Supplementary material for Probabilistic n-Choose-k Models for Classification and Ranking

Proofs

We introduce shorthand notation $\#_r(\boldsymbol{y}_c) = \sum_{d \in c} \mathbf{1}\{y_d = r\}$ and $\#_{\leq r}(\boldsymbol{y}_c) = \sum_{d \in c} \mathbf{1}\{y_d \leq r\}$ to represent the number of variables in $\boldsymbol{y}_c$ that take on value r or value less than or equal to r. We will also use the notation $\boldsymbol{y}_{-i}$ and $\boldsymbol{y}_{c-i}$ as shorthand for all variables except for i, and all variables with indices in c except for i, respectively.

Proof of Proposition 1

In order for a variable to be given value R, it must be chosen in the first step of the generative process that assigns values to variables. For $k_R = 0$, the statement holds with equality (both are 0). For $k_R > 0$, we have,

$$p(y_i = R \mid k_R) \propto \sum_{\boldsymbol{y} \mid (y_i = R) \wedge (\#_R(\boldsymbol{y}_{-i}) = k_R - 1)} \exp\left\{\theta_i + \sum_{d \neq i} \theta_d \mathbf{1}\{y_d = R\}\right\} \tag{10}$$

$$p(y_j = R \mid k_R) \propto \sum_{\boldsymbol{y} \mid (y_j = R) \wedge (\#_R(\boldsymbol{y}_{-j}) = k_R - 1)} \exp\left\{\theta_j + \sum_{d \neq j} \theta_d \mathbf{1}\{y_d = R\}\right\}. \tag{11}$$

The idea is to split these sums into a common component and a disjoint component e.g.,

$$p(y_i = R \mid k_R) \propto \sum_{\boldsymbol{y} \mid (y_i = R \wedge y_j = R) \wedge (\#_R(\boldsymbol{y}_{-i}) = k_R - 1)} \exp\left\{\theta_i + \theta_j + \sum_{d \neq i,j} \theta_d \mathbf{1}\{y_d = R\}\right\} \tag{12}$$

$$+ \sum_{\boldsymbol{y} \mid (y_i = R \wedge y_j \neq R) \wedge (\#_R(\boldsymbol{y}_{-i}) = k_R - 1)} \exp\left\{\theta_i + \sum_{d \neq i,j} \theta_d \mathbf{1}\{y_d = R\}\right\} \tag{13}$$

Both $p(y_i = R)$ and $p(y_j = R)$ will share the first term, so it suffices to compare second terms, which are disjoint. Here, we can see the summations are identical, except that one will have a sum involving θ_i, and the other will have a sum involving θ_j, so clearly the claim holds for any k_R. The full probability $p(y_i = R)$ is a sum $\sum_{k_R} p(y_i = R \mid k_R)$, so given that the relation holds for each component in the sum, it also holds for the full sum.

Proof of Proposition 2

We begin by dividing event space into a 3×3 matrix of possibilities: $\{y_i \geq r, y_i = r-1, y_i < r-1\} \times \{y_j \geq r, y_j = r-1, y_j < r-1\}$. The inductive assumption tells us that the sum of probabilities across the first "row", $p(y_i \geq r) = p(y_i \geq r \wedge y_j \geq r) + p(y_i \geq r \wedge y_j = r-1) + p(y_i \geq r \wedge y_j < r-1)$ is greater than or equal to the sum of probabilities across the first "column", $p(y_j \geq r) = p(y_j \geq r \wedge y_i \geq r) + p(y_j \geq r \wedge y_i = r-1) + p(y_j \geq r \wedge y_i < r-1)$. Our goal is to prove that the sum of probabilities across the first two rows is greater than or equal to the sum of probabilities across the first two columns. The central element of this matrix, which corresponds to $y_i = r-1 \wedge y_j = r-1$ is included in both sums, so it suffices to show that $p(y_i = r \wedge y_j < r) \geq p(y_j = r \wedge y_i < r)$.

As before, we begin by showing that this holds for any particular choice of $\boldsymbol{k}$, which then implies that it holds for the summation over all possible $\boldsymbol{k}$. Similarly, we can assume that we are given an arbitrary choice of subset $c_{\geq r}$ of $\boldsymbol{y}_{-i,-j}$ to take on labels $\geq r$. The desired property will hold for all choices, so when we sum over all the choices, it will also still hold.

Given $\boldsymbol{k}$ and $\boldsymbol{y}_{c_{\geq r}}$, the argument follows similarly to Proposition 1. The probability of choosing y_i to be in level r is

$$p(y_i = r-1 \wedge y_j < r-1) \propto \sum_{\boldsymbol{y}|(y_i=r-1 \wedge y_j \neq r-1) \wedge (\#_{r-1}(\boldsymbol{y}_{\bar{c}_{\geq r}})=k_{r-1}-1)} \exp\left\{\theta_i + \sum_{d \in \bar{c}_{\geq r}, d \neq i, d \neq j} \theta_d \mathbf{1}\left\{y_d = r\right\}\right\}. \quad (14)$$

The expression for $p(y_i = r-1 \wedge y_j < r-1)$ will be identical, but θ_i will be replaced with θ_j. Using the assumption that $\theta_i > \theta_j$ completes the proof.

Proof of Proposition 3

We can rewrite g_d by regrouping the summation (assuming we have defined $f(0) = 0$):

$$g_d = \sum_{r=1}^{R} f(r) p(y_d = r) = \sum_{r=1}^{R} (f(r) - f(r-1)) p(y_d \leq r). \quad (15)$$

We then consider the difference between g_i and g_j:

$$g_i - g_j = \sum_{r=1}^{R} (f(r) - f(r-1))(p(y_i \leq r) - p(y_i \leq r)). \quad (16)$$

Due to the monotonicity of f, each $f(r) - f(r-1)$ term will be non-negative. By Lemma 1, the $p(y_i \leq r) - p(y_i \leq r)$ terms are also all non-negative, so the total sum is non-negative, and we get $g_i - g_j \geq 0$.

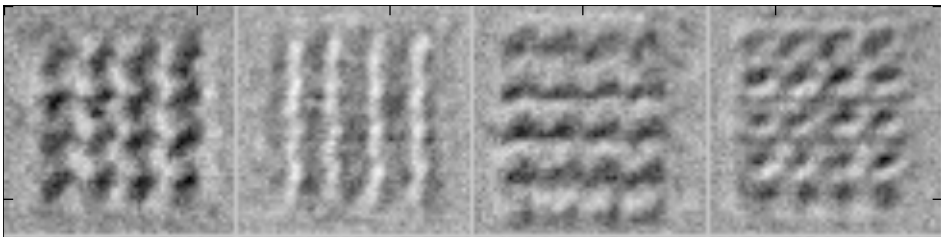
Proof of Proposition 4

$$\begin{aligned} & a_i(b_i - b_j) + a_j(b_j - b_i) \geq a_i(b_i - b_j) + a_i(b_j - b_i) = 0 \\ \Leftrightarrow & a_i b_i - a_i b_j + a_j b_j - a_j b_i \geq 0 \\ \Leftrightarrow & a_i b_i + a_j b_j \geq a_i b_j + a_j b_i \end{aligned}$$

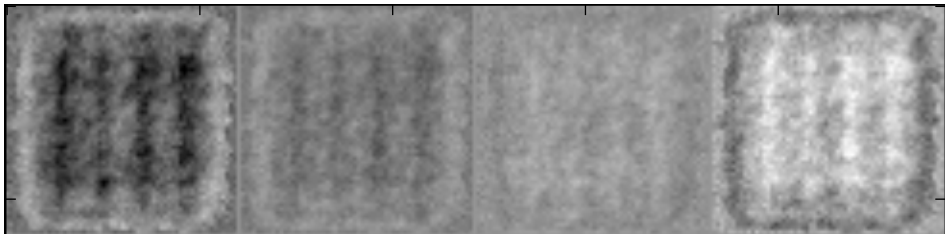
Visualization of the learned parameters from embedded MNIST

Figure 3 shows a visualization of the parameters learned by the binary n-choose-k model on the embedded MNIST dataset. The likelihood parameters form a 1000×10 matrix, where each column corresponds to a different class. We take these and multiply them by the 3600×1000 RBM weights that generated the features in order to project them to pixel-space. We then reshape each column to form a 60×60 image. The same can be done for the prior parameters, except now the 10 columns correspond to counts instead of classes.

For the likelihood parameters, we show the first four classes corresponding to the digits 0 to 3. Clearly the parameters recognize the 4×4 grid in which the digits were embedded (before adding a slight jitter). For the count parameters, we also visualize the first four, corresponding to the counts 1 to 4. Note that the count parameters became extremely strongly negative after 4, suggesting that the model correctly learned that there can be at most 4 digits embedded in an image. In logistic regression, the likelihood parameters look similar to the ones shown, however note that they must also be used to simultaneously model the prior over counts.



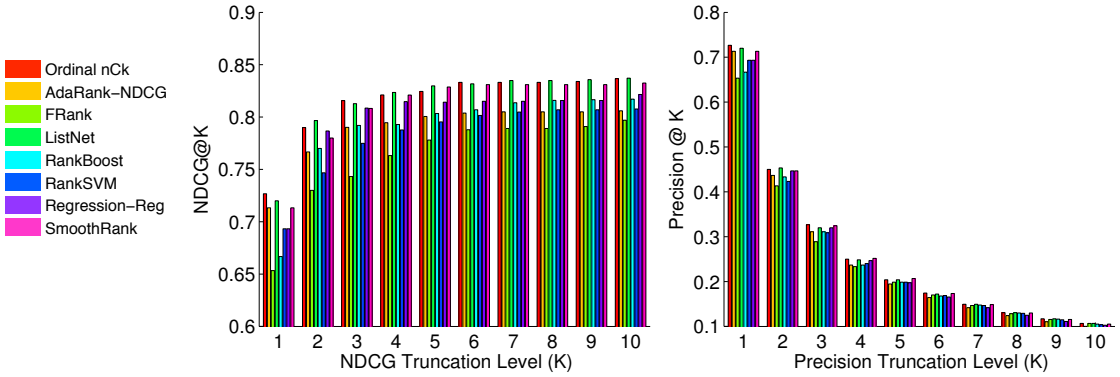
(a) Likelihood parameters



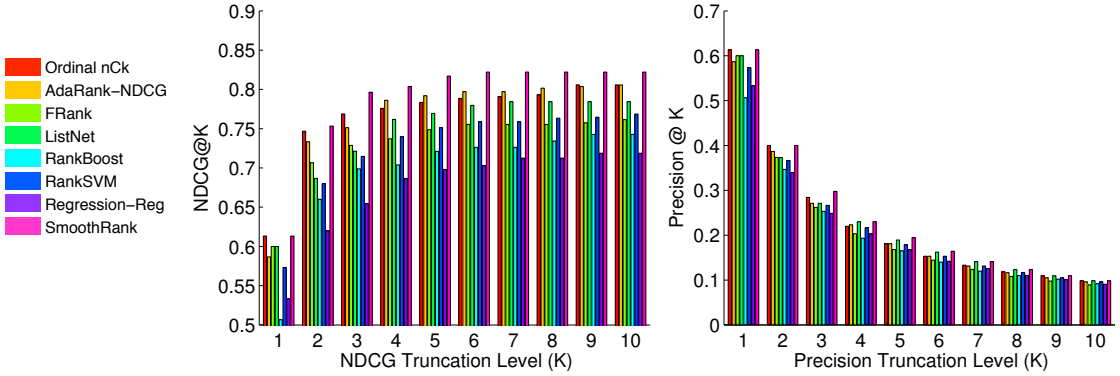
(b) Prior parameters

Figure 3: A visualization of the parameters learned by the binary n-choose-k model on the embedded MNIST dataset. (a) corresponds to the parameters connecting the inputs to the first 4 classes (out of 10), while (b) corresponds to the input-dependent prior over counts. (also out of 10). White pixels correspond to large, positive parameters while black pixels correspond to large, negative parameters.

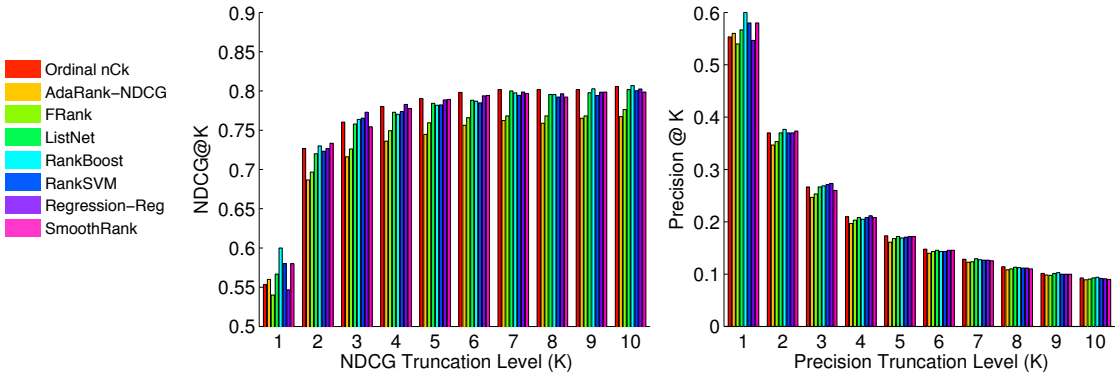
More LETOR Results



(d) HP2003

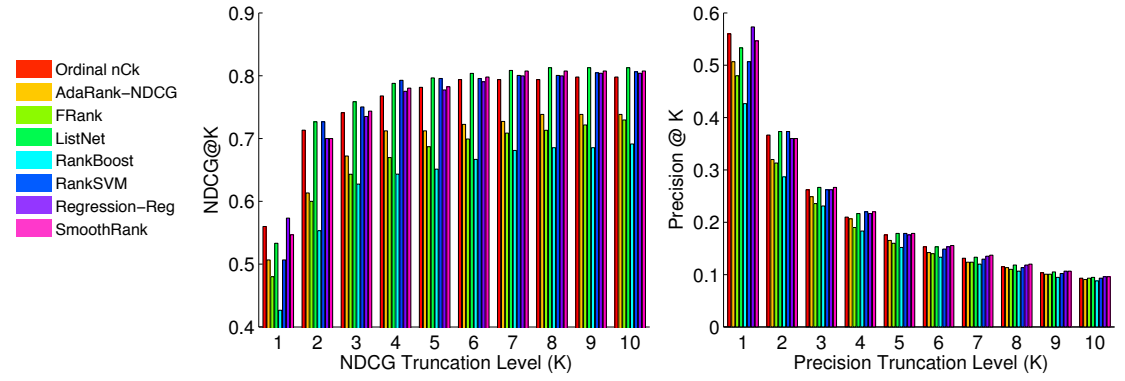


(h) HP2004

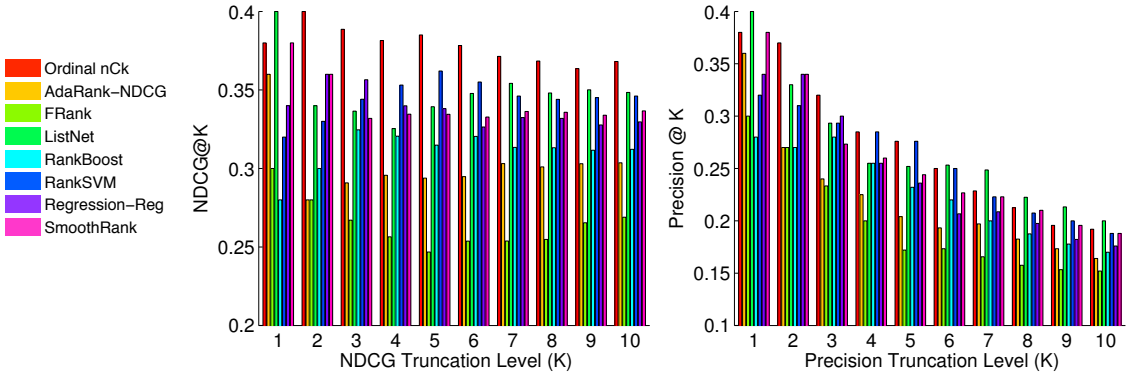


(l) NP2003

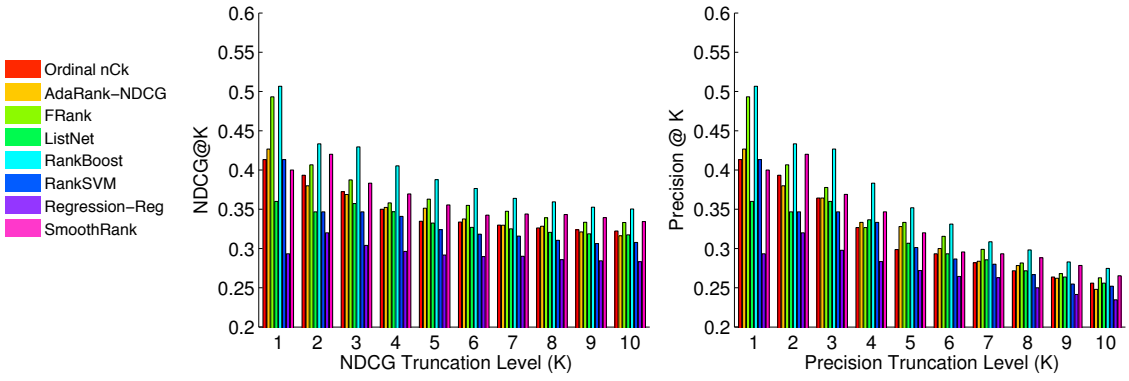
Figure 4: NDCG and Precision for the Ordinal n-Choose-k Model and other benchmark methods on the LETOR 3 datasets.



(p) NP2004

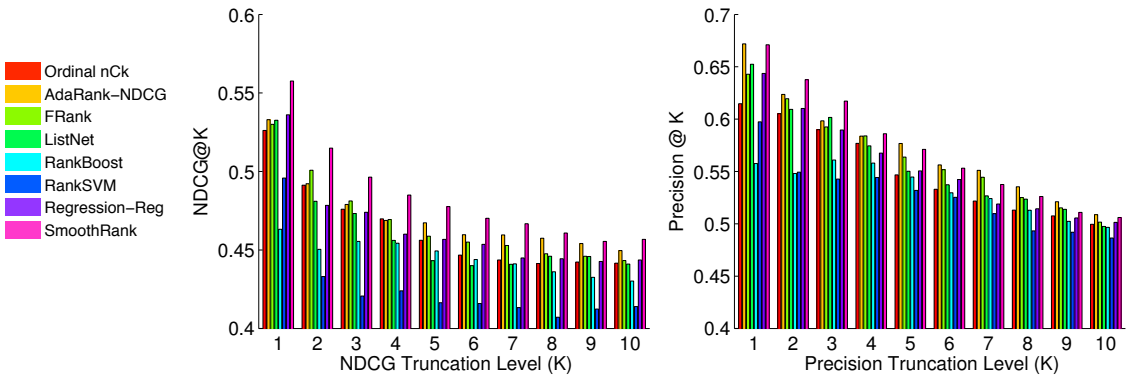


(t) TD2003



(x) TD2004

Figure 4: NDCG and Precision for the Ordinal n-Choose-k Model and other benchmark methods on the LETOR 3 datasets.



(d) Ohsumed

Figure 5: NDCG and Precision for the Ordinal n-Choose-k Model and other benchmark methods on the LETOR 3 datasets.