# An expectation-maximization method to estimate a rank-based choice model of demand

## Appendix

**Garrett van Ryzin**
Graduate School of Business, Columbia University, New York, NY 10027
e-mail: gjv1@columbia.edu

**Gustavo Vulcano**
School of Business, Torcuato di Tella University, Buenos Aires, Argentina,
Leonard N. Stern School of Business, New York University, New York, NY 10012,
e-mail: gvulcano@stern.nyu.edu

## A1 Pseudocodes of the EM method

### A1.1 Uncensored demand case

We next summarize the entire EM algorithm for the uncensored demand case using pseudocode.

**EM algorithm for estimating the rank-based choice model**

*[Input data]:* Number of compatible customer types $N$, availability information represented by sets $S_t$ and transaction data $j_t, t = 1, \ldots, \hat{T}$.

*[Preprocessing]:* Set initial proportions $x_i$. Based on the availability information and transaction data, build the sets $\mathcal{M}_t(j_t, S_t) := \{i : \sigma^{(i)}(j_t) < \sigma^{(i)}(k), \forall k \in S_t, k \neq j_t\}$.

```
Repeat
  Set $m_i := 0, x_{it} := 0, i = 1, \ldots, N, t = 1, \ldots, \hat{T}$.
  [E-step]:
  For $t := 1, \ldots, \hat{T}$ do
      [Update probabilities of customer types]
    For $i \in \mathcal{M}_t(j_t, S_t)$ do
        Set $x_{it} := x_i / (\sum_{h \in \mathcal{M}_t(j_t,S_t)} x_h)$.
    Endfor
  Endfor
  [Compute estimates for $m_i$]
  For $i := 1$ to $N$ do
    For $t := 1$ to $\hat{T}$ do
      Set $m_i := m_i + x_{it}$
    Endfor
  Endfor
  [M-step]:
  For $i := 1$ to $N$ do
    Set $x_i := m_i / \sum_{k=1}^{N} m_k$.
  Endfor
Until Stopping criterion is met.
```

A few remarks on the implementation are in order. The initialization of $x_i$, $i = 1, \ldots, N$, is arbitrary (as long as the set of types is compatible with the transactions); we merely need starting values different from zero. The stopping criteria can be based on various measures of numerical convergence, e.g., that the Euclidean norm of the difference between two consecutive vectors $\hat{\boldsymbol{x}}$ is less than $\epsilon$. Alternatively, we could set a maximum number of iterations. In all of our experiments we observed very fast convergence, so it would appear that the stopping criterion is not critical.

## A1.2 Censored demand case

We next summarize the entire EM algorithm for the censored demand case using pseudocode.

**EM algorithm for estimating the rank-based choice model**
*[Input data]:* Number of customer types $N$, availability information represented by sets $S_t$ and transaction data $j_t, t = 1, \ldots, T$.
*[Initialization]:* Set $x_i := 1/N$, and $\lambda = 0.5$. Based on the availability information and transaction data, build the sets $\mathcal{M}_t(j_t, S_t) := \{i : \sigma^{(i)}(j_t) < \sigma^{(i)}(k), \forall k \in S_t, k \neq j_t\}$. Set $a_t := 0$, $t = 1, \ldots, T$.

```
Repeat
  Set $m_i := 0, x_{it} := 0, i = 1, \ldots, N, t = 1, \ldots, T$.
  [E-step]:
  For $t := 1, \ldots, T$ do
      [Update probabilities of customer types]
      For $i \in \mathcal{M}_t(j_t, S_t)$ do
          Set $x_{it} := x_i / (\sum_{h \in \mathcal{M}_t(j_t, S_t)} x_h)$.
      Endfor
      [Update estimates $a_t$]
      If $j_t > 0$
          Set $a_t := 1$,
      Else (i.e., $j_t = 0$)
          If $\mathcal{M}_t(0, S_t) = \emptyset$
              Set $a_t := 0$,
          Else
              Set $a_t := \lambda \sum_{i \in \mathcal{M}_t(0, S_t)} x_i / (\lambda \sum_{i \in \mathcal{M}_t(0, S_t)} x_i + (1 - \lambda))$
          Endif
      Endif
  Endfor
  [Compute estimates for $m_i$]
  For $i := 1$ to $N$ do
      For $t := 1$ to $T$ do
          Set $m_i := m_i + a_t x_{it}$
      Endfor
  Endfor
  [M-step]:
  For $i := 1$ to $N$ do
      Set $x_i := m_i / \sum_{t=1}^{T} a_t$.
  Endfor
```

```
    Set $\lambda := \sum_{t=1}^{T} a_t/T$.
Until Stopping criterion is met.
```

We point out that as in the uncensored demand case, the initialization of $\lambda$ and $x_i$, $i = 1, \ldots, N$, can be set arbitrarily.

## A2 Supplement to numerical examples

In the figures below we report the relative errors of Direct Max (left column) and EM (right column) with respect to the true underlying model parameters, for the case where $N = 10$ and the number of periods increases from $T = 10,000$ (top) to $T = 50,000$ and $T = 100,000$ (bottom). Since the results obtained by both versions of Direct Max are (almost) identical, we just report results for Direct Max V2 (the version which runs faster).

Nonsurprisingly, when $T$ increases, the errors of both methods become more target oriented around zero. The distributions of errors for both methods are also very similar. We also verify that the log-likelihoods of the three proportions (based on the generating demand model, estimated via Direct Max, and estimated via EM) are very close (up to three decimal numbers), even when the estimation error is more significant (i.e., when $T = 10,000$).

### A2.1 Uncensored demand case

Figure A1 shows the relative errors for the uncensored demand case when we set $\lambda = 1$ (which is common knowledge).

### A2.2 Censored demand case

Figure A2 reports estimation errors for a low volume demand scenario (with $\lambda = 0.2$), and Figure A3 does it for a high volume demand scenario (with $\lambda = 0.8$).

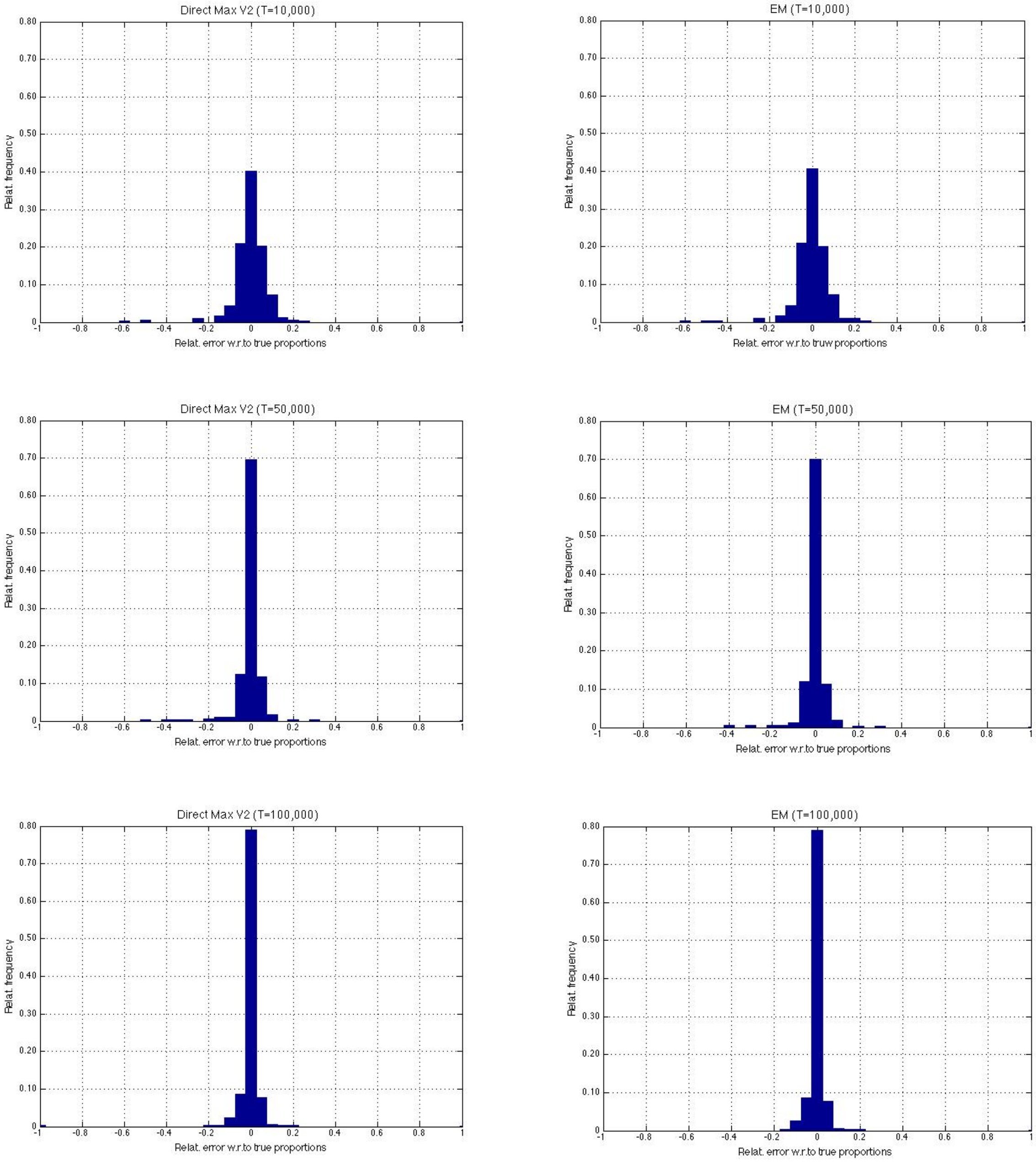

Figure A1: Uncensored demand case. Relative errors of Direct Max V2 (left) and EM (right) with respect to the true underlying proportions $\boldsymbol{x}$ of customer types in the market, for $n = 15$, $N = 10$, and $T \in \{10,000,\ 50,000,\ 100,000\}$. There are between 2 and 10 products available per period.

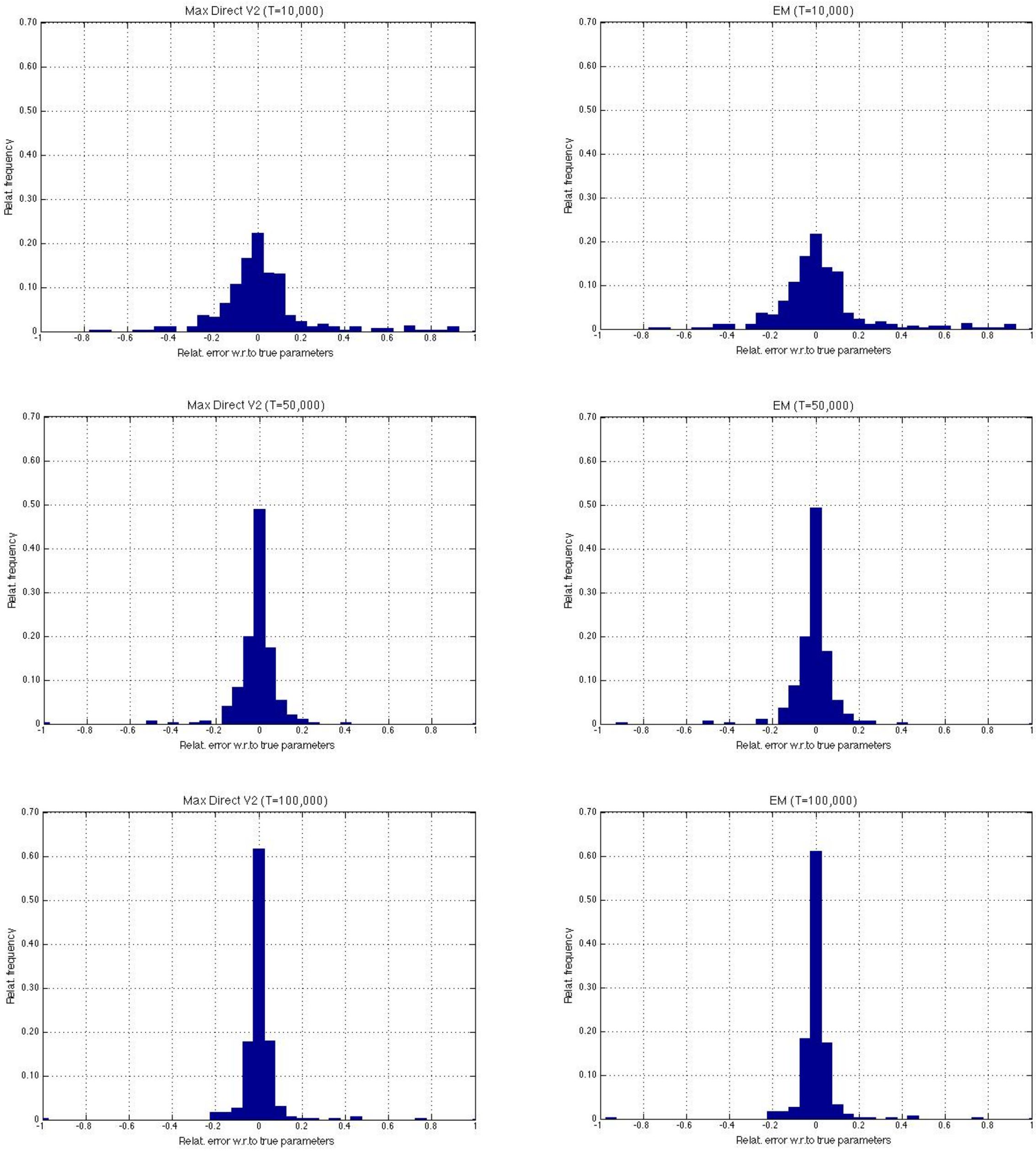

Figure A2: Relative errors of Direct Max V2 (left) and EM (right) with respect to the true underlying proportions $\boldsymbol{x}$ and arrival rate $\lambda$ of the model, for $\lambda = 0.2$, $n = 15$, $N = 10$, and $T \in \{10,000,\ 50,000\}$. There are between 2 and 10 products available per period.

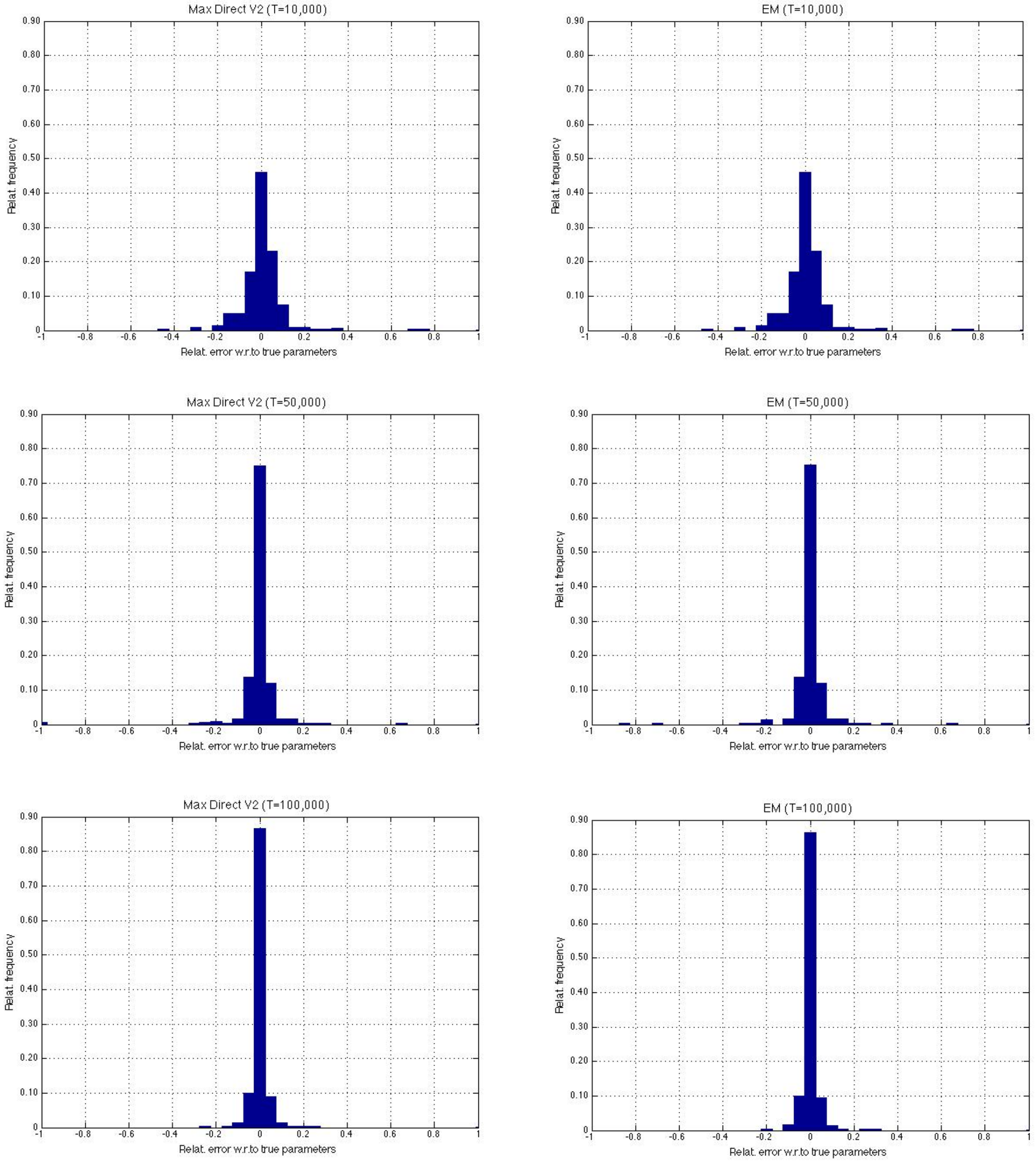

Figure A3: Relative errors of Direct Max V2 (left) and EM (right) with respect to the true underlying proportions $\boldsymbol{x}$ and arrival rate $\lambda$ of the model, for $\lambda = 0.8$, $n = 15$, $N = 10$, and $T \in \{10,000,\, 50,000\}$. There are between 2 and 10 products available per period.