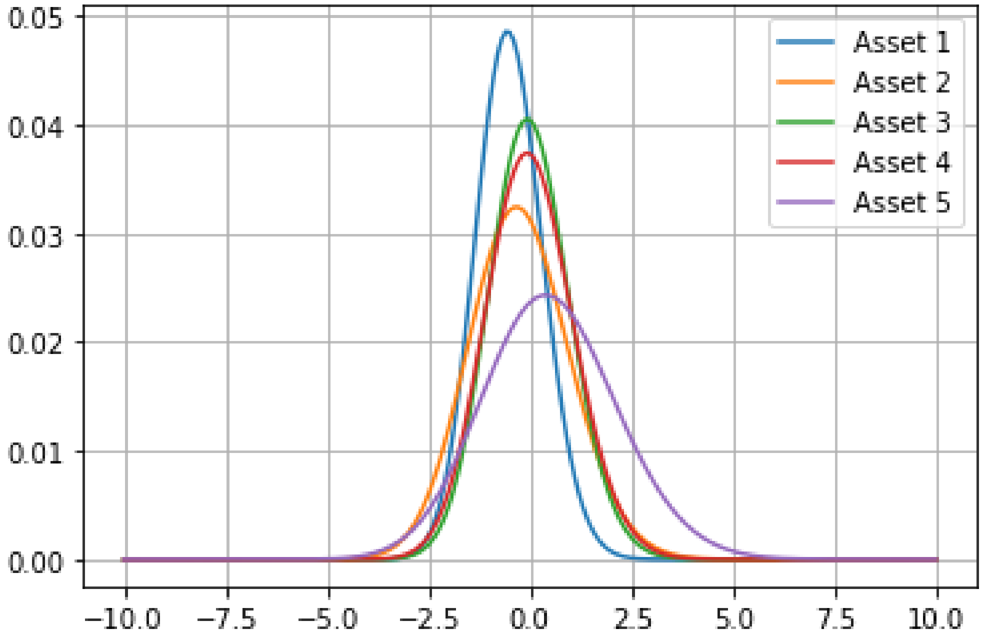

Figure 1: Skew-normal return distributions for five assets, used as an example for rank probability calculations.

Another example is provided by the M6 Financial Forecasting Duathlon. Participants are invited to submit five-way rank probabilities for stock returns. But it is certainly plausible that some would-be participants might shy away from this task, yet possess valuable information.

Similarly, published market share statistics for trading venues represent partial information, from which the aggressiveness of participants in their bidding and offering might need to be be inferred. An estimate of whether a particular bid or offer will be competitive can lean on the same calculus, as discussed in Cotton (2021).

Similarly, betting markets for winning events are often more liquid, and might contain more information, then relatively illiquid markets for lower placings. A market might exist for the winner of an election, for example, leaving open the task of forecasting relative vote counts.

In some of these settings, imputation can exploit the fast algorithm provided. To illustrate with an example, Figure 4 depicts a partial rank probability table. Here an opinion has been expressed as to the probability that a stock will be the worst performing of the five, but no other information is supplied.

Suppose, for illustration, that we assume all five assets have similar volatilities and, for that matter, identical return profiles up to a translation. The imputation proceeds by calibrating the

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Asset 1	0.25	?	?	?	?
Asset 2	0.15	?	?	?	?
Asset 3	0.20	?	?	?	?
Asset 4	0.22	?	?	?	?
Asset 5	0.18	?	?	?	?

Table 4: A partially completed competition entry showing only worst-performer probabilities.

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Assumed 1	Discrepancy
Asset 1	**0.252**	0.218	0.197	0.178	0.156	**0.250**	0.002
Asset 2	**0.149**	0.176	0.198	0.222	0.255	**0.150**	-0.001
Asset 3	**0.200**	0.203	0.202	0.200	0.195	**0.200**	0.000
Asset 4	**0.220**	0.210	0.201	0.191	0.178	**0.220**	0.000
Asset 5	**0.179**	0.193	0.202	0.209	0.216	**0.180**	-0.001

Table 5: Imputed five-way rank probabilities taking Table 4 as a starting point. A gaussian correlation of $\rho = 0.25$ has been assumed, and this leads to a small differential in Rank 1 probabilities.

relative location parameters of the return distributions. Figure 2 depicts return distributions consistent with Table 4.

In other context, those distributions might represent the "ability" of an item to attract the attention of a consumer.

Using the procedure provided in Section 2 the corresponding five-way rank probabilities (state prices, to be pedantic) are reported in Table 5. Again, purely for illustration, the parameter $\rho = 0.25$ has been used. The penultimate column repeats the assumed probabilities taken from Table 4 and the last column reports the discrepancy after calibration.

The small discrepancy noted arises from two sources. One is due to numerical issues discussed in Cotton (2021) that, in our example, modify the losing probabilities by roughly one part in a million.

By far the largest component in the last column is due to inconsistency in the calibration approach. Technically speaking, one should calibrate assuming $\rho = 0.25$ whereas we have calibrated using $\rho = 0$ implicitly. However the errors are still quite small, as can be seen, and so a more

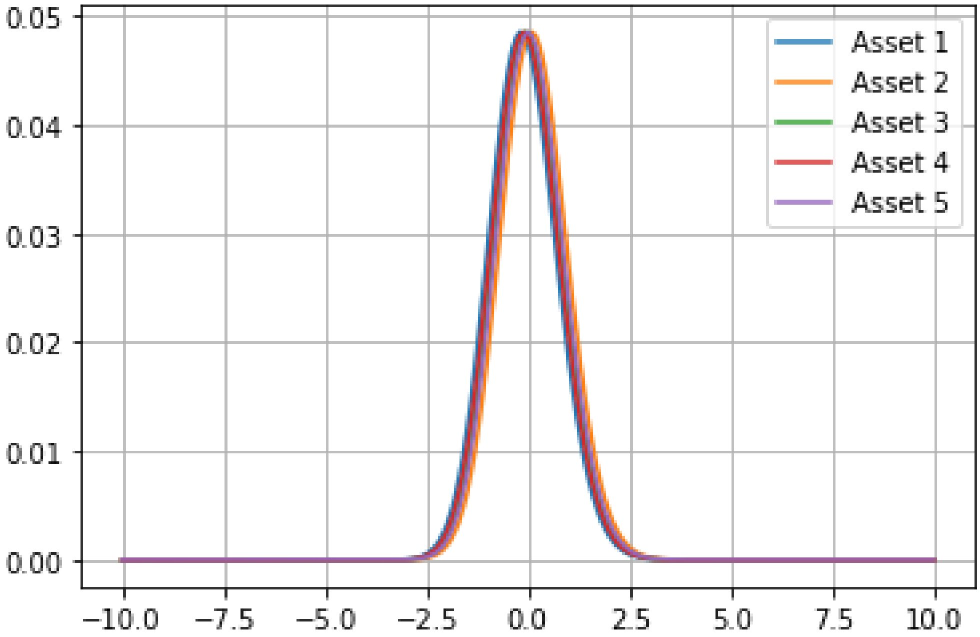

Figure 2: Implied relative return distributions consistent with Table 4.

complex procedure might not be warranted.

3.2. Coherence

A related issue in the crowd-sourcing of forecasts is coherence. Some respondents to forecasting surveys might provide rank probabilities, but the use of these might benefit from additional processing.

One class of approach takes submitted rank probabilities, removes k randomly chosen entries from the table, and then calibrates (using a derivative free optimizer) numerous choices of generative model and correlation parameter. By seeking to infer a performance model that is consistent with the supplied submission, one might hope to form a prior opinion as to the reasonableness of the supplied rank probabilities, and also suggest changes to some of those values supplied.

Another hypothesis is that overly simplistic assumptions about rank-probabilities can be discerned in some submissions, and that this alone might help predict their performance in the contest. For instance some entries might, contrary to the approach suggested herein, reinvent the Harville model Harville (1973) and, in all likelihood, suffer from its dubious empirical properties.

4. Summary

When k items are chosen from n, a fast computational method has been provided for discerning the probability that the k are a pre-specified group. These probabilities can be consistent with an extremely broad class of models for underlying latent variables presumed to represent relative attractiveness of the items.

A fast computation of rank probabilities is a corollary, for moderate k and very large n. As a special case, a numerical procedure for pricing quinella bets has been presented - the first of its kind that makes no assumption about performance distributions.

One obvious application is to the M6 Forecasting Competition. Due to the generality of the numerical approach, it may be considered a projection from the space of "full information" models (i.e. a joint distribution of all assets) into the incomplete information that is required to be submitted.

References

Cotton, P. (2021). Inferring Relative Ability from Winning Probability in Multientrant Contests. *SIAM Journal on Financial Mathematics*, *12*. doi:10.1137/19M1276261.

Harville, D. A. (1973). Assigning probabilities to the outcomes of multi-entry competitions. *Journal of the American Statistical Association*, . doi:10.1080/01621459.1973.10482425.