On the closure of relational models

Anna Klimova
Institute of Science and Technology (IST) Austria
aklimova@ist.ac.at

Tamás Rudas
Eötvös Loránd University, Budapest, Hungary
rudas@tarki.hu

Abstract
Relational models for contingency tables are generalizations of log-linear models, allowing effects associated with arbitrary subsets of cells in the table, and not necessarily containing the overall effect, that is, a common parameter in every cell. Similarly to log-linear models, relational models can be extended to non-negative distributions, but the extension requires more complex methods. An extended relational model is defined as an algebraic variety, and it turns out to be the closure of the original model with respect to the Bregman divergence. In the extended relational model, the MLE of the cell parameters always exists and is unique, but some of its properties may be different from those of the MLE under log-linear models. The MLE can be computed using a generalized iterative scaling procedure based on Bregman projections.

Keywords: algebraic variety, Bregman divergence, contingency table, extended MLE, relational model

1 Introduction
The existence of the maximum likelihood estimates under log-linear models for contingency tables has been thoroughly studied, see [Haberman, 1974], [Andersen, 1974], [Barndorff-Nielsen, 1978], [Lauritzen, 1996], among others. It was established that the maximum likelihood estimates of the cell parameters always exist if the observed table has only positive cell counts, but may not exist if some of the observed counts are zeros. The patterns of zero cells that lead to the non-existence of the MLE were described in several ways [cf. Haberman, 1974, Fienberg and Rinaldo, 2012b]. Extended log-linear models provide the framework in which all data sets irrespective of the pattern of zeros have an MLE. The extended log-linear model is the closure of the original model in the topology of pointwise convergence [cf. Lauritzen, 1996]. This extension coincides with the closure of the original model with respect to the Kullback-Leibler divergence [cf. Csiszár and Matúš, 2003] and with the “aggregate” exponential family derived from the original model [Brown, 1988].
This paper introduces extensions of relational models for distributions on contingency tables. A relational model is generated by a class of non-empty subsets of cells and can be specified in the form:

\[
\log \delta = A^t \beta.
\] (1)

Here, \(\delta\) denotes the vector of cell parameters, probabilities or intensities, and \(A\) is the 0-1 matrix whose rows are the indicators of generating subsets. A hierarchical log-linear model [cf. Bishop, Fienberg, and Holland, 1975] applies to a table which is a Cartesian product, and the model is generated by a collection of cylinder sets corresponding to marginals of the table and thus is a special case of a relational model. If the row space of \(A\) contains the vector \(1^t = (1, \ldots, 1)\), as in the case of hierarchical log-linear models, then the model is said to include the overall effect. A model with the overall effect can be parameterized to include a common parameter in every cell, often called the normalizing constant. The models without the overall effect cannot be parameterized in such a way. The characteristics, including the properties of the MLE, of relational models with and without the overall effect are fairly different. If the observed frequency distribution is \(y\), then, when the overall effect is not present, the MLE does not preserve \(Ay\) for probabilities and \(1'y\) for intensities. An IPF-type algorithm based on Bregman projections, to compute the MLEs, was given in Klimova and Rudas [2014a]. The Bregman divergence between two distributions is a generalization of the Kullback-Leibler divergence, but, unlike the latter, stays non-negative whether or not the two distributions have the same total. This property is essential for relational model for intensities without the overall effect as these models may include distributions with distinct totals.

A necessary and sufficient condition for the existence of the maximum likelihood estimates of the cell parameters under relational models is obtained in Section 2. The MLE for \(y\) exists if and only if there is a positive vector \(\psi\) such that \(A\psi = Ay\). This is literally the same condition as the one that applies to log-linear models.

In Section 3, the extension of relational models is studied. The extended relational model is defined as the set of distributions parameterized by the elements of an algebraic variety associated with the model matrix of the original relational model. It is shown that this set is equal to the closure of the original model with respect to both the pointwise convergence and the Bregman divergence.

In Section 4, a polyhedral condition for the existence of the MLE in the original or the extended relational model is formulated. If the vector of the sufficient statistics, \(Ay\), of the observed distribution is not contained in any of the faces of the polyhedral cone associated with the model matrix, the MLE exists in the original model, and otherwise, it does in the extended model. This condition is the same as for the log-linear case, but the proof is very different. The multiplicative representation of the distributions in the extended model and the existence of the MLEs of the model parameters are also discussed in this section. Finally, the generalized iterative proportional fitting procedure suggested in Klimova and Rudas [2014a] is extended to the case of observed zeros.

While the conditions of the existence of the MLE in the generality considered in this paper may be formulated to coincide with the known conditions for the case of log-linear models, the proofs turn out to be more involved. Also, the algorithm to obtain that the MLEs is more complex. The additional complications come from properties of the MLE when the overall effect is not present. In fact, Lauritzen [1996, p. 75] mentioned the existence of models
without the overall effect, which he called the “constant function”, but to avoid difficulties
did not consider them. On the other hand, such models have been used in practice, see
Klimova et al. [2012], Klimova and Rudas [2014a].

2 MLE under relational models

Let \( Y_1, \ldots, Y_K \) be discrete random variables with finite ranges, and the vector \( I \) be their
joint sample space. Here, \( I \) may also be a proper subset of the Cartesian product of the
ranges of the variables. A distribution \( P = P_\delta \) on \( I \) is parameterized by the cell parameters
\( \delta = \{ \delta_i, \text{ for } i \in I \} \), where the components of \( \delta \) are either probabilities:
\( \delta_i \equiv p_i \in (0, 1) \), with \( \sum_{i \in I} p_i = 1 \), or intensities: \( \delta_i \equiv \lambda_i > 0 \), for all \( i \in I \). Let \( P \) denote the set of positive
distributions on \( I \).

Let \( A \) be a 0-1 matrix of size \( J \times |I| \), which is interpreted as the indicator matrix of
\( J \) subsets generating the model. Assume that \( A \) has no zero column. A relational model
\( RM_\delta(A) \) is the following set of distributions:

\[
RM_\delta(A) = \{ P_\delta \in P : \delta_i = \prod_{j=1}^{J} \theta_j^{a_{ji}}, i \in I, \text{ for some } \theta \in \mathbb{R}_{>0}^J \},
\]

(2)

where \( \theta = (\theta_1, \ldots, \theta_J) \in \mathbb{R}_{>0}^J \) denotes the vector of parameters associated with the generating
subsets. Under the model, the cell parameters are equal to the products of the parameters
\( \theta \) corresponding to the subsets to which the cell belongs. In the sequel, the components of
\( \theta \) are referred to as the multiplicative parameters, and \( A \) is assumed to be of full row rank.
In fact, the model \( RM_\delta(A) \) is uniquely determined by the row space of its model matrix,
\( \mathcal{R}(A) \). The relational models for which \( 1' \in \mathcal{R}(A) \) are said to include the overall effect.

A dual representation of a relational model \( RM_\delta(A) \) can be obtained using the kernel
basis matrix \( D \), whose rows, \( d_1, \ldots, d_K \), are a basis of \( \text{Ker}(A) \). In this representation, any
distribution in the model satisfies

\[
D \log \delta = 0,
\]

(3)

which can be re-written using the generalized odds ratios:

\[
\delta^{d_i^+} / \delta^{d_i^-} = 1, \quad \delta^{d_2^+} / \delta^{d_2^-} = 1, \quad \ldots \quad \delta^{d_K^+} / \delta^{d_K^-} = 1.
\]

(4)

Here, \( d^+ \) and \( d^- \) denote, respectively, the positive and negative parts of a vector \( d \) [Klimova et al.,
2012].

The properties of the maximum likelihood estimators under relational models are re-
viewed next. Let \( Y = (Y_1, \ldots, Y_K) \) be a random variable that has a multivariate Poisson
distribution parameterized by \( \delta \equiv \lambda \) or a multinomial distribution parameterized by \( N \) and
\( \delta \equiv p \). Let \( y \) be a realization of \( Y \), and

\[
\tau = \begin{cases} y, & \text{if } \delta \equiv \lambda, \\ y / (1'y), & \text{if } \delta \equiv p. \end{cases}
\]

(5)
If the MLE $\hat{\delta}_y$ of the cell parameters under the model $RM_\delta(A)$ exists, it is the unique solution to the system of equations:

$$\begin{align*}
A\delta &= \gamma A\tau, \\
D \log \delta &= 0, \\
1'\delta &= 1 \quad \text{(only for } \delta \equiv p). \\
\end{align*}$$

The value of $\gamma$ is called the adjustment factor. If $RM_\delta(A)$ is a model for probabilities with the overall effect or a model for intensities, then $\gamma = 1$ for every $y$. If $RM_\delta(A)$ is a model for probabilities without the overall effect, then the value of $\gamma$ depends on $y$ [Klimova et al., 2012]. A necessary and sufficient condition for the existence of the MLE under relational models is given next.

**Lemma 2.1.** Let $y$ be the vector of observed frequencies under Poisson or multinomial sampling, and let $RM_\delta(A)$ be a relational model. If $y > 0$, the MLE $\hat{\delta}_y$ under the model exists.

**Proof.** A relational model for intensities is a regular exponential family [Klimova et al., 2012], and the standard proof applies [cf. Andersen, 1974].

In the case of probabilities, $\delta \equiv p$, the MLE, if exists, is the unique solution to (6). Klimova and Rudas [2014a, Lemma 3.5] showed that there exist $\gamma_1, \gamma_2 > 0$ such that the adjustment factor $\gamma \in [\gamma_1, \gamma_2]$. Since $\gamma y > 0$, the MLE $\hat{\lambda}_{\gamma y}$ under the model for intensities $RM_\lambda(A)$ exists for every $\gamma \in [\gamma_1, \gamma_2]$, and, by Lemma 3.6 in Klimova and Rudas [2014a], one can find a unique $\gamma^*$ such that $1'\hat{\lambda}_{\gamma^* y} = 1$. Since $\hat{\lambda}_{\gamma^* y}$ satisfies (6), $\hat{p}_y = \hat{\lambda}_{\gamma^* y}$.

**Theorem 2.2.** Let $y$ be the vector of observed frequencies under Poisson or multinomial sampling, and let $RM_\delta(A)$ be a relational model. The MLE $\hat{\delta}_y$ under the model exists if and only if there is a $\tau^* > 0$, such that $A\tau^* = A\tau$, with $\tau$ defined in (2).

**Proof.** In the case of intensities, $\delta \equiv \lambda$, the standard proof for regular exponential families [cf. Andersen, 1974] applies.

The case of probabilities, $\delta \equiv p$, is considered next. Suppose $\hat{p}_y > 0$ exists. By Corollary 4.2 in Klimova et al. [2012], $A\hat{p}_y = \gamma A\tau$ for some $\gamma > 0$. Therefore, $\hat{p}_y = \gamma \tau + d$, for some $d \in Ker(A)$. Consider

$$\tau^* = \frac{1}{\gamma} \hat{p}_y = \tau + \frac{1}{\gamma} d > 0.$$ 

Since $\frac{1}{\gamma} d \in Ker(A)$, $A\tau^* = A\tau$, as required.

To prove the converse, assume that there exists a $\tau^* > 0$, such that $A\tau^* = A\tau$. Thus, $\tau^* = \tau + d$ for some $d \in Ker(A)$. Let

$$d_1 = \frac{1}{1 + 1'd} d,$$

and note that $1 + 1'd = 1'\tau + 1'd = 1'\tau^* > 0$. Next, consider $q = (1 - 1'd_1)\tau + d_1$. Since $1'q = (1 - 1'd_1) + 1'd_1 = 1$, and

$$q = (1 - 1'd_1)\tau + d_1 = \frac{1}{1 + 1'd} \tau + \frac{1}{1 + 1'd} d = \frac{1}{1 + 1'd} (\tau + d) > 0,$$
the vector $q$ parameterizes a (positive) probability distribution, and, by Lemma \ref{lem:MLEexists}, the MLE $\hat{p}_q$ exists. As $Aq = (1 - 1'd_1)A\tau$, the likelihoods of $q$ and of $\tau$ under $RM_{\delta}(A)$ are proportional. Therefore, the same parameter maximizes both: $\hat{p}_y = \hat{p}_\tau = \hat{p}_q$. \hfill \Box

The next example illustrates a situation when the MLE under a relational model does not exist.

**Example 2.1.** Let $RM_p(A)$ be the model for probabilities generated by

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \end{pmatrix}.$$ 

Let $\tau = (3/7, 3/7, 0, 1/7, 0)'$ be the parameter of the observed probability distribution. With $\theta_1 = \frac{3}{3n+4}$, $\theta_2 = \frac{n}{2}$, $\theta_3 = \frac{1}{4}$,

the distribution

$$p^{(n)} = \left( \frac{3n}{8(3n+4)}, \frac{3n}{2(3n+4)}, \frac{3n}{3n+4}, \frac{1}{4}, \frac{3n}{8(3n+4)} \right)'$$

is in the model, for $n \in \mathbb{Z}_{>0}$. Then, the kernel of the log-likelihood,

$$l(\tau, p^{(n)}) = \frac{3}{7} \log \frac{3n}{8(3n+4)} + \frac{3}{7} \log \frac{3n}{2(3n+4)} + \frac{1}{7} \log \frac{1}{4}$$

is strictly monotone increasing in $n$. On the other hand,

$$\lim_{n \to \infty} p^{(n)} = (1/8, 1/2, 0, 1/4, 1/8)'$$

is not in $RM_p(A)$, and, therefore, $\tau$ has no MLE in the model. \hfill \Box

The next section discusses the extension of relational models to include pointwise limits of sequences of distributions in the model.

## 3 Extended relational models

There is a literature on the extensions of log-linear models in terms of the polynomial varieties which are determined from the sets of odds ratios specifying the model \cite[cf. Geiger et al., 2006]{Geiger06}. A similar extension of relational models, based on their dual representation, is defined next.

Re-write the dual representation in \ref{eq:dual1} and \ref{eq:dual2} in terms of the cross-product differences:

$$\delta^{d_1} - \delta^{d_2} = 0, \quad \delta^{d_2} - \delta^{d_3} = 0, \quad \ldots \quad \delta^{d_k} - \delta^{d_k} = 0,$$
with the convention that $0^0 = 1$. This representation links the relational model to the algebraic variety associated with the matrix $A$:

$$X_A = \left\{ \delta \in \mathbb{R}_{\geq 0}^I : \delta^{d^+} = \delta^{d^-}, \forall d \in \text{Ker}(A) \right\}. \quad (8)$$

Let $\mathcal{P}$ be the set of non-negative distributions on $I$.

**Definition 3.1.** The extended relational model for intensities, $\overline{RM}_\lambda(A)$, is the subset of $\overline{\mathcal{P}}$ parameterized by

$$\lambda \in X_A. \quad (9)$$

The extended relational model for probabilities, $\overline{RM}_p(A)$, is the subset of $\overline{\mathcal{P}}$ parameterized by

$$p \in X_A \cap \Delta_{|I|-1}, \quad (10)$$

where $\Delta_{|I|-1}$ is the $(|I| - 1)$-dimensional simplex.

Note that every positive vector in $X_A$ satisfies (3) and (4). Thus, if for some $\delta > 0$, $P_\delta \in \overline{RM}_\delta(A)$ then also $P_\delta \in \overline{RM}_{\delta^+}(A)$.

Let $a_1, \ldots, a_{|I|}$ denote the columns of $A$, and let $C_A$ be the set of all non-negative linear combinations of these columns:

$$C_A = \{ t \in \mathbb{R}^J_{\geq 0} : \exists \delta \in \mathbb{R}^I_{\geq 0} \ t = A\delta \}. \quad (11)$$

The relative interior of $C_A$, $\text{relint}(C_A)$, comprises such $t \in \mathbb{R}^J_{> 0}$, for which there exists a (strictly) positive $\delta$ that satisfies $t = A\delta$.

The set $C_A$ is a polyhedral cone in $\mathbb{R}^J$. If an affinely independent set $a_{i_1}, a_{i_2}, \ldots, a_{i_f}$ of columns of $A$ spans a proper face of $C_A$, the set of indices $F = \{i_1, i_2, \ldots, i_f\}$ is called facial [cf. Grünbaum, 2003; Geiger et al., 2006]. The facial sets of $A$ are determined by its row space [cf. Fienberg and Rinaldo, 2012b]. If $t \in C_A \setminus \text{relint}(C_A)$, then $t$ is said to lie on a face of $C_A$. In that case, there is a facial set $F = F(t)$, such that

$$t = s_1 a_{i_1} + \cdots + s_f a_{i_f}. \quad (12)$$

The properties of facial sets which will be used in the proofs of the theorems to follow are stated next.

**Lemma 3.1.** Let $A$ be the model matrix of a relational model, and let $F$ be a facial set of $A$. Then:

(i) There exists a $c \in \mathbb{R}^J$, such that $c^t a_i = 0$ for any $i \in F$ and $c^t a_i > 0$ for any $i \notin F$.

(ii) For any $d \in \text{Ker}(A)$, either both $\text{supp}(d^+) \subseteq F$ and $\text{supp}(d^-) \subseteq F$ or both $\text{supp}(d^+) \nsubseteq F$ and $\text{supp}(d^-) \nsubseteq F$.

(iii) For any $\delta \in X_A$, either $\text{supp}(\delta) = I$ or $\text{supp}(\delta)$ is a facial set of $A$.

(iv) If $F$ is a facial set of $A$, there exists a $\delta \in X_A$, such that $\text{supp}(\delta) = F$. 

The statements of the lemma were proved by Geiger et al. [2006] and Rauh, Kahle, and Ay [2011] for models of type (1) when the overall effect is present. Their proofs do not rely on the latter characteristic and thus apply here.

The condition of Theorem 2.2 is related to the concept of a facial set.

**Lemma 3.2.** There exists a \( \tau^* > 0 \), such that \( A \tau^* = A \tau \), if and only if \( \text{supp}(\tau) \) is not contained in any facial set of \( A \).

**Proof.** Suppose there exists a \( \tau^* > 0 \), such that \( A \tau^* = A \tau \), and thus \( d = \tau^* - \tau \in \text{Ker}(A) \) and \( \tau + d > 0 \).

Let \( F \) be a facial set of \( A \). If both \( d^+ \subseteq F \) and \( d^- \subseteq F \), then \( d_i = 0 \) for all \( i \notin F \). Since \( \tau + d > 0 \), \( \tau_i + d_i = \tau_i > 0 \) for all \( i \notin F \). Therefore, \( \text{supp}(\tau) \) is not contained in \( F \). Otherwise, see Lemma 3.1, both \( d^+ \nsubseteq F \) and \( d^- \nsubseteq F \), and there exists an \( i \notin F \) such that \( d_i < 0 \). If \( \tau_i \) was zero, then \( \tau_i + d_i \) would be negative, which contradicts the initial assumption \( \tau + d > 0 \). Therefore, \( \tau_i \) has to be positive, which implies that \( \text{supp}(\tau) \) is not contained in \( F \).

To prove the converse, assume that \( \text{supp}(\tau) \) is not contained in any facial set \( F \). Suppose the equation \( A \tau = A \tau^* \) has no (strictly) positive solution in \( \tau^* \), and therefore, \( A \tau \notin \text{relint}(CA) \). A non-negative solution always exists, and thus \( A \tau \) belongs to a face of \( CA \). Then (12) holds for \( t = A \tau \) for some facial set \( F \); without loss of generality, \( F = \{1, \ldots, f\} \):

\[
A \tau = s_1a_1 + \cdots + s_f a_f.
\]

Hence,

\[
(\tau_1 - s_1)a_1 + \cdots + (\tau_f - s_f)a_f + \tau_{f+1}a_{f+1} + \cdots + \tau_{|I|}a_{|I|} = 0. \tag{13}
\]

Multiplying both sides of (13) by a vector \( c \), such that \( c' a_i = 0 \) for \( i \in F \) and \( c' a_i > 0 \) for \( i \notin F \), leads to:

\[
\tau_{f+1} = 0, \ldots, \tau_{|I|} = 0,
\]

which means that \( \text{supp}(\tau) \subset F \). This contradicts the initial assumption that \( \text{supp}(\tau) \) is not contained in any facial set.

The following theorem describes the structure of the parameter set of the extended relational model.

**Theorem 3.3.** The extended relational model \( \overline{RM}_\delta(A) \) is the closure of the relational model \( RM_\delta(A) \) in the topology of pointwise convergence: \( \overline{RM}_\delta(A) = \text{cl}(RM_\delta(A)) \).

**Proof.** The proof extends the arguments given by Geiger et al. [2006] and Rauh et al. [2011]. It will be shown first that for any distribution in \( RM_\delta(A) \) there exists a sequence of distributions in \( RM_\delta(A) \) that converges to it pointwise.

Let \( P_{\delta^*} \in \overline{RM}_\delta(A) \). By Lemma 3.1, as \( \delta^* \in X_A \), \( F = \text{supp}(\delta^*) \) is either \( I \) or a facial set of \( A \). If \( F = I \), then \( \delta^* > 0 \), and the statement holds with \( P^{(n)} \equiv P_{\delta^*} \). Assume that \( F \subsetneq I \). For simplicity of exposition, let \( F = \{1, \ldots, f\} \), and then \( \delta^* = (\delta_1^*, \ldots, \delta_f^*, 0, \ldots, 0) \).

First, find \( \eta_1, \ldots, \eta_J > 0 \) that satisfy:

\[
\prod_{j=1}^{J} q_j^{\eta_i} = \delta_i^* \quad \text{for } i \in F.
\]
Further, for \( i \in \mathbb{Z}_{>0} \),

\[
\lambda_i^{(n)} = \prod_{j=1}^{J} (n^{-c_j} \eta_j)^{a_{ji}}, \quad i \in \mathcal{I}.
\]

The existence of such \( \theta \)'s can be proved using the same argument as [Geiger et al. 2006, p.28] gave for the case of extended log-linear models. By Lemma 3.1 there exists a \( \mathbf{c} = (c_1, \ldots, c_J) \in \mathbb{R}^J \), such that \( \mathbf{c}' \mathbf{a}_i = 0 \) for all \( i \in F \) and \( \mathbf{c}' \mathbf{a}_i > 0 \) for any \( i \notin F \). Order the columns of \( \mathbf{A} \) so that \( c_1 > 0 \), and then order the rows of \( \mathbf{A} \) so that \( a_{11} = 1 \).

If \( \delta \equiv \lambda \), set, for \( n \in \mathbb{Z}_{>0} \),

\[
\lim_{n \to \infty} \lambda_i^{(n)} = \lim_{n \to \infty} n^{-\mathbf{c}' \mathbf{a}_i} \prod_{j=1}^{J} (n^{-c_j} \eta_j)^{a_{ji}} = \left\{ \begin{array}{ll}
\delta_i^*, & \text{if } i \in F, \\
0, & \text{if } i \notin F,
\end{array} \right.
\]

thus \( \mathbf{P}^{(n)} \to \mathbf{P}_{\delta^*} \) pointwise, as \( n \to \infty \).

If \( \delta \equiv \mathbf{p} \), take

\[
\eta_1^{(n)} = \frac{1 - \sum_{a_{11}=0}^{\infty} \prod_{j=1}^{J} (n^{-c_j} \eta_j)^{a_{ji}}}{\sum_{a_{11}=0}^{\infty} \prod_{j=1}^{J} (n^{-c_j} \eta_j)^{a_{ji}}},
\]

and set

\[
\mathbf{P}_i^{(n)} = (\eta_1^{(n)})^{a_{1i}} \prod_{j=2}^{J} (n^{-c_j} \eta_j)^{a_{ji}}, \quad i \in \mathcal{I}.
\]

The choice of \( \eta_1^{(n)} \) implies that \( \mathbf{P}_1^{(n)} = 1 \). As \( \mathbf{p}^{(n)} = (\mathbf{P}_1^{(n)}, \ldots, \mathbf{P}_i^{(n)})' \) is positive and satisfies (2) with \( \theta_1 = \eta_1^{(n)} \), \( \theta_j = n^{-c_j} \eta_j \), for \( j = 2, \ldots, J \), the distribution \( \mathbf{P}^{(n)} \) parameterized by \( \mathbf{p}^{(n)} \) is in \( RM_{\mathbf{P}}(\mathbf{A}) \). Next, because \( \mathbf{c}' \mathbf{a}_i = 0 \) if \( i \in F \),

\[
\lim_{n \to \infty} n^{c_{1i}} \eta_1^{(n)} = \lim_{n \to \infty} \frac{n^{c_{1i}} (1 - \sum_{a_{11}=0}^{\infty} \prod_{j=2}^{J} \eta_j^{a_{ji}} - \sum_{a_{11}=0}^{\infty} \prod_{j=2}^{J} \eta_j^{a_{ji}})}{n^{c_{1i}} \sum_{a_{11}=0}^{\infty} \prod_{j=2}^{J} \eta_j^{a_{ji}}} = \frac{1 - \sum_{i \in F: a_{11}=0}^{\infty} \prod_{j=2}^{J} \eta_j^{a_{ji}}}{\sum_{i \in F: a_{11}=1}^{\infty} \prod_{j=2}^{J} \eta_j^{a_{ji}}} = \eta_1. \quad (14)
\]

Further, for \( i \in \mathcal{I} \), using (14),

\[
\lim_{n \to \infty} \mathbf{P}_i^{(n)} = \lim_{n \to \infty} n^{a_{1i}c_{1i} - \mathbf{c}' \mathbf{a}_i} (\eta_1^{(n)})^{a_{1i}} \prod_{j=2}^{J} \eta_j^{a_{ji}} = \lim_{n \to \infty} n^{-\mathbf{c}' \mathbf{a}_i} (n^{c_{1i}} \eta_1^{(n)})^{a_{1i}} \prod_{j=2}^{J} \eta_j^{a_{ji}}
\]

\[
= \lim_{n \to \infty} n^{-\mathbf{c}' \mathbf{a}_i} (\eta_1)^{a_{1i}} \prod_{j=2}^{J} \eta_j^{a_{ji}} = \lim_{n \to \infty} n^{-\mathbf{c}' \mathbf{a}_i} \prod_{j=1}^{J} \eta_j^{a_{ji}} = \left\{ \begin{array}{ll}
\delta_i^*, & i \in F, \\
0, & i \notin F.
\end{array} \right.
\]
Hence, $P^{(n)} \rightarrow P^*$ pointwise, as $n \rightarrow \infty$.

Therefore, $\overset{RM}_\delta(A) \subset cl(RM_\delta(A))$.

To prove the converse, choose a $P^* \in cl(RM_\delta(A))$. Then, $P^*$ is a pointwise limit of a sequence of distributions in $RM_\delta(A)$, and $\delta^*$ is the pointwise limit of a sequence in $\mathcal{X}_A$. As $\mathcal{X}_A$ is closed in the topology of pointwise convergence [cf. Geiger et al., 2006], $\delta^* \in \mathcal{X}_A$. If $\delta \equiv p$, both $\delta^*$ and the sequence converging to it belong to the simplex $\Delta_{|\mathcal{Z}|-1}$. Therefore, $P^* \in \overset{RM}_\delta(A)$, and the proof is complete.

The following example illustrates the construction given in the proof.

**Example 2.1** (revisited):

Let $p = (p_1, p_2, 0, p_4, p_5)'$, with $p_1, p_2, p_4, p_5 > 0$, be the parameter of a distribution in $\overset{RM}(A)$. The support of $p$ is a facial set, with $c = (c_1, c_2, c_3)' = (1, -1, 0)'$. Find some $\eta_1, \eta_2, \eta_3 > 0$ that satisfy:

$$\eta_1 \eta_2 \eta_3 = p_1, \quad \eta_1 \eta_2 = p_2, \quad \eta_3 = p_4, \quad \eta_1 \eta_2 \eta_3 = p_5.$$

Set

$$\eta_1^{(n)} = \frac{1 - \eta_3 n^{-c_3}}{1 + 2\eta_2 n^{-c_2} \eta_3 n^{-c_3} + \eta_2 n^{-c_2}}, \quad \eta_2^{(n)} = \eta_2 n^{-c_2}, \quad \eta_3^{(n)} = \eta_3 n^{-c_3}.$$

The distribution $P^{(n)}$, parameterized by

$$P^{(n)} = \left(\theta_1^{(n)} \theta_2^{(n)} \theta_3^{(n)}, \quad \theta_1^{(n)} \theta_2^{(n)}, \quad \theta_1^{(n)}, \quad \theta_2^{(n)}, \quad \theta_3^{(n)}\right)'$$

is positive, satisfies the multiplicative structure of (2), and thus is in $RM_p(A)$. Finally, using that $c_1 > 0, c_1 + c_2 + c_3 = 0$, and $c_1 + c_2 = 0$:

$$\lim_{n \rightarrow \infty} \eta_1^{(n)} \eta_2^{(n)} \eta_3^{(n)} = \lim_{n \rightarrow \infty} \frac{(1 - \eta_3 n^{-c_3}) \eta_2 n^{-c_2} \eta_3 n^{-c_3}}{1 + 2\eta_2 n^{-c_2} \eta_3 n^{-c_3} + \eta_2 n^{-c_2}} = \lim_{n \rightarrow \infty} \frac{n^{c_1} (1 - \eta_3) \eta_2 \eta_3}{1 + 2n^{c_1} \eta_2 \eta_3 + n^{c_1} \eta_2}$$

$$= \lim_{n \rightarrow \infty} \frac{(1 - \eta_3) \eta_2 \eta_3}{n^{-c_1} + 2\eta_2 \eta_3 + \eta_2} = \eta_1 \eta_2 \eta_3,$$

$$\lim_{n \rightarrow \infty} \eta_1^{(n)} \eta_2^{(n)} = \lim_{n \rightarrow \infty} \frac{(1 - \eta_3 n^{-c_3}) \eta_2 n^{-c_2}}{1 + 2\eta_2 n^{-c_2} \eta_3 n^{-c_3} + \eta_2 n^{-c_2}} = \lim_{n \rightarrow \infty} \frac{n^{c_1} (1 - \eta_3) \eta_2}{1 + 2n^{c_1} \eta_2 \eta_3 + n^{c_1} \eta_2} = \eta_1 \eta_2,$$

$$\lim_{n \rightarrow \infty} \eta_1^{(n)} = \lim_{n \rightarrow \infty} \frac{1 - \eta_3 n^{-c_3}}{1 + 2\eta_2 n^{-c_2} \eta_3 n^{-c_3} + \eta_2 n^{-c_2}} = 0,$$

$$\lim_{n \rightarrow \infty} \eta_3^{(n)} = \lim_{n \rightarrow \infty} \eta_3 n^{-c_3} = \eta_3.$$

Other ways of extending exponential families have also been considered in the literature. The closure of exponential families using the Kullback-Leibler divergence was described for
regular families by Brown [1988], among others, and for full families by Csiszár and Matúš [2003]. However, both of these approaches rely on the presence of the overall effect, which implies, through the possibility of normalization, that the Kullback-Leibler divergence is non-negative and Pinsker’s inequality [cf. Csiszár, 1975] holds. In the generality considered in the present paper, the approach does not apply, and the Bregman divergence is used to extend relational models.

Let \( D(\cdot||\cdot) \) denote the Bregman divergence between two vectors \( t, u \in \mathbb{R}_{\geq 0}^{|I|} \), associated with the function \( F(x) = \sum_{i \in I} x(i) \log x(i) \):

\[
D(t||u) = \sum_{i \in I} t(i) \log \left( \frac{t(i)}{u(i)} \right) + \left( \sum_{i \in I} u(i) - \sum_{i \in I} t(i) \right).
\]

(15)

Under the convention \( 0 \log 0 = 0 \), \( D(t||u) \) is also defined for non-negative \( t \) and \( u \) if \( \text{supp}(t) \subseteq \text{supp}(u) \). The function \( D(t||u) \) is non-negative, and \( D(t||u) = 0 \) if and only if \( t = u \). For any \( u^* \in \mathbb{R}_{\geq 0}^{|I|} \) and for any convex set \( S \subseteq \mathbb{R}_{\geq 0}^{|I|} \) there exists a unique \( u^* \in \mathbb{R}_{\geq 0}^{|I|} \), such that

\[
D(u^*||u) = \min_{z \in S} D(z||u),
\]

(16)

see Bregman [1967]. This \( u^* \) is called the \( D \)-projection, or the Bregman projection, of \( u \) on \( S \). If \( P_p \) and \( Q_q \) are probability distributions, then \( D(p||q) = I(P_p||Q_q) \), the Kullback-Leibler divergence between \( P_p \) and \( Q_q \).

Let \( \overline{RM}_\delta(A) \) be the closure of \( RM_\delta(A) \) with respect to the Bregman divergence:

\[
\overline{RM}_\delta(A) = \left\{ P_\delta \in \mathcal{P} : \exists P_{\delta(n)} \in RM_\delta(A), n \in \mathbb{N}, \text{ such that } D(\delta||\delta^{(n)}) \to 0 \text{ as } n \to \infty \right\}.
\]

**Theorem 3.4.** The extended relational model \( \overline{RM}_\delta(A) \) is identical to the closure of the relational model \( RM_\delta(A) \) with respect to the Bregman divergence: \( \overline{RM}_\delta(A) = \overline{RM}_\delta(A) \).

**Proof.** Let \( P_{\delta^*} \in \overline{RM}_\delta(A) \). Then, there exists a sequence \( P_{\delta(n)} \in RM_\delta(A) \) such that \( P_{\delta(n)} \to P_{\delta^*} \) pointwise, as \( n \to \infty \). The function \( D(\delta^*||\delta^{(n)}) \) is defined and continuous for \( \delta^{(n)} > 0 \), even if some of the components of \( \delta^* \) are zero. Therefore, \( D(\delta^*||\delta^{(n)}) \to 0 \), as \( n \to \infty \).

Suppose \( P_{\delta^*} \in \overline{RM}_\delta(A) \), and, thus, there exists a sequence \( P_{\delta(n)} \in RM_\delta(A) \), such that:

\[
D(\delta^*||\delta^{(n)}) \to 0 \text{ as } n \to \infty.
\]

Therefore, \( D(\delta^*||\delta^{(n)}) \leq 1 \) for all large enough \( n \). Since the set \( \{ \delta \geq 0 : D(\delta^*||\delta) \leq 1 \} \) is compact in \( \mathbb{R}_{|I|} \) [Bregman, 1967], there exists a subsequence \( \delta^{(nk)} \) that converges pointwise to \( \delta^* \), as \( k \to \infty \). \( \Box \)

A relational model \( RM_\delta(A) \) is a multiplicativity family of distributions; the conditions under which the extended model \( \overline{RM}_\delta(A) \) is also a multiplicativity family are studied next.

A distribution \( P_\delta \in \mathcal{P} \) is said to factor according to a matrix \( A \) if it has a representation given in (2), with \( \theta = (\theta_1, \ldots, \theta_J)' \geq 0 \). Every distribution in a relational model factors according to the model matrix \( A \) but, as the next example demonstrates, this is not necessary the case for a distribution in an extended model.
Example 2.1 (revisited):
The distribution \( P \in \mathbb{R}M_p(A) \) parameterized by \( p = (p_1, p_2, 0, p_4, p_5)' \), with \( p_1, p_2, p_4, p_5 > 0 \), does not factor according to \( A \). However, it does factor according to the matrix

\[
A_1 = \begin{pmatrix}
0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 1
\end{pmatrix},
\]

which leads to the same polynomial variety: \( X_A = X_{A_1} \), because \( \text{Ker}(A) = \text{Ker}(A_1) \).

A necessary and sufficient condition of the existence of such a factorization for a distribution in an extended relational model is given next.

Theorem 3.5. A distribution \( P_\delta \in \overline{\mathbb{R}M}_\delta(A) \) factors according to \( A \) if and only if for any \( i_0 \notin \text{supp}(\delta) \) there exists an index \( j = j(i_0) \in \{1, \ldots, J\} \) such that \( a_{ji} = 0 \) for all \( i \in \text{supp}(\delta) \).

The condition of the theorem, called the \( A \)-feasibility of \( \text{supp}(\delta) \), means that a generating subset which contains a zero cell of the distribution does not include any positive cell. For extended log-linear models, this condition was proved in Geiger et al. [2006] and Rauh et al. [2011]. The proofs given did not actually rely on the presence of the overall effect and thus apply here.

Maximum likelihood estimation in the extended relational model is studied next.

4 MLE in the extended model

Let \( F \) be a facial set, and let \( A_F \) denote the sub-matrix of \( A \) comprising the columns with indices in \( F \), and \( \delta_F \) denote the sub-vector of \( \delta \) with indices in \( F \). The following result extends Theorem 9 in Fienberg and Rinaldo [2012a].

Theorem 4.1. Let \( y \) be the vector of observed frequencies under Poisson or multinomial sampling, and let \( \mathbb{R}M_\delta(A) \) be a relational model. Consider \( \tau \) defined in (3), and assume that \( \text{supp}(\tau) \subset \mathcal{I} \).

(i) If for all facial sets \( F \), \( \text{supp}(\tau) \not\subset F \), then the MLE \( \hat{\delta}_y \) under the model \( \mathbb{R}M_\delta(A) \) exists, and is also the MLE under \( \mathbb{R}M_\delta(A) \): \( \hat{\delta}_y = \tilde{\delta}_y \). Otherwise,

(ii) Let \( F \) be the smallest facial set such that \( \text{supp}(\tau) \subset F \). Then the MLE \( \hat{\delta}_{y,F} \) of \( \delta_F \) under the model \( \mathbb{R}M_\delta(A_F) \) exists, and \( \hat{\delta}_y = (\hat{\delta}_{y,F}, 0_{\mathcal{I}\setminus F}) \) is the MLE under the model \( \overline{\mathbb{R}M}_\delta(A) \).

(iii) The MLE \( \tilde{\delta}_y \) under \( \overline{\mathbb{R}M}_\delta(A) \) always exists and is the unique point of \( X_A \) which satisfies:

\[
A\delta = \gamma A\tau, \quad \text{for some } \gamma > 0; \tag{17}
\]

\[
1'\delta = 1 \quad (\text{only for } \delta \equiv p).
\]

11
The vector \( \tilde{\delta}_y \) is called the extended MLE of \( \delta \) under the relational model. The proof is given in the Appendix and uses the following lemma:

**Lemma 4.2.** If \( F \) is a facial set of \( A \), then, for any \( \delta_F \in X_{A_F} \), \( \delta = (\delta_F, 0_{I \setminus F}) \in X_A \). 

**Proof.** Take an arbitrary \( d \in \text{Ker}(A) \). As \( F \) is a facial set of \( A \), by Lemma 3.1(ii), exactly one of the following holds:

- \( \text{supp}(d^+) \subseteq F \) and \( \text{supp}(d^-) \subseteq F \), or
- \( \text{supp}(d^+) \not\subseteq F \) and \( \text{supp}(d^-) \not\subseteq F \).

In the first case, there exists a \( d_F \in \text{Ker}(A_F) \), such that \( d = (d_F, 0_{I \setminus F}) \). Since \( \delta_F \in X_{A_F} \),

\[
(\delta)^{d^+} = (\delta_F)^{d^+_F} \cdot (0_{I \setminus F})^{0_{I \setminus F}} = (\delta_F)^{d^+_F} \cdot (0_{I \setminus F})^{0_{I \setminus F}} = (\delta)^{d^-}.
\]

In the second case, there exist such \( i_1, i_2 \notin F \) that \( d_{i_1} > 0 \) and \( d_{i_2} < 0 \), and thus,

\[
(\delta)^{d^+} = (\delta_F)^{d^+_F} \cdot 0 = (\delta_F)^{d^+_F} \cdot 0 = (\delta)^{d^-}.
\]

As \( (\delta)^{d^+} = (\delta)^{d^-} \) for any \( d \in \text{Ker}(A), \delta \in X_A \). \( \square \)

The next theorem establishes a condition under which the maximum likelihood estimates of the model parameters under an extended relational model exist:

**Theorem 4.3.** Assume that the MLE \( \hat{\delta} \) under the extended relational model \( \text{RM}_{\delta}(A) \) exists. The maximum likelihood estimates of the model parameters \( \theta \) exist if and only if \( \text{supp}(\hat{\delta}) \) is \( A \)-feasible.

**Proof.** By Theorem 3.3, the distribution parameterized by \( \hat{\delta} \) factors according to \( A \) if and only if \( \text{supp}(\hat{\delta}) \) is \( A \)-feasible. In this case \( \hat{\delta}(i) = \prod_{j=1}^J \hat{\theta}_{i,j} \) for all \( i \in I \), and, by uniqueness, \( \hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_J)' \) are the maximum likelihood estimates of the model parameters. \( \square \)

If \( \text{supp}(\hat{\delta}) \) is not \( A \)-feasible, then the distribution parameterized by \( \hat{\delta} \) is the limit of a sequence of the positive distributions in the model which factor according to \( A \). Although the cell parameters of these distributions can be factored using some model parameters \( \theta^{(n)} > 0 \), the limits of individual components of \( \theta^{(n)} \), as \( n \to 0 \), may not exist. In the case of the log-linear models this fact was illustrated by [Rinaldo 2006]. The same situation occurs in the construction of Example 2.1 where \( \theta^{(n)}_2 \to \infty \) as \( n \to \infty \).

As Theorem 4.1 implies, the MLE in the extended relational model can be obtained using the MLE in a non-extended model. [Klimova and Rudas 2014a] proposed a generalized iterative scaling procedure, called G-IPF, for computing the MLE under (non-extended) relational models. The algorithm relies on the condition that \( A \tau > 0 \). Every iteration of this procedure implements the following algorithm, IPF(\( \gamma \)), for a specific value of \( \gamma \).

**IPF(\( \gamma \)) Algorithm:**

Set \( n = 0; \quad \delta^{(0)}_{\gamma}(i) = 1 \) for all \( i \in I \), and proceed as follows.
Step 1: Find \( j \in \{1, 2, \ldots, J\} \), such that \( n + 1 \equiv j \mod J \);

Step 2: Compute

\[
\delta_\gamma^{(n+1)}(i) = \delta_\gamma^{(n)}(i) \left( \frac{A_j \tau}{A_j \delta_\gamma^{(n)}} \right)^{a_{ji}} \text{ for all } i \in \mathcal{I}.
\] (18)

Step 3: While \( \gamma A_j \tau \neq A_j \delta_\gamma^{(n+1)} \) for at least one \( j \), set \( n = n + 1 \), go to Step 1.

Step 4: Set \( \delta_\gamma^* = \delta_\gamma^{(n)} \), and finish. \( \square \)

The G-IPF algorithm commences with executing IPF(\( \gamma \)) for \( \gamma = 1 \), which is sufficient to compute the MLE in the case of probabilities with the overall effect and in the case of intensities. If in the case of probabilities the overall effect is not present, G-IPF updates \( \gamma \) and calls IPF(\( \gamma \)) again. The procedure is repeated until, for some \( \gamma \), the limit vector \( \delta_\gamma^* \) sums to 1, and thus is a parameter of a non-negative probability distribution. The variant of G-IPF, which employs the bisection method to update \( \gamma \), is described in the following.

**G-IPF Algorithm:**

If \( \delta \equiv \lambda \), compute \( \hat{\lambda} \) using IPF(1), and finish.

If \( \delta \equiv p \), compute \( p^* \) using IPF(1).
If \( 1p^* = 1 \), set \( \hat{p} = p^* \), and finish. Otherwise, compute \( \gamma_L = (1'A\tau)^{-1}, \gamma_R = \min \{1/A_1\tau, \ldots, 1/A_J\tau\} \), and proceed as follows:

Step 1: Find \( \delta_{(\gamma_L+\gamma_R)/2}^* \) using IPF(\( \gamma \)).

Step 2: While \( 1\delta_{(\gamma_L+\gamma_R)/2}^* \neq 1 \),

if \( 1\delta_{(\gamma_L+\gamma_R)/2}^* < 1 \), set \( \gamma_L = \frac{\gamma_L+\gamma_R}{2} \),
else set \( \gamma_R = \frac{\gamma_L+\gamma_R}{2} \),
go to Step 1.

Step 3: Set \( \hat{p} = \delta_{(\gamma_L+\gamma_R)/2}^* \), and finish. \( \square \)

If \( A\tau > 0 \), the G-IPF algorithm applies to the extended case directly.

**Theorem 4.4.** Let \( y \) be the vector of observed frequencies under Poisson or multinomial sampling, with \( \tau \) defined in (5), and let \( \text{RM}_\delta(A) \) be a relational model. Assume that \( A\tau > 0 \). The G-IPF algorithm converges to the MLE \( \hat{\delta}_y \) under \( \text{RM}_\delta(A) \).
Proof. Since $A\tau > 0$, the IPF-sequence $\delta^{(n)}$ defined in (18) is positive, and the proof of its convergence in Klimova and Rudas [2014a, Theorem 3.2] applies. In particular, the limit of the sequence, $\delta^{*}_{\gamma}$, satisfies $A\delta^{*}_{\gamma} = \gamma A\tau$, and, for an arbitrary kernel basis matrix $D$, $\text{Dlog } \delta^{(n)} = 0$ for all $n \in \mathbb{Z}_{\geq 0}$. The latter implies that $\delta^{(n)} \in X_{A}$ for all $n$, and, as $X_{A}$ is a closed set in $\mathbb{R}_{\geq 0}$, $\delta^{*} \in X_{A}$.

Let $\delta_{1}^{*}$ be the limit vector obtained from IPF(1), and thus $\delta_{1}^{*} \in X_{A}$ and $A\delta_{1}^{*} = A\tau$.

Suppose $\delta \equiv \lambda$. Then, as (17) holds for $\delta_{1}^{*}$ with $\gamma = 1$, Theorem 4.1(iii) implies that $\delta_{1}^{*}$ is equal to the extended MLE: $\delta_{y}^{*} = \delta_{1}^{*}$.

Suppose $\delta \equiv p$. First, assume that the overall effect is present, and thus there exists a $k \in \mathbb{R}_{\geq 0}^{J}$, such that $1' = k' A$. The latter yields that $1' \delta_{1}^{*} = k' A \delta_{1}^{*} = k' A \tau = 1' \tau = 1$. Therefore, (17) holds for $\delta_{1}^{*}$ with $\gamma = 1$. By Theorem 4.1(iii), $\delta_{y}^{*} = \delta_{1}^{*}$.

Now, assume that the overall effect is not present. In this situation, G-IPF updates $\gamma$ and calls IPF($\gamma$); and this procedure is repeated until a $\gamma^{*}$ for which the IPF-limit $\delta_{\gamma}^{*}$ sums to 1 is found. Then, $\delta_{\gamma}^{*}$ satisfies (17) with $\gamma = \gamma^{*}$. By Theorem 4.1(iii), $\delta_{y}^{*} = \delta_{1}^{*}$.

Next, it is shown how G-IPF can be used if the condition $A\tau > 0$ does not hold. Let $\mathcal{J}_{0} = \{j \in \{1, \ldots, J\}: A_{j} \tau = 0\}$, and assume that $\mathcal{J}_{0} \neq \emptyset$. Further, let $\mathcal{I}_{0} = \{i \in \mathcal{I}: \exists j \in \mathcal{J}_{0} \ a_{ji} = 1\}$, and let $\mathcal{I}_{s} = \mathcal{I} \setminus \mathcal{I}_{0}$. Denote by $A_{s}$ the matrix obtained from $A$ by removing the columns with indices in $\mathcal{I}_{0}$ and by removing the zero rows, if such occur afterwards, and by $\delta_{s}$, $y_{s}$, and $\tau_{s}$ the corresponding sub-vectors of $\delta$, $y$, and $\tau$. By Theorem 4.1(iii), the MLE $\delta_{y_{s}}^{*}$ of $y_{s}$ under $\widetilde{RM}_{\delta}(A_{s})$ exists and is unique. Since $A_{s}\tau_{s} > 0$, $\delta_{y_{s}}^{*}$ can be computed using G-IPF, see Theorem 4.1 and the following holds:

**Theorem 4.5.** The MLE of $y$ under $\widetilde{RM}_{\delta}(A)$ is equal to $\delta_{y} = (\delta_{y_{s}}^{*}, 0_{\mathcal{I}_{0}})$.

Proof. In order to show that $\delta_{y} \in X_{A}$, it will first be verified that $\mathcal{I}_{s}$ is a facial set of $A$. Let $a_{i}$ be the $i$-th column of $A$, then, with $c = (0_{\mathcal{J}_{0} \setminus \mathcal{J}_{0}}, 1_{\mathcal{J}_{0}})'$, $c' a_{i} = 0$ for any $i \in \mathcal{I}_{s}$. If $i \notin \mathcal{I}_{s}$, then $a_{ji} = 1$ for some $j \in \mathcal{J}_{0}$, and thus $c' a_{i} > 0$. Therefore, $\mathcal{I}_{s}$ is a facial set of $A$. Then, by Lemma 4.1, $\delta_{y} \in X_{A}$.

Next, in the case of probabilities, the normalization condition $1' \delta_{y_{s}} = 1$ implies that $1' \delta_{y} = 1$. Further, $A_{s} \delta_{y_{s}} = \gamma A_{s} \tau_{s}$ implies that $A \delta_{y} = \gamma A \tau$.

Finally, by Theorem 4.1(iii), $\delta_{y}^{*}$ is the MLE of $y$ under $\widetilde{RM}_{\delta}(A)$.

**Appendix**

**Proof of Theorem 4.1:**

The statement (i) follows from Theorem 2.2 and Lemma 3.2.

In order to prove (ii), notice first that, the smallest facial set $F$ of $A$ which contains $\text{supp}(\tau)$ is uniquely defined. In this case, $A_{F} \tau_{F} \in \text{relint}(C_{A_{F}})$, and, therefore, $\text{supp}(\tau)$ is not contained in any facial set of $A_{F}$. By part (i) of this theorem, the MLE $\delta_{y_{F}}^{*}$ under $RM_{\delta_{F}}(A_{F})$ exists.
Let \( \tilde{\delta}_y = (\tilde{\delta}_{yp},0_{I \setminus F}) \). By Lemma 4.2, \( \tilde{\delta}_y \in \mathcal{X}_A \). If \( \delta \equiv p \), \( 1'\hat{p}_y = 1 \), and thus \( \hat{p}_y \) satisfies the normalization condition \( 1'\hat{p}_y = 1 \). It will be shown next that \( \tilde{\delta}_y \) maximizes the full log-likelihood of \( y \).

Let \( \delta \equiv \lambda \). The log-likelihood under the model \( RM_{\lambda_F}(A_F) \) is equal to

\[
l_F(\tau_F, \lambda_F) = \sum_{i \in F} \tau_i \log \lambda_{Fi} - \sum_{i \in F} \lambda_{Fi},
\]

and for any \( \lambda_F > 0 \), \( l_F(\tau_F, \lambda_F) \leq l_F(\tau_F, \hat{\lambda}_{y_F}) \).

Let \( \lambda = (\lambda'_F, 0)' \), and let \( \lambda^{(n)} \) be the sequence that was described in the proof of Theorem 3.3. The full log-likelihood of the elements of this sequence is

\[
l(\tau, \lambda^{(n)}) = \sum_{i \in I} \tau_i \log \lambda_i^{(n)} - \sum_{i \in I} \lambda_i^{(n)} = \sum_{i \in F} \tau_i \log \lambda_i^{(n)} - \sum_{i \in I} \lambda_i^{(n)}
= \sum_{i \in F} \tau_i \log \left\{ n^{-ca_i} \prod_{j=1}^J \theta_{a_{ji}} \right\} - \sum_{i \in I} n^{-ca_i} \prod_{j=1}^J \theta_{a_{ji}}
= \sum_{i \in F} \tau_i \log \left\{ \prod_{j=1}^J \theta_{a_{ji}} \right\} - \sum_{i \in I} \prod_{j=1}^J \theta_{a_{ji}}
= l_F(\tau_F, \lambda_F) - \sum_{i \in I} n^{-ca_i} \prod_{j=1}^J \theta_{a_{ji}}.
\]

Therefore,

\[
l(\tau, \lambda^{(n)}) \leq l_F(\tau_F, \lambda_F) \leq l_F(\tau_F, \hat{\lambda}_{y_F}). \quad (19)
\]

Let \( \delta \equiv p \). The log-likelihood under the model \( RM_{p_F}(A_F) \) is equal to

\[
l_F(\tau_F, p_F) = \sum_{i=1}^f q_{Fi} \log p_{Fi},
\]

and for any \( p_F > 0 \), such that \( 1'p_F = 1 \), \( l_F(\tau_F, p_F) \leq l_F(\tau_F, \hat{p}_y) \).

Let \( p = (p'_F, 0)' \), and let \( p^{(n)} \) be the sequence that was described in the proof of Theorem 3.3. The full log-likelihood of the elements of this sequence is

\[
l(\tau, p^{(n)}) = \sum_{i \in I} \tau_i \log p_i^{(n)} = \sum_{i \in F} \tau_i \log p_i^{(n)}
= \sum_{i \in F} \tau_i \log \left\{ (\theta_1^{(n)})_{a_{i1}} \prod_{j=2}^J (n^{-ca_{j}} \theta_{j})_{a_{ji}} \right\} = \sum_{i \in F} \tau_i \log \left\{ (\theta_1^{(n)})_{a_{i1}} n^{a_{1c1} - c'a_{i1}} \prod_{j=2}^J \theta_{a_{ji}} \right\}
= \sum_{i \in F: a_{i1} = 1} \tau_i \log \theta_1^{(n)} n^{c_1} \prod_{j=2}^J \theta_{a_{ji}} + \sum_{i \in F: a_{i1} = 0} \tau_i \log \prod_{j=2}^J \theta_{a_{ji}}
= \sum_{i \in F: a_{i1} = 1} \tau_i \log \prod_{j=1}^J \theta_{a_{ji}} + \sum_{i \in F: a_{i1} = 0} \tau_i \log \prod_{j=1}^J \theta_{a_{ji}} - \sum_{i \in F: a_{i1} = 1} \tau_i \log \left\{ \theta_1 / (\theta_1^{(n)} n^{c_1}) \right\}
= l_F(\tau_F, p_F) - \log \left\{ \theta_1 / (\theta_1^{(n)} n^{c_1}) \right\} \cdot \sum_{i \in F: a_{i1} = 1} q_i.
\]

15
It will be shown next that $\frac{\theta_1}{\theta_1 (n^{c_1})} > 1$.

$$\frac{\theta_1}{\theta_1 (n^{c_1})} = \frac{1 - \sum_{i \in F: a_{i1} = 0} \prod_{j=2}^{J_i} \theta_{j}^{a_{ij}}}{\sum_{i \in F: a_{i1} = 1} \prod_{j=2}^{J_i} \theta_{j}^{a_{ij}}} = \frac{n^{c_1} (\sum_{a_{i1} = 1, i \in F} \prod_{j=2}^{J_i} \theta_{j}^{a_{ij}} + \sum_{a_{i1} = 1, i \notin F} n^{-c} \prod_{j=2}^{J_i} \theta_{j}^{a_{ij}})}{n^{c_1} (1 - \sum_{a_{i1} = 0, i \in F} \prod_{j=2}^{J_i} \theta_{j}^{a_{ij}} - \sum_{a_{i1} = 0, i \notin F} n^{-c} \prod_{j=2}^{J_i} \theta_{j}^{a_{ij}})} \geq 1.$$ 

Therefore, 

$$l(\tau, p^{(n)}) \leq l_F(\tau_F, p_F) \leq l_F(\tau_F, \hat{p}_{y_F}).$$

Combining (19) and (20),

$$l(\tau, \delta^{(n)}) \leq l_F(\tau_F, \delta_F) \leq l_F(\tau_F, \hat{y}_F),$$

and

$$\sup_n l(\tau, \delta^{(n)}) \leq l_F(\tau_F, \hat{y}_F).$$

Hence, whenever $\delta^{(n)} \rightarrow \delta$ as $n \rightarrow \infty$, $l(\tau, \delta^{(n)}) \rightarrow l_F(\tau_F, \hat{y}_F)$.

Therefore, $l(\tau, \hat{y}_F) = \sup l(\tau, \delta) = l_F(\tau_F, \hat{y}_F)$, which concludes the proof of (iii).

The uniqueness claim in (iii) follows from the convexity of the log-likelihood function. The proof is similar to the one given by [Lauritzen 1996, Proposition 4.7] for the case of extended log-affine models, and is thus omitted. In order to prove the second claim, suppose first that there exists a facial set $F$ such that $\text{supp}(\tau) \subseteq F$. Let $F$ be the minimal of such sets. As shown in the proof of (ii), the MLE $\hat{y}_F$ under $RM_{\delta}(\mathbf{A}_F)$ exists, and, from (6),

$$\mathbf{A}_F \hat{y}_F = \gamma \mathbf{A}_F \tau_F,$$

for some $\gamma > 0$, and, if $\delta \equiv p$, $1' \hat{y}_F = 1$.

The MLE under $RM_{\delta}(\mathbf{A})$ is equal to $\hat{y} = (\hat{y}_F, 0_{I \setminus F})$. As $\hat{y}_{F,i} = 0$ for $i \notin F$, $\mathbf{A} \hat{y} = \gamma \mathbf{A} \tau$, and, in the case of probabilities, $1' \hat{y}_F = 1$.

If, for all facial sets $F$, $\text{supp}(\tau) \not\subseteq F$, then the MLE $\hat{y}_F$ under the extended model exists and is also the MLE under $RM_{\delta}(\mathbf{A})$. In this case, (6) holds and is the same as (17), which completes the proof.

Acknowledgments

The second author was supported in part by Grant K-106154 from the Hungarian National Scientific Research Fund (OTKA).
References

A. H. Andersen. Multidimensional contingency tables. *Scand. J. Statist.*, 1:115–127, 1974.

O. E. Barndorff-Nielsen. *Information and exponential families*. Wiley, New York, 1978.

Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete multivariate analysis: theory and practice*. MIT, 1975.

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

L. D. Brown. *Fundamentals of statistical exponential families*. Institute of Mathematical Statistics, Hayward, Calif., 1988.

I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, 3:146–158, 1975.

I. Csiszár and F. Matúš. Information projections revisited. *IEEE Trans. Inform. Theory*, 49:1474–1490, 2003.

S. E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. *Ann. Statist.*, 40:996–1023, 2012a.

S. E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. *Ann. Statist.*, 40:996–1023, 2012b.

D. Geiger, C. Meek, and B. Sturmfels. On the toric algebra of graphical models. *Ann. Statist.*, 34:1463–1492, 2006.

B. Grünbaum. *Convex polytopes*. Springer, 2003.

S. J. Haberman. *The analysis of frequency data*. The University of Chicago Press, 1974.

A. Klimova and T. Rudas. Iterative scaling in curved exponential families. arXiv:1307.3282, 2014a.

A. Klimova and T. Rudas. *gIPFrm: Generalized Iterative Proportional Fitting for Relational Models*, 2014b. URL [http://CRAN.R-project.org/package=gIPFrm](http://CRAN.R-project.org/package=gIPFrm). R package version 2.0.

A. Klimova, T. Rudas, and A. Dobra. Relational models for contingency tables. *J. Multivariate Anal.*, 104:159–173, 2012.

S. L. Lauritzen. *Graphical models*. Oxford Statistical Science Series. Oxford University Press, 1996.

F. Rapallo. Toric statistical models: parametric and binomial representations. *AISM*, 59:727–740, 2007.
J. Rauh, T. Kahle, and N. Ay. Support sets in exponential families and oriented matroid theory. *International Journal of Approximate Reasoning*, 52:613–626, 2011.

A. Rinaldo. On maximum likelihood estimation in log-linear models. Technical Report 833, Carnegie Mellon Univ., 2006.