Smoothness of marginal log-linear parameterizations

Robin J. Evans

Department of Statistics
University of Oxford
Oxford, OX1 3TG, UK
e-mail: evans@stats.ox.ac.uk

Abstract: We provide results demonstrating the smoothness of some marginal log-linear parameterizations for distributions on multi-way contingency tables. First we give an analytical relationship between log-linear parameters defined within different margins, and use this to prove that some parameterizations are equivalent to ones already known to be smooth. Second we construct an iterative method for recovering joint probability distributions from marginal log-linear pieces, and prove its correctness in particular cases. Finally we use Markov chain theory to prove that certain cyclic conditional parameterizations are also smooth. These results are applied to show that certain conditional independence models are curved exponential families.

MSC 2010 subject classifications: Primary 62H17; secondary 62H20.
Keywords and phrases: Conditional independence, contingency table, curved exponential family, log-linear parameter, marginal parameterization.

1. Introduction

Models for multi-way contingency tables may include restrictions on various marginal or conditional distributions, especially in the context of longitudinal or causal models [see, for example, 8, 1, 5, and references therein]. Such models can often be parameterized by combining log-linear parameters from within different marginal tables. The resulting *marginal log-linear* parameterizations, introduced by Bergsma and Rudas [2], provide an elegant and flexible way to parameterize a multivariate discrete probability distribution.

Setting these marginal log-linear parameters to zero can be used to define arbitrary conditional independence models [11, 7], including those corresponding to undirected graphical models or Bayesian networks. If these zero parameters can be embedded into a larger smooth parameterization of the joint distribution, then the model defined by the conditional independence constraints is a curved exponential family, and therefore possesses good statistical properties. This approach is applied by Rudas et al. [11] and Evans and Richardson [5] to classes of graphical models.

Unfortunately, there exist models of conditional independence which—though believed to be curved exponential families—cannot be embedded into parame-

terizations currently known to be smooth. Forcina [6] studies examples of models defined by 'loops' of conditional independences, such as

$$X_1 \perp\!\!\!\perp X_2 \,|\, X_3, \qquad X_1 \perp\!\!\!\perp X_3 \,|\, X_4, \qquad X_1 \perp\!\!\!\perp X_4 \,|\, X_2, \tag{1}$$

which can be defined by constraints on the conditional distributions $p_{2|13}$, $p_{3|14}$ and $p_{4|12}$ respectively. However it is not clear whether a smooth parameterization of the joint distribution can be constructed using these conditionals. The model can also be defined by setting a particular collection of marginal log-linear parameters to zero (see Section 5 for details), but there is no way to embed these parameters into a smooth parameterization of the kind studied by Bergsma and Rudas [2], so their results do not apply. Forcina [6] gives a numerical test for this model which is highly suggestive of smoothness, but no formal proof is available.

The contribution of this paper is to show that the class of smooth discrete parameterizations which can be constructed using marginal log-linear (MLL) parameters is considerably larger than had previously been known, and that models such as (1) can indeed be embedded into these parameterizations. We give three different methods for demonstrating smoothness in this context. First we provide an analytical expression for the relationship between log-linear parameters defined within different marginal distributions; this allows us to prove the equivalence of various parameterizations. Second we show that particular fixed point maps relating different parameters are contractions, and hence can be used to uniquely recover the joint probability distribution. Lastly we use Markov chain theory to show that we can smoothly recover joint probability distributions from 'cyclic' conditional distributions; this is used to show that certain conditional independence models, including the one above, are curved exponential families of distributions.

The rest of the paper is organized as follows: Section 2 reviews marginal log-linear parameters and their properties. Section 3 specifies the relationship between log-linear parameters defined within different margins, enabling certain parameterizations to be proven equivalent. Section 4 extends this by constructing fixed point methods that smoothly recover a joint distribution. Section 5 further extends the results of Section 3 using Markov chain theory, and demonstrates that certain conditional independence models are curved exponential families. Section 6 contains discussion, and a conjecture on the precise characterization of smooth MLL parameterizations.

2. Marginal log-linear parameters

We consider multivariate distributions over a finite collection of binary random variables $X_v \in \{0, 1\}$, for $v \in V$; we denote their joint distribution by $p_V \equiv p(x_V) \equiv P(X_V = x_V)$. All the results herein also hold (or have analogues) in the case of general finite discrete variables, but the notation becomes more cumbersome. For $M \subseteq V$ we denote the marginal distribution over $X_M = (X_v, v \in M)$ by $p_M \equiv p(x_M) \equiv P(X_M = x_M)$, and for disjoint $A, B \subseteq V$ we

denote the relevant conditional distribution by $p_{A|B} \equiv p(x_A \,|\, x_B) \equiv P(X_A = x_A \,|\, X_B = x_B)$. Distributions are assumed to be strictly positive: $p_V > 0$.

Definition 2.1. Let $\Delta^k \equiv \{p_V > 0\}$ be the strictly positive probability simplex of dimension $k = 2^{|V|} - 1$. We say that a homeomorphism $\theta : \Delta^k \to \Theta \subseteq \mathbb{R}^k$ onto an open set Θ is a *smooth parameterization* of Δ^k if θ is twice continuously differentiable, and its Jacobian has full rank k everywhere.

The canonical smooth parameterization of Δ^k is via *log-linear parameters* η_L, defined by the Möbius expansion

$$\log p_V(x_V) = \sum_{L \subseteq V} (-1)^{|x_L|} \eta_L;$$

here $|x_L| = \sum_{v \in L} x_v$ is the number of 1s in $x_L \in \{0,1\}^L$. It follows by Möbius inversion that

$$\eta_L = 2^{-|V|} \sum_{x_V \in \{0,1\}^{|V|}} (-1)^{|x_L|} \log p_V(x_V); \tag{2}$$

see, for example, Lauritzen [9]. For example, if $V = \{1, 2, 3\}$,

$$\eta_{13} = \frac{1}{8} \log \frac{p(0,0,0)\, p(0,1,0)\, p(1,0,1)\, p(1,1,1)}{p(1,0,0)\, p(1,1,0)\, p(0,0,1)\, p(0,1,1)}.$$

It is well known that the collection $\boldsymbol{\eta} \equiv (\eta_L,\, \emptyset \neq L \subseteq V)$ provides a smooth parameterization of the joint distribution p_V with $\Theta = \mathbb{R}^k$.

Definition 2.2. We define a *marginal log-linear parameter* by analogy with (2), as the ordinary log-linear parameter for a particular marginal distribution. Let $L \subseteq M \subseteq V$; then

$$\lambda_L^M = 2^{-|M|} \sum_{x_M \in \{0,1\}^{|M|}} (-1)^{|x_L|} \log p_M(x_M)$$

is the marginal log-linear parameter associated with the *margin* M and the *effect* L. See Bergsma and Rudas [2].

Clearly $\lambda_L^V = \eta_L$ and, for example,

$$\lambda_{13}^{13} = \frac{1}{4} \log \frac{p_{13}(0,0)\, p_{13}(1,1)}{p_{13}(1,0)\, p_{13}(0,1)},$$

which is the log-odds ratio between X_1 and X_3. In order to fit a model with the constraint $X_1 \perp\!\!\!\perp X_3$ we could choose a parameterization that includes λ_{13}^{13}, and fix it to be zero.

One way to characterize the main idea of Bergsma and Rudas [2] is as follows: given some arbitrary margins $p_{M_1}, \ldots, p_{M_k}$ of a joint distribution p_V, what additional information does one need to smoothly reconstruct the full joint

distribution p_V? They show that one possibility is to take the collection of log-linear parameters $\eta_L = \lambda_L^V$ where $L \nsubseteq M_i$ for any $i = 1, \ldots, k$.

It follows that given any inclusion-respecting sequence of margins $M_1, \ldots, M_k = V$ (i.e. $M_i \subseteq M_j$ only if $i < j$), we can smoothly parameterize p_V with marginal log-linear parameters of the form $\lambda_L^{M_i}$, where $L \subseteq M_i$ but $L \nsubseteq M_j$ for any $j < i$.

Example 2.3. Take the inclusion-respecting sequence of margins $\{1, 2\}$, $\{2, 3\}$, $\{1, 2, 3\}$. This gives us the smooth parameterization consisting of the vector $\boldsymbol{\lambda}_{\mathcal{P}}$ below. The pairs (L, M) are summarized (grouped by margin) in the adjacent table.[1]

$$
\mathcal{P}: \quad
\begin{array}{c|c}
M_i & L \\
\hline
12 & 1,\, 2,\, 12 \\
23 & 3,\, 23 \\
123 & 13,\, 123.
\end{array}
\qquad
\boldsymbol{\lambda}_{\mathcal{P}} = (\lambda_1^{12}, \lambda_2^{12}, \lambda_{12}^{12}, \lambda_3^{23}, \lambda_{23}^{23}, \lambda_{13}^{123}, \lambda_{123}^{123})^T.
$$

Now, let $\mathcal{P}$ be an **arbitrary** collection of effect-margin pairs (L, M) such that $\emptyset \neq L \subseteq M \subseteq V$. Define

$$
\boldsymbol{\lambda}_{\mathcal{P}} = \boldsymbol{\lambda}_{\mathcal{P}}(\boldsymbol{p}) = (\lambda_L^M : (L, M) \in \mathcal{P})
$$

to be the corresponding vector of marginal log-linear parameters. The main question considered by this paper is: under what circumstances does $\boldsymbol{\lambda}_{\mathcal{P}}$ constitute a smooth parameterization of p_V?

2.1. Existing results

We say that $\mathcal{P}$ is *complete* if every non-empty subset of V appears as an effect in $\mathcal{P}$ exactly once. If, in addition, the margins can be ordered so that each effect appears with the first margin of which it is a subset, we say that $\mathcal{P}$ is *hierarchical*. Parameterizations that can be constructed from an inclusion-respecting sequence of margins in the manner of Example 2.3 correspond precisely to hierarchical $\mathcal{P}$. Bergsma and Rudas [2] show that if $\mathcal{P}$ is complete and hierarchical then $\boldsymbol{\lambda}_{\mathcal{P}}$ gives a smooth parameterization of the joint distribution; in addition, they show that completeness is necessary for smoothness. Forcina [6] shows that if $\mathcal{P}$ is complete and contains only two distinct margins M, then $\boldsymbol{\lambda}_{\mathcal{P}}$ is smooth.

To our knowledge, these are the only existing results on the smoothness of marginal log-linear parameterizations. No example has been provided of a complete parameterization which is non-smooth. In Sections 3, 4 and 5 we will show that, in fact, many more complete parameterizations are smooth than had previously been known.

The issue of smoothness in non-hierarchical models was raised by Forcina [6] in the context of loop models of conditional independence, and expanded upon

---

[1]Note that here, and in the sequel, we abbreviate sets of integers by omitting the braces and commas in order to avoid overburdened notation: so, for example, 23 means $\{2, 3\}$.

by Colombi and Forcina [3] for models of context-specific conditional independence; the latter consider a more general class of models than we do, but there is no overlap in the theoretical results. Examples of ordinary conditional independences models that require non-complete parameterizations (and therefore are not curved exponential families) are found in Drton [4].

3. An analytical map between margins

To parameterize a marginal distribution p_M we can use the marginal log-linear parameters $\{\lambda_L^M : \emptyset \neq L \subseteq M\}$. An analogous result holds for conditional distributions: for disjoint A, B define

$$\boldsymbol{\lambda}_{A|B} \equiv (\lambda_L^{AB} \mid L \subseteq A \cup B, L \cap A \neq \emptyset);$$

in other words, all the MLL parameters for the margin $A \cup B$ whose effect contains some element of A. Then $\boldsymbol{\lambda}_{A|B}$ constitutes a smooth parameterization of the conditional distribution $X_A \mid X_B$

A consequence of this is to aid us in understanding the relationship between log-linear parameters defined within different margins. Theorem 3 of Bergsma and Rudas [2] shows that distinct MLL parameters corresponding to the same effect in different margins (i.e. λ_L^M and λ_L^N with $M \neq N$) are linearly dependent at certain points in the parameter space, and that therefore no smooth parameterization can include two such parameters. The following theorem elucidates the exact relationship between such parameters, and will later be used to demonstrate the smoothness of certain non-hierarchical parameterizations.

Theorem 3.1. *Let A, M be disjoint subsets of V. The log-linear parameter λ_L^{MA} may be decomposed as*

$$\lambda_L^{MA} = \lambda_L^M + f(\boldsymbol{\lambda}_{A|M}), \tag{3}$$

for a smooth function f, which vanishes whenever $X_A \perp\!\!\!\perp X_v \mid X_{M \setminus \{v\}}$ for some $v \in L$. In addition, if $K \cap (V \setminus M) \neq \emptyset$

$$\frac{\partial \lambda_L^M}{\partial \eta_K} = 2^{-|M|} \sum_{x_V} (-1)^{|x_{K \triangle L}|} p(x_{V \setminus M} \mid x_M) \tag{4}$$

(where $(\eta_J : J \neq K)$ are held fixed).

Proof. We have

$$\lambda_L^{MA} = 2^{-|AM|} \sum_{x_{MA}} (-1)^{|x_L|} \log p(x_{MA})$$

$$= 2^{-|AM|} \sum_{x_{MA}} (-1)^{|x_L|} \left[\log p(x_M) + \log p(x_A \mid x_M)\right]$$

$$= 2^{-|M|} \sum_{x_M} (-1)^{|x_L|} \log p(x_M) + 2^{-|AM|} \sum_{x_{MA}} (-1)^{|x_L|} \log p(x_A \mid x_M)$$

$$= \lambda_L^M + 2^{-|AM|} \sum_{x_{MA}} (-1)^{|x_L|} \log p(x_A \,|\, x_M).$$

Since the second term is a smooth function of the conditional probabilities $p(x_A \,|\, x_M)$, it follows that it is also a smooth function of the claimed parameters. The implication of independence follows from Lemma 2.9 of Evans and Richardson [5].

Now,

$$\frac{\partial}{\partial \eta_K} p(x_V) = \frac{\partial}{\partial \eta_K} \exp\left(\sum_{J \subseteq V} (-1)^{|x_J|} \eta_J \right) = (-1)^{|x_K|} p(x_V),$$

and similarly

$$\frac{\partial}{\partial \eta_K} p(x_M) = \frac{\partial}{\partial \eta_K} \sum_{y_{V \setminus M}} p(y_{V \setminus M}, x_M)$$

$$= (-1)^{|x_{K \cap M}|} \sum_{y_{V \setminus M}} (-1)^{|y_{K \setminus M}|} p(y_{V \setminus M}, x_M).$$

Hence the derivative of (3) in the case $A = V \setminus M$ becomes

$$\frac{\partial f}{\partial \eta_K} = 2^{-|V|} \sum_{x_V} (-1)^{|x_L|} \left\{ (-1)^{|x_K|} \frac{p(x_V)}{p(x_V)} \right.$$

$$\left. - \frac{(-1)^{|x_{K \cap M}|}}{p(x_M)} \sum_{y_{V \setminus M}} (-1)^{|y_{K \setminus M}|} p(y_{V \setminus M}, x_M) \right\}$$

$$= 0 - 2^{-|V|} \sum_{x_V} (-1)^{|x_L| + |x_{K \cap M}|} \sum_{y_{V \setminus M}} (-1)^{|y_{K \setminus M}|} p(y_{V \setminus M} \,|\, x_M)$$

and, since there is no dependence upon $x_{V \setminus M}$, this is the same as

$$= -2^{-|V|} 2^{|V \setminus M|} \sum_{x_V} (-1)^{|x_L| + |x_{K \cap M}| + |x_{K \setminus M}|} p(x_{V \setminus M} \,|\, x_M).$$

Then note that $|x_L| + |x_{K \cap M}| + |x_{K \setminus M}| = |x_L| + |x_K|$ simply counts the number of 1s in L and in K, so $|x_{L \triangle K}|$ is even if and only if $|x_L| + |x_K|$ is. Hence

$$\frac{\partial f}{\partial \eta_K} = -2^{-|M|} \sum_{x_V} (-1)^{|x_{L \triangle K}|} p(x_{V \setminus M} \,|\, x_M),$$

which gives the required result.   □

Remark 3.2. We have shown that if the conditional distribution of X_A given X_M is fixed the relationship between λ_L^M and λ_L^{MA} (and indeed any parameter of the form λ_L^{MB} for $B \subseteq A$) is linear. In particular, if we *know* $p_{A|M}$, then λ_L^{MA} and λ_L^M become interchangeable as part of a parameterization, preserving smoothness and (when relevant) variation independence.

3.1. Constructing smooth parameterizations

The following example shows how Theorem 3.1 can be used to prove the smoothness of a parameterization.

Example 3.3. Consider the complete collections $\mathcal{P}$ and $\mathcal{Q}$ below.

$$
\mathcal{P}: \quad
\begin{array}{c|l}
M_i & L \\
\hline
3 & 3 \\
23 & 23 \\
123 & 1,\,2,\,12,\,13,\,123.
\end{array}
\qquad
\mathcal{Q}: \quad
\begin{array}{c|l}
M_i & L \\
\hline
3 & 3 \\
23 & 2,\,23 \\
123 & 1,\,12,\,13,\,123
\end{array}
\quad.
$$

$\mathcal{P}$ is not hierarchical because in any inclusion-respecting ordering the margin 23 must precede 123, in which case the effect 2 (contained in the pair $(2, 123)$) is not associated with the first margin of which it is a subset. Existing results therefore cannot tell us whether or not $\boldsymbol{\lambda}_{\mathcal{P}}$ is smooth. However, by fixing the parameters $\boldsymbol{\lambda}_{1|23} = (\lambda_1^{123}, \lambda_{12}^{123}, \lambda_{13}^{123}, \lambda_{123}^{123})$ Theorem 3.1 shows that λ_2^{123} and λ_2^{23} are interchangeable. Hence $\boldsymbol{\lambda}_{\mathcal{P}}$ is smooth if and only if $\boldsymbol{\lambda}_{\mathcal{Q}}$ is also smooth which, since $\mathcal{Q}$ satisfies the conditions of a hierarchical parameterization, it is. In addition, $\boldsymbol{\lambda}_{\mathcal{P}}$ and $\boldsymbol{\lambda}_{\mathcal{Q}}$ are both variation independent parameterizations (i.e. any $\boldsymbol{\lambda}_{\mathcal{P}} \in \mathbb{R}^7$ corresponds to a valid probability distribution).

We generalize the approach used in the preceding example with the following definition and proposition.

Definition 3.4. Let $\mathcal{P}$ be a collection of MLL parameters, and define

$$\mathcal{P}_{-v} = \{(L, M \setminus \{v\}) \,|\, (L, M) \in \mathcal{P},\, v \notin L\}.$$

That is, all effects involving v are removed, and any margins M containing v are replaced by $M \setminus \{v\}$.

Proposition 3.5. *Let $\mathcal{P}$ be a complete collection of marginal log-linear parameters over V such that the variable v is not in any margin except V. Then $\boldsymbol{\lambda}_{\mathcal{P}}$ is a smooth parameterization of X_V if and only if $\boldsymbol{\lambda}_{\mathcal{P}_{-v}}$ is a smooth parameterization of $X_{V \setminus v}$. In addition, $\boldsymbol{\lambda}_{\mathcal{P}}$ is variation independent if and only if $\boldsymbol{\lambda}_{\mathcal{P}_{-v}}$ is.*

Proof. Since V is the only margin containing v and the parameterization is complete, we have the parameters $\boldsymbol{\lambda}_{v|V \setminus v} = (\lambda_L^V : v \in L)$. Hence we can smoothly parameterize the distribution of $X_v \,|\, X_{V \setminus v}$ with these parameters.

By Theorem 3.1, any other parameter λ_L^V such that $v \notin L$ is (having fixed the distribution of $X_v \,|\, X_{V \setminus v}$) a smooth function of $\lambda_L^{V \setminus v}$. It follows that we have a smooth map between $\boldsymbol{\lambda}_{\mathcal{P}}$ and $(\boldsymbol{\lambda}_{\mathcal{P}_{-v}}, \boldsymbol{\lambda}_{v|V \setminus v})$. Since $\boldsymbol{\lambda}_{\mathcal{P}_{-v}}$ is a function of $p_{V \setminus v}$, and $\boldsymbol{\lambda}_{v|V \setminus v}$ smoothly parameterizes $p_{v|V \setminus v}$, it follows that $\boldsymbol{\lambda}_{\mathcal{P}}$ smoothly parameterizes p_V if and only if $\boldsymbol{\lambda}_{\mathcal{P}_{-v}}$ smoothly parameterizes $p_{V \setminus v}$.

Lastly, the two pieces $\boldsymbol{\lambda}_{\mathcal{P}_{-v}}$ and $\boldsymbol{\lambda}_{v|V \setminus v}$ are variation independent of one another because this is a parameter cut, and parameters within $\boldsymbol{\lambda}_{v|V \setminus v}$ are

all variation independent since they are just ordinary log-linear parameters; therefore $\boldsymbol{\lambda}_{\mathcal{P}_{-v}}$ is variation independent if and only if $\boldsymbol{\lambda}_{\mathcal{P}}$ is. □

Corollary 3.6. *Any complete parameterization in which the margins are strictly nested ($M_1 \subset M_2 \subset \cdots \subset M_k = V$) is smooth and variation independent.*

Lemma 6 of Forcina [6] deals with the special case $k = 2$, which to our knowledge was the only prior result showing that a non-hierarchical MLL parameterization may be smooth.

Example 3.7. Consider

$$
\mathcal{P} : \quad
\begin{array}{c|l}
M_i & L \\
\hline
13 & 3 \\
23 & 23 \\
123 & 1, 2, 12, 13, 123
\end{array}
\qquad
\mathcal{Q} : \quad
\begin{array}{c|l}
M_i & L \\
\hline
13 & 3 \\
123 & 1, 2, 12, 23, 13, 123
\end{array}.
$$

$\mathcal{P}$ does not satisfy the conditions of Proposition 3.5; however, applying Theorem 3.1 shows that λ_{23}^{23} is just a linear function of λ_{23}^{123} after fixing the other parameters in the margin 123, so $\mathcal{P}$ is smooth if and only if $\mathcal{Q}$ is. Applying Corollary 3.6 shows that $\boldsymbol{\lambda}_{\mathcal{Q}}$ (and therefore $\boldsymbol{\lambda}_{\mathcal{P}}$) is smooth.

Proposition 3.8. *Let $\mathcal{P}$ be a complete parameterization, and suppose that for some $v \in V$, and every $A \subseteq V \setminus \{v\}$, the sets $A \cup \{v\}$ and A appear as effects within the same margin in $\mathcal{P}$.*

Then $\boldsymbol{\lambda}_{\mathcal{P}}$ is a smooth parameterization of X_V if and only if $\boldsymbol{\lambda}_{\mathcal{P}_{-v}}$ is a smooth parameterization of $X_{V \setminus v}$. In addition, $\boldsymbol{\lambda}_{\mathcal{P}}$ is variation independent if and only if $\boldsymbol{\lambda}_{\mathcal{P}_{-v}}$ is variation independent.

Proof. Since $A \subseteq V \setminus \{v\}$ and $A \cup \{v\}$ appear in the same margin, say $M \cup \{v\}$, set

$$
\begin{aligned}
\kappa_A^M(x_v) &= \lambda_A^{Mv} + (-1)^{|x_v|} \lambda_{Av}^{Mv} \\
&= 2^{-|Mv|} \sum_{y_{Mv}} (-1)^{|y_A|} \left[(-1)^{|y_A|} + (-1)^{|y_{Av}| + |x_v|} \right] \log p_{M|v}(y_M, y_v)
\end{aligned}
$$

which is zero unless $x_v = y_v$, leaving

$$
\begin{aligned}
&= 2^{-|M|} \sum_{y_M} (-1)^{|y_A|} \log p_{M|v}(y_M, x_v) \\
&= 2^{-|M|} \sum_{y_M} (-1)^{|y_A|} \log p_{M|v}(y_M \mid x_v).
\end{aligned}
$$

But notice this is of the same form as an MLL parameter for the pair (A, M) over the conditional distribution $p_{V \setminus v | v}(\cdot \mid x_v)$. It follows that for fixed x_v the parameters $\{\kappa_A^M(x_v) : (A, M) \in \mathcal{P}, v \notin A\}$ form a complete MLL collection of the form $\mathcal{P}_{-v}$ for the conditional distribution of $X_{V \setminus v} \mid X_v = x_v$. If $\boldsymbol{\lambda}_{\mathcal{P}_{-v}}$

is smooth then we can smoothly recover the conditional distribution $p_{V \setminus v | v}$. Furthermore, if the effect $\{v\}$ is in a margin $N \cup \{v\}$, then using (3) we obtain

$$\lambda_v^{Nv} = \lambda_v^v + f(p_{N|v}),$$

and smoothly recover λ_v^v. In addition λ_v^v is variation independent of $p_{N|v}$ (since p_v, $p_{N|v}$ constitutes a parameter cut) and has range $\mathbb{R}$, so the same is true of λ_v^{Nv}.

Conversely if $\boldsymbol{\lambda}_{\mathcal{P}}$ is smooth, then given parameters $\boldsymbol{\lambda}_{\mathcal{P}_{-v}}$ we can set up a dummy distribution on p_V in which $\kappa_A^M(x_v) = \lambda_A^M$ for each x_v, and $\lambda_v^N = 0$, thus smoothly recovering $p_{V \setminus v}$. $\qquad \square$

Example 3.9. As an example, consider

$$\mathcal{P} : \begin{array}{c|l} M_i & L \\ \hline 12 & 2,\ 12 \\ 13 & 3,\ 13 \\ 123 & 1,\ 23,\ 123 \end{array}$$

which is not hierarchical, and nor does it satisfy the conditions of Proposition 3.5. However it does satisfy the the conditions of Proposition 3.8 for $v = 1$, and $\mathcal{P}_{-1} = \{(2,2),(3,3),(23,23)\}$, which is hierarchical and so certainly smooth. Hence $\mathcal{P}$ represents a smooth parameterization.

4. Fixed point mappings

The previous section gives analytical maps between some parameterizations, but Propositions 3.5 and 3.8 only apply directly to a relatively small number of cases. In this section we build on these results by presenting conditions for the existence of a smooth map, even without a closed form expression.

Given a particular complete MLL parameterization $\boldsymbol{\lambda}_{\mathcal{P}}$, the identity (3) in Theorem 3.1 can be written in vector form as

$$\boldsymbol{\eta} = \boldsymbol{\lambda} + \boldsymbol{f}(\boldsymbol{\eta}).$$

For a given $\boldsymbol{\lambda}$ this suggests that $\boldsymbol{\eta}$ might be recovered using fixed point methods; the identity (4) gives us information about the Jacobian of $\boldsymbol{f}$.

Example 4.1. Consider the parameterization based on

$$\mathcal{P} : \begin{array}{c|l} M_i & L \\ \hline 23 & 2,\ 23 \\ 13 & 1 \\ 123 & 12,\ 3,\ 13,\ 123. \end{array}$$

If we can smoothly recover η_1, η_2 and η_{23} from $\boldsymbol{\lambda}_{\mathcal{P}}$ then it follows that $\boldsymbol{\lambda}_{\mathcal{P}}$ is a smooth parameterization. From (3) we have

$$\begin{pmatrix} \eta_2 \\ \eta_{23} \end{pmatrix} = \begin{pmatrix} \lambda_2^{23} \\ \lambda_{23}^{23} \end{pmatrix} + \boldsymbol{f}(\eta_1, \eta_{12}, \eta_{13}, \eta_{123});$$

since η_{12}, η_{13} and η_{123} are given in the parameterization we can assume these to be fixed, so abusing notation slightly

$$\begin{pmatrix} \eta_2 \\ \eta_{23} \end{pmatrix} = \begin{pmatrix} \lambda_2^{23} \\ \lambda_{23}^{23} \end{pmatrix} + \boldsymbol{f}(\eta_1) = \begin{pmatrix} \lambda_2^{23} + f_1(\eta_1) \\ \lambda_{23}^{23} + f_2(\eta_1) \end{pmatrix}.$$

Similarly, $\eta_1 = \lambda_1^{13} + g(\eta_2, \eta_{23})$ for some smooth g, so η_1 is a solution to the equation

$$\begin{aligned} x &= \lambda_1^{13} + g(\lambda_2^{23} + f_1(x), \lambda_{23}^{23} + f_2(x)) \\ &\equiv \Psi(x). \end{aligned}$$

If Ψ can be shown to be a contraction mapping, then we are guaranteed to find a unique solution, and therefore recover the joint distribution. In addition, if Ψ is a contraction for all $\boldsymbol{\eta}$, then since it varies smoothly in $\boldsymbol{\eta}$ we will have shown that $\boldsymbol{\lambda}_{\mathcal{P}}$ is a smooth parameterization.

Define $\epsilon = \min_{x_V} p(x_V)$ to be the smallest amount of probability assigned to any cell in our joint distribution, and $\Delta_\epsilon = \{p : \min_{x_V} p(x_V) \geq \epsilon\}$ to be the probability simplex consisting of such distributions. The Jacobian of an otherwise smooth parameterization can become singular on the boundary of the probability simplex, so it is useful to have control over this quantity.

The next result allows us to control the magnitude of the columns (or rows) of the Jacobian of Ψ in certain examples. The proof is given in the appendix.

Lemma 4.2. *Let $J \subseteq M$, and $\emptyset \neq K \subseteq V \setminus M$. Then*

$$\sum_{\emptyset \neq C \subseteq M} \left| \frac{\partial \lambda_C^M}{\partial \eta_{JK}} \right|^2 \leq 1 - \epsilon.$$

Alternatively, if $\emptyset \neq C \subseteq M$, then

$$\sum_{J \subseteq M} \left| \frac{\partial \lambda_C^M}{\partial \eta_{JK}} \right|^2 \leq 1 - \epsilon.$$

Example 4.3. Returning to the parameterization in Example 4.1, the derivative of Ψ is

$$\Psi'(x) = \frac{\partial \lambda_1^{13}}{\partial \eta_2} \frac{\partial \lambda_2^{23}}{\partial \eta_1} + \frac{\partial \lambda_1^{13}}{\partial \eta_{23}} \frac{\partial \lambda_{23}^{23}}{\partial \eta_1},$$

which is the dot product of the vectors

$$\left(\frac{\partial \lambda_1^{13}}{\partial \eta_2}, \frac{\partial \lambda_1^{13}}{\partial \eta_{23}} \right)^T \qquad\qquad \left(\frac{\partial \lambda_2^{23}}{\partial \eta_1}, \frac{\partial \lambda_{23}^{23}}{\partial \eta_1} \right)^T.$$

By applying the two parts of Lemma 4.2, these vectors each have magnitude at most $1 - \epsilon$. Hence $|\Psi'(x)| \leq 1 - \epsilon$, and Ψ is a contraction on Δ_ϵ for every $\epsilon > 0$.

It follows that the equation has a unique solution among all positive probability distributions (and this can be found by iteratively applying Ψ to any initial distribution), and by the inverse function theorem it is a smooth function of $\boldsymbol{\lambda}$. Hence $\boldsymbol{\lambda}_{\mathcal{P}}$ is indeed smooth.

Remark 4.4. Forcina [6] also uses fixed point methods to recover distributions from marginal log-linear parameters, but that approach involves computing probabilities directly. We discuss those methods in Section 5.

Lemma 4.2 enables us to formulate the following generalization of the idea used in the example above.

Lemma 4.5. *Let $\mathcal{P}$ be complete and such that for any $(L, M) \in \mathcal{P}$ with $M \subset V$, there is at most one other margin $N \subset V$ in $\mathcal{P}$ with $L \cap (V \setminus N) \neq \emptyset$. Then $\boldsymbol{\lambda}_{\mathcal{P}}$ is smooth.*

Proof. By Theorem 3.1,

$$\eta_L = \lambda_L^M + f(\boldsymbol{\lambda}_{V \setminus M \mid M}).$$

Since N is the only margin in $\mathcal{P}$ such that $L \cap (V \setminus N) \neq \emptyset$, it follows that all the parameters in $\boldsymbol{\lambda}_{V \setminus M \mid M}$ are known and fixed except for $(\lambda_K^V : K \in \mathbb{L}_N)$, where $\mathbb{L}_N$ is the set of effects contained in the margin N. Hence

$$\eta_L = \lambda_L^M + f(\boldsymbol{\eta}_{\mathbb{L}_N}). \tag{5}$$

Now, consider the vector equation obtained by stacking (5) over all pairs $(L, M) \in \mathcal{P}$. This defines a fixed point equation whose solution is $\boldsymbol{\eta}$, and the column of the Jacobian corresponding to L has non-zero entries

$$\frac{\partial \lambda_C^N}{\partial \eta_L}, \qquad C \in \mathbb{L}_N.$$

From Lemma 4.2, each column has magnitude at most $1 - \epsilon$, and therefore the mapping is a contraction on Δ_ϵ for each $\epsilon > 0$. It follows that the fixed point equation has a unique solution which, by the inverse function theorem, is a smooth function of $\boldsymbol{\lambda}$. $\square$

From this result we obtain the following corollary, the conditions of which are easy to verify.

Corollary 4.6. *Any complete parameterization with at most three margins is smooth.*

Proof. Since one of the margins must be V, it is clear that the conditions of Lemma 4.5 hold. $\square$

Example 4.7. Consider $\mathcal{P}$ below.

$$\mathcal{P}: \quad \begin{array}{c|l} M_i & L \\ \hline 1 & 1 \\ 12 & 2 \\ 13 & 3 \\ 123 & 12, 13, 23, 123 \end{array} \qquad\qquad \mathcal{Q}: \quad \begin{array}{c|l} M_i & L \\ \hline 1 & 1 \\ 123 & 2, 12, 3, 13, 23, 123 \end{array}$$

Although it does not satisfy the conditions of Lemma 4.5 directly, one can use the basic idea to set up a smooth contraction mapping from $\boldsymbol{\lambda}_{\mathcal{P}}$ to $\boldsymbol{\lambda}_{\mathcal{Q}}$; since $\mathcal{Q}$ is hierarchical, both parameterizations are smooth.

5. Cyclic parameterizations

This section takes a third approach to determining smoothness, by using Markov chain theory to recover certain marginal distributions. This method allows us to demonstrate the smoothness of certain conditional independence models.

Forcina [6, Example 2] considers the model defined (up to some relabelling) by the conditional independences

$$X_1 \perp\!\!\!\perp X_2 \,|\, X_3, \qquad X_1 \perp\!\!\!\perp X_3 \,|\, X_4, \qquad X_1 \perp\!\!\!\perp X_4 \,|\, X_2, \qquad (6)$$

which is equivalent to setting the parameters

M_i	L
123	12, 123
134	13, 134
124	14, 124

$$(7)$$

to zero. Note that we cannot embed these parameters into a larger hierarchical parameterization, because each pairwise effect will 'belong' to a margin preceding it; for example, 12 is a subset of 124, so for hierarchy the margin 123 must precede 124; by a similar argument, 124 must precede 134 which must precede 123. We therefore have a cyclic parameterization, referred to as a 'loop' by Forcina. None of the methods used in the previous sections seem well suited to dealing with this situation.

Forcina [6] presents an algorithm for recovering joint distributions given parameterizations of this kind, together with a condition under which it is guaranteed to converge to the unique solution. However, this condition is on the spectral radius of a complicated Jacobian, and is difficult to verify except in a few special cases: a numerical test is suggested, but this does not constitute a proof of smoothness. Here we show that, at least in some cases, Forcina's algorithm can be recast as a Markov chain whose stationary distribution is some margin of the relevant probability distribution.

Theorem 5.1. *Let $A_1, \ldots, A_k$ be a disjoint sequence of sets with $k \geq 2$ such that the conditional distributions $p(x_{A_i} \,|\, x_{A_{i-1}}) > 0$ for $i = 2, \ldots, k$ are known, together with $p(x_{A_1} \,|\, x_{A_k})$. Then the marginal distributions $p(x_{A_i})$ are smoothly recoverable.*

Proof. Define a $|X_{A_1}| \times |X_{A_1}|$ matrix M with entries

$$M(x'_{A_1}, x_{A_1}) = \sum_{x_{A_k}} \cdots \sum_{x_{A_2}} p(x_{A_1} \,|\, x_{A_k}) p(x_{A_k} \,|\, x_{A_{k-1}}) \cdots p(x_{A_2} \,|\, x'_{A_1}).$$

This is a (right) stochastic matrix with strictly positive entries, and the marginal distribution $p(x_{A_1})$ satisfies

$$p(x_{A_1}) = \sum_{x'_{A_1}} p(x'_{A_1}) M(x'_{A_1}, x_{A_1}).$$

In other words, $p(x_{A_1})$ is an invariant distribution for the Markov chain with transition matrix defined by M. Since M has a finite state-space and all transition probabilities are positive, the chain is positive recurrent and the equations have a unique solution [see, e.g. 10]. Hence $p(x_{A_1})$ is defined by the kernel of the matrix $I - M^T$, and this is a smooth function of the original conditional probabilities. □

Remark 5.2. The Markov chain corresponding to M is that which would be obtained by picking some X_{A_1}, and then evolving X_{A_i} using $p(x_{A_i} \mid x_{A_{i-1}})$ until we get back to $i = 1$. The equations can be solved iteratively by repeatedly right multiplying any positive vector by M, so that it converges to the stationary distribution of the chain; this corresponds precisely to Forcina's algorithm.

Example 5.3 (Forcina [6], Example 9)**.** Consider the cyclic parameterization $\mathcal{P}$.

$$\mathcal{P}: \quad \begin{array}{c|c} M_i & L \\ \hline 12 & 1,\ 12 \\ 23 & 2,\ 23 \\ 13 & 3,\ 13 \\ 123 & 123 \end{array} \qquad\qquad \mathcal{Q}: \quad \begin{array}{c|c} M_i & L \\ \hline 3 & 3 \\ 23 & 2,\ 23 \\ 12 & 1,\ 12 \\ 13 & 13 \\ 123 & 123 \end{array}.$$

The parameters corresponding to the first three margins in $\mathcal{P}$ are equivalent to the conditional distributions $p_{1|2}$, $p_{2|3}$ and $p_{3|1}$. Using the conditionals in the manner suggested by Theorem 5.1, we can smoothly recover (for example) the margin p_3 (or equivalently λ_3^3), and consequently $\mathcal{P}$ is equivalent to the hierarchical parameterization $\mathcal{Q}$.

Example 5.4. The parameters (7) can be embedded in the complete parameterization $\mathcal{P}$ below.

$$\mathcal{P}: \quad \begin{array}{c|c} M_i & L \\ \hline 123 & 2,\ 23,\ 12,\ 123 \\ 134 & 3,\ 34,\ 13,\ 134 \\ 124 & 4,\ 24,\ 14,\ 124 \\ 1234 & \text{(other subsets)} \end{array}, \qquad \mathcal{P}_{-1}: \quad \begin{array}{c|c} M_i & L \\ \hline 23 & 2,\ 23 \\ 34 & 3,\ 34 \\ 24 & 4,\ 24 \\ 234 & 234 \end{array}.$$

$\mathcal{P}$ satisfies Proposition 3.8 with $v = 1$ and reduces to $\mathcal{P}_{-1}$, which is isomorphic to the smooth parameterization in Example 5.3. Hence $\mathcal{P}$ is smooth, and the conditional independence model (6) is a curved exponential family.

Example 5.5. Consider the model defined by

$$X_1 \perp\!\!\!\perp X_2 \mid X_3, \qquad X_2 \perp\!\!\!\perp X_4 \mid X_1, \qquad X_1 \perp\!\!\!\perp X_3 \mid X_4, \qquad X_3 \perp\!\!\!\perp X_4 \mid X_2;$$

it consists of setting the parameters in $\mathcal{P}$ below to zero.

M_i	L
14	1, 4, 14
23	2, 3, 23
123	12, 123
124	24, 124
134	13, 134
234	34, 234
1234	1234

$\mathcal{P}:$

M_i	L
123	12, 123
124	24, 124
134	13, 134
234	34, 234

$\mathcal{Q}:$ (table at right).

We can embed $\mathcal{P}$ in the complete parameterization $\mathcal{Q}$. Note that using $\lambda_4^{14}, \lambda_{14}^{14}$ and the fact that $X_4 \perp\!\!\!\perp X_2 \mid X_1$, means we can construct the conditional distribution $p(x_4 \mid x_1, x_2)$. Similarly we have $p(x_3 \mid x_2, x_4)$, $p(x_1 \mid x_3, x_4)$ and $p(x_2 \mid x_1, x_3)$. In a manner analogous to the previous example, we can set up a Markov chain whose stationary distribution is the marginal $p(x_1, x_2)$ as follows. First pick $x_1^{(0)}, x_2^{(0)}$. Now, for $i > 0$

- draw $x_4^{(i)}$ from the distribution $p(x_4 \mid x_1^{(i-1)}, x_2^{(i-1)})$;
- draw $x_3^{(i)}$ from the distribution $p(x_3 \mid x_2^{(i-1)}, x_4^{(i)})$;
- draw $x_1^{(i)}$ from the distribution $p(x_1 \mid x_3^{(i)}, x_4^{(i)})$;
- draw $x_2^{(i)}$ from the distribution $p(x_2 \mid x_1^{(i)}, x_3^{(i)})$.

Then the distribution of $(x_1^{(i)}, x_2^{(i)})$ converges to p_{12}. We can therefore smoothly recover a distribution satisfying the conditional independence constraints from the 7 free parameters. The dimension of the model is full, so we have a smooth parameterization of the model, which is therefore a curved exponential family Lauritzen [9].

Note that the construction of the Markov chain in Example 5.5 is only possible when the conditional independence constraints hold, so—unlike in Examples 5.3 and 5.4—we have not actually demonstrated that $\boldsymbol{\lambda}_\mathcal{Q}$ is generally smooth, only that the model defined by setting $\boldsymbol{\lambda}_\mathcal{P} = \mathbf{0}$ is a curved exponential family.

Remark 5.6. Some conditional independence models are non-smooth: e.g. the model defined by $X_1 \perp\!\!\!\perp X_2, X_4$ and $X_2 \perp\!\!\!\perp X_4 \mid X_1, X_3$ [4]. This is essentially because it requires that $\lambda_{124}^{124} = \lambda_{124}^{1234} = 0$, and setting repeated (non-redundant) effects to zero always leads to non-smooth parameterizations.

We remark that all discrete conditional independence models on four variables either require repeated effects to be constrained in different margins, or can be shown to be smooth using the results of this section. However, the next example shows that for five variables the picture is incomplete.

Example 5.7. The conditional independence model defined by

$$X_1 \perp\!\!\!\perp X_2 \mid X_3, \qquad X_1 \perp\!\!\!\perp X_5 \mid X_2, \qquad X_1 \perp\!\!\!\perp X_3 \mid X_4,$$
$$X_3 \perp\!\!\!\perp X_5 \mid X_1, \qquad X_3 \perp\!\!\!\perp X_4 \mid X_2, X_5$$

contains no repeated effects, and yet does not appear to be approachable using the methods outlined above. Empirically, Forcina's algorithm seems to converge to the correct solution, which suggests that the model is indeed smooth.

6. Discussion

We have presented three new approaches to demonstrating that complete but non-hierarchical marginal log-linear parameterizations are smooth, although a general result eludes us. Note that each of the approaches provides an explicit algorithm for obtaining the probabilities from the parameterization, either using the map in Section 3, the fixed point iteration in Section 4, or the Markov chain in Section 5.

There are 104 complete MLL parameterizations on three variables, of which 23 are hierarchical and a further 4 consist of only two margins, so are smooth by the results of Bergsma and Rudas [2] and Forcina [6] respectively. These 27 were the only ones known to be smooth prior to this paper.

A further 5 can be shown smooth using Proposition 3.5, and one using Proposition 3.8 (Example 3.9). Another 26 can be dealt with using Lemma 4.5 in combination with other methods, and the approach in Example 4.7 can be applied to three more. Example 5.3 brings the total number of known smooth models to 63.

In addition, of the remaining 41 complete parameterizations, there are smooth mappings between a group of four and a group of three, so it remains to establish the smoothness (or otherwise) of at most 36 distinct parameterizations. As an example of a parameterization whose smoothness is still not established, consider:

$$
\mathcal{P}: \quad
\begin{array}{c|c}
M_i & L \\
\hline
12 & 1,\,2 \\
13 & 3,\,13 \\
23 & 23 \\
123 & 12,\,123
\end{array}
\quad .
$$

We conjecture that any complete parameterization is smooth, a result which would enable us to show that models such as that given in Example 5.7 are curved exponential families of distributions.

Conjecture 6.1. *Any complete MLL parameterization is smooth.*

Acknowledgements

Appendix A: Technical proofs

A.1. Proof of Lemma 4.2

Lemma A.1. *Let $\boldsymbol{d} = (d_A)$ be a vector indexed by subsets $A \subseteq \{1, \ldots, k\}$. Then $\|\boldsymbol{d}\| < 1$ if and only if for any $B \subseteq [k] \equiv \{1, \ldots, k\}$,*

$$
\left| \sum_{A \subseteq [k]} (-1)^{|A \cap B|} d_A \right| < 1.
$$

Proof. The $2^k \times 2^k$-matrix M with (B, A)th entry $M_{B,A} = 2^{-k/2}(-1)^{|A \cap B|}$ is orthogonal, and therefore preserves vector lengths. Then the vector $M\boldsymbol{d}$ has entries with magnitude at most $2^{-k/2}$, and therefore has total magnitude at most 1. The same is therefore true of $\boldsymbol{d}$. □

Proof of Lemma 4.2. For $C \subseteq M$, define

$$d_C \equiv -2^{-|M|} \sum_{x_V} (-1)^{|x_{C \triangle (J \cup K)}|} p(x_{V \setminus M} \mid x_M)$$

$$= -2^{-|M|} \sum_{x_V} (-1)^{|x_{C \triangle J}| + |x_K|} p(x_{V \setminus M} \mid x_M)$$

so that $d_C = \frac{\partial \lambda_C^M}{\partial \eta_{JK}}$ for $C \neq \emptyset$. Given y_M,

$$\sum_{C \subseteq M} (-1)^{|y_C|} d_C = -2^{-|M|} \sum_{C \subseteq M} (-1)^{|y_C|} \sum_{x_V} (-1)^{|x_{C \triangle J}| + |x_K|} p(x_{V \setminus M} \mid x_M),$$

and note that

$$\sum_{C \subseteq M} (-1)^{|y_C|} (-1)^{|x_{C \triangle J}|}$$

$$= \sum_{C \subseteq M \setminus \{v\}} (-1)^{|y_C|} (-1)^{|x_{C \triangle J}|} + (-1)^{|y_{Cv}|} (-1)^{|x_{Cv \triangle J}|}$$

$$= \sum_{C \subseteq M \setminus \{v\}} (-1)^{|y_C|} (-1)^{|x_{C \triangle J}|} \left\{ 1 + (-1)^{|y_v|} (-1)^{|x_v|} \right\}$$

where the expression in braces is 2 if $x_v = y_v$ or 0 otherwise, so

$$= \sum_{C \subseteq M \setminus \{v\}} 2(-1)^{|y_C|} (-1)^{|x_{C \triangle J}|} \mathbb{1}_{\{x_v = y_v\}}$$

$$= 2^{|M|} \mathbb{1}_{\{x_M = y_M\}}.$$

Hence

$$\sum_{C \subseteq M} (-1)^{|y_C|} d_C = - \sum_{x_{V \setminus M}} (-1)^{|x_K|} p(x_{V \setminus M} \mid y_M),$$

which is an alternating sum of probabilities which sum to one, so has absolute value at most $1 - \epsilon$. The result follows from Lemma A.1. The second result is essentially identical, due to the symmetry between L, K in (4). □

References

[1] BERGSMA, W., CROON, M. A., and HAGENAARS, J. A., *Marginal Models: For Dependent, Clustered, and Longitudinal Categorical Data.* Springer Science & Business Media, 2009.

[2] BERGSMA, W. P. and RUDAS, T., Marginal models for categorical data. *Ann. Stat.*, 30(1):140–159, 2002. MR1892659

[3] COLOMBI, R. and FORCINA, A., A class of smooth models satisfying marginal and context specific conditional independencies. *Journal of Multivariate Analysis*, 126:75–85, 2014. MR3173082

[4] DRTON, M., Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009. MR2555197

[5] EVANS, R. J. and RICHARDSON, T. S., Marginal log-linear parameterizations for graphical Markov models. *Journal of Royal Statistical Society, Series B*, 75:743–768, 2013. MR3091657

[6] FORCINA, A., Smoothness of conditional independence models for discrete data. *Journal of Multivariate Analysis*, 106:49–56, 2012. MR2887679

[7] FORCINA, A., LUPPARELLI, M., and MARCHETTI, G. M., Marginal parameterizations of discrete models defined by a set of conditional independencies. *Journal of Multivariate Analysis*, 101:2519–2527, 2010. MR2719878

[8] LANG, J. B. and AGRESTI, A., Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, 89(426):625–632, 1994.

[9] LAURITZEN, S. L., *Graphical Models.* Clarendon Press, Oxford, UK, 1996. MR1419991

[10] NORRIS, J. R., *Markov Chains.* Cambridge University Press, 1997. MR1600720

[11] RUDAS, T., BERGSMA, W. P., and NÉMETH, R., Marginal log-linear parameterization of conditional independence models. *Biometrika*, 94:1006–1012, 2010. MR2746171