Biometrika (2014), **101**, 1, *pp.* 189–204
Printed in Great Britain

doi: 10.1093/biomet/ast050
Advance Access publication 8 December 2013

Retrospective-prospective symmetry in the likelihood and Bayesian analysis of case-control studies

By SIMON P. J. BYRNE

Department of Statistical Science, University College, London WC1E 6BT, U.K.
simon.byrne@ucl.ac.uk

AND A. PHILIP DAWID

Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge CB3 0WB, U.K.
apd@statslab.cam.ac.uk

Summary

Prentice & Pyke (1979) established that the maximum likelihood estimate of an odds ratio in a case-control study is the same as would be found by fitting a logistic regression; in other words, for this specific target the incorrect prospective model is inferentially equivalent to the correct retrospective model. Similar results have been obtained for other models, and conditions have also been identified under which the corresponding Bayesian property holds, namely that the posterior distribution of the odds ratio is the same whether it is computed using the prospective or the retrospective likelihood. In this article we demonstrate how these results follow directly from certain parameter independence properties of the models and priors, and identify prior laws that support such reverse analysis, for both standard and stratified designs.

Some key words: Case-control study; Conditional independence; Hyper Markov law; Logistic regression; Retrospective likelihood.

1. Introduction

In order to estimate the effects of risk factors on a binary outcome, such as a disease, there are two basic experimental approaches: a prospective or cohort study, in which subjects are selected from the population, possibly based on their risk factors, and observed to determine if the disease arises; and a retrospective or case-control study, in which random samples are taken from both the subpopulation with the disease, the cases, and the subpopulation without the disease, the controls, and the relative frequencies of the risk factors in the two samples are recorded. Case-control studies are often desirable or unavoidable, particularly where the disease is relatively rare or the time to diagnosis is long, since the costs of obtaining a sufficiently large sample for a prospective study are then likely to be prohibitive.

Let Y be the outcome variable, taking values 0 or 1 corresponding to the absence or presence of disease, respectively. Let X be the vector of covariates, or risk factors, taking values in $\mathcal{X} \subseteq \mathbb{R}^k$. In a prospective study we are sampling from the conditional distribution of Y given X. Under a proportional odds assumption, the model is that of a logistic regression:

$$p(y \mid x, \alpha, \beta) = \frac{\exp\{y(\alpha + \beta^{\mathrm{T}} x)\}}{1 + \exp(\alpha + \beta^{\mathrm{T}} x)} \quad (\alpha \in \mathbb{R},\ \beta \in \mathbb{R}^k).$$

A case-control study, however, will result in observations generated from the conditional distribution of X given Y. In this case, specifying and analysing the probabilistic model become much more difficult, particularly if $\mathcal{X}$ is large or infinite. But Prentice & Pyke (1979) showed that the maximum likelihood estimator of the log odds ratio parameter β, as well as its asymptotic covariance matrix, can be computed from a logistic regression; in other words, we can use the incorrect but simpler prospective model to analyse data gathered retrospectively. This result has been widely applied in epidemiology and other areas. Other models have since been identified that satisfy this property, notably the multinomial logistic model (Baker, 1994), the stereotype model (Greenland, 1994), and the multiplicative intercept model (Weinberg & Wacholder, 1993).

There exist analogous results for Bayesian analysis, showing that, for an appropriately chosen prior distribution, the posterior distribution of β can be computed using the incorrect prospective likelihood instead of the true retrospective likelihood. Zelen & Parker (1986), Nurminen & Mutanen (1987), Marshall (1988) and Ashby et al. (1993) developed this analysis for the case of a single binary covariate; this involves computing the posterior distribution of the log odds ratio of a 2×2 contingency table under a Dirichlet prior. For the case of categorical covariates, where $\mathcal{X}$ is finite, Seaman & Richardson (2004) identified a class of improper priors that satisfy the desired properties; this class was extended to include proper priors by Staicu (2010). Extensions to stratified and general multinomial designs have been studied by Ghosh et al. (2006, 2012).

With the advent of computational tools such as Markov chain Monte Carlo simulation, direct analysis of the retrospective likelihood need no longer present an obstacle. Müller & Roeder (1997), Seaman & Richardson (2001) and Gustafson et al. (2002) have pursued this approach, which is reviewed in Mukherjee et al. (2005). Nevertheless, for complicated models the retrospective likelihood can still be computationally prohibitive, so that use of the prospective approach remains widespread.

In this paper we observe that these likelihood and Bayesian results are all consequences of certain properties of independence between parameters. In § 3 we show that the results for maximum likelihood estimation hold whenever we have a strong meta Markov model, embodying properties of variation independence in the parameter space. In § 4 we show that the corresponding Bayesian result holds when, in addition, we use an overall prior distribution that is a strong hyper Markov law, exhibiting analogous probabilistic independence between parameters. In § 5, we derive parametric classes of strong hyper Markov laws that can be used for such an analysis, and show that these encompass the proper prior laws mentioned above. These results are further extended to stratified designs in § 6.

2. Notation and definitions

Throughout the paper, (X, Y) will denote a single joint observation from the specified model, and $(X^{(n)}, Y^{(n)})$ a sequence of n such observations; p will denote density with respect to an appropriate measure, with variables indicating the context.

We recall the notation and definitions of Dawid & Lauritzen (1993). If θ denotes a joint probability distribution for (X, Y), then θ_X and θ_Y will denote the corresponding marginal distributions of X and Y, respectively. We use $\theta_{Y|X=x}$ to denote the conditional distribution of Y given $X = x$, and $\theta_{Y|X} = (\theta_{Y|X=x} : x \in \mathcal{X})$ the family of all such conditional distributions, labelled by x; we define $\theta_{X|Y=y}$ and $\theta_{X|Y}$ similarly.

A model is a set Θ of joint probability distributions θ. A parameter in this model is a function defined on Θ. We use the relation $\phi \simeq \psi$ to signify the existence of a bijective function between the parameters ϕ and ψ. For example, we have $\theta \simeq (\theta_X, \theta_{Y|X}) \simeq (\theta_Y, \theta_{X|Y})$.

For two parameters ϕ and τ, we define the conditional range of ϕ given $\tau = t$ to be $\{\phi(\theta) : \theta \in \Theta,\ \tau(\theta) = t\}$. We say that ϕ is variation independent of τ, and write $\phi \ddagger \tau$, when this conditional range is constant for all possible values t of τ or, equivalently, when (ϕ, τ) takes values in a product space. In a similar manner we can define the conditional variation independence $\phi \ddagger \tau \mid \psi$ (Dawid & Lauritzen, 1993).

A model is said to be strong meta Markov if

$$\theta_X \ddagger \theta_{Y|X}, \quad \theta_Y \ddagger \theta_{X|Y}. \tag{1}$$

In a Bayesian setting, we use the term law to mean a probability distribution over the model Θ for the parameter variable $\tilde{\theta}$. We say that a law $\mathcal{L}$ is strong hyper Markov if the variation independence of (1) is replaced with probabilistic independence, denoted by $\perp\!\!\!\perp$, under $\mathcal{L}$:

$$\tilde{\theta}_X \perp\!\!\!\perp \tilde{\theta}_{Y|X}, \quad \tilde{\theta}_Y \perp\!\!\!\perp \tilde{\theta}_{X|Y} \quad [\mathcal{L}].$$

A necessary, but not sufficient, condition for a law to be strong hyper Markov is that its support be a strong meta Markov model.

3. Maximum likelihood estimation in strong meta Markov models

The saturated model, consisting of all probability distributions on the product space $\mathcal{X} \times \mathcal{Y}$, is trivially strong meta Markov. We now investigate some other meta Markov models.

Example 1. Let ν_X and ν_Y be measures over $\mathcal{X}$ and $\mathcal{Y}$, respectively. The family of all probability distributions which have positive densities with respect to $\nu_X \times \nu_Y$ is strong meta Markov.

In particular, if $\mathcal{X}$ and $\mathcal{Y}$ are finite, with ν_X and ν_Y being counting measures, this is the family of two-way $|\mathcal{X}| \times |\mathcal{Y}|$ contingency tables without structural zeros.

Example 2. Let Θ be the family of bivariate normal distributions for (X, Y):

$$\theta = N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \sigma_{XX} & \sigma_{XY} \\ \sigma_{XY} & \sigma_{YY} \end{bmatrix} \right).$$

Then $\theta_X = N(\mu_X, \sigma_{XX})$ and $\theta_{Y|X=x} = N(\mu_{Y|X} + \beta_{Y|X}x, \sigma_{Y|X})$, where

$$\mu_{Y|X} = \mu_Y - \frac{\sigma_{XY}\mu_Y}{\sigma_{XX}}, \quad \beta_{Y|X} = \frac{\sigma_{XY}}{\sigma_{XX}}, \quad \sigma_{Y|X} = \sigma_{YY} - \frac{\sigma_{XY}^2}{\sigma_{XX}}.$$

It is straightforward to establish that $(\mu_X, \sigma_{XX}) \ddagger (\mu_{Y|X}, \beta_{Y|X}, \sigma_{Y|X})$ and hence $\tilde{\theta}_X \ddagger \tilde{\theta}_{Y|X}$, with parallel results when X and Y are interchanged. Therefore this family is a strong meta Markov model. This property extends to higher dimensions.

Definition 1. *Suppose that the model Θ consists of a set of joint distributions θ for (X, Y) having positive joint density $p(x, y \mid \theta)$. The odds ratio parameter $\lambda = \lambda(\theta)$ is defined to be the labelled collection*

$$\left(\frac{p(x, y \mid \theta)\, p(x', y' \mid \theta)}{p(x, y' \mid \theta)\, p(x', y \mid \theta)} : x, x' \in \mathcal{X};\ y, y' \in \mathcal{Y} \right). \tag{2}$$

As an example, in the bivariate normal model, elements of (2) are of the form $\exp\{-\Lambda_{XY}(x - x')(y - y')\}$, where $\Lambda_{XY} = -\sigma_{XY}/(\sigma_{XX}\sigma_{YY} - \sigma_{XY}^2)$ is the off-diagonal term of the precision matrix. Therefore $\lambda \simeq \Lambda_{XY}$.

The parameter λ has been well studied in the context of contingency tables. Altham (1970) demonstrated that it has certain desirable properties as a measure of association between X and Y. We note that λ also characterizes such dependence for more general models.

Lemma 1. *For a given joint distribution θ, $\lambda(\theta) \equiv 1$ if and only if X and Y are independent under θ.*

Proof. By definition, $\lambda \equiv 1$ if and only if

$$p(x, y \mid \theta)\, p(x', y' \mid \theta) = p(x, y' \mid \theta)\, p(x', y \mid \theta) \tag{3}$$

for all x, y, x', y'. If (3) holds, then upon integrating over x' and y' we obtain $p(x, y \mid \theta) = p(x \mid \theta)\, p(y \mid \theta)$. Conversely, if $p(x, y \mid \theta)$ factorizes in this manner, (3) must hold. □

Our particular interest in λ is due to its being a common parameter of both the prospective and retrospective models.

Lemma 2. *The odds ratio λ can be expressed as a function of $\theta_{Y\mid X}$ and also of $\theta_{X\mid Y}$.*

Proof. Elements of (2) can be written as

$$\frac{p(y \mid x, \theta_{Y\mid X})\, p(y' \mid x', \theta_{Y\mid X})}{p(y' \mid x, \theta_{Y\mid X})\, p(y \mid x', \theta_{Y\mid X})} = \frac{p(x \mid y, \theta_{X\mid Y})\, p(x' \mid y', \theta_{X\mid Y})}{p(x \mid y', \theta_{X\mid Y})\, p(x' \mid y, \theta_{X\mid Y})}.$$

□

As we shall see below, it is this shared parameter property that makes it possible to use retrospective data to make inferences about the prospective model.

By constraining λ, we can construct new strong meta Markov models.

Lemma 3. *Let Θ be a strong meta Markov model for (X, Y), and for a given function f define $\Theta' = \{\theta \in \Theta : f(\lambda) = 0\}$. Then Θ' is strong meta Markov.*

Proof. Since $\theta_{Y\mid X} \ddagger \theta_X$ and $f(\lambda)$ is a function of $\theta_{Y\mid X}$, it follows from the separoid properties of variation independence (Dawid, 2001a,b) that $\theta_{Y\mid X} \ddagger \theta_X \mid f(\lambda)$. Similarly, $\theta_{X\mid Y} \ddagger \theta_Y \mid f(\lambda)$. □

Example 3. Let $\mathcal{Y} = \{0, 1\}$, and let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ whose affine span is $\mathbb{R}^d$. Let the model Θ comprise all distributions with positive densities on $\mathcal{X} \times \mathcal{Y}$. By the affine condition, there exist $x_1, \ldots, x_{d+1} \in \mathcal{X}$ such that $(1, x_1), \ldots, (1, x_{d+1})$ are linearly independent. We can then write $\theta_{Y\mid X} \simeq (\alpha, \beta, \eta)$, where

$$p(y \mid x, \alpha, \beta, \eta) = \frac{\exp\{y(\alpha + \beta^{\mathrm{T}} x + \eta_x)\}}{1 + \exp(\alpha + \beta^{\mathrm{T}} x + \eta_x)}$$

with $\eta_x = 0$ for $x = x_1, \ldots, x_{d+1}$. The odds ratios are then

$$\frac{p(1 \mid x, \alpha, \beta, \eta)\, p(0 \mid x', \alpha, \beta, \eta)}{p(1 \mid x', \alpha, \beta, \eta)\, p(0 \mid x, \alpha, \beta, \eta)} = \exp\{\beta^{\mathrm{T}}(x - x') + \eta_x - \eta_{x'}\},$$

and hence $\lambda \simeq (\beta, \eta)$. The logistic model is then obtained upon constraining η to be 0. As η is a function of λ, by Lemma 3 it is strong meta Markov. Moreover, $\lambda \simeq \beta$ in this model.

Example 4. We can generalize to let $\mathcal{Y}$ be a finite set. Using essentially the same argument yields the multinomial logistic model:

$$p(y \mid x, \alpha, \beta) = \begin{cases} \dfrac{\exp(\alpha_y + \beta_y^{\mathrm{T}} x)}{1 + \sum_{y' \neq y^*} \exp(\alpha_{y'} + \beta_{y'}^{\mathrm{T}} x)}, & y \neq y^*, \\ \dfrac{1}{1 + \sum_{y' \neq y^*} \exp(\alpha_{y'} + \beta_{y'}^{\mathrm{T}} x)}, & y = y^*, \end{cases}$$

for some reference element $y^* \in \mathcal{Y}$. We then have $\lambda \simeq \beta = (\beta_y : y \neq y^*)$.

The cumulative logit model (McCullagh, 1980), which is widely used for ordinal data, is not strong meta Markov. However, there is an alternative model that can be used in this setting.

Example 5. The stereotype model (Anderson, 1984) is obtained by constraining the multinomial logistic model so that $\beta_y = \beta\gamma_y$, where $\beta \in \mathbb{R}^d$ and $\gamma_y \in \mathbb{R}$. Then $\lambda \simeq (\beta, \gamma)$. This model can be made more general by allowing β to take values in $\mathbb{R}^{d \times k}$ and γ_y to take values in $\mathbb{R}^k$, where $k < |\mathcal{Y}| - 1$. Several authors have proposed this model for ordinal data; in particular, Greenland (1994) noted its validity for analysing retrospective data, as we demonstrate below.

Example 6. The multiplicative intercept model (Hsieh et al., 1985; Weinberg & Wacholder, 1993) is a general strong meta Markov model for binary response data. Its density has the form

$$p(y \mid x, \alpha, \beta) = \frac{[\exp\{\alpha + f(x, \beta)\}]^y}{1 + \exp\{\alpha + f(x, \beta)\}}.$$

This model can be obtained by constraining the odds ratios (2) to be of the form $f(x, \beta) - f(x', \beta)$. It has $\lambda \simeq \beta$.

For the logistic model, Prentice & Pyke (1979) showed that the maximum likelihood odds ratio estimators obtained from a case-control study have the same values and asymptotic distribution as those arising from a prospective study. The following result states that this property holds for any strong meta Markov model.

THEOREM 1. *Let Θ be a strong meta Markov model for (X, Y). Then the profile likelihood function for any function of λ is the same, up to proportionality, under the joint model Θ, the retrospective model $\Theta_{X\mid Y}$ and the prospective model $\Theta_{Y\mid X}$.*

Proof. The argument is similar to that of Lemma 4.10 in Dawid & Lauritzen (1993). The joint density under the model θ can be written as $p(x, y \mid \theta) = p(x \mid \theta_X)\, p(y \mid x, \theta_{Y\mid X})$. Therefore the profile likelihood $L_{\mathrm{p}}^{\text{joint}}(\lambda)$ for the joint model is

$$L_{\mathrm{p}}^{\text{joint}}(\lambda) = \max_{\theta : \lambda(\theta) = \lambda} p(x \mid \theta_X)\, p(y \mid x, \theta_{Y\mid X}). \tag{4}$$

Since we have the conditional variation independence $\theta_X \ddagger \theta_{Y\mid X} \mid \lambda$, the maximization in (4) can be performed separately for each factor; hence

$$L_{\mathrm{p}}^{\text{joint}}(\lambda) = \max_{\theta_X : \lambda(\theta_X) = \lambda} p(x \mid \theta_X) \times \max_{\theta_{Y\mid X} : \lambda(\theta_{Y\mid X}) = \lambda} p(y \mid x, \theta_{Y\mid X}).$$

Moreover, since $\theta_X \perp\!\!\!\perp \theta_{Y|X}$ and λ is a function of $\theta_{Y|X}$, we have $\theta_X \perp\!\!\!\perp \lambda$, so that the first term is constant for all λ, giving

$$L_{\text{p}}^{\text{joint}}(\lambda) \propto \max_{\theta_{Y|X}:\lambda(\theta_{Y|X})=\lambda} p(y \mid x, \theta_{Y|X}) = L_{\text{p}}^{\text{pro}}(\lambda),$$

where $L_{\text{p}}^{\text{pro}}$ denotes the profile likelihood of the prospective model. An identical argument shows that $L_{\text{p}}^{\text{joint}}(\lambda) \propto L_{\text{p}}^{\text{ret}}(\lambda)$. This argument can be extended to any function of λ. □

From this we obtain the following result, generalizing that of Prentice & Pyke (1979).

Corollary 1. *Suppose that Θ is a strong meta Markov model parameterized by a finite-dimensional parameter. Then, for data observed under retrospective sampling, the maximum likelihood estimator of any function of the parameter λ, as well as its asymptotic covariance matrix, can be computed as if the data were observed prospectively.*

Proof. The maximum likelihood estimator is a function of the profile likelihood, as is its asymptotic covariance matrix when θ is finite-dimensional (Patefield, 1985). □

We emphasize that it is necessary for this result that the parameter of interest be a function of λ; it is not sufficient that the parameter be variation independent of the marginals. In the bivariate normal example, the correlation coefficient $\rho = \sigma_{XY}/(\sigma_{XX}\sigma_{YY})^{1/2}$ is variation independent of both θ_X and θ_Y, but it cannot be expressed as a function of either $\theta_{Y|X}$ or $\theta_{X|Y}$ and cannot be estimated from a regression.

The above argument can also be applied to the value, but not the covariance matrix, of a penalized maximum likelihood estimator of λ, when the penalty term is a function of λ only, for example in the case of estimating β in a logistic regression by maximizing $\log p(y \mid x, \alpha, \beta) - \phi(\beta)$ over α and β. Examples of such estimators include ridge regression, where $\phi(\beta) \propto \|\beta\|_2$, and lasso, where $\phi(\beta) \propto \|\beta\|_1$. Such methods have proven successful in genome-wide association studies, which involve case-control data with extremely high-dimensional covariates (Park & Hastie, 2008; Wu et al., 2009).

4. Bayesian analysis of retrospective studies

We now extend the results of the previous section to Bayesian analysis. Let $\mathcal{L}$ be a prior law for the parameter variable $\tilde{\theta} \in \Theta$, and let $\mathcal{L}_{\text{pro}}$ and $\mathcal{L}_{\text{ret}}$ denote the induced marginal priors for $\tilde{\theta}_{Y|X}$ and $\tilde{\theta}_{X|Y}$, respectively. For observations $(X^{(n)}, Y^{(n)}) = (x^{(n)}, y^{(n)})$, denote by $\mathcal{L}^{\text{joint}}$ the posterior law for $\tilde{\theta}$, based on prior $\mathcal{L}$ and the joint likelihood $p(x^{(n)}, y^{(n)} \mid \theta)$; denote by $\mathcal{L}^{\text{pro}}$ the posterior law for $\tilde{\theta}_{Y|X}$, based on the prior law $\mathcal{L}_{\text{pro}}$ and the prospective likelihood $p(y^{(n)} \mid x^{(n)}, \theta_{Y|X})$; and denote by $\mathcal{L}^{\text{ret}}$ the posterior law for $\tilde{\theta}_{X|Y}$, based on the prior law $\mathcal{L}_{\text{ret}}$ and the retrospective likelihood $p(x^{(n)} \mid y^{(n)}, \theta_{X|Y})$.

We now present the key result of this section.

Theorem 2. *Let $\mathcal{L}$ be a strong hyper Markov prior law over the joint model Θ for (X, Y). Then the posterior marginal law of $\tilde{\lambda} = \lambda(\tilde{\theta})$ is the same whether it is computed from $\mathcal{L}^{\text{joint}}$, from $\mathcal{L}^{\text{pro}}$, or from $\mathcal{L}^{\text{ret}}$.*

Proof. The posterior law for $\tilde{\lambda}$ under the joint analysis is determined by its Radon–Nikodym derivative with respect to the prior law:

$$\frac{d\mathcal{L}^{\text{joint}}}{d\mathcal{L}}(\lambda) \propto \int \prod_{i=1}^{n} p(y_i \mid x_i, \theta_{Y|X}) p(x_i \mid \theta_X) \, d\mathcal{L}(\theta \mid \lambda). \tag{5}$$

By the strong hyper Markov property, $\tilde{\theta}_{Y|X} \perp\!\!\!\perp \tilde{\theta}_X \mid \tilde{\lambda}$, so the right-hand side of (5) factorizes as

$$\int \prod_{i=1}^{n} p(y_i \mid x_i, \theta_{Y|X}) \, d\mathcal{L}(\theta_{Y|X} \mid \lambda) \int \prod_{i=1}^{n} p(x_i \mid \theta_X) \, d\mathcal{L}(\theta_X \mid \lambda).$$

Also, $\tilde{\theta}_X \perp\!\!\!\perp \tilde{\lambda}$, so only the first of these terms is a function of λ. Therefore

$$\frac{d\mathcal{L}^{\text{joint}}}{d\mathcal{L}}(\lambda) \propto \int \prod_{i=1}^{n} p(y_i \mid x_i, \theta_{Y|X}) \, d\mathcal{L}(\theta_{Y|X} \mid \lambda) \propto \frac{d\mathcal{L}^{\text{pro}}}{d\mathcal{L}_{\text{pro}}}(\lambda).$$

Since the distribution of $\tilde{\lambda}$ is the same under the priors $\mathcal{L}$ and $\mathcal{L}_{\text{pro}}$, the posteriors for $\tilde{\lambda}$ under $\mathcal{L}^{\text{joint}}$ and $\mathcal{L}^{\text{pro}}$ are proportional and hence identical. A parallel argument shows the identity of the joint and the retrospective analyses. □

Several authors have obtained similar results. Müller & Roeder (1997) almost identified these conditions for the logistic regression model, but then incorrectly claimed that the 'argument about the retrospective likelihood only carries over to posterior inference on β if α and β are independent and θ_X is not otherwise constrained'. This misconception appears to be due to the fact that, although there is a one-to-one mapping between α and θ_Y, this mapping is itself dependent on β. Unfortunately, this means that Müller and Roeder's proposed Dirichlet process mixture law does not satisfy the required properties.

For the case of the logistic regression model where the covariate space $\mathcal{X}$ is finite, conditions equivalent to the strong hyper Markov property were shown to be sufficient in a 2007 University of Bristol technical report by A.-M. Staicu.

The converse result to Theorem 2 does not strictly hold. For instance, if $\tilde{\lambda}$ is almost surely constant under the prior law, then so must it be under any of the posterior laws, irrespective of whether or not the strong hyper Markov property holds. However, we conjecture that, with the addition of suitable technical conditions to exclude such special cases, the identity of the joint, prospective and retrospective analyses for $\tilde{\lambda}$ will hold only when the joint prior law for $\tilde{\theta}$ is strong hyper Markov.

It follows immediately from Theorem 2 that, with the stated conditions and definitions, the posterior for $\tilde{\lambda}$ that we would obtain by combining the true retrospective likelihood with prior law $\mathcal{L}_{\text{ret}}$ for its parameter $\tilde{\theta}_{X|Y}$ could also be obtained by combining the incorrect prospective likelihood with prior law $\mathcal{L}_{\text{pro}}$ for its parameter $\tilde{\theta}_{Y|X}$. Here we wish to emphasize a constraint that previous authors have not always made clear: in order to invoke this result, we must be using a prior law $\mathcal{L}_{\text{ret}}$ for the retrospective parameter $\tilde{\theta}_{X|Y}$ that can arise as the marginal of some strong hyper Markov law $\mathcal{L}$ for $\tilde{\theta}$. Only then is one justified in using instead the prospective likelihood in conjunction with a suitable prior law for its parameter $\tilde{\theta}_{Y|X}$, which law we can take to be that derived from $\mathcal{L}$.

The problem of model comparison for case-control studies has received relatively little attention in the literature, particularly for Bayesian analyses. However, we can approach it through a result similar to Theorem 2.

Theorem 3. *Let $\mathcal{L}_1(\tilde{\theta})$ and $\mathcal{L}_2(\tilde{\theta})$ be strong hyper Markov laws whose marginal laws for $\tilde{\theta}_X$ are identical, as are those for $\tilde{\theta}_Y$. Then the Bayes factors between $\mathcal{L}_1$ and $\mathcal{L}_2$ computed under the prospective, retrospective and joint likelihoods are all equal.*

Proof. Define a joint law $\mathcal{L}^*$ for $(\tilde{M}, \tilde{\theta})$ such that $\tilde{M}$ takes values 1 and 2 each with probability 1/2 and, given $\tilde{M} = j$, the conditional law of $\tilde{\theta}$ is $\mathcal{L}_j$. The strong hyper Markov condition implies that $\tilde{\theta}_X \perp\!\!\!\perp \tilde{\theta}_{Y|X} \mid \tilde{M}$ $[\mathcal{L}^*]$, while the condition of the equality of marginals can be expressed as $\tilde{\theta}_X \perp\!\!\!\perp \tilde{M}$ $[\mathcal{L}^*]$. These properties are together equivalent to $\tilde{\theta}_X \perp\!\!\!\perp (\tilde{\theta}_{Y|X}, \tilde{M})$ $[\mathcal{L}^*]$, and similarly $\tilde{\theta}_Y \perp\!\!\!\perp (\tilde{\theta}_{X|Y}, \tilde{M})$ $[\mathcal{L}^*]$. An argument similar to that of Theorem 2 now shows that the posterior distributions for $\tilde{M}$, and hence the Bayes factors, must be the same, whether they are computed using the joint, prospective or retrospective analyses. □

5. Strong hyper Markov laws

We now investigate known families of strong hyper Markov laws, as well as methods for deriving new families. As noted in § 2, strong hyper Markov laws exist only for strong meta Markov models, so we shall focus on the same models discussed in § 3.

Dawid & Lauritzen (1993) identified two strong hyper Markov laws.

Example 7. For discrete X and Y, the saturated model comprises all multinomial distributions, which can be parameterized by their joint probabilities $\theta = (\theta_{x,y} : x \in \mathcal{X}, y \in \mathcal{Y})$. The standard conjugate prior is a Dirichlet law, $\mathcal{L}(\tilde{\theta}) = \mathcal{D}(a_{xy} : x \in \mathcal{X}, y \in \mathcal{Y})$, with hyperparameters $a_{xy} > 0$, having density proportional to

$$\prod_{x\in\mathcal{X}, y\in\mathcal{Y}} \theta_{xy}^{a_{xy}-1}.$$

The posterior is of the same form, with updated hyperparameters $a^*_{xy} = a_{xy} + n_{xy}$, where n_{xy} is the number of cases having $X = x$ and $Y = y$.

By the aggregation properties of Dirichlet laws (see, e.g., Dawid & Lauritzen, 1993, Lemma 7.2),

$$\tilde{\theta}_X \sim \mathcal{D}(a_{x+} : x \in \mathcal{X}), \quad \tilde{\theta}_{Y|X=x^*} \sim \mathcal{D}(a_{x^*y} : y \in \mathcal{Y}) \quad (x^* \in \mathcal{X}),$$

all independently, where $a_{x+} = \sum_y a_{xy}$; similarly for $\tilde{\theta}_Y$ and $\tilde{\theta}_{X|Y}$. Thus this law is strong hyper Markov. Because it is continuous, it also works for the restricted model without structural zeros of Example 1.

The Dirichlet law has been widely used for the analysis of case-control studies with a single binary covariate, corresponding to a 2 × 2 table (Zelen & Parker, 1986; Nurminen & Mutanen, 1987; Marshall, 1988; Ashby et al., 1993). The distribution of the odds ratio parameter $\tilde{\lambda}$ has been explored by Altham (1969).

Example 8. Consider the bivariate normal model of Example 2, restricted for simplicity to have zero means. The standard conjugate prior is the inverse Wishart distribution for the dispersion matrix Σ, having density proportional to

$$|\Sigma|^{a} \exp\left\{-\tfrac{1}{2}\operatorname{tr}(A\Sigma)\right\}.$$

Then the posterior is of the same form, with updated hyperparameters a^{*} and A^{*}. The inverse Wishart distribution determines a strong hyper Markov law, with similar marginalization properties to those of the Dirichlet law (Dawid & Lauritzen, 1993, Lemma 7.4). Similar results hold for the nonzero means model, where the conjugate normal-inverse Wishart distribution determines a strong hyper Markov law.

The independence of the odds ratio $\tilde{\lambda}$ from each of the marginal distributions $\tilde{\theta}_X$ and $\tilde{\theta}_Y$ allows us to construct further families of strong hyper Markov laws from existing ones.

Theorem 4. *If $\mathcal{L}$ is a strong hyper Markov law, then any law $\mathcal{L}'$ having Radon–Nikodym derivative of the form*

$$\frac{\mathrm{d}\mathcal{L}'}{\mathrm{d}\mathcal{L}}(\theta) = h(\lambda)$$

is also strong hyper Markov. Furthermore, the marginal laws for $\tilde{\theta}_X$ and $\tilde{\theta}_Y$ are the same under $\mathcal{L}'$ as under $\mathcal{L}$.

Proof. Let A be an element of the σ-algebra generated by $\tilde{\theta}_{Y|X}$. Since $\tilde{\theta}_{Y|X} \perp\!\!\!\perp \tilde{\theta}_X$ under $\mathcal{L}$,

$$\mathcal{L}'(A \mid \tilde{\theta}_X) = E_{\mathcal{L}}\left[h\{\lambda(\tilde{\theta}_{Y|X})\}\, 1_A(\tilde{\theta}_{Y|X}) \mid \tilde{\theta}_X\right] = E_{\mathcal{L}}\left\{h(\tilde{\lambda})\, 1_A(\tilde{\theta}_{Y|X})\right\} = \mathcal{L}'(A),$$

and hence $\tilde{\theta}_{Y|X} \perp\!\!\!\perp \tilde{\theta}_X$ under $\mathcal{L}'$. Similarly, $\tilde{\theta}_{X|Y} \perp\!\!\!\perp \tilde{\theta}_Y$ under $\mathcal{L}'$.

Now let B be an element of the σ-algebra generated by $\tilde{\theta}_X$. Then

$$\mathcal{L}'(B) = E_{\mathcal{L}}\left[h\{\lambda(\tilde{\theta}_{Y|X})\} 1_B(\tilde{\theta}_X)\right] = E_{\mathcal{L}}\left[h\{\lambda(\tilde{\theta}_{Y|X})\}\right] E_{\mathcal{L}}\left\{1_B(\tilde{\theta}_X)\right\} = \mathcal{L}(B),$$

and similarly for $\tilde{\theta}_Y$. □

We can also extend the constraint procedure of Lemma 3 to construct strong hyper Markov laws on the resulting submodel Θ'.

Theorem 5. *Let $\mathcal{L}(\tilde{\theta})$ be a strong hyper Markov law, and let f be a function of λ. Then the law $\mathcal{L}'(\tilde{\theta}) = \mathcal{L}(\tilde{\theta} \mid \tilde{f} = 0)$ is strong hyper Markov for the submodel Θ' specified by $f = 0$. Furthermore, the marginal laws for $\tilde{\theta}_X$ and $\tilde{\theta}_Y$ are the same under $\mathcal{L}'$ as under $\mathcal{L}$.*

Proof. As $\tilde{\theta}_X \perp\!\!\!\perp \tilde{\theta}_{Y|X}$ and $\tilde{f}$ is a function of $\tilde{\theta}_{Y|X}$, we have

$$\tilde{\theta}_X \perp\!\!\!\perp \tilde{\theta}_{Y|X} \mid \tilde{f} \quad [\mathcal{L}], \tag{6}$$

$$\tilde{\theta}_X \perp\!\!\!\perp \tilde{f} \quad [\mathcal{L}]. \tag{7}$$

Parallel results hold with X and Y interchanged. Then (6) shows that $\mathcal{L}(\tilde{\theta})$ remains strong hyper Markov under conditioning on $\tilde{f} = 0$, while (7) shows that this conditioning does not affect the marginal laws. □

Remark 1. Together, Theorems 4 and 5 can be paraphrased as saying that if $\mathcal{L}$ is a strong hyper Markov law for $\tilde{\theta}$ and the law $\mathcal{L}'$ has the same conditional distribution for $\tilde{\theta}$ given $\tilde{\lambda}$ as $\mathcal{L}$

does, then $\mathcal{L}'$ is strong hyper Markov, with unchanged marginal laws for $\tilde{\theta}_X$ and $\tilde{\theta}_Y$. In particular, this construction allows $\tilde{\lambda}$ to be assigned any distribution whatsoever under $\mathcal{L}'$.

Example 9. For a two-way contingency table, any law with density of the form

$$h\left(\frac{\theta_{xy}\theta_{x'y'}}{\theta_{xy'}\theta_{x'y}}\right)_{x,y,x',y'} \prod_{(x,y)} \theta_{xy}^{a_{xy}-1}$$

will be strong hyper Markov. Geiger & Heckerman (1997, Equation (10)) noted that all strong hyper Markov laws for 2 × 2 tables must have a density of this form.

Example 10. For the zero-means bivariate normal model, any law with density of the form

$$h\left(\frac{\sigma_{XY}}{\sigma_{XX}\sigma_{YY}-\sigma_{XY}^{2}}\right)|\Sigma|^{a}\exp\left\{-\tfrac{1}{2}\,\mathrm{tr}(A\Sigma)\right\}$$

will be strong hyper Markov. Geiger & Heckerman (2002, Theorem 12) showed that all strong hyper Markov laws for the bivariate normal model must have a density of this form.

The construction of laws for nested models by conditioning on specific parameters has been proposed in Dawid & Lauritzen (2001, § 4). Laws constructed by this procedure will also satisfy the conditions of Theorem 3.

Example 11. Consider a logistic model for finite covariate space $\mathcal{X}$, as generated by the conditioning procedure of Example 3.

We start with a generalized Dirichlet law $\mathcal{L}(\tilde{\theta})$ for the saturated model. Then the law for $\tilde{\theta}_{Y|X}$ has density of the form

$$h(\lambda)\prod_{x\in\mathcal{X}}\theta_{0|x}^{a_{x0}-1}\theta_{1|x}^{a_{x1}-1}.$$

The Jacobian determinant of the transformation to the logistic parameterization is

$$\left|\frac{\mathrm{d}\theta_{Y|X}}{\mathrm{d}(\alpha,\beta,\eta)}\right| \propto \prod_{x\in\mathcal{X}}\frac{\exp(\alpha+\beta^{\mathrm{T}}x+\eta_x)}{\{1+\exp(\alpha+\beta^{\mathrm{T}}x+\eta_x)\}^{2}},$$

and hence the density for $\mathcal{L}(\tilde{\alpha},\tilde{\beta},\tilde{\eta})$ is of the form

$$g(\beta,\eta)\prod_{x\in\mathcal{X}}\frac{\exp\{(\alpha+\beta^{\mathrm{T}}x+\eta_x)a_{x1}\}}{\{1+\exp(\alpha+\beta^{\mathrm{T}}x+\eta_x)\}^{a_{x+}}},$$

where $a_{x+}=a_{x0}+a_{x1}$. By conditioning on $\tilde{\eta}=0$, we obtain the density of $\mathcal{L}'(\tilde{\alpha},\tilde{\beta})$, which is of the form

$$g(\beta)\prod_{x\in\mathcal{X}}\frac{\exp\{(\alpha+\beta^{\mathrm{T}}x)a_{x1}\}}{\{1+\exp(\alpha+\beta^{\mathrm{T}}x)\}^{a_{x+}}}. \tag{8}$$

The Jacobian of the transformation in terms of the retrospective parameters is

$$\left|\frac{\mathrm{d}(\alpha,\beta,\theta_X)}{\mathrm{d}(\theta_{X|0},\beta,\theta_{Y=1})}\right| = \frac{(1-\theta_{Y=1})^{|\mathcal{X}|-1}}{\theta_{Y=1}}\prod_{x\in\mathcal{X}}\left\{1+\exp(\alpha+\beta^{\mathrm{T}}x)\right\}.$$

Therefore, using a prior law with density (8) for the prospective analysis of retrospective data is justified when the true retrospective prior law is

$$g(\beta)\frac{\prod_{x\in\mathcal{X}}\theta_{x|0}^{a_{x+}-1}\exp(a_{x1}\beta^{\mathrm{T}}x)}{\{\sum_{x\in\mathcal{X}}\theta_{x|0}\exp(\beta^{\mathrm{T}}x)\}^{a_{+1}}}.$$

Priors of this form have previously appeared in the literature. The prior of Staicu (2010, Example 2) is obtained by rewriting (8) as

$$g^{*}(\beta)\exp(\alpha a_{+1})\prod_{x\in\mathcal{X}}\{1+\exp(\alpha+\beta^{\mathrm{T}}x)\}^{-a_{x+}},$$

where $g^{*}(\beta)=g(\beta)\exp(\sum_{x\in\mathcal{X}}a_{x1}\beta^{\mathrm{T}}x)$. The improper prior of Seaman & Richardson (2004) and Staicu (2010, Example 1) can be obtained by further taking the limit as $a_{+1}\rightarrow 0$. However, we argue that the form of (8) is more easily interpreted: it can be thought of as the product of an improper prior with density element $g(\beta)\,\mathrm{d}\beta\,\mathrm{d}\alpha$ and a logistic likelihood function, where the a_{xy} represent pseudo-counts. This has the further benefit of being easily adaptable to existing computational methods; for example, a Laplace approximation can be found using standard logistic regression software.

Although x appears in the density (8), we disagree with Staicu (2010) that this constitutes a covariate-dependent prior, like the g-priors of Zellner (1986); it is only dependent on the a priori expected frequencies of the covariates, not on their observed frequencies in the data.

The logistic generalized Dirichlet law can similarly be extended to the multinomial model of Example 4, yielding a density of the form

$$g(\beta)\prod_{x\in\mathcal{X}}\frac{\prod_{y\neq y^{*}}\exp\{(\alpha_y+\beta_y^{\mathrm{T}}x)a_{xy}\}}{\{1+\sum_{y\neq y^{*}}\exp(\alpha_y+\beta_y^{\mathrm{T}}x)\}^{a_{x+}}}. \tag{9}$$

By further conditioning, this can be applied to the stereotype model of Example 5, using a prior density of the form

$$g(\beta,\gamma)\prod_{x\in\mathcal{X}}\frac{\prod_{y\neq y^{*}}\exp\{(\alpha_y+\gamma_y\beta^{\mathrm{T}}x)a_{xy}\}}{\{1+\sum_{y\neq y^{*}}\exp(\alpha_y+\gamma_y\beta^{\mathrm{T}}x)\}^{a_{x+}}}. \tag{10}$$

An analogous construction for the multiplicative intercept model of Example 6 uses a prior density of the form

$$g(\beta)\prod_{x\in\mathcal{X}}\frac{\exp[\{\alpha+f(x,\beta)\}a_{x1}]}{[1+\exp\{\alpha+f(x,\beta)\}]^{a_{x+}}}. \tag{11}$$

The improper priors of Ghosh et al. (2012, Theorem 1) can be obtained from (9), (10) and (11) by taking the limit as $a_{x+}\rightarrow 0$. However, their claim that these priors can also be used for link functions other than the logistic one, such as the probit, skew-symmetric or cumulative logit cases, is incorrect, as these models are not strong meta Markov and hence cannot support strong hyper Markov laws.

The form of the generalized logistic Dirichlet law allows for easy implementation in generic Bayesian Markov chain Monte Carlo packages such as WinBUGS, OpenBUGS and JAGS, which accept noninteger values for binomial counts. Furthermore, arbitrary functions g can be included by use of the zero Poisson trick; see Lunn et al. (2013, § 9.5). Unfortunately, this method is somewhat impractical for large numbers of covariates, since the size of $\mathcal{X}$ increases exponentially with its dimensionality k. Furthermore, as $\mathcal{X}$ increases, $\tilde{\beta}$ will tend to concentrate around 0. To compensate for this, the values of the a_{xy} can be chosen closer to 0, but the above software packages do not work well for very small values.

6. Stratified models

A more complicated analysis is that of stratified or matched case-control studies, in which participants are selected by both the outcome Y and an additional stratum variable S, taking values in $\mathcal{S}$. Such a design can often estimate the odds ratio of interest with much greater efficiency than an unstratified study.

It is enough to consider sampling schemes that condition on S, so that the parameter of the joint likelihood is $\theta_{XY|S}$. The prospective parameter of interest is $\theta_{Y|XS}$, but data may be observed under the retrospective regime, only allowing estimation of $\theta_{X|YS}$. In this case the parameter λ that is a function of both $\theta_{Y|XS}$ and $\theta_{X|YS}$ is the set of all odds ratios of the form

$$\frac{p(x, y \mid s, \theta)\, p(x', y' \mid s, \theta)}{p(x, y' \mid s, \theta)\, p(x', y \mid s, \theta)} \quad (x, x' \in \mathcal{X};\ y, y' \in \mathcal{Y};\ s \in \mathcal{S}).$$

Example 12. The stratified logistic model is similar to Example 3 but with an intercept parameter that varies by stratum, so that the prospective model is

$$p(y \mid x, s, \alpha, \beta) = \frac{\exp(\alpha_s + \beta^{\mathrm{T}} x)}{1 + \exp(\alpha_s + \beta^{\mathrm{T}} x)}.$$

As in the unstratified case, $\lambda \simeq \beta$.

This additional complication can make estimation more difficult. The number of strata will typically increase with sample size, with the result that the maximum likelihood estimator is inconsistent. An alternative under the classical approach is to maximize the conditional likelihood

$$L_{\mathrm{c}}(\beta) = \prod_{s \in \mathcal{S}} \frac{\prod_{i \in I_s} \exp(y_i \beta^{\mathrm{T}} x_x)}{\sum_{\rho} \prod_{i \in I_s} \exp(y_{\rho(i)} \beta^{\mathrm{T}} x_x)},$$

where $I_s = \{i : s_i = s\}$ and the summation in the denominator is over all possible permutations of $(y_i)_{i \in I_s}$. If there are a cases and b controls in each stratum, called a:b matching, the sum in the denominator will have $(a + b)!/(a!\, b!)$ terms. In order to keep this computationally tractable, most studies use 1:1 or 1:m matching.

The conditional likelihood does not have a direct Bayesian interpretation. Rice (2004, Theorem 1) showed that there exists a law such that the marginal retrospective likelihood $\bar{p}(x \mid y, s, \beta)$ is proportional to the conditional likelihood. However, this law depends on the matching scheme; for example, a 1:1 matched design and a 1:2 matched design will require different laws.

Alternatively, Theorem 2 can be extended to support use of the prospective likelihood.

Theorem 6. *Let $\mathcal{L}$ be a prior law for the parameter $\tilde{\theta}_{XY|S}$ of a stratified model, with the property that*

$$\tilde{\theta}_{Y|XS} \perp\!\!\!\perp \tilde{\theta}_{X|S}, \quad \tilde{\theta}_{X|YS} \perp\!\!\!\perp \tilde{\theta}_{Y|S} \quad [\mathcal{L}].$$

Then the posterior marginal law for the odds ratios $\tilde{\lambda}$ is the same under the prospective, the retrospective and the joint likelihoods.

The argument is essentially the same as that for Theorem 2.

Laws satisfying Theorem 6 can be constructed from a collection of strong hyper Markov laws $\mathcal{L}_s(\tilde{\theta}_{XY|S=s})$ on the individual strata. A simple example is the product law

$$\mathcal{L}(\tilde{\theta}_{XY|S}) = \prod_{s\in\mathcal{S}} \mathcal{L}_s(\tilde{\theta}_{XY|S=s}),$$

which is equivalent to fitting a separate model for each stratum, each having its individual odds ratio parameter. The opposite case is that of a law $\mathcal{L}$ that constrains $\tilde{\theta}_{XY|S=s} = \tilde{\theta}_{XY|S=s'}$ almost surely, thus ignoring stratification altogether. However, neither of these extreme cases is able to exploit the key advantage of stratification, namely that it allows for fitting a model with both common and stratum-specific parameters, such as the logistic model in Example 12, where all strata share a common odds ratio. This can be effected as follows.

THEOREM 7. *Let $\{\mathcal{L}_s(\tilde{\theta}_{XY|S=s}) : s \in \mathcal{S}\}$ be a collection of strong hyper Markov laws such that the marginal laws for the odds ratios are equal; that is,*

$$\mathcal{L}_s(\tilde{\lambda}_s) = \mathcal{L}_{s'}(\tilde{\lambda}_{s'})$$

for all $s, s' \in \mathcal{S}$. Then there exists a unique joint law $\mathcal{L}(\tilde{\theta}_{XY|S})$ such that $\mathcal{L}(\tilde{\theta}_{XY|S=s}) = \mathcal{L}_s(\tilde{\theta}_{XY|S=s})$, $\tilde{\lambda}_s = \tilde{\lambda}_{s'}$ almost surely, and the $\tilde{\theta}_{XY|S=s}$ $(s \in \mathcal{S})$ are conditionally independent given $\tilde{\lambda}$. Moreover, this law satisfies the conditions of Theorem 6.

Proof. The existence and uniqueness of $\mathcal{L}$ are given by the Markov combination construction of Dawid & Lauritzen (1993, Lemma 2.5). It remains to show that the conditions of Theorem 6 are satisfied for $\mathcal{L}$.

The mutual independence of all the $\tilde{\theta}_{XY|S=s}$ conditional on $\tilde{\lambda}$, combined with the strong hyper Markov properties of the $\mathcal{L}_s$, implies the mutual independence, given $\tilde{\lambda}$, of all terms of the form $\tilde{\theta}_{Y|X,S=s}$, $\tilde{\theta}_{X|S=s'}$. In particular,

$$\tilde{\theta}_{Y|XS} \perp\!\!\!\perp \tilde{\theta}_{X|S} \mid \tilde{\lambda}, \tag{12}$$

$$\coprod_{s\in\mathcal{S}} \{\tilde{\theta}_{X|S=s}\} \mid \tilde{\lambda}. \tag{13}$$

Also, since $\mathcal{L}_s$ is strong hyper Markov, we have, for each s,

$$\tilde{\theta}_{X|S=s} \perp\!\!\!\perp \tilde{\lambda}. \tag{14}$$

An easy application of the rules of conditional independence shows that (13) and (14) together imply $\tilde{\theta}_{X|S} \perp\!\!\!\perp \tilde{\lambda}$, which combined with (12) gives $\tilde{\theta}_{Y|XS} \perp\!\!\!\perp \tilde{\theta}_{X|S}$, since $\tilde{\lambda}$ is a function of $\tilde{\theta}_{Y|XS}$. Similarly, $\tilde{\theta}_{X|YS} \perp\!\!\!\perp \tilde{\theta}_{Y|S}$. □

Example 13. For the stratified logistic model in Example 12, suppose that each law $\mathcal{L}_s$ is specified by a density for $(\tilde{\alpha}_s, \tilde{\beta})$ of the form

$$g_s(\beta) \prod_{x\in\mathcal{X}} \frac{\exp\{(\alpha_s + \beta^{\mathrm{T}} x) a_{x1s}\}}{\{1 + \exp(\alpha_s + \beta^{\mathrm{T}} x)\}^{a_{x+s}}},$$

such that the marginal density for $\tilde{\beta}$ is $p(\beta)$ in each stratum s. By Theorem 5, this can be achieved by choosing

$$g_s(\beta) = \frac{p(\beta)}{\int_{\mathbb{R}} \prod_{x\in\mathcal{X}} \frac{\exp\{(\alpha_s+\beta^{\mathrm{T}}x)a_{x1s}\}}{\{1+\exp(\alpha_s+\beta^{\mathrm{T}}x)\}^{a_{x+s}}}\, \mathrm{d}\alpha_s}.$$

The corresponding joint density for $(\tilde{\alpha}, \tilde{\beta})$ is then

$$g(\beta) \prod_{(x,s)\in\mathcal{X}\times\mathcal{S}} \frac{\exp\{(\alpha_s+\beta^{\mathrm{T}}x)a_{x1s}\}}{\{1+\exp(\alpha_s+\beta^{\mathrm{T}}x)\}^{a_{x+s}}},$$

where $g(\beta) = \{\prod_{s\in\mathcal{S}} g_s(\beta)\}/\{p(\beta)^{|\mathcal{S}|-1}\}$.

This is of the same form as the density (8), where the strata are treated as an additional categorical covariate in the model. As with the unmatched case, the improper laws of Ghosh et al. (2006, 2012) can be obtained by taking the limit as $a_{xys} \to 0$, though again the claims in Ghosh et al. (2012) regarding the use of different link functions are incorrect. Similar priors can be obtained for the multinomial and stereotype models in the previous section.

Again, we emphasize that using such a law for the prospective analysis of retrospective data requires that the prior law $\mathcal{L}(\tilde{\theta}_{X|YS})$ be the marginal of a joint law such that $\tilde{\theta}_{Y|XS} \perp\!\!\!\perp \tilde{\theta}_{X|S}$ and $\mathcal{L}(\tilde{\theta}_{X|S=s}) = \mathcal{D}(a_{xs})$.

We have not specified a model for the stratum variable S, as we have assumed that all data are observed conditional on S. However, under the additional assumption that $\tilde{\theta}_{XY|S} \perp\!\!\!\perp \tilde{\theta}_S \ [\mathcal{L}]$, the data can be treated as if they were randomly sampled from the population, as would be the case for a cross-sectional study.

7. Discussion

We have outlined a broad framework with necessary assumptions for the analysis of retrospective data using a prospective likelihood or Bayesian approach.

Our Bayesian analysis requires the existence of a joint strong hyper Markov law of which the prospective and retrospective laws are its margins. Because of the difficulties in defining and handling marginalization for improper priors (Dawid et al., 1973), our arguments do not readily extend to improper priors, whose use in this context may require a different justification.

These results apply only to functions of the odds ratio. Other quantities such as an intercept parameter α cannot be inferred using this approach, nor does it account for more recent developments such as case-cohort designs and incorporation of population incidence data.

Many analyses (e.g., de Vocht et al., 2012) have used multivariate normal prior laws for the logistic log odds parameter $\tilde{\beta} \simeq \tilde{\lambda}$; but the overall laws used are not strong hyper Markov, and the resulting prospective and retrospective posterior laws for $\tilde{\beta}$ are not equal. However, Remark 1 shows that it is indeed possible to construct a strong hyper Markov law such that $\tilde{\beta}$ is multivariate normal, and the previously suggested prior laws might possibly be interpretable as approximating such a strong hyper Markov law. There could nevertheless be considerable difficulty in determining the precise form of the implied law for the retrospective parameters.

Similar properties and techniques arise in other contexts. A recent example is the development of inverse regression techniques used for dimension reduction (Cook & Li, 2009; Taddy, 2013). These methods exploit the existence of low-dimensional representations of the odds ratio λ, termed a sufficient reduction, and utilize a similar method of obtaining estimates by fitting the wrong inverse model to the data.

Another example arises in the computation of graphical lasso estimators for high-dimensional covariance matrices (Banerjee et al., 2008; Friedman et al., 2008). These are shrinkage estimators which penalize off-diagonal elements of the precision matrix. Due to the strong meta Markov property of the multivariate normal model and the penalized terms being functions of the odds ratio, a similar argument to the proof of Theorem 1 can be used to show that the solution to the optimization problem is equivalent to a set of penalized regression problems of each covariate against all the others. As a result, the estimate can be computed by an iterative scheme of lasso regressions.

Acknowledgement

The research of the first author was supported by an EPSRC Postdoctoral Fellowship.

References

Altham, P. M. E. (1969). Exact Bayesian analysis of a 2 × 2 contingency table, and Fisher's "exact" significance test. *J. R. Statist. Soc.* B **31**, 261–9.

Altham, P. M. E. (1970). The measurement of association of rows and columns for an $r \times s$ contingency table. *J. R. Statist. Soc.* B **32**, 63–73.

Anderson, J. A. (1984). Regression and ordered categorical variables. *J. R. Statist. Soc.* B **46**, 1–30.

Ashby, D., Hutton, J. L. & McGee, M. A. (1993). Simple Bayesian analyses for case-control studies in cancer epidemiology. *Statistician* **42**, 385–97.

Baker, S. G. (1994). The multinomial-Poisson transformation. *Statistician* **43**, 495–504.

Banerjee, O., El Ghaoui, L. & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9**, 485–516.

Cook, R. D. & Li, L. (2009). Dimension reduction in regressions with exponential family predictors. *J. Comp. Graph. Statist.* **18**, 774–91.

Dawid, A. P. (2001a). Separoids: A mathematical framework for conditional independence and irrelevance. *Ann. Math. Artif. Intel.* **32**, 335–72.

Dawid, A. P. (2001b). Some variations on variation independence. In *Artificial Intelligence and Statistics 2001*, T. Jaakkola & T. Richardson, eds. San Francisco: Morgan Kaufmann, pp. 187–91.

Dawid, A. P. & Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21**, 1272–317.

Dawid, A. P. & Lauritzen, S. L. (2001). Compatible prior distributions. In *Bayesian Methods with Applications to Science, Policy and Official Statistics*, E. I. George, ed. Luxembourg: Office for Official Publications of the European Communities, pp. 109–18.

Dawid, A. P., Stone, M. & Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. R. Statist. Soc.* B **35**, 189–233.

de Vocht, F., Cherry, N. & Wakefield, J. (2012). A Bayesian mixture modeling approach for assessing the effects of correlated exposures in case-control studies. *J. Expos. Sci. Envir. Epidemiol.* **22**, 352–60.

Friedman, J., Hastie, T. J. & Tibshirani, R. J. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.

Geiger, D. & Heckerman, D. (1997). A characterization of the Dirichlet distribution through global and local parameter independence. *Ann. Statist.* **25**, 1344–69.

Geiger, D. & Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.* **30**, 1412–40.

Ghosh, M., Song, J., Forster, J., Mitra, R. & Mukherjee, B. (2012). On the equivalence of posterior inference based on retrospective and prospective likelihoods: Application to a case-control study of colorectal cancer. *Statist. Med.* **31**, 2196–208.

Ghosh, M., Zhang, L. & Mukherjee, B. (2006). Equivalence of posteriors in the Bayesian analysis of the multinomial-Poisson transformation. *Metron* **64**, 19–28.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statist. Med.* **13**, 1665–77.

Gustafson, P., Le, N. D. & Vallée, M. (2002). A Bayesian approach to case-control studies with errors in covariables. *Biostatistics* **3**, 229–43.

Hsieh, D. A., Manski, C. F. & McFadden, D. (1985). Estimation of response probabilities from augmented retrospective observations. *J. Am. Statist. Assoc.* **80**, 651–62.

Lunn, D., Jackson, C., Best, N., Thomas, A. & Spiegelhalter, D. (2013). *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Boca Raton: CRC Press.

Marshall, R. J. (1988). Bayesian analysis of case-control studies. *Statist. Med.* **7**, 1223–30.

McCullagh, P. (1980). Regression models for ordinal data (with Discussion). *J. R. Statist. Soc.* B **42**, 109–42.

Mukherjee, B., Sinha, S. & Ghosh, M. (2005). Bayesian analysis of case-control studies. In *Bayesian Thinking: Modeling and Computation*, D. K. Dey & C. R. Rao, eds., vol. 25 of *Handbook of Statistics*. Amsterdam: Elsevier/North-Holland, pp. 793–819.

Müller, P. & Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika* **84**, 523–37.

Nurminen, M. & Mutanen, P. (1987). Exact Bayesian analysis of two proportions. *Scand. J. Statist.* **14**, 67–77.

Park, M. Y. & Hastie, T. J. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50.

Patefield, W. M. (1985). Information from the maximized likelihood function. *Biometrika* **72**, 664–8.

Prentice, R. L. & Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

Rice, K. M. (2004). Equivalence between conditional and mixture approaches to the Rasch model and matched case-control studies, with applications. *J. Am. Statist. Assoc.* **99**, 510–22.

Seaman, S. R. & Richardson, S. (2001). Bayesian analysis of case-control studies with categorical covariates. *Biometrika* **88**, 1073–88.

Seaman, S. R. & Richardson, S. (2004). Equivalence of prospective and retrospective models in the Bayesian analysis of case-control studies. *Biometrika* **91**, 15–25.

Staicu, A.-M. (2010). On the equivalence of prospective and retrospective likelihood methods in case-control studies. *Biometrika* **97**, 990–6.

Taddy, M. (2013). Multinomial inverse regression for text analysis. *J. Am. Statist. Assoc.* **108**, 755–70.

Weinberg, C. R. & Wacholder, S. (1993). Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika* **80**, 461–5.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. & Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–21.

Zelen, M. & Parker, R. A. (1986). Case-control studies and Bayesian inference. *Statist. Med.* **5**, 261–9.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In *Bayesian Inference and Decision Techniques*, P. K. Goel & A. Zellner, eds., vol. 6 of *Studies in Bayesian Econometrics and Statistics*. Amsterdam: North-Holland, pp. 233–43.

[*Received May* 2013. *Revised August* 2013]