Biometrics xx, 1–21 DOI: xxx/xxxx.x
June 2018

Fast Likelihood-based Inference for Latent Count Models using the Saddlepoint Approximation

W. Zhang[1,*], M. V. Bravington[2], and R. M. Fewster[1]

[1]Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand

[2]CSIRO Data61, Hobart, Tasmania 7000, Australia

*email: wzha217@aucklanduni.ac.nz

Summary: Latent count models constitute an important modeling class in which a latent vector of counts, $\boldsymbol{z}$, is summarized or corrupted for reporting, yielding observed data $\boldsymbol{y} = \boldsymbol{T}\boldsymbol{z}$ where $\boldsymbol{T}$ is a known but non-invertible matrix. The observed vector $\boldsymbol{y}$ generally follows an unknown multivariate distribution with a complicated dependence structure. Latent count models arise in diverse fields, such as estimation of population size from capture-recapture studies; inference on multi-way contingency tables summarized by marginal totals; or analysis of route flows in networks based on traffic counts at a subset of nodes. Currently, inference under these models relies primarily on stochastic algorithms for sampling the latent vector $\boldsymbol{z}$, typically in a Bayesian data-augmentation framework. These schemes involve long computation times and can be difficult to implement. Here, we present a novel maximum-likelihood approach using likelihoods constructed by the saddlepoint approximation. We show how the saddlepoint likelihood may be maximized efficiently, yielding fast inference even for large problems. For the case where $\boldsymbol{z}$ has a multinomial distribution, we validate the approximation by applying it to a specific model for which an exact likelihood is available. We implement the method for several models of interest, and evaluate its performance empirically and by comparison with other estimation approaches. The saddlepoint method consistently gives fast and accurate inference, even when $\boldsymbol{y}$ is dominated by small counts.

Key words: Capture-recapture; Contingency table; Latent count model, Multi-list method; Population size estimate; Saddlepoint approximation.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

We consider a class of latent count models, in which observed counts $\boldsymbol{y}$ arise from an underdetermined linear system $\boldsymbol{y} = \boldsymbol{T}\boldsymbol{z}$. Here, $\boldsymbol{T}$ is a known non-invertible matrix, and $\boldsymbol{z}$ is a latent vector of counts, for example from a multinomial distribution. Modeling $\boldsymbol{z}$ is natural, but $\boldsymbol{z}$ itself is unobservable and only a summary or corruption $\boldsymbol{y}$ is available. Our task is to infer the parameters underlying $\boldsymbol{z}$ from the observed data $\boldsymbol{y}$.

Latent count models have numerous applications in diverse fields. Examples include estimation of population size in epidemiology or ecology; inference on partially-reported contingency tables in social sciences; and network analysis for road traffic engineering or communications systems. The first of these relates to capture-recapture estimation, in which a latent count structure arises when individual identity is not fully observable. For example, when estimating disease prevalence from multiple lists of patient records, a single patient might generate two unmatched records if there are two lists that do not share a common identifier (Sutherland and Schwarz, 2005). This creates an unknown number of patients that are counted twice in the observed data $\boldsymbol{y}$. A similar latent-count structure arises in analysis of contingency tables when only a subset of marginal totals is made available, perhaps to preserve confidentiality (Dobra, Tebaldi, and West, 2006). In network models, a latent linear structure emerges when seeking inference on mean route flows, given traffic counts at a subset of nodes (Hazelton, 2015).

A likelihood function for these models is easily formulated as $\mathcal{L}(\boldsymbol{\theta}) = \sum_{\boldsymbol{z}\in\mathcal{A}} \Pr(\boldsymbol{z} \mid \boldsymbol{\theta})$, where $\mathcal{A} = \{\boldsymbol{z} \mid \boldsymbol{y} = \boldsymbol{T}\boldsymbol{z}\}$ and $\boldsymbol{\theta}$ parametrizes the model for the latent vector $\boldsymbol{z}$. Numerous authors have commented that evaluation of this likelihood is computationally infeasible, due to the typically enormous cardinality of the set $\mathcal{A}$ (e.g. Dobra et al., 2006; Link et al., 2010). Alternative estimation methods proposed include method-of-moments (Vardi, 1996), quasi-likelihood (Lee, 2002; Sutherland and Schwarz, 2005), and least-squares (Yoshizaki et al.,

2011). Recently, attention has focused on stochastic estimation, based on sampling from the set $\mathcal{A}$ rather than enumerating it. A substantial literature has grown around the design of samplers for the latent vector $\boldsymbol{z}$ under the constraint $\boldsymbol{T}\boldsymbol{z} = \boldsymbol{y}$, enabling inference under a Bayesian MCMC framework or via the stochastic EM algorithm (e.g. Tebaldi and West, 1998; Chen et al., 2005; Dobra et al., 2006; Link et al., 2010; Hazelton, 2015). However, this sampling problem is difficult, and later work has uncovered problems with earlier algorithms such as arbitrarily poor mixing or non-irreducibility of the sampler (Schofield and Bonner, 2015; Hazelton, 2015). Proposed solutions include use of dynamic Markov bases (Diaconis and Sturmfels, 1998; Dobra, 2012; Bonner et al., 2016) or iterative re-partitioning of the matrix $\boldsymbol{T}$ (Hazelton, 2015); the latter requires $\boldsymbol{T}$ to satisfy a unimodularity property. While these stochastic methods share considerable ingenuity and versatility, they typically involve long computation times and require expert implementation in practice.

Here, we propose a new solution for latent models of the form $\boldsymbol{y} = \boldsymbol{T}\boldsymbol{z}$, using a saddlepoint approximation to the likelihood $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y})$. The saddlepoint approximation is a general method for approximating a probability density function when the associated moment generating function is known. It offers a promising alternative to simulation-based inference in cases where exact densities of random variables are intractable, but their moment generating functions are readily available. Such cases include sums or transformations of latent variables, especially where Gaussian approximations are not suitable due to small samples or integer-valued variables. In our case, the saddlepoint method enables us to construct a likelihood based directly on $\Pr(\boldsymbol{y} \mid \boldsymbol{\theta})$, without involving the latent variable $\boldsymbol{z}$.

The saddlepoint approximation was first proposed by Daniels (1954), and was developed by subsequent authors including Lugannani and Rice (1980), Reid (1988), Barndorff-Nielsen and Cox (1989), and Goutis and Casella (1999). Butler (2007) gives an accessible introduction with practical applications. The approximation is known for its accuracy, even in the tails

of a distribution (Butler, 2007; Brazzale and Davison, 2008), but it is surprisingly underused, especially for estimation applications. A recent exception is Pedeli, Davison, and Fokianos (2015), who demonstrated excellent performance of the saddlepoint approximation for maximum likelihood inference on integer-valued autoregressive processes. Saddlepoint probability densities are not normalized, which may create problems in some contexts, but these may be resolvable by more advanced adaptations (Kleppe and Skaug, 2008).

Although the saddlepoint approach applies to any model for $\boldsymbol{z}$, discrete or continuous, we focus here on the discrete case, in particular where $\boldsymbol{z}$ has a multinomial distribution. We describe such models as *latent multinomial models* (Link et al., 2010). The latent multinomial class is noteworthy because an exact likelihood is available for one model in the class, namely model $M_{t,\alpha}$ for modeling misidentification in capture-recapture studies (Vale et al., 2014). The exact computation is based on a combinatorial reformulation of the likelihood, and does not generalize to other models. However, it creates a significant opportunity to evaluate the multinomial saddlepoint approximation in an authentic setting. Multinomial models are suitable for most of the aforementioned applications, although other models may be favored in some contexts (Hazelton, 2015). Focusing on the latent multinomial class, we conduct extensive empirical tests to validate the saddlepoint approach using model $M_{t,\alpha}$ and other models. We also give an efficient implementation for maximizing the saddlepoint likelihood using customary R software (R Core Team, 2018). This implementation is suitable for any model for $\boldsymbol{z}$.

We derive the saddlepoint approximation and its implementation in Section 2. In Section 3, we explore its performance for model $M_{t,\alpha}$ by comparison with the exact likelihood of Vale et al. (2014). In Section 4, we apply the method to multi-list capture-recapture models for estimating disease prevalance, and show that the approximation gives fast and accurate inference. An application to multi-way contingency tables summarized by subsets

of marginal totals is given in Web Appendix A. Although we focus on maximum likelihood estimation, we note that the saddlepoint likelihood could equally well be used in a Bayesian framework, to enable fast inference without needing to sample the latent vector $\boldsymbol{z}$. Compared with simulation-intensive alternatives, we propose that the saddlepoint methodology is more general, easier to implement, and yields much faster computation times.

2. Likelihood Inference for Latent Multinomial Models

2.1 *Notation*

We define the random vectors underlying $\boldsymbol{y}$ and $\boldsymbol{z}$ to be $\boldsymbol{Y} = (Y_1, \ldots, Y_I)^\top$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_J)^\top$. Then $\boldsymbol{T}$ is a known $I \times J$ matrix such that $\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Z}$. Suppose $\boldsymbol{Z}$ follows a multinomial distribution with index N and cell probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)^\top$. We use $\boldsymbol{\theta}$ for any parameters of $\boldsymbol{Z}$ to be estimated, including those underlying $\boldsymbol{\pi}$ and N as required. In the context of latent multinomial models, we aim to estimate $\boldsymbol{\theta}$ from observed data $\boldsymbol{y}$.

To illustrate notation, consider a 2×3 contingency table whose six entries constitute the latent J-vector $\boldsymbol{Z} = (Z_{11}, Z_{12}, Z_{13}, Z_{21}, Z_{22}, Z_{23})^\top$. We model $\boldsymbol{Z}$ as multinomial with known index N and cell probabilities $\boldsymbol{\pi}(\boldsymbol{\theta})$. Our interest is in the case where information on $\boldsymbol{Z}$ is released only via marginal totals, perhaps to satisfy concurrent standards of participant privacy and open data (Dobra et al., 2006). The five observed marginal totals constitute the I-vector $\boldsymbol{Y} = (Z_{1+}, Z_{2+}, Z_{+1}, Z_{+2}, Z_{+3})^\top$. The 5×6 matrix $\boldsymbol{T}$ is readily derived such that $\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Z}$. However, the distribution of $\boldsymbol{Y}$ is not multinomial, and the vector $\boldsymbol{Z}$ is not recoverable from $\boldsymbol{Y}$. For example, the observation $\boldsymbol{y} = (6, 15, 5, 7, 9)^\top$ could arise from either $\boldsymbol{z} = (1, 2, 3, 4, 5, 6)^\top$ or $\boldsymbol{z} = (3, 2, 1, 2, 5, 8)^\top$. Our aim is to derive a likelihood $\mathcal{L}(\boldsymbol{\theta} \,|\, \boldsymbol{y})$ that does not involve knowledge of $\boldsymbol{z}$.

2.2 *Likelihood factorization and row-reduction of the link matrix*

Given the observed vector $\boldsymbol{y}$, the latent vector $\boldsymbol{z}$ generally has a large number of feasible solutions; however, for some models, it can happen that some components of $\boldsymbol{z}$ are determined

by the vector $\boldsymbol{y}$. For example, if $T_{ij} = 1$ is the only nonzero entry in the ith row of the matrix $\boldsymbol{T}$, we have $z_j = y_i$. If $\boldsymbol{y}$ has a component $y_k = z_l + z_m$ observed to be zero, then the corresponding components z_l and z_m of $\boldsymbol{z}$ are also known to be zero, since z_l and z_m are non-negative counts. The zero components of $\boldsymbol{y}$ are not the same in different data realizations, so the components of $\boldsymbol{z}$ known to be zero will differ according to the data. For cases where some components of $\boldsymbol{z}$ are determined by $\boldsymbol{y}$, we first apply a likelihood factorization step. This is necessary for the validity of the saddlepoint method if $\boldsymbol{y}$ contains zeros, and can also improve efficiency in general. Reasons for this step are given in Section 2.3. If the factorization step is not needed, we proceed directly to the penultimate paragraph of this section.

Suppose we can find unique solutions for R elements of $\boldsymbol{z}$ from the observed vector $\boldsymbol{y}$. We reorder the elements of $\boldsymbol{z}$ such that these R values appear first, and write $\boldsymbol{z} = \left(\boldsymbol{v}^\top, \boldsymbol{u}^\top\right)^\top$ to distinguish the R known values $\boldsymbol{v} = (z_1, \ldots, z_R)^\top$ from the unknown values $\boldsymbol{u} = (z_{R+1}, \ldots, z_J)^\top$. Accordingly, we also reorder the elements of $\boldsymbol{\pi}$, and the columns of $\boldsymbol{T}$. We continue to use the equation $\boldsymbol{y} = \boldsymbol{T}\boldsymbol{z}$, although $\boldsymbol{z}$ and $\boldsymbol{T}$ have been reordered.

We partition matrix $\boldsymbol{T}$ as $\boldsymbol{T} = (\boldsymbol{B}, \boldsymbol{A})$, where $\boldsymbol{B}$ is an $I \times R$ matrix that contains the first R columns of $\boldsymbol{T}$, and $\boldsymbol{A}$ is an $I \times (J - R)$ matrix that contains the remaining $J - R$ columns. It follows that $\boldsymbol{y} = \boldsymbol{T}\boldsymbol{z} = \boldsymbol{B}\boldsymbol{v} + \boldsymbol{A}\boldsymbol{u}$, and thus the latent vector $\boldsymbol{u}$ satisfies $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{y} - \boldsymbol{B}\boldsymbol{v}$. For convenience, let $\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{B}\boldsymbol{v}$. Since $\boldsymbol{y}$, $\boldsymbol{v}$, and $\boldsymbol{B}$ are all known, $\boldsymbol{x}$ is a known vector. However, the equation $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{x}$ still has a large number of feasible solutions for $\boldsymbol{u}$, so the original problem still remains.

The likelihood factorization is formulated as follows:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \sum_{\boldsymbol{z}:\boldsymbol{T}\boldsymbol{z}=\boldsymbol{y}} \Pr(\boldsymbol{Z} = \boldsymbol{z}) &= \sum_{\boldsymbol{u}:\boldsymbol{A}\boldsymbol{u}=\boldsymbol{x}} \Pr(\boldsymbol{Z}_{1:R} = \boldsymbol{v} \cap \boldsymbol{Z}_{R+1:J} = \boldsymbol{u}) \\ &= \Pr(\boldsymbol{Z}_{1:R} = \boldsymbol{v}) \sum_{\boldsymbol{u}:\boldsymbol{A}\boldsymbol{u}=\boldsymbol{x}} \Pr(\boldsymbol{Z}_{R+1:J} = \boldsymbol{u} \mid \boldsymbol{Z}_{1:R} = \boldsymbol{v}), \end{aligned} \tag{1}$$

where the random vector $\boldsymbol{Z}_{1:R}$ contains the first R components of the multinomial vector $\boldsymbol{Z}$, and $\boldsymbol{Z}_{R+1:J}$ contains all remaining components of $\boldsymbol{Z}$. By the multinomial marginal property,

we have

$$\Pr\left(\boldsymbol{Z}_{1:R}=\boldsymbol{v}\right)=\frac{N!}{z_1!\ldots z_R!\,(N-v^*)!}\left(\prod_{j=1}^{R}\pi_j^{z_j}\right)\left(1-\sum_{j=1}^{R}\pi_j\right)^{N-v^*}, \tag{2}$$

where $v^*=\sum_{j=1}^{R}z_j$ is the sum of all elements of vector $\boldsymbol{v}$.

The problem of evaluating the likelihood $\mathcal{L}\left(\boldsymbol{\theta}\mid\boldsymbol{y}\right)$ now reduces to finding

$$\sum_{\boldsymbol{u}:\boldsymbol{A}\boldsymbol{u}=\boldsymbol{x}}\Pr\left(\boldsymbol{Z}_{R+1:J}=\boldsymbol{u}\mid\boldsymbol{Z}_{1:R}=\boldsymbol{v}\right)=\Pr\left(\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{v}}=\boldsymbol{x}\right), \tag{3}$$

where $\boldsymbol{U}_{\boldsymbol{v}}$ is a random variable following the conditional distribution of $\boldsymbol{Z}_{R+1:J}\mid\boldsymbol{Z}_{1:R}=\boldsymbol{v}$. By the multinomial conditional property, we have $\boldsymbol{U}_{\boldsymbol{v}}\sim\text{Multinomial}\left(\widetilde{N};\widetilde{\boldsymbol{\pi}}\right)$, where $\widetilde{N}=N-v^*$ and $\widetilde{\boldsymbol{\pi}}=\left(\pi_{R+1},\ldots,\pi_J\right)^\top/\sum_{j=R+1}^{J}\pi_j$. Let $\boldsymbol{X}=\boldsymbol{A}\boldsymbol{U}_{\boldsymbol{v}}$. The problem of computing (3) is solved if we can find the probability mass function of $\boldsymbol{X}$ given a known matrix $\boldsymbol{A}$ and a multinomial distribution $\boldsymbol{U}_{\boldsymbol{v}}$ of dimension $J-R$. For brevity, we use $\boldsymbol{U}$ to replace $\boldsymbol{U}_{\boldsymbol{v}}$ hereafter, and $\widetilde{\boldsymbol{\pi}}=\left(\widetilde{\pi}_1,\ldots,\widetilde{\pi}_H\right)^\top$ to replace its original formulation, where $H=J-R$.

The matrix $\boldsymbol{A}$ is of dimension $I\times H$, but it is not necessarily of full row-rank. Lower-rank matrices generate observed vectors $\boldsymbol{x}$ with redundant information, in the sense that some elements of $\boldsymbol{x}$ are determined by the other elements. For cases where we skip the factorization step and work with the original formulation $\boldsymbol{y}=\boldsymbol{T}\boldsymbol{z}$, we may find similarly that $\boldsymbol{T}$ is not of full row-rank. The proposed saddlepoint method requires a matrix of full row-rank for reasons described in Section 2.3. In these cases we must reduce $\boldsymbol{A}$ or $\boldsymbol{T}$ to a submatrix consisting of a maximal set of linearly independent rows. This procedure does not affect estimation results, as explained below.

We illustrate the row-reduction procedure using the notation $\boldsymbol{x}=\boldsymbol{A}\boldsymbol{u}$. Suppose matrix $\boldsymbol{A}$ has row-rank $L<I$. We start by using any non-zero row of matrix $\boldsymbol{A}$ as the submatrix, and attempt to add other rows of $\boldsymbol{A}$ one by one. If a new row increases the row-rank of the submatrix by one, we accept it and update the submatrix; otherwise we reject it. We finally obtain a submatrix $\boldsymbol{A}_1$ of row-rank L, and a corresponding subvector $\boldsymbol{x}_1$ of $\boldsymbol{x}$ such that

$\boldsymbol{x}_1 = \boldsymbol{A}_1\boldsymbol{u}$. Estimation results are not affected by using $\boldsymbol{x}_1$ instead of $\boldsymbol{x}$ because the missing data points are determined by those in $\boldsymbol{x}_1$. See Web Appendix B for proof that estimation using $\boldsymbol{x}_1$ is equivalent to estimation using $\boldsymbol{x}$, and is invariant to the choice of submatrix $\boldsymbol{A}_1$. For brevity, we still use the notations $\boldsymbol{A}$ and $\boldsymbol{x}$ to denote the submatrix and the subvector. Assume matrix $\boldsymbol{A}$ is now of dimension $L \times H$ and vector $\boldsymbol{x}$ is of dimension L, where $L \leqslant I$.

2.3 *Likelihood approximation using the saddlepoint method*

To approximate a probability density $f(\boldsymbol{x})$, the saddlepoint method uses a Taylor expansion of the log-integrand in the integral inversion of the moment generating function, known as the Bromwich integral. The Taylor expansion is taken about a 'saddlepoint', $\hat{s}(\boldsymbol{x})$, specific to the value $\boldsymbol{x}$ and chosen to eliminate the linear term of the expansion. This leaves a Gaussian formulation for the leading terms of the integrand, engendering accuracy in the Laplace approximation which is then used to evaluate the integral. The result is a simple formula for the approximated $\tilde{f}(\boldsymbol{x})$. The derivation and analysis of the method are complicated by the oscillatory nature of the integrand over the complex plane. While the use of a saddlepoint customized to each $\boldsymbol{x}$ makes the approximation very accurate, it creates the computational challenge of finding the saddlepoint for each evaluation.

The joint moment generating function of $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{U}$ is

$$M_{\boldsymbol{X}}(\boldsymbol{s}) = \mathrm{E}\left\{\exp\left(\boldsymbol{s}^\top \boldsymbol{X}\right)\right\} = \mathrm{E}\left[\exp\left\{\left(\boldsymbol{A}^\top \boldsymbol{s}\right)^\top \boldsymbol{U}\right\}\right] = M_{\boldsymbol{U}}\left(\boldsymbol{A}^\top \boldsymbol{s}\right),$$

where $M_{\boldsymbol{U}}$ is the moment generating function of $\boldsymbol{U}$, and $\boldsymbol{s} = (s_1, \ldots, s_L)^\top$ takes values in $\mathbb{R}^L$ for which the expectation of $\exp\left(\boldsymbol{s}^\top \boldsymbol{X}\right)$ exists. Let $\boldsymbol{t} = \boldsymbol{A}^\top \boldsymbol{s} = (t_1, \ldots, t_H)^\top \in \mathbb{R}^H$. Since $\boldsymbol{U}$ follows a multinomial distribution in the latent multinomial framework, we have

$$M_{\boldsymbol{X}}(\boldsymbol{s}) = M_{\boldsymbol{U}}(\boldsymbol{t}) = \left\{\sum_{h=1}^{H} \widetilde{\pi}_h \exp(t_h)\right\}^{\widetilde{N}}.$$

The cumulant generating function of $\boldsymbol{X}$, which is defined as the logarithm of $M_{\boldsymbol{X}}(\boldsymbol{s})$, is:

$$K_{\boldsymbol{X}}(\boldsymbol{s}) = \log M_{\boldsymbol{X}}(\boldsymbol{s}) = \widetilde{N} \log \left\{ \sum_{h=1}^{H} \widetilde{\pi}_h \exp(t_h) \right\}.$$

Following Butler (2007), the probability mass function of any L-dimensional random variable $\boldsymbol{X}$ can be approximated by the saddlepoint density,

$$\tilde{f}_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{L/2} |K''_{\boldsymbol{X}}(\hat{\boldsymbol{s}})|^{1/2}} \exp\left\{ K_{\boldsymbol{X}}(\hat{\boldsymbol{s}}) - \hat{\boldsymbol{s}}^{\top} \boldsymbol{x} \right\}, \tag{4}$$

where the L-vector $\hat{\boldsymbol{s}}$ solves the saddlepoint equation for the first derivative of $K_{\boldsymbol{X}}$,

$$K'_{\boldsymbol{X}}(\boldsymbol{s}) = \boldsymbol{x}, \tag{5}$$

and $|K''_{\boldsymbol{X}}(\hat{\boldsymbol{s}})|$ denotes the determinant of the Hessian matrix of $K_{\boldsymbol{X}}(\boldsymbol{s})$ evaluated at $\hat{\boldsymbol{s}}$. In most cases, the saddlepoint equation (5) cannot be solved analytically. In practice we regard it as an optimization problem, and find $\hat{\boldsymbol{s}}$ to minimize $K_{\boldsymbol{X}}(\boldsymbol{s}) - \boldsymbol{s}^{\top}\boldsymbol{x}$ by numerical methods such as the Newton–Raphson method.

Substituting equations (2) to (4) into (1) generates an approximate likelihood for latent multinomial models,

$$\widetilde{\mathcal{L}}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \Pr(\boldsymbol{Z}_{1:R} = \boldsymbol{v})\, \tilde{f}_{\boldsymbol{X}}(\boldsymbol{x}). \tag{6}$$

For latent multinomial models, the saddlepoint equation (5) to be solved for $\hat{\boldsymbol{s}}$ expands to

$$\left.\frac{\partial K_{\boldsymbol{X}}(\boldsymbol{s})}{\partial s_l}\right|_{\boldsymbol{s}=\hat{\boldsymbol{s}}} = \frac{\widetilde{N} \sum_{h=1}^{H} A_{lh} \widetilde{\pi}_h \exp(\hat{t}_h)}{\sum_{h=1}^{H} \widetilde{\pi}_h \exp(\hat{t}_h)} = x_l, \quad l = 1, \ldots, L, \tag{7}$$

where $\hat{\boldsymbol{t}} = (\hat{t}_1, \ldots, \hat{t}_H)^{\top} = \boldsymbol{A}^{\top}\hat{\boldsymbol{s}}$. Equation (7) reveals the reasons for the likelihood factorization and row-reduction steps described in Section 2.2. Firstly, the matrix $\boldsymbol{A}$ consists only of non-negative entries for all models considered here, so the middle term of (7) is strictly positive for all finite $\boldsymbol{s}$. Consequently, if any observation $x_l = 0$ in the rightmost term of (7), there is no finite solution to (7) for the saddlepoint $\hat{\boldsymbol{s}}$, and therefore no valid saddlepoint approximation $\tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})$. The likelihood factorization described in Section 2.2 is

therefore necessary to ensure that there are no zeros in the observed data vector $\boldsymbol{x}$ to which the saddlepoint approximation is applied.

Secondly, Butler (2007) pointed out that the cumulant generating function $K_{\boldsymbol{X}}(\boldsymbol{s})$ must be strictly convex, to ensure that the saddlepoint equation has a solution and the square root of $|K''_{\boldsymbol{X}}(\hat{\boldsymbol{s}})|$ in (4) is defined and strictly positive. It can be seen from (7) that if $\boldsymbol{A}$ is not of full row-rank, then $\partial K_{\boldsymbol{X}}(\boldsymbol{s})/\partial s_l$ for $l = 1, \ldots, L$ are not linearly independent, and consequently the Hessian matrix $K''_{\boldsymbol{X}}(\boldsymbol{s})$ is also not of full rank. Thus the row-reducing exercise in Section 2.2 is needed to ensure $\boldsymbol{A}$ has full row-rank, because otherwise $|K''_{\boldsymbol{X}}(\hat{\boldsymbol{s}})| = 0$ appears in the denominator of (4) and the saddlepoint density is not defined.

2.4 *Implementation in TMB*

Approximate maximum likelihood estimates of $\boldsymbol{\theta}$ are found by minimizing the negative logarithm of (6) in the usual manner. To compute confidence intervals, we use a lognormal distribution for N, and a normal distribution for other parameters (Vale et al., 2014). Gradient-based minimization is complicated because each evaluation of the likelihood involves an inner optimization to obtain $\hat{\boldsymbol{s}}(\boldsymbol{\theta})$ in (5). In Web Appendix C, we show how this may be tackled by an adaptation of the R package TMB (Template Model Builder: Kristensen et al., 2016). Using automatic differentiation, the arg-min value $\hat{\boldsymbol{s}}(\boldsymbol{\theta})$ can be incorporated into a machine-precision gradient function for $\boldsymbol{\theta}$ suitable for the outer optimization, without the need for any symbolic derivatives to be supplied.

3. Validation of the Saddlepoint Method

3.1 *Model $M_{t,\alpha}$ for misidentification in capture-recapture studies*

We introduce model $M_{t,\alpha}$ briefly, following Link et al. (2010) and Vale et al. (2014). Model $M_{t,\alpha}$ is of interest because it is the only latent multinomial model for which an exact likelihood is efficiently computable, based on a combinatorial reformulation of the problem (Vale et al., 2014). Suppose we wish to estimate the size N of a closed animal population, using an error-

prone scheme such as DNA samples or photographs to identify individuals. Assume that animals are independent and each animal has probability p_t of being captured on capture occasion $t = 1, \ldots, K$. Denote α to be the probability that a captured animal is correctly identified on each occasion, for example that its genotype is obtained without error. For model $M_{t,\alpha}$, we have $\boldsymbol{\theta} = (N, \alpha, p_1, \ldots, p_K)$.

Under model $M_{t,\alpha}$, there are three possible events for each animal on occasion t, namely, it was not captured, it was captured and identified correctly, or it was captured but misidentified. We use codes 0, 1, and 2 to represent the three events, occurring with probabilities $1 - p_t$, αp_t, and $(1 - \alpha)\, p_t$. For example, an individual with latent history 1012 was captured and correctly identified on occasions 1 and 3, captured but misidentified on occasion 4, and not captured on occasion 2. There are $J = 3^K$ possible latent histories. Observable histories for model $M_{t,\alpha}$, by contrast, consist of only two codes 0 and 1, representing non-capture and capture. For example, the history 1001 denotes an animal that was observed at times 1 and 4 and whose samples from these occasions were correctly matched. Excluding the null history 0...0, we have $I = 2^K - 1$ possible observable histories.

Latent histories containing code 2 cannot be observed due to identification errors. Model $M_{t,\alpha}$ assumes that the same identification error never occurs twice, and that an individual is never misidentified as other captured individuals. Following these assumptions, an animal with latent history 1221 necessarily produces three observed histories: 1001, 0100, and 0010. The vectors $\boldsymbol{Z}$ and $\boldsymbol{Y}$ comprise frequencies of the latent and observable histories respectively. The matrix $\boldsymbol{T}$ connecting $\boldsymbol{Y}$ and $\boldsymbol{Z}$ for model $M_{t,\alpha}$ can be derived according to the known relationships between the observable and latent histories (Link et al., 2010).

The latent vector $\boldsymbol{Z}$ follows a multinomial distribution with index N and cell probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)^\top$, where for $j = 1, \ldots, J$,

$$\pi_j = \prod_{t=1}^{K} \left[p_t^{\mathcal{I}\{\lambda_{jt}>0\}} (1 - p_t)^{\mathcal{I}\{\lambda_{jt}=0\}} \alpha^{\mathcal{I}\{\lambda_{jt}=1\}} (1 - \alpha)^{\mathcal{I}\{\lambda_{jt}=2\}} \right],$$

where λ_{jt} is the capture code for latent history λ_j on occasion t and $\mathcal{I}\{\cdot\}$ is the usual indicator function. The parameter of interest is the population size, N.

3.2 *Comparison of saddlepoint and exact MLEs*

We simulated data sets in R using every combination of parameter values chosen from $K \in \{4, 6, 8, 10\}$, $N \in \{400, 1000\}$, $\alpha \in \{0.8, 0.9, 0.95, 0.97\}$, and $p_1 = \cdots = p_K \in \{0.1, 0.2, 0.3, 0.4\}$. For each setting we generated 500 data sets, and for each data set we calculated maximum likelihood estimates and 95% confidence intervals using both the saddlepoint method and the exact computation of Vale et al. (2014).

[Figure 1 about here.]

Figure 1 shows parameter estimates obtained from the two methods, using settings $N = 400$, $\alpha = 0.97$, and $p_1 = \cdots = p_8 = 0.1$. The saddlepoint method consistently gives almost identical estimation results to the exact likelihood function. Estimates of p_t for $t = 1, \ldots, 8$, and α from the two approaches are typically identical to four or five decimal places, while estimates of N are typically identical to one or two decimal places. The two methods also produce almost the same variance estimates, so confidence intervals for all parameters from the two methods are almost identical.

For all other settings listed above, as well as many others outside this range, scatterplots maintained the same pattern of agreement seen in Figure 1. This applies even for N as small as 50; we did not explore lower values of N because maximum likelihood inference itself becomes inherently inaccurate for very small N (Vale et al., 2014). The two methods continue to agree when α or p_t take values close to their boundaries. We did not observe any case where the two approaches yielded noticeably different estimates or confidence intervals.

Average fitting time using the saddlepoint method for one simulated data set using the settings in Figure 1 was roughly 10 seconds on a customary laptop with a clock speed of 1.3 GHz, contrasting with approximately 4 seconds using the exact likelihood on the same

machine. The slight loss of efficiency arises because every evaluation of the saddlepoint likelihood involves numerical solution of the inner optimization problem.

3.3 *Comparison with Bayesian estimation*

We used a data set generated by Link et al. (2010) under a specific setting of $M_{t,\alpha}$ with $K = 5$. As usual, we obtained almost identical parameter estimates and confidence intervals to those reported by Vale et al. (2014), which are extremely close to the results of Link et al. (2010). The saddlepoint method cost 0.9 seconds on a 1.3 GHz laptop, while Link et al. (2010) indicated that their Bayesian method cost over 30 minutes on a 3.8 GHz machine.

3.4 *Comparison of saddlepoint and exact likelihood curves*

We investigate the performance of the saddlepoint approximation in reproducing likelihood curves for model $M_{t,\alpha}$, which are available from the computation of Vale et al. (2014). We generated data sets by simulation and plotted exact and saddlepoint log-likelihoods, each expressed in terms of the parameter N while $\alpha, p_1, \ldots, p_K$ are fixed at their maximum likelihood estimates under the exact method. A suite of examples with various settings is shown in Figure 2.

From Figure 2, and many other similar plots, we see that the log-likelihoods obtained using the two methods do not match one another perfectly; however, the two functions differ by an almost constant value. This presumably reflects the missing normalization constant in the saddlepoint formulation, the value of which differs for different settings. The discrepancy does not affect either the position of the maximum or the curvature near the maximum, allowing the methods to deliver indistinguishable estimates and confidence intervals.

[Figure 2 about here.]

The difference between the saddlepoint and exact log-likelihoods appears to decrease when p_t increases, or when the number of capture occasions K decreases while other parameters

remain the same. Each of these scenarios leads to an increase in the proportion of relatively large frequencies in the vector $\boldsymbol{x}$. Inspection of numerous results indicates that if most components of the vector $\boldsymbol{x}$ are at least five, the saddlepoint method yields an extremely good approximation to the exact log-likelihood function. For example, the second and third panels in the top row of Figure 2 present two scenarios where all components of $\boldsymbol{x}$ are over five. When the number of capture occasions increases, it is more difficult to observe a vector $\boldsymbol{x}$ with most components larger than five, and accordingly differences between the two log-likelihoods are larger for the three scenarios shown in the bottom row.

4. Applications

4.1 *Multi-list models for capture-recapture in social sciences*

Capture-recapture methods are used in social sciences to estimate the size of a human population based on records distributed across several administrative lists. Each list uses tags such as name or health insurance number to identify individuals. Here, we consider the case where some lists do not share a common tag. For a K-list problem, the latent history λ of an individual is defined as $\lambda_1 \dots \lambda_K$, where λ_k is 1 if the individual is on list k, otherwise 0, for $k = 1, \dots, K$. There are $J = 2^K$ latent histories. For example, when $K = 4$, an individual with latent history 1010 is on lists 1 and 3, but not on lists 2 and 4. We assume individuals are matched correctly between lists when a common tag is available.

[Figure 3 about here.]

The set of observable histories differs for different list structures. Following Sutherland and Schwarz (2005), we use a graph such as that shown in Figure 3a to illustrate the procedure of finding observable histories. Lists are represented by vertices, and are joined by edges if they share a common tag. The example in Figure 3a requires at least two different tags, one for matching records on lists 1, 2, and 3, and another for matching records on lists 3 and 4. Latent histories with code 1 for list 3 are fully observed for this list structure. Other latent

histories are not observable: for example, the latent history 1101 is observed as two vague histories 110· and ··01, where "·" means that it is unknown whether or not an individual is recorded on a list. Sutherland and Schwarz (2005) show that there are 12 observable histories, namely {100·, 010·, 110·, 1010, 1011, 0110, 0111, 0010, 0011, 1110, 1111, ··01}, and they present the 12×16 matrix $\boldsymbol{T}$ such that $\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Z}$, where $\boldsymbol{Z}$ and $\boldsymbol{Y}$ comprise counts of latent and observed histories respectively.

In the modeling framework of Sutherland and Schwarz (2005), it is assumed that

$$\boldsymbol{Z} \sim \text{Poisson}\left(\boldsymbol{\mu_Z}\right), \;\; \log\left(\boldsymbol{\mu_Z}\right) = \boldsymbol{W}\boldsymbol{\beta}, \tag{8}$$

where $\boldsymbol{W}$ is a design matrix that typically consists of zero and one entries, and $\boldsymbol{\beta}$ is a vector of parameters. The vector $\boldsymbol{\beta}$ typically consists of an intercept β_0, main effects β_k for lists $k = 1, \ldots, K$, and interaction effects between some pairs of the lists, such as 2-way interactions β_{kl} between lists k and l for $k, l \in \{1, \ldots, K\}$ with $k < l$.

Previous authors have noted that the distribution of $\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Z}$ resulting from (8) is unknown, with dependence among components, and have employed a quasi-likelihood approach for inference on $\boldsymbol{\beta}$ based on data $\boldsymbol{Y}$ (Sutherland and Schwarz, 2005; Lee, 2002). A separate step is used for estimating N given $\hat{\boldsymbol{\beta}}$. In our analysis, we instead apply a multinomial model to describe the latent vector $\boldsymbol{Z}$. This provides a natural way of estimating N together with cell probabilities that match those of the Poisson model (8), and properly accommodates dependence between cells of $\boldsymbol{Y}$:

$$\boldsymbol{Z} \sim \text{Multinomial}\left(N; \boldsymbol{\pi}\right), \;\; \boldsymbol{\pi} = \frac{\exp\left(\boldsymbol{W}\boldsymbol{\beta}\right)}{\sum \exp\left(\boldsymbol{W}\boldsymbol{\beta}\right)}. \tag{9}$$

The probability vector $\boldsymbol{\pi}$ does not depend on the parameter β_0, because the first column of the matrix $\boldsymbol{W}$ consists entirely of one-entries, causing $\exp\left(\beta_0\right)$ to vanish from (9). The parameters $\boldsymbol{\theta}$ in our model therefore include all components of $\boldsymbol{\beta}$ except for the intercept β_0, as well as the extra parameter N, so the two models have the same number of parameters.

4.2 *Example: Auckland diabetes study*

We consider multi-list data from a study for estimating the prevalence of diabetes in Auckland, New Zealand (Sutherland and Schwarz, 2005). The list structure considered by Sutherland and Schwarz (2005) is given in Figure 3b. There are $J = 16$ latent capture histories, written as $\boldsymbol{\lambda} = (\lambda_G, \lambda_P, \lambda_O, \lambda_D)$, where $\lambda_i = 1$ if the individual is on list i and $\lambda_i = 0$ otherwise, for $i \in \{G, P, O, D\}$. The observable histories are {01··, 0·1·, 0··1, 1000, 1001, 1010, 1011, 1100, 1101, 1110, 1111}. The observed data vector of counts is $\boldsymbol{y} = (1183, 12265, 3276, 654, 51, 366, 91, 40, 4, 5, 14)^\top$ (Sutherland and Schwarz, 2005).

Sutherland and Schwarz (2005) selected the Poisson log-linear model $[GP = OD][GO = GD = PO = PD]$ to describe the latent vector $\boldsymbol{Z}$, because it yielded the lowest QIC_u statistic (Pan, 2001) among their candidate models. QIC is an analog of the Akaike Information Criterion (AIC) (Akaike, 1974) when parameters are estimated using estimating functions. The model selected includes all possible 2-way interactions, but some of these are set to be equal in the interests of parsimony. For example, the notation $[GP = OD]$ means that the interaction effect between lists G and P is set equal to that between lists O and D, i.e. $\beta_{GP} = \beta_{OD}$. Thus the parameter vector for this model is $\boldsymbol{\beta} = (\beta_0, \beta_G, \beta_P, \beta_O, \beta_D, \beta_{GP}, \beta_{GO})^\top$.

[Table 1 about here.]

The estimate $\hat{N}$ of the number of diabetes sufferers from the Poisson quasi-likelihood approach of Sutherland and Schwarz (2005) is 45,853. Associated with the estimate are three standard errors, 4530, 4343, and 4008, calculated in different ways. Here, we apply the multinomial saddlepoint method using the same interaction effects and obtain a smaller estimate 43,422, with a similar standard error 4303. Computation time was less than one second on a 1.3 GHz laptop. The two methods yield slightly different estimates of the population size; however, they give almost identical estimates and standard errors for other model parameters as shown in Table 1. This suggests that the difference in $\hat{N}$ is due to the

ad-hoc step required under the Poisson formulation to estimate N as a derived parameter (Sutherland and Schwarz, 2005). The multinomial formulation favored here incorporates estimation of N, as well as non-independence of cell entries, in a natural fashion.

4.3 *Model selection*

We use the Auckland diabetes data to investigate performance of AIC for model selection. Because AIC is based on the absolute value of the maximized log-likelihood, caution is needed when applying it to saddlepoint likelihoods due to the missing normalization constant shown in Figure 2. Model comparisons by AIC may be invalid if the normalization constant differs substantially between different models; but in general this constant is not known or computable. However, the investigations in Section 3.4 suggest that the missing normalization constant may be negligible if most components of the observed data vector $\boldsymbol{y}$ are moderately large: for example if most $y_i \geqslant 5$. Since this is the case for the Auckland diabetes data, we proceed with AIC for this dataset. If counts in $\boldsymbol{y}$ were not sufficiently large, model selection could instead be based on score tests, which rely only on the derivatives of the log-likelihood function and not on its absolute value (McCrea and Morgan, 2011).

The model $[GP=OD][GO=GD=PO=PD]$ in Table 1 was selected by Sutherland and Schwarz (2005) from 15 candidate models. These models include two with 3-way interactions: $[GPO][GPD][POD]$ and $[GPO=GPD=POD]$, each with all 2-way interactions equal; two with main effects only; and the remaining 11 with various 2-way interactions in different combinations. Sutherland (2003) lists the models examined together with their QIC statistics. We applied saddlepoint estimation to the same 15 models. The saddlepoint AIC results closely mirror the QIC results, with both criteria producing very similar rankings of the 15 models. The only exception of note was the most complex 3-way interaction model, which ranked better according to AIC than to QIC. Both routines yielded the same selected model $[GP=OD][GO=GD=PO=PD]$ with the lowest AIC and QIC. We call this Model 1.

Since Model 1 retains all six 2-way interactions, we also conducted an exhaustive search of all 203 models which partition these six interactions into different equality classes. We fitted all 203 models using the saddlepoint likelihood. The top-ranked model was Model 2: $[GP = OD = PO][GO = GD = PD]$, for which AIC was 5.3 units lower than Model 1. We used a chi-square test on the elements of $\boldsymbol{y}$ to confirm that the fit under Model 2 ($\chi^2_4 = 1.68$; $p = 0.79$) was better than that under Model 1 ($\chi^2_4 = 7.12$; $p = 0.13$). We conclude that model selection using AIC has proved effective for this dataset. Our final estimate of N using Model 2 is $\hat{N} = 37{,}467$ with 95% confidence interval (30,482, 46,051).

4.4 *Simulation study*

Here we show simulations of the saddlepoint method applied to the list structure shown in Figure 3a, and explore different types of list dependence following Sutherland and Schwarz (2005). In the first simulation, we assumed the four lists to be independent and fitted $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4)^\top$. Secondly, we investigated scenarios with simple list interactions using $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4, \beta_{12}, \beta_{13})^\top$. Thirdly, we explored scenarios with interactions between every pair of lists. The parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_4, \beta_{12}, \ldots, \beta_{34})^\top$ contains 10 parameters. We compare results with the quasi-likelihood method of Sutherland and Schwarz (2005).

[Figure 4 about here.]

Figure 4 shows results for one setting of the third scenario. For every parameter except N, the two methods each generate estimation results with negligible bias and almost identical values for mean confidence interval width and confidence interval coverage. However, while the saddlepoint method also gives estimates for the parameter N with no discernable bias and close to nominal confidence interval coverage, the ad-hoc estimator of N from Sutherland and Schwarz (2005) is positively biased with very low confidence interval coverage. This is consistent with the outcome seen in the Auckland diabetes study. Similar results were obtained from the other two simulation scenarios.

4.5 *Application to incomplete contingency tables*

Web Appendix A presents a further application of the saddlepoint method to analysis of multi-way contingency tables that are partially reported using subsets of marginal totals. We analyse epidemiological data on risk factors for coronary thrombosis among Czech factory workers (Dobra et al., 2006), and demonstrate by simulation that the saddlepoint method yields accurate inference.

5. Discussion

The saddlepoint method delivered uniformly fast and accurate inference across numerous parameter settings in all of the latent multinomial models we explored. We did not encounter any case in which the approximation broke down. Using model $M_{t,\alpha}$, for which the exact likelihood is known, we found that saddlepoint inference remained accurate for N as small as 50, suggesting that the saddlepoint approximation holds across the range of application for which maximum likelihood is itself suitable. We have also provided an efficient implementation for maximizing saddlepoint likelihoods using the R package TMB. This implementation is suitable for any saddlepoint likelihood, not just for latent multinomials.

It is straightforward to derive saddlepoint likelihoods for latent models of the form $\boldsymbol{y} = \boldsymbol{T}\boldsymbol{z}$, or joint likelihoods for independent vectors $(\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}) = (\boldsymbol{T_1}\boldsymbol{z_1}, \ldots, \boldsymbol{T_n}\boldsymbol{z_n})$, as long as the moment generating functions of the latent vectors $\boldsymbol{z}$ are known. Empirical work may be needed to verify the approximation for new distributions of $\boldsymbol{z}$. It is anticipated that saddlepoint behavior will generally be good when distributions do not deviate too far from Gaussian forms. If performance is found to be inadequate, saddlepoint approximations with non-Gaussian leading terms offer a promising alternative (Kleppe and Skaug, 2008).

Acknowledgements

We thank the editor, associate editor, and two referees for their constructive and insightful comments. We also thank S. J. Bonner and M. R. Schofield for comments and suggestions

on earlier versions of this work. This work was funded by the China Scholarship Council, and by the Royal Society of New Zealand through Marsden grant 14-UOA-155.

Supplementary Materials

R code for fitting the models in this paper, and Web Appendices A to C referenced in Sections 1, 2 and 4, are available with this paper at the *Biometrics* website on Wiley Online Library.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19,** 716–723.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic techniques for use in statistics*. London: Chapman and Hall.

Bonner, S. J., Schofield, M. R., Noren, P., and Price, S. J. (2016). Extending the latent multinomial model with complex error processes and dynamic Markov bases. *The Annals of Applied Statistics* **10,** 246–263.

Brazzale, A. R. and Davison, A. C. (2008). Accurate parametric inference for small samples. *Statistical Science* **23,** 465–484.

Butler, R. W. (2007). *Saddlepoint approximations with applications*. Cambridge: Cambridge University Press.

Chen, Y., Diaconis, P., Holmes, S. P., and Liu, J. S. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association* **100,** 109–120.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics* **25,** 631–650.

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *The Annals of Statistics* **26,** 363–397.

Dobra, A. (2012). Dynamic Markov bases. *Journal of Computational and Graphical Statistics* **21,** 496–517.

Dobra, A., Tebaldi, C., and West, M. (2006). Data augmentation in multi-way contingency tables with fixed marginal totals. *Journal of Statistical Planning and Inference* **136,** 355–372.

Goutis, C. and Casella, G. (1999). Explaining the saddlepoint approximation. *The American Statistician* **53,** 216–224.

Hazelton, M. L. (2015). Network tomography for integer-valued traffic. *The Annals of Applied Statistics* **9,** 474–506.

Kleppe, T. S. and Skaug, H. J. (2008). Building and fitting non-Gaussian latent variable models via the moment-generating function. *Scandinavian Journal of Statistics* **35,** 664–676.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. J., and Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software* **70,** 1–21.

Lee, A. (2002). Effect of list errors on the estimation of population size. *Biometrics* **58,** 185–191.

Link, W. A., Yoshizaki, J., Bailey, L. L., and Pollock, K. H. (2010). Uncovering a latent multinomial: Analysis of mark–recapture data with misidentification. *Biometrics* **66,** 178–185.

Lugannani, R. and Rice, S. (1980). Saddlepoint approximation for the distribution of the sum of independent random variables. *Advances in Applied Probability* **12,** 475–490.

McCrea, R. S. and Morgan, B. J. (2011). Multistate mark–recapture model selection using score tests. *Biometrics* **67,** 234–241.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biomet-*

rics **57,** 120–125.

Pedeli, X., Davison, A. C., and Fokianos, K. (2015). Likelihood estimation for the INAR(p) model by saddlepoint approximation. *Journal of the American Statistical Association* **110,** 1229–1238.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reid, N. (1988). Saddlepoint methods and statistical inference. *Statistical Science* **3,** 213–227.

Schofield, M. R. and Bonner, S. J. (2015). Connecting the latent multinomial. *Biometrics* **71,** 1070–1080.

Sutherland, J. (2003). *Multi-list methods in closed populations with stratified or incomplete information*. PhD thesis, Simon Fraser University.

Sutherland, J. and Schwarz, C. J. (2005). Multi-list methods using incomplete lists in closed populations. *Biometrics* **61,** 134–140.

Tebaldi, C. and West, M. (1998). Bayesian inference on network traffic using link count data. *Journal of the American Statistical Association* **93,** 557–573.

Vale, R. T. R., Fewster, R. M., Carroll, E. L., and Patenaude, N. J. (2014). Maximum likelihood estimation for model $M_{t,\alpha}$ for capture–recapture data with misidentification. *Biometrics* **70,** 962–971.

Vardi, Y. (1996). Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association* **91,** 365–377.

Yoshizaki, J., Brownie, C., Pollock, K. H., and Link, W. A. (2011). Modeling misidentification errors that result from use of genetic tags in capture–recapture studies. *Environmental and Ecological Statistics* **18,** 27–55.

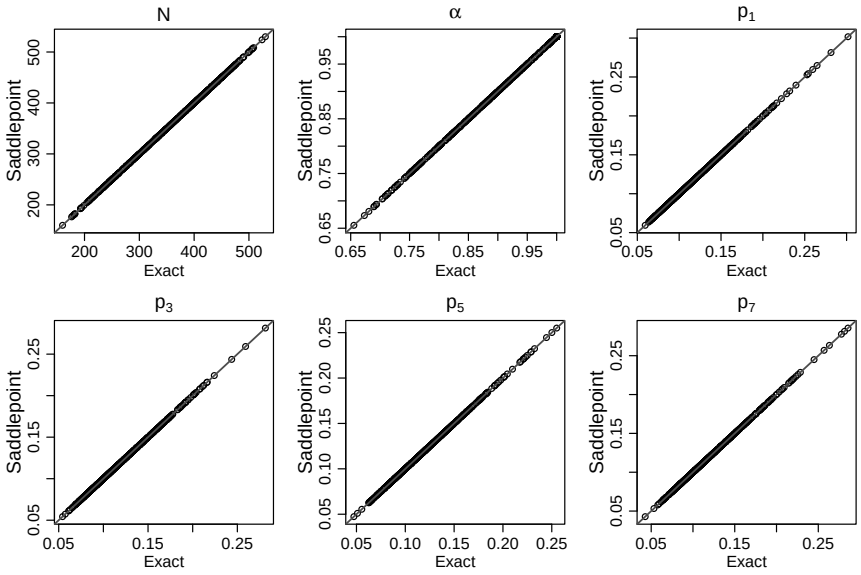


Figure 1. Parameter estimates from 500 simulations using the saddlepoint approximation method and the exact likelihood method of Vale et al. (2014), using the setting $N = 400$, $\alpha = 0.97$, and $p_1 = \cdots = p_8 = 0.1$ under model $M_{t,\alpha}$. Points on straight lines across the plots indicate that the estimates from the two approaches are almost identical.

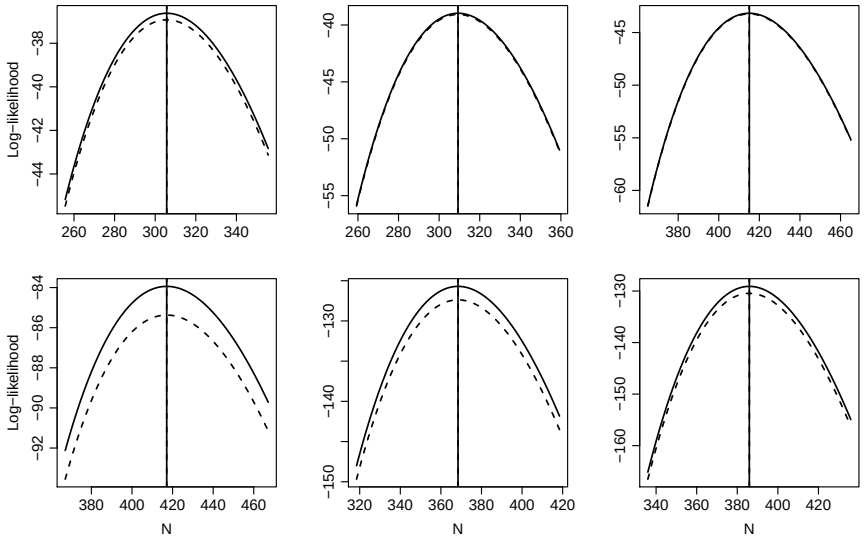


Figure 2. Saddlepoint log-likelihood curves (solid) shown against exact log-likelihood curves (dashed) for model $M_{t,\alpha}$. The number of capture occasions is $K = 4$ in the top row and $K = 6$ in the bottom row. All six panels have $N = 400$ and $\alpha = 0.97$, while $p_1 = \cdots = p_K$ are 0.2, 0.3, and 0.4 for left, center, and right panels respectively. Vertical lines show the positions of maxima under the saddlepoint computation (solid) and the exact computation (dashed). These cannot be distinguished because the saddlepoint estimates and the exact estimates of N are extremely close to one another.

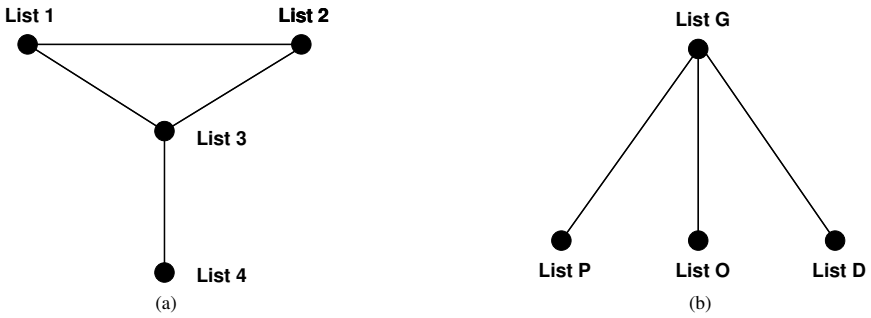


Figure 3. List structures for: (a) a four-list two-tag example; and (b) the Auckland diabetes study. Lists in (b) are: general practitioner records (G); pharmacy records (P); outpatient records (O); and inpatient discharge records (D).

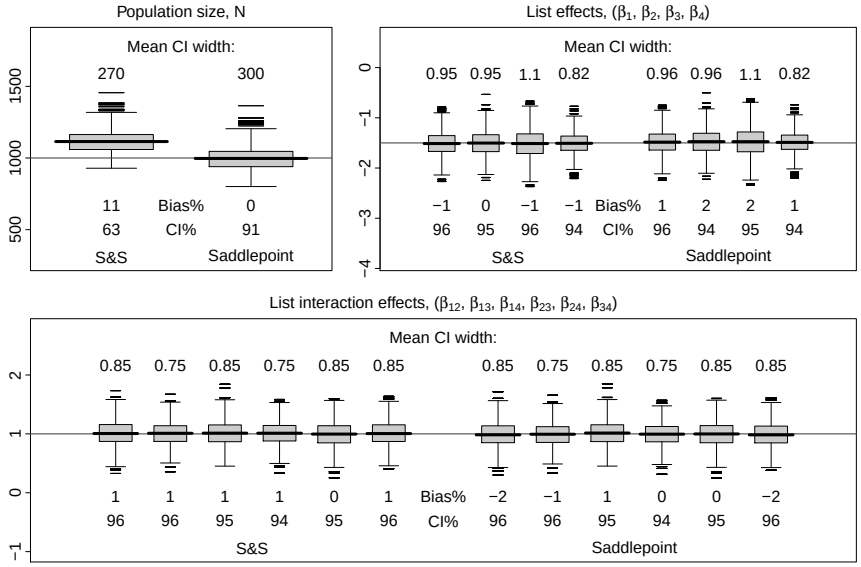


Figure 4. Boxplots of parameter estimates from 1000 simulations using the saddlepoint approximation method (right-hand boxes) and the quasi-likelihood approach of Sutherland and Schwarz (2005), given true parameter values $N = 1000$, $\beta_1 = \cdots = \beta_4 = -1.5$, and $\beta_{12} = \cdots = \beta_{34} = 1$, indicated by thin horizontal lines across the plots. Bold horizontal lines across the boxes indicate the means of the 1000 estimates. Quantities above the boxes show the mean width of 95% confidence intervals. Percentages below the boxes show percentage bias and coverage of nominal 95% confidence intervals.

Table 1
Parameter estimates and standard errors for the Auckland diabetes study obtained by the saddlepoint method and the method of Sutherland and Schwarz (2005).

	β_G	β_P	β_O	β_D	β_{GP}	β_{GO}
Sutherland and Schwarz (2005)						
Estimate	−3.76	−3.74	−1.01	−2.95	1.13	0.45
Standard error	0.14	0.14	0.14	0.11	0.10	0.10
Saddlepoint approximation						
Estimate	−3.76	−3.74	−1.00	−2.94	1.13	0.44
Standard error	0.14	0.14	0.14	0.11	0.10	0.10

Web-based Supplementary Materials for "Fast Likelihood-based Inference for Latent Count Models using the Saddlepoint Approximation"
by W. Zhang, M. V. Bravington, and R. M. Fewster

Web Appendix A

Application to multi-way contingency tables

Consider a K-way table of counts over a K-dimensional discrete random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)$. For each $k \in \{1, \ldots, K\}$, the random variable ξ_k has I_k possible values, denoted by integers $1, \ldots, I_k$ for convenience. Let $i_1 \ldots i_K$ denote a single cell of the table, where i_k takes values from $\{1, \ldots, I_k\}$. Each cell therefore represents one unique combination of the K variables. The total number of cells in the table is $m = \prod_{k=1}^{K} I_k$. The cell $i_1 \ldots i_K$ has a non-negative integer cell entry $Z_{i_1 \ldots i_K}$ that represents the frequency of the random vector $\boldsymbol{\xi}$ being observed as $(i_1, \ldots, i_K)$, i.e. the number of participants with this combination of observations.

To illustrate notation, consider a simple table with $K = 3$ binary variables, for example male/female; smoker/non-smoker; employed/unemployed. There are $I_k = 2$ possible values for each variable $k = 1, 2, 3$. The 3-way table has $m = \prod_{k=1}^{3} I_k = 8$ cells, each corresponding to one combination of the three variables. The cell 111 corresponds to the first level of each of the three variables (male, smoker, employed) and the count Z_{111} is the number of participants with this combination of attributes.

For simplicity, let Z_j denote the cell entry of the jth cell of the table for $j = 1, \ldots, m$. The

contingency table is an m-dimensional vector consisting of all cell entries:

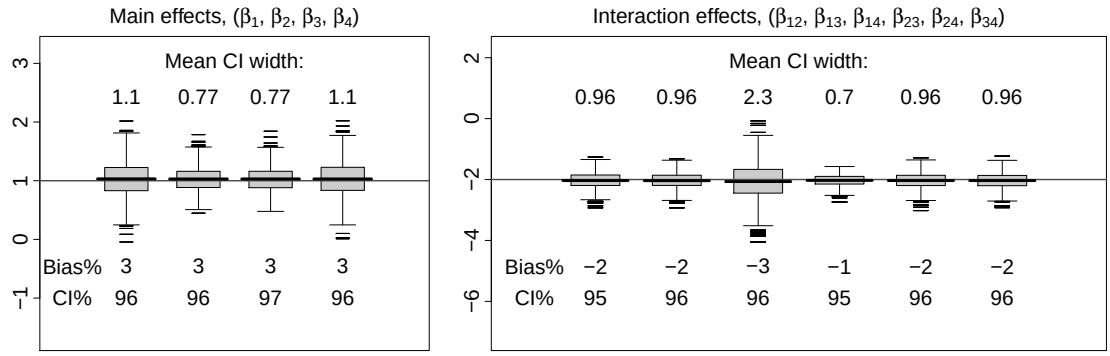
$$\boldsymbol{Z} = (Z_{1\ldots1}, \ldots, Z_{I_1\ldots I_K})^\top = (Z_1, \ldots, Z_m)^\top \ .$$

A marginal table is a vector of summary statistics of the full table, obtained by summing over a subset of the K variables. Consider a subset $D = \{\xi_d \mid d \in \Omega\}$ with $\Omega \subseteq \{1, \ldots, K\}$. The vector $\boldsymbol{Y}_D$ is obtained by summing over all variables not included in D, and it corresponds to the marginal table with dimension $\prod_{d\in\Omega} I_d$. Continuing the example with $K = 3$ for illustration, suppose $D = \{\xi_1, \xi_2\}$ corresponds to gender and smoking status. Then $\boldsymbol{Y}_D$ corresponds to a two-way marginal table, $\boldsymbol{Y}_D = \boldsymbol{Y}_{\{\xi_1,\xi_2\}} = (Z_{11+}, Z_{12+}, Z_{21+}, Z_{22+})^\top$, where Z_{11+} is the total number of male smokers, summed over the two employment categories, and so on. In general, we would have $Z_{i_1 i_2 +} = \sum_{i_3=1}^{I_3} Z_{i_1 i_2 i_3}$ for all combinations of $i_1 \in \{1, \ldots, I_1\}$ and $i_2 \in \{1, \ldots, I_2\}$. If $D' \subseteq D$, the marginal table $\boldsymbol{Y}_{D'}$ can be obtained directly from $\boldsymbol{Y}_D$.

Any marginal table can be expressed as a linear transformation of the full table. If we consider several marginal tables $\boldsymbol{Y}_{D_1}, \ldots, \boldsymbol{Y}_{D_n}$ over subsets $D_1, \ldots, D_n$ of the K variables, then $\boldsymbol{Y} = \left(\boldsymbol{Y}_{D_1}^\top, \ldots, \boldsymbol{Y}_{D_n}^\top\right)^\top$ is a linear transformation of the full table $\boldsymbol{Z}$. Thus $\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{Z}$, where $\boldsymbol{T}$ is a known matrix of zero and one entries. We wish to draw inference on the parameters underlying $\boldsymbol{Z}$, given observed data $\boldsymbol{Y}$.

As with multi-list studies, contingency tables can be modeled using Poisson or multinomial formulations via equations (8) and (9) in the main text. The number of participants N is no longer a parameter to be estimated, as it can be obtained by summing any of the n marginal tables. Inference to date has focused on Bayesian MCMC approaches (e.g. Dobra et al., 2006; Dobra, 2012).

We first investigate a real example in the form of a six-way contingency table over six binary variables $\{A, B, C, D, E, F\}$ as shown in Table 1 of Dobra et al. (2006). This study examined risk factors for coronary thrombosis among 1841 workers in a Czech car factory.



Web Figure 1: Distributions of parameter estimates obtained using the saddlepoint method for a four-way table with true parameter values: $N = 5000, \beta_1 = \cdots = \beta_4 = 1$, and $\beta_{12} = \cdots = \beta_{34} = -2$, indicated by thin horizontal lines across the plots. Bold horizontal lines across the boxes indicate the means of the 1000 estimates. Quantities above the boxes show the mean width of 95% confidence intervals. Percentages below the boxes show percentage bias and coverage of nominal 95% confidence intervals.

The six binary variables are A: smoking status; B: strenuous mental work; C: strenuous physical work; D: high systolic blood pressure; E: high ratio of β and α lipoproteins; F: family anamnesis of coronary heart disease.

Assume the information we know about the table is three five-way marginal tables, $\boldsymbol{y} = \left(\boldsymbol{y}_{\{A,B,C,D,F\}}^\top, \boldsymbol{y}_{\{A,B,C,E,F\}}^\top, \boldsymbol{y}_{\{A,B,D,E,F\}}^\top\right)^\top$. These data do not determine $\boldsymbol{z}$ because the three-way table $\boldsymbol{y}_{\{C,D,E\}}$ is omitted. We fitted a model using the saddlepoint likelihood involving all possible 2-way and 3-way interactions and six main effects. One possible output of interest is to use the fitted model to draw inference on a particular cell entry. We use the entry $\boldsymbol{Z}_{122112}$ considered by Dobra et al. (2006). The saddlepoint multinomial method gives estimate 1.42 with 95% confidence interval $[0, 3]$, which is consistent with the true count of 1 in the entry.

For a simulation study, we consider a four-way table over four binary variables $\xi_1, \ldots, \xi_4$ with an underlying model consisting of four main effects and all six 2-way interactions. The data are the marginal totals $\boldsymbol{y} = \left(\boldsymbol{y}_{\{\xi_1,\xi_2,\xi_3\}}^\top, \boldsymbol{y}_{\{\xi_2,\xi_3,\xi_4\}}^\top\right)^\top$ that omit the two-way marginal table $\boldsymbol{y}_{\{\xi_1,\xi_4\}}$. Results are shown in Web Figure 1. The saddlepoint method produces roughly unbiased estimates with approximately nominal confidence interval coverage for all model

parameters. Fitting one data set cost less than one second. The mean width of confidence intervals for β_{14} is much higher than that for the other parameters, due to the omission of the marginal table $\boldsymbol{y}_{\{\xi_1,\xi_4\}}$. All parameters related to variables ξ_1 and ξ_4 are estimated with lower precision than their analogs that rely only on ξ_2 and ξ_3.

Web Appendix B

Estimation invariance from maximally independent rows of matrix $\boldsymbol{A}$

Let $\boldsymbol{A}$ be an $I \times H$ matrix with row-rank $L < I$. Recall that $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{u}$ where $\boldsymbol{x}$ is $I \times 1$ and $\boldsymbol{u}$ is $H \times 1$. Define $\boldsymbol{A}_1$ to be an $L \times H$ submatrix of $\boldsymbol{A}$ consisting of maximally independent rows of $\boldsymbol{A}$, and let $\boldsymbol{x}_1 = \boldsymbol{A}_1\boldsymbol{u}$ be the corresponding $L \times 1$ subvector of $\boldsymbol{x}$. We aim to show that estimation based on $\boldsymbol{x}_1$ is equivalent to estimation based on $\boldsymbol{x}$.

Let $\boldsymbol{A}_1^c$ be the remaining $I-L$ rows of $\boldsymbol{A}$ formed by removing matrix $\boldsymbol{A}_1$, and let $\boldsymbol{x}_1^c = \boldsymbol{A}_1^c\boldsymbol{u}$ be the corresponding $(I-L) \times 1$ subvector of $\boldsymbol{x}$.

Since $\boldsymbol{A}_1$ comprises maximally independent rows of $\boldsymbol{A}$, the rows of matrix $\boldsymbol{A}_1$ span the row-space of $\boldsymbol{A}$. Because each row of matrix $\boldsymbol{A}_1^c$ also lies in the row-space of $\boldsymbol{A}$, it follows that $\boldsymbol{A}_1^c = \boldsymbol{C}\boldsymbol{A}_1$, where $\boldsymbol{C}$ is a fixed $(I-L) \times L$ matrix of coefficients. Consequently, $\boldsymbol{x}_1^c = \boldsymbol{C}\boldsymbol{x}_1$.

Thus,

$$\Pr(\boldsymbol{x}) = \Pr(\boldsymbol{x}_1, \boldsymbol{x}_1^c) = \Pr(\boldsymbol{x}_1)\Pr(\boldsymbol{x}_1^c \mid \boldsymbol{x}_1) = \Pr(\boldsymbol{x}_1)\Pr(\boldsymbol{C}\boldsymbol{x}_1 \mid \boldsymbol{x}_1) = \Pr(\boldsymbol{x}_1)\,.$$

This proves that estimation based on $\boldsymbol{x}_1$ is equivalent to estimation based on $\boldsymbol{x}$.

Estimation is also invariant to the choice of submatrix $\boldsymbol{A}_1$. Suppose we select a different $L \times H$ submatrix $\boldsymbol{A}_2$ comprising maximally independent rows of $\boldsymbol{A}$. The data vector after reduction is $\boldsymbol{x}_2 = \boldsymbol{A}_2\boldsymbol{u}$. Both $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ have rank L, and the rows of each matrix span the row-space of $\boldsymbol{A}$. Because $\boldsymbol{A}_1$ lies in the row-space spanned by $\boldsymbol{A}_2$, there is an $L \times L$ matrix $\boldsymbol{G}$ such that $\boldsymbol{A}_1 = \boldsymbol{G}\boldsymbol{A}_2$. Since the rank of a matrix product is less than or equal to the rank

of each constituent matrix, we have:

$$L = \text{rank}(\boldsymbol{A}_1) = \text{rank}(\boldsymbol{G}\boldsymbol{A}_2) \leq \text{rank}(\boldsymbol{G})\,.$$

Since G has dimension $L \times L$, it also holds that $\text{rank}(G) \leq L$. It follows that $\text{rank}(\boldsymbol{G}) = L$, so $\boldsymbol{G}$ is invertible.

We therefore have a bijection between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ such that $\boldsymbol{x}_1 = \boldsymbol{G}\boldsymbol{x}_2$ and $\boldsymbol{x}_2 = \boldsymbol{G}^{-1}\boldsymbol{x}_1$. Consequently, $\Pr(\boldsymbol{x}_2) = \Pr(\boldsymbol{x}_1)$, so it is equivalent to use either $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$ for inference.

Web Appendix C

Maximizing saddlepoint-based likelihoods using TMB

Our efficient implementation of saddlepoint-based likelihood maximization relies upon the theoretical connection between the saddlepoint approximation and the Laplace approximation, which leads to similarities in the corresponding optimization problems. The R package TMB is designed to deliver highly efficient optimization for random-effect models, using the Laplace approximation in tandem with automatic differentiation software (Kristensen et al., 2016). We show how we can adapt their formulation to maximize saddlepoint likelihoods, maintaining the advantages of automatic differentiation for speed and accuracy. We first give a brief introduction to the Laplace approximation following Skaug and Fournier (2006) and Kristensen et al. (2016).

Suppose $l\left(\boldsymbol{\mu}, \boldsymbol{\gamma}\right)$ denotes the joint negative log-likelihood function of a statistical model with random effects $\boldsymbol{\mu} \in \mathbb{R}^m$ and fixed parameters $\boldsymbol{\gamma} \in \mathbb{R}^n$. The maximum likelihood estimate of $\boldsymbol{\gamma}$ can be obtained by maximizing the marginal likelihood:

$$\mathscr{L}\left(\boldsymbol{\gamma}\right) = \int_{\mathbb{R}^m} \exp\left\{-l\left(\boldsymbol{\mu}, \boldsymbol{\gamma}\right)\right\} d\boldsymbol{\mu},$$

which is a function of only the fixed model parameters after the random effects are integrated out. However, calculating this integral is prohibitively difficult when the dimension m of the random effects is high.

Define $\hat{\boldsymbol{\mu}}(\boldsymbol{\gamma})$ to be the minimizer of $l(\boldsymbol{\mu}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\mu}$, so that

$$\hat{\boldsymbol{\mu}}(\boldsymbol{\gamma}) = \arg\min_{\boldsymbol{\mu}} l(\boldsymbol{\mu}, \boldsymbol{\gamma}).$$

This minimization is treated as an inner problem in TMB, and can be handled by the classical Newton method. Then the Laplace approximation to $\mathscr{L}(\boldsymbol{\gamma})$ is

$$\mathscr{L}^*(\boldsymbol{\gamma}) = (2\pi)^{\frac{m}{2}} \det\{H(\boldsymbol{\gamma})\}^{-\frac{1}{2}} \exp\{-l(\hat{\boldsymbol{\mu}}, \boldsymbol{\gamma})\},$$

where $H(\boldsymbol{\gamma}) = H(\hat{\boldsymbol{\mu}}, \boldsymbol{\gamma}) = l''_{\boldsymbol{\mu}\boldsymbol{\mu}}(\hat{\boldsymbol{\mu}}, \boldsymbol{\gamma})$ is the Hessian matrix of $l(\boldsymbol{\mu}, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\mu}$ and evaluated at $\hat{\boldsymbol{\mu}}$. Maximizing $\mathscr{L}^*(\boldsymbol{\gamma})$ is equivalent to minimizing the negative logarithm of $\mathscr{L}^*(\boldsymbol{\gamma})$, which is

$$-\log \mathscr{L}^*(\boldsymbol{\gamma}) = -\frac{m}{2}\log(2\pi) + \frac{1}{2}\log\det\{H(\boldsymbol{\gamma})\} + l(\hat{\boldsymbol{\mu}}, \boldsymbol{\gamma}). \tag{1}$$

To fit a random effects model using TMB, the user codes the joint negative log-likelihood function $l(\boldsymbol{\mu}, \boldsymbol{\gamma})$ in a `C++` function template. When $\boldsymbol{\mu}$ is declared as a vector of random effects, the `MakeADFun` function from TMB is formulated to return the objective (1) and its gradient function with respect to $\boldsymbol{\gamma}$, where the gradient incorporates the computation of the arg-min value $\hat{\boldsymbol{\mu}}(\boldsymbol{\gamma})$ (see Kristensen et al., 2016, for more details). Thus, TMB has specific functionality for minimizing expressions of the form (1), using automatic differentiation to generate a gradient function with respect to $\boldsymbol{\gamma}$ that encapsulates the inner optimization needed to find $\hat{\boldsymbol{\mu}}(\boldsymbol{\gamma})$. Automatic differentiation generates gradients that are as accurate as symbolic differentiation, without requiring any analytical derivatives to be supplied (Fournier et al., 2012). Availability of a machine-precision gradient function ensures very high accuracy

and efficiency in the outer optimization with respect to $\boldsymbol{\gamma}$, using standard gradient-based R optimizers such as `nlminb` or `nlm`.

To use TMB to maximize saddlepoint-based likelihoods, we consider

$$\tilde{f}_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{L/2}\left|K''_{\boldsymbol{X}}(\hat{\boldsymbol{s}})\right|^{1/2}} \exp\left\{K_{\boldsymbol{X}}(\hat{\boldsymbol{s}}) - \hat{\boldsymbol{s}}^{\top}\boldsymbol{x}\right\},$$

which is the general expression for a saddlepoint likelihood, and is not confined to latent multinomial models. Some algebra gives:

$$-\log \tilde{f}_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{L}{2}\log(2\pi) + \frac{1}{2}\log\det\left\{K''_{\boldsymbol{X}}(\hat{\boldsymbol{s}})\right\} - h(\hat{\boldsymbol{s}}), \tag{2}$$

where $h(\boldsymbol{s}) = K_{\boldsymbol{X}}(\boldsymbol{s}) - \boldsymbol{s}^{\top}\boldsymbol{x}$, and $\hat{\boldsymbol{s}}$ is the solution to $h'(\boldsymbol{s}) = K'_{\boldsymbol{X}}(\boldsymbol{s}) - \boldsymbol{x} = 0$. The Hessian matrix $K''_{\boldsymbol{X}}(\boldsymbol{s})$ with respect to $\boldsymbol{s}$ is positive definite because $K_{\boldsymbol{X}}$ is strictly convex, so finding $\hat{\boldsymbol{s}}$ is equivalent to minimizing $h(\boldsymbol{s})$ with respect to $\boldsymbol{s}$. In the specific case of latent multinomial models, there is an additional factor in the likelihood $\widetilde{\mathcal{L}}(\boldsymbol{\theta} \mid \boldsymbol{y}) = \Pr(\boldsymbol{Z}_{1:R} = \boldsymbol{v})\,\tilde{f}_{\boldsymbol{X}}(\boldsymbol{x})$, which creates the following saddlepoint objective function for minimization:

$$\begin{aligned} -\log \widetilde{\mathcal{L}}(\boldsymbol{\theta} \mid \boldsymbol{y}) &= -\log\Pr(\boldsymbol{Z}_{1:R} = \boldsymbol{v}) + \frac{L}{2}\log(2\pi) + \frac{1}{2}\log\det\left\{K''_{\boldsymbol{X}}(\hat{\boldsymbol{s}})\right\} - h(\hat{\boldsymbol{s}}) \\ &= \frac{L}{2}\log(2\pi) + \frac{1}{2}\log\det\left\{K''_{\boldsymbol{X}}(\hat{\boldsymbol{s}}, \boldsymbol{\theta})\right\} - g(\hat{\boldsymbol{s}}, \boldsymbol{\theta}), \end{aligned} \tag{3}$$

where $g(\hat{\boldsymbol{s}}, \boldsymbol{\theta}) = \log\Pr(\boldsymbol{Z}_{1:R} = \boldsymbol{v}) + h(\hat{\boldsymbol{s}}, \boldsymbol{\theta})$. Here, we write $K_{\boldsymbol{X}}(\boldsymbol{s}) = K_{\boldsymbol{X}}(\boldsymbol{s}, \boldsymbol{\theta})$ and $h(\boldsymbol{s}) = h(\boldsymbol{s}, \boldsymbol{\theta})$ to emphasize dependence on the parameters $\boldsymbol{\theta}$. Clearly $g''_{\boldsymbol{ss}}(\hat{\boldsymbol{s}}, \boldsymbol{\theta}) = h''_{\boldsymbol{ss}}(\hat{\boldsymbol{s}}, \boldsymbol{\theta}) = K''_{\boldsymbol{X}}(\hat{\boldsymbol{s}})$ and $\hat{\boldsymbol{s}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{s}} h(\boldsymbol{s}, \boldsymbol{\theta}) = \arg\min_{\boldsymbol{s}} g(\boldsymbol{s}, \boldsymbol{\theta})$.

Equation (3) is an exact analog of equation (1), with $\hat{\boldsymbol{s}}$ replacing $\hat{\boldsymbol{\mu}}$, $\boldsymbol{\theta}$ replacing $\boldsymbol{\gamma}$, and $g(\hat{\boldsymbol{s}}, \boldsymbol{\theta})$ replacing $l(\hat{\boldsymbol{\mu}}, \boldsymbol{\gamma})$, except for two sign changes in the first and third terms of equation (3). Motivated by this similarity, we implement maximum saddlepoint likelihood by copying the source code of the function `MakeADFun` in TMB, and changing the requisite two

signs. By declaring vector $\boldsymbol{s}$ to be a vector of random effects, and defining our objective function to be $g(\boldsymbol{s}, \boldsymbol{\theta})$, we can deploy the modified version of TMB to generate an efficient optimization of (3) that includes a gradient function calculated using automatic differentiation. The inner optimization problem for $\hat{\boldsymbol{s}}$ is taken care of by TMB, whereas the outer problem for $\boldsymbol{\theta}$ is dealt with efficiently by customary gradient-based optimizers.

Extending this method to likelihoods outside of the latent multinomial framework is straightforward. The formulation above minimizes the general saddlepoint objective (2) simply by setting $g(\boldsymbol{s}, \boldsymbol{\theta}) = h(\boldsymbol{s}, \boldsymbol{\theta})$.

References

Dobra, A. (2012). Dynamic Markov bases. *Journal of Computational and Graphical Statistics* **21,** 496–517.

Dobra, A., Tebaldi, C., and West, M. (2006). Data augmentation in multi-way contingency tables with fixed marginal totals. *Journal of Statistical Planning and Inference* **136,** 355–372.

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., and Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* **27,** 233–249.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. J., and Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software* **70,** 1–21.

Skaug, H. J. and Fournier, D. A. (2006). Automatic approximation of the marginal likelihood in non-Gaussian hierarchical models. *Computational Statistics & Data Analysis* **51,** 699–709.