Biometrika (2016), **103**, 4, *pp.* 971–983
Printed in Great Britain
doi: 10.1093/biomet/asw053

Bayesian inference on quasi-sparse count data

By JYOTISHKA DATTA

Department of Mathematical Sciences, University of Arkansas, Fayetteville, Arkansas 72701, U.S.A.
jd033@uark.edu

and DAVID B. DUNSON

Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A.
dunson@duke.edu

Summary

There is growing interest in analysing high-dimensional count data, which often exhibit quasi-sparsity corresponding to an overabundance of zeros and small nonzero counts. Existing methods for analysing multivariate count data via Poisson or negative binomial log-linear hierarchical models with zero-inflation cannot flexibly adapt to quasi-sparse settings. We develop a new class of continuous local-global shrinkage priors tailored to quasi-sparse counts. Theoretical properties are assessed, including flexible posterior concentration and stronger control of false discoveries in multiple testing. Simulation studies demonstrate excellent small-sample properties relative to competing methods. We use the method to detect rare mutational hotspots in exome sequencing data and to identify North American cities most impacted by terrorism.

Some key words: Count data; High-dimensional data; Local-global shrinkage; Rare variant; Shrinkage prior; Zero-inflation.

1. Introduction

In this article we consider the modelling of high-dimensional quasi-sparse count data $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ having an excess of values near zero. In many applications, the rates of event occurrence are very small at a majority of locations, with substantially higher rates at a subset of locations. For example, y_i may represent the number of mutations observed at location i in the genome or the number of terrorist activities in city i. There is a rich literature on the modelling of count data focused on accommodating overdispersion relative to the Poisson distribution and zero-inflation, but such models are insufficiently flexible for quasi-sparse data with an abundance of very small nonzero counts. We propose a novel shrinkage prior for the rate parameters, which simultaneously accommodates quasi-sparse and heavy-tailed signals θ, while maintaining computational tractability and theoretical guarantees. The proposed shrinkage prior is built upon the Gauss hypergeometric distribution proposed by Armero & Bayarri (1994) for modelling the traffic intensity of an $M/M/1$ queue in equilibrium.

For simplicity, we focus on the model $y_i \sim \mathrm{Po}(\theta_i)$, independently for $i = 1, \ldots, n$, with $\theta = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$. However, our proposed approach can be used directly for more elaborate models that let $y_i \sim \mathrm{Po}(\theta_i \eta_i)$, where θ_i represents a random effect and η_i a structured part characterizing dependence on covariates, hierarchical designs, or spatial or temporal structure. In the simple case, estimating θ is commonly referred to as the Poisson compound decision problem.

Empirical Bayes approaches have been popular in this context, dating back to the formula of Robbins (1956). In Robbins's approach, the θ_i are assumed to be independent draws from a distribution $G(\cdot)$, and the goal is to estimate $\theta = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$ depending on the observations $y = (y_1, \ldots, y_n)^{\mathrm{T}}$. The accuracy of an estimator $\hat{\theta} = \delta(y) = \{\delta_1(y_1), \ldots, \delta_n(y_n)\}^{\mathrm{T}}$ is assessed by the risk $W(\delta) = E_\theta\{\|\delta(y) - \theta\|^2\}$. The Bayes estimator that minimizes $W(\delta)$ assumes a simple form for the Poisson kernel, with

$$\delta_i(y_i) = \frac{\int \theta_i p(y_i \mid \theta_i)\, \mathrm{d}G(\theta_i)}{\int p(y_i \mid \theta_i)\, \mathrm{d}G(\theta_i)} = \frac{(y_i + 1)P_y(y_i + 1)}{P_y(y_i)} \quad (i = 1, \ldots, n),$$

where $P_y(\cdot)$ is the marginal distribution of y. Robbins's frequency ratio estimator uses the empirical frequencies $\hat{P}_y(z) = n^{-1} \sum_{i=1}^n I(y_i = z)$ to estimate P_y. Brown et al. (2013) showed that slow convergence of $\hat{P}_y$ to P_y deteriorates performance, and proposed a three-stage smoothing adjustment that substantially improves the total Bayes risk $nW(\delta)$ in simulation studies. Koenker & Mizera (2014) proposed a computationally efficient approximation to estimating θ by nonparametrically maximizing the likelihood with respect to the unknown distribution G (Kiefer & Wolfowitz, 1956).

In this article we develop a hierarchical Bayesian model that allows for quasi-sparsity while maintaining the ability to capture large signals. The proposed prior is adaptive to the degree of quasi-sparsity in the data, and is inspired by local-global shrinkage priors for sparse Gaussian means and linear regression (Carvalho et al., 2010; Armagan et al., 2011, 2013; Bhattacharya et al., 2015). Such priors are structured as scale mixtures of Gaussian densities for computational convenience, with corresponding theoretical support when the true mean or regression vector is mostly zero (van der Pas et al., 2014; Bhattacharya et al., 2015). Naïvely, one could apply such priors to the coefficients in Poisson log-linear models, but such formulations lack the computational advantages afforded in the Gaussian case and fail to represent quasi-sparsity.

Our proposed model induces inflation of small counts in a continuous manner, which has important advantages over zero-inflated Poisson models and their many variants, such as the zero-inflated generalized Poisson and zero-inflated negative binomial distributions (Yang et al., 2009). Under a zero-inflated model, the θ_i are set to zero with probability p or sampled from a simple parametric distribution, typically either a degenerate distribution at a single θ value or a gamma distribution. This restrictive parametric form limits performance, as we shall illustrate. The two-component mixture form also leads to computational instability in quasi-sparse count examples due to dependence between the parameters p and θ.

2. SHRINKAGE PRIORS FOR COUNT DATA

2·1. *Motivation*

If $\theta_i \sim \mathrm{Ga}(\alpha, \beta_i)$, the marginal distribution of y_i is negative binomial with variance $\alpha\beta_i^{-1}(1 + \beta_i^{-1})$, which is higher than the mean $\alpha\beta_i^{-1}$. To allow for zero-inflation, the usual approach would be to mix a Poisson or negative binomial distribution with a degenerate distribution at zero. We instead choose a prior for θ_i with a pole at zero, leading to a large probability mass in the marginal distribution for y_i at zero. Our Poisson-gamma hierarchical model can be expressed as

$$y_i \sim \mathrm{Po}(\theta_i), \quad \theta_i \sim \mathrm{Ga}(\alpha, \lambda_i^2\tau^2), \quad \lambda_i \sim p(\lambda_i^2), \quad \tau^2 \sim p(\tau^2) \quad (\lambda_i, \tau > 0),$$

where $p(\lambda_i^2)$ and $p(\tau^2)$ are densities for λ_i^2 and τ^2, respectively. Marginalizing out θ_i and writing $\kappa_i = 1/(1+\lambda_i^2\tau^2)$, the hierarchical relationship can be rewritten as

$$p(y_i \mid \lambda_i, \tau) \propto \left(\frac{\lambda_i^2\tau^2}{1+\lambda_i^2\tau^2}\right)^{y_i}\left(\frac{1}{1+\lambda_i^2\tau^2}\right)^{\alpha},$$
$$p(y_i \mid \kappa_i) \propto (1-\kappa_i)^{y_i}\kappa_i^{\alpha} \quad (\alpha > 0). \tag{1}$$

This implies that marginally y_i follows a negative binomial distribution with size α and probability of success $1-\kappa_i$. The posterior density and mean of θ_i given y_i and κ_i are, respectively,

$$p(\theta_i \mid y_i, \kappa_i) \sim \mathrm{Ga}(y_i+\alpha, 1-\kappa_i), \quad E(\theta_i \mid y_i, \kappa_i) = (1-\kappa_i)(y_i+\alpha) \quad (\alpha > 0).$$

Hence, κ_i can be interpreted as a shrinkage factor pulling the posterior mean towards zero.

Priors on shrinkage factors that have a U-shaped distribution are appealing in shrinking small signals to zero while avoiding shrinkage of larger signals. In normal linear models, such priors have been widely used and include the horseshoe (Carvalho et al., 2010), generalized double Pareto (Armagan et al., 2013), three-parameter beta (Armagan et al., 2011) and Dirichlet–Laplace (Bhattacharya et al., 2015) priors. In quasi-sparse count applications, we require additional flexibility in the mass of the shrinkage parameter κ_i around 0 and 1, due to the occurrences of very low counts in addition to zeros. For example, small counts can arise from measurement errors and should be separated from true rare events to the extent possible. This requires careful treatment of the prior for κ_i.

Consider the three-parameter beta prior (Armagan et al., 2011)

$$p(\kappa_i \mid a, b, \phi) \propto (1-\kappa_i)^{a-1}\kappa_i^{b-1}\{1-(1-\phi)\kappa_i\}^{-(a+b)} \quad (0 \leqslant \kappa_i \leqslant 1;\ a, b, \phi, \gamma > 0). \tag{2}$$

Armagan et al. (2011) recommend $a, b \in (0, 1)$, leading to both Cauchy-like tails and a kink at zero. For example, $a = b = 1/2$ and $\phi = 1$ in (2) leads to $\kappa_i \sim \mathrm{Be}(1/2, 1/2)$, where $\mathrm{Be}(a, b)$ denotes the beta distribution with parameters a and b. The horseshoe-shaped $\mathrm{Be}(1/2, 1/2)$ prior combined with the likelihood in (1) produces a $(1-\kappa_i)^{y_i-1/2}$ term in the posterior, which leaves all nonzero y_i unshrunk. For $a = b = 1/2$ and $\phi \to 0$, the corresponding term in the posterior would be $(1-\kappa_i)^{y_i-3/2}$, which shrinks $y_i \leqslant 1$.

To extend the flexibility of the prior on κ_i, while retaining the heavy-tailed property of the induced marginal prior on θ_i, we make the exponent in the final term in (2) a general nonnegative parameter γ. Higher values of γ imply shrinkage of larger observations in the posterior density. The proposed prior density on κ_i has the form

$$\mathrm{GH}(\kappa_i \mid a, b, \phi, \gamma) = C\kappa_i^{a-1}(1-\kappa_i)^{b-1}\{1-(1-\phi)\kappa_i\}^{-\gamma} \quad (0 \leqslant \kappa_i \leqslant 1;\ a, b, \phi, \gamma > 0), \tag{3}$$

where $C^{-1} = B(a, b)\,{}_2F_1(\gamma, a, a+b, 1-\phi)$ is the norming constant, with $B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$ being the beta function and ${}_2F_1$ the Gauss hypergeometric function, i.e.,

$${}_2F_1(a, b, c, z) = \sum_{k=0}^{\infty} \frac{(a)_k(b)_k}{(c)_k}\frac{z^k}{k!} \quad (|z| < 1),$$

where $(q)_k$ denotes the rising Pochhamer symbol, defined as $(q)_k = q(q+1)\cdots(q+k-1)$ for $k > 0$ and $(q)_0 = 1$. Expression (3) corresponds to the Gauss hypergeometric distribution

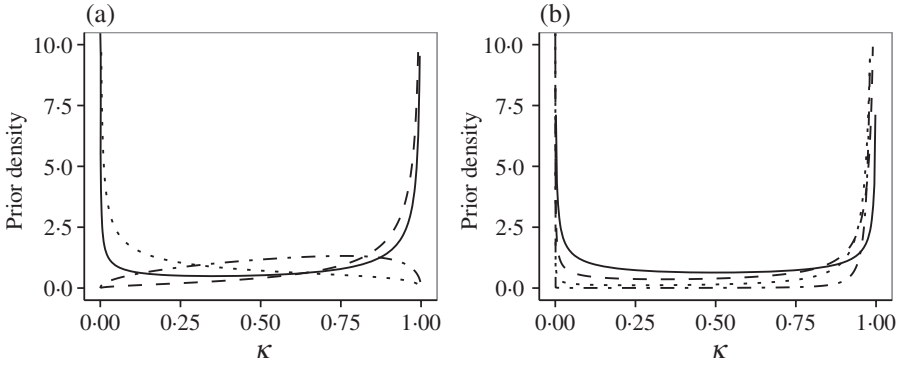

Fig. 1. Comparison of the density of κ_i under the Gauss hypergeometric prior for different values of the hyperparameters: (a) effect of a and b on the prior density of κ_i, for $a = b = 0{\cdot}5$ (solid), $a = 0{\cdot}5$ and $b = 1{\cdot}5$ (dashed), $a = 1{\cdot}5$ and $b = 0{\cdot}5$ (dotted), and $a = b = 1{\cdot}5$ (dot-dash); (b) effect of γ on the prior density of κ_i, for $\gamma = 0$ (solid), $\gamma = 0{\cdot}5$ (dashed), $\gamma = 1$ (dotted), and $\gamma = 2$ (dot-dash).

introduced by Armero & Bayarri (1994) for a specific queuing application. We fix $a = b = 1/2$ and $\phi = \tau^2$ for our prior on κ_i, where τ^2 is a global shrinkage parameter that adjusts to the level of quasi-sparsity in the data. This prior density is conjugate to the negative binomial likelihood in (1) and yields the posterior density

$$p(\kappa_i \mid y_i, \alpha, \tau, \gamma) = \frac{\kappa_i^{\alpha-1/2}(1-\kappa_i)^{y_i-1/2}\{1-(1-\tau^2)\kappa_i\}^{-\gamma}}{B(\alpha+1/2, y_i+1/2)\,{}_2F_1(\gamma, \alpha+1/2, y_i+\alpha+1, 1-\tau^2)},$$
$$\kappa_i \mid y_i, \alpha, \tau, \gamma \sim \mathrm{GH}\big(\alpha+1/2,\, y_i+1/2,\, \tau^2,\, \gamma\big) \quad (\alpha, \tau, \gamma > 0). \tag{4}$$

Plots of the prior density for different values of the hyperparameters a, b and γ for $\tau^2 = 0{\cdot}01$ are displayed in Fig. 1. Panel (a) shows the effect of different choices of the parameters a and b when $\gamma = 1/2$, and panel (b) shows the effect of γ on $p(\kappa_i)$ for $a = b = 1/2$.

As Fig. 1 shows, the Gauss hypergeometric prior results in a U-shaped prior density for κ_i when $a = b = 1/2$ with a small τ^2 for different values of γ. This is a general class that includes the horseshoe prior (Carvalho et al., 2010) as a default shrinkage prior for the sparse normal means problem. The horseshoe special case is outperformed by better default Gauss hypergeometric prior specifications in terms of estimation and misclassification error, as we will show in § 4.

The kth posterior moment for κ_i given y_i, α, τ and γ can be written as

$$E(\kappa_i^k \mid y_i, \alpha, \tau, \gamma) = \frac{B(k+\alpha+1/2, y_i+1/2)\,{}_2F_1(\gamma, k+\alpha+1/2, y_i+\alpha+1+k, 1-\tau^2)}{B(\alpha+1/2, y_i+1/2)\,{}_2F_1(\gamma, \alpha+1/2, y_i+\alpha+1, 1-\tau^2)}. \tag{5}$$

The posterior mean $E(\kappa_i \mid y_i, \alpha, \tau, \gamma)$ can be rapidly calculated using (5) by exploiting the fast convergence of the Gauss hypergeometric function ${}_2F_1(a, b, c, z)$ for $|z| < 1$, where τ and γ are chosen by empirical Bayes or crossvalidation. We show in § 3 that the posterior distribution for κ_i will concentrate near 0 or 1 for large or small observations, respectively.

2·2. *Impact of hyperparameters*

The three hyperparameters α, γ and τ in (4) determine the shape of the posterior density on κ_i given y_i and control the rate of concentration of the posterior. Small values of α cause the

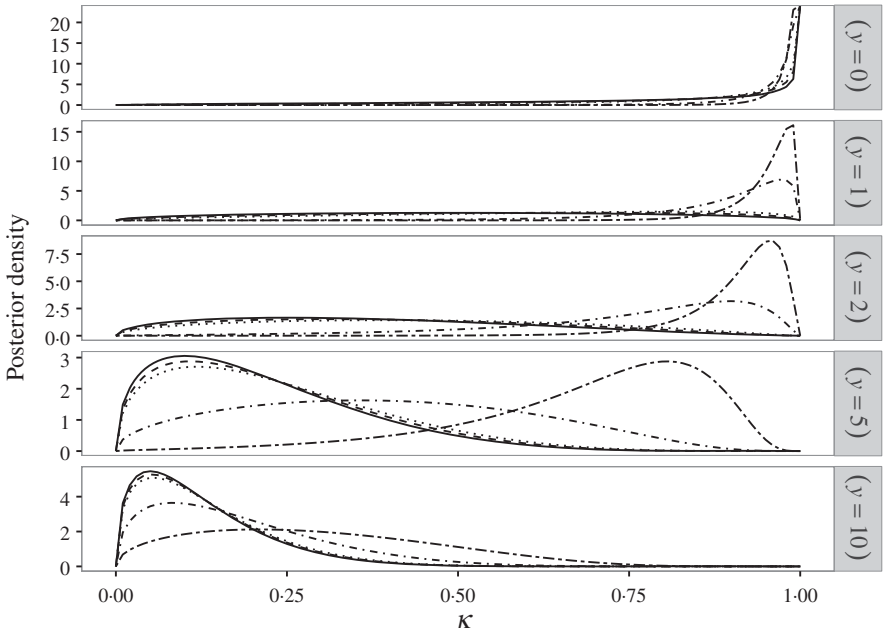

Fig. 2. Posterior distribution of the shrinkage parameter κ_i under the Gauss hypergeometric prior for different values of the hyperparameter γ: $\gamma = 0$ (solid), $\gamma = 0{\cdot}5$ (dashed), $\gamma = 1$ (dotted), $\gamma = 5$ (dot-dash), and $\gamma = 10$ (long-short dashed). Each row corresponds to a different value of $y_i \in \{0, 1, 2, 5, 10\}$.

mode near $\kappa_i = 0$ to have relatively large mass, minimizing shrinkage of large y_i. On the other hand, the parameter γ can be construed as a pseudo-negative observation in (4) that shrinks small counts, acting as a threshold when $\tau \to 0$. The parameter τ is a global shrinkage parameter that controls the rate of concentration of the posterior density of κ_i, as Proposition 2 shows. The thresholding parameter γ and y_i act antagonistically to determine the posterior concentration of κ_i near 0 and 1. In particular, if y_i is larger than γ, the posterior density concentrates near $\kappa_i = 0$, indicating no shrinkage. However, a large value of γ relative to y_i reinforces the shrinkage by τ and the posterior density concentrates near $\kappa_i = 1$, allowing for shrinkage of low counts. We plot the posterior density of κ_i for different combinations of γ and y_i values in Fig. 2. The extra flexibility that comes with γ results in stronger control of the Type I error probability, as shown in Proposition 3 below.

3. Theoretical properties

3·1. *Flexible posterior concentration*

The posterior mean for θ_i under our proposed prior can be written as $(1 - \hat{\kappa}_i)(y_i + \alpha)$, where $\hat{\kappa}_i$ denotes the posterior mean $E(\kappa_i \mid y_i, \alpha, \tau, \gamma)$. Hence, it is natural to expect that the posterior distribution of κ_i puts increasing mass at zero as y_i becomes large relative to the hyperparameter γ. On the other hand, the posterior distribution of κ_i concentrates near 1 for values of y_i that are small relative to γ. Indeed, the plots in Fig. 2 show the concentration of the posterior density at either extreme of the κ_i scale, depending on the magnitude of γ. There is great flexibility in the shrinkage profile through the posterior mean $\hat{\theta}_i = (1 - \hat{\kappa}_i)(y_i + \alpha)$, with differential shrinkage depending on the value of γ. We prove this formally in the two propositions that follow. Proposition 1 states that the posterior probability of an interval around 1 for κ_i approaches 0 as $y_i \to \infty$. Proposition 2

establishes that the posterior distribution of κ_i concentrates near 1 when $y_i < \gamma - 1/2$ and $\tau \to 0$.

PROPOSITION 1. *Suppose that* $y_i \sim \mathrm{Po}(\theta_i)$ *and let* $p(\kappa_i \mid y_i, \alpha, \tau, \gamma)$ *denote the posterior density of* κ_i *given* y_i *and fixed* α, τ, γ *for the Gauss hypergeometric prior* $(\kappa_i \mid \tau, \gamma) \sim \mathrm{GH}(1/2, 1/2, \tau^2, \gamma)$. *Then, as* $\tau \to 0$,

$$\mathrm{pr}(\kappa_i > \eta \mid y_i, \alpha, \tau, \gamma) \leqslant \{\Gamma(\alpha + 1/2)\}^{-1}(1-\eta)^{y+1/2-\gamma}(y + 1/2 - \gamma)^{\alpha-1/2}.$$

For a small τ, $p(\kappa_i \mid y_i, \alpha, \tau, \gamma) \to \delta_{\{0\}}$ *as* $y_i \to \infty$, *where* $\delta_{\{0\}}$ *denotes the point mass at* 0.

PROPOSITION 2. *Suppose that* $y_i \sim \mathrm{Po}(\theta_i)$, *let* $p(\kappa_i \mid y_i, \alpha, \tau, \gamma)$ *denote the posterior density of* κ_i *given* y_i *and fixed* α, τ, γ *for the Gauss hypergeometric prior* $(\kappa_i \mid \tau, \gamma) \sim \mathrm{GH}(1/2, 1/2, \tau^2, \gamma)$, *and let* $d_i = \gamma - 1/2 - y_i > 0$. *Then*

$$\mathrm{pr}(\kappa_i < \eta \mid y_i, \alpha, \tau, \gamma) \leqslant \left(\frac{\tau^2}{1-\eta}\right)^{d_i}, \quad y_i \leqslant \gamma - 1/2.$$

Hence, for a fixed y_i, $p(\kappa_i \mid y_i, \alpha, \tau, \gamma) \to \delta_{\{1\}}$ *as* $\tau \to 0$, *where* $\delta_{\{1\}}$ *denotes the point mass at* 1.

3·2. *Tighter control on false discoveries*

A two-group model provides a natural framework for incorporating quasi-sparsity, with the θ_i being independent and identical draws from a scale mixture of two gamma distributions

$$\theta_i \sim (1-p)\,\mathrm{Ga}(\alpha, \beta) + p\,\mathrm{Ga}(\alpha, \beta + \delta) \quad (0 \leqslant p \leqslant 1). \tag{6}$$

We are interested in testing $H_{0i} : \theta_i \sim \mathrm{Ga}(\alpha, \beta)$ against $H_{1i} : \theta_i \sim \mathrm{Ga}(\alpha, \beta + \delta)$ $(i = 1, \ldots, n)$. We set the shape and scale parameters of the null distribution, α and β, to small values to ensure higher concentration near zero under H_{0i}, and we set δ to a large value relative to β so that the prior density becomes more flat under H_{1i}. The posterior mean of θ_i can be written as

$$E(\theta_i \mid y_i) = \left\{(1-\omega_i)\frac{\beta}{1+\beta} + \omega_i \frac{\beta+\delta}{1+\beta+\delta}\right\}(y_i + \alpha) = \omega_i^*(y_i + \alpha), \tag{7}$$

where ω_i and ω_i^* denote the posterior probability $\mathrm{pr}(H_{1i} \mid y_i)$ and the observation-specific shrinkage weight, respectively. For a fixed $\beta > 0$, if $\delta \to \infty$, ω_i^* converges to $(\omega_i + \beta)(1+\beta)^{-1}$, which is an increasing function of ω_i obtained by redistributing the probability mass in $[\beta(1+\beta)^{-1}, 1]$. A multiple testing procedure can be constructed by applying an appropriate thresholding rule to the shrinkage weights, which could be obtained by clustering the ω_i^* into two classes and using the decision boundary as the threshold. For sparse θ, $\beta \to 0$, $\omega_i^* \to \omega_i$, and the testing rule will reject H_{0i} if $\omega_i^* > 1/2$.

The Gauss hypergeometric prior directly models the shrinkage weight through the hierarchy: $y_i \sim \mathrm{Po}(\theta_i)$, $\theta_i \sim \mathrm{Ga}(\alpha, \kappa_i^{-1} - 1)$ and $\kappa_i \sim \mathrm{GH}(1/2, 1/2, \tau^2, \gamma)$. Since the posterior mean of θ_i is $E(\theta_i \mid y_i, \alpha, \tau, \gamma) = \{1 - E(\kappa_i \mid y_i, \alpha, \tau, \gamma)\}(y_i + \alpha)$, a comparison with (7) suggests that the term $1 - E(\kappa_i \mid y_i, \alpha, \tau, \gamma)$ mimics the shrinkage factor ω_i^* and induces a multiple testing rule that rejects H_{0i} if $1 - E(\kappa_i \mid y_i, \alpha, \tau, \gamma) > \xi$ $(i = 1, \ldots, n)$, where ξ is a suitably chosen threshold, calculated by clustering the shrinkage weights into two groups as described above. The resulting multiple testing rule yields excellent performance in terms of Type I error and misclassification

rates in our simulation studies. We now establish that the probability of Type I error for the multiple testing rule induced by the Gauss hypergeometric prior decreases exponentially with γ.

PROPOSITION 3. *Suppose we have n independent observations* $y_1, \ldots, y_n$ *such that each* $y_i \sim \mathrm{Po}(\theta_i)$ *and the* θ_i *are drawn from the two-group mixture distribution in* (6)*. Under the assumption* $\tau \to 0$ *as* $n \to \infty$*, the Type I error probability for the multiple testing rule induced by the Gauss hypergeometric prior* $(\kappa_i \mid \tau, \gamma) \sim \mathrm{GH}(1/2, 1/2, \tau^2, \gamma)$ *is*

$$t_{1i} = t_1 = \mathrm{pr}_{H_{0i}}\{E(\kappa_i \mid y_i, \alpha, \tau, \gamma) < 1 - \xi\} \leqslant \frac{\left(\frac{\beta}{1+\beta}\right)^{\gamma+1/2}\left(\frac{1}{1+\beta}\right)^{\alpha-1}}{(\gamma + 1/2)B(\gamma + 1/2, \alpha)}.$$

Proposition 3 shows the importance of the additional parameter γ in controlling the Type I error probability when the null distribution of θ_i is positive and nondegenerate. In particular, the Type I error rate for the Gauss hypergeometric prior would be lower than that of the three-parameter beta prior corresponding to the special case of $\gamma = 1$, since it has a spike at zero and a heavy tail but no mechanism for flexible thresholding. The proof is given in the Supplementary Material.

4. SIMULATION STUDIES

4·1. *Quasi-sparse count data*

In this section, we present two simulation studies to compare the performance of different estimators for a quasi-sparse Poisson mean vector. We compare our Gauss hypergeometric estimator with the horseshoe estimator, the Kiefer–Wolfowitz nonparametric maximum likelihood estimator, Robbins's frequency ratio estimator, a Bayesian zero-inflated Poisson estimator, and a global shrinkage Bayes estimator. The horseshoe prior is

$$\theta_i \sim \mathrm{Ga}(\alpha, \lambda_i^2\tau^2), \quad \lambda_i \sim C^+(0,1), \quad \tau \sim C^+(0,1) \quad (\lambda_i, \tau > 0),$$

where $C^+(0,1)$ denotes a standard half-Cauchy distribution. For the Bayesian zero-inflated Poisson model, we use a gamma hyperprior on the Poisson mean and a beta prior on the zero-occurrence probability, where the hyperparameters are estimated from the data. We use average Bayes risk, $\mathrm{ABR}(\theta) = n^{-1}E_\Theta(\|\hat{\theta} - \theta\|^2)$, as the estimation performance criterion.

The global shrinkage estimator is obtained by putting a standard conjugate gamma prior on the Poisson means. The parameters θ_i and the observations are drawn from the model

$$y_i \sim \mathrm{Po}(\theta_i), \quad \theta_i \sim (1-\omega)\delta_{\{0\}} + \omega|t_3| \quad (i = 1, \ldots, n),$$

with $|t_3|$ denoting a folded t-distribution with three degrees of freedom. We generate 1000 different datasets from the above model for each combination of multiplicity $n = 200, 500$ and proportion of nonzero parameters $\omega = 0{\cdot}1, 0{\cdot}15, 0{\cdot}2$. For each of the datasets, we estimate θ and report the mean and standard deviation of the squared error loss for each of the estimators based on these 1000 simulations.

The results are reported in Table 1, with boxplots for $\omega = 0{\cdot}2$ and $n = 200$ or 500 shown in Fig. 3. The Gauss hypergeometric estimator outperforms its competitors in this quasi-sparse simulation set-up across different values of multiplicity and sparsity. The Kiefer–Wolfowitz method also performs well, and is a close runner-up in terms of accuracy, while the Bayesian zero-inflated Poisson model comes in third, with better performance in sparser situations. The

Table 1. *Average Bayes risk* $\mathrm{ABR}(\theta) = n^{-1}E_\Theta(\|\hat{\theta} - \theta\|^2)$ *and corresponding standard deviation (in parentheses) for different Bayes and empirical Bayes procedures, with* $\theta_i \sim (1-\omega)\delta_{\{0\}} + \omega|t_3|$ $(i = 1, \ldots, n)$ *over* 1000 *replicates*

		HS	KW	GH	Robbins	Global	ZIP	Naïve
$n = 200$	$w = 0{\cdot}1$	10·2 (3·7)	11·5 (7·1)	8·2 (6·1)	15·6 (11·4)	29·6 (2·7)	11·2 (3·3)	12·5 (6·0)
	$w = 0{\cdot}15$	14·9 (6·3)	6·3 (2·5)	5·5 (2·8)	14·1 (10·0)	9·4 (0·8)	8·4 (2·2)	12·7 (4·7)
	$w = 0{\cdot}2$	19·6 (7·1)	14·0 (7·5)	11·4 (5·9)	17·8 (12·4)	28·1 (3·3)	17·9 (4·5)	19·9 (7·1)
$n = 500$	$w = 0{\cdot}1$	11·1 (3·4)	8·7 (3·2)	7·4 (3·1)	16·6 (9·5)	20·2 (1·2)	7·4 (1·5)	9·2 (2·4)
	$w = 0{\cdot}15$	18·6 (4·7)	8·0 (1·8)	8·8 (2·5)	18·4 (10·7)	15·2 (0·9)	12·7 (1·9)	16·3 (4·4)
	$w = 0{\cdot}2$	22·8 (5·5)	13·8 (3·3)	13·1 (3·6)	27·4 (14·9)	26·1 (1·9)	22·5 (2·8)	24·8 (4·4)

HS, horseshoe; GH, Gauss hypergeometric; KW, Kiefer–Wolfowitz; ZIP, zero-inflated Poisson.

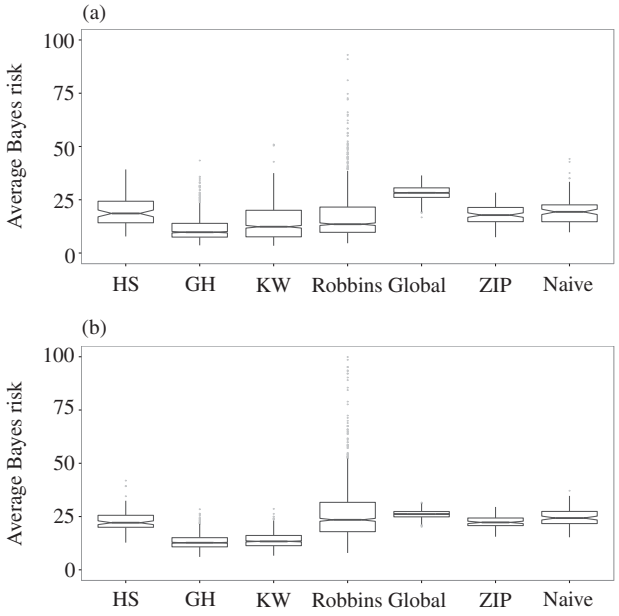

Fig. 3. Boxplots of the average Bayes risk $\mathrm{ABR}(\theta) = n^{-1}E_\Theta(\|\hat{\theta} - \theta\|^2)$ for the competing estimators, namely the horseshoe (HS), Gauss hypergeometric (GH), Kiefer–Wolfowitz (KW), Robbins, global shrinkage, zero-inflated Poisson (ZIP), and naïve estimators, for $\omega = 0{\cdot}2$ and (a) $n = 200$, (b) $n = 500$.

frequency ratio estimator does poorly, and the difference is more prominent for higher multiplicity values. The global shrinkage prior seems to have the lowest accuracy, as it was not designed to handle quasi-sparsity. We also report the naïve risk $n^{-1}\hat{E}(\|y - \theta_0\|^2)$ as a baseline to highlight the poor performance of Robbins's estimator in this situation.

4·2. *Multiple testing*

As argued in § 3·2, thresholding the shrinkage weights $1 - \hat{\kappa}_i$ under the Gauss hypergeometric prior induces a multiple testing rule for the θ_i. We show below that the Gauss hypergeometric decision rule dominates those induced by the Kiefer–Wolfowitz estimator and the three-parameter

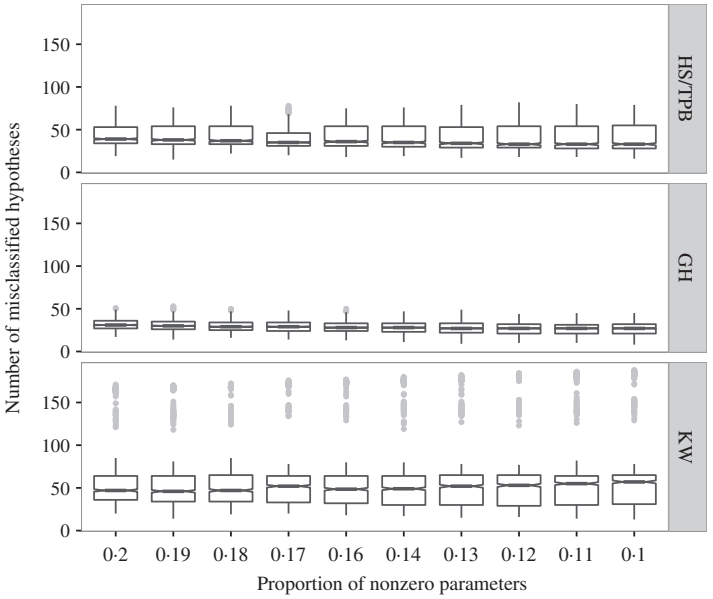

Fig. 4. Number of misclassified hypotheses out of 1000 tests under competing decision rules, namely the horseshoe/three-parameter beta (HS/TPB) prior, Gauss hypergeometric (GH) prior and Kiefer–Wolfowitz (KW) method, for different values of the proportion of nonnull effects.

beta prior (Armagan et al., 2011). For the Kiefer–Wolfowitz estimator, we threshold the shrinkage weights $\omega_i = \hat{P}_G(y_i)^{-1}\hat{P}_G(y_i+1)$, where $\hat{P}_G(\cdot)$ is the estimated probability mass function. We choose $a = b = 1/2$ and $\phi = \tau^2$ as the hyperparameter values for the three-parameter beta prior, which makes the prior density on κ identical to that induced by the horseshoe prior of Carvalho et al. (2010).

We generate $n = 200$ observations from a contaminated zero-inflated model, where y_i is either zero with probability $1-\omega$ or drawn from a Po(4) distribution with probability ω. We contaminated the data by setting a proportion p of the zeros equal to 1. Our goal is to detect the nonnull parameters. The nonnull parameter value $\lambda = 4$ is chosen to be of the order of the maximum order statistics $X_{(n)}$ of n independent and identically distributed Po(1) random variables. The motivation here is similar to that for the hard-thresholding estimate of Donoho & Johnstone (1994) for Gaussian sparse signal detection, where an observation is treated as a signal only if it exceeds $E(X_{(n)}) = 2\log n$, the expected value of the maximum among n pure noises. For Poisson counts, $X_{(n)}$ lies inside the interval $[I_n, I_n+1]$ with probability 1 as $n \to \infty$, where $I_n \sim \log n/(\log\log n)$ (Kimber, 1983), and the midpoint of the interval is $I_{200} + 0{\cdot}5 = 3{\cdot}67$. Following § 3·2, the decision rule induced by the shrinkage priors is to reject the ith null hypothesis if $1 - E(\kappa_i \mid y_i, \alpha, \tau, \gamma) > \xi$ for some fixed threshold ξ. We calculated this threshold by applying the k-means clustering algorithm to the shrinkage weights with number of clusters $k = 2$ and setting ξ to be the mean of the cluster centres. To compare the shrinkage priors, we calculate the number of misclassified hypotheses over 1000 simulated datasets for 10 equidistant values of the proportion of sparsity $\omega \in [0{\cdot}1, 0{\cdot}3]$. We fixed the value of the contamination proportion p at 0·1. The boxplots of the number of misclassification errors, shown in Fig. 4, and the mean number of misclassified hypotheses, reported in Table 2, suggest that the Gauss hypergeometric decision rule outperforms the methods induced by the three-parameter beta prior and by the Kiefer–Wolfowitz method for these quasi-sparse examples. Further simulation studies in the Supplementary Material show that the Gauss hypergeometric prior is superior to its competitors

Table 2. *Mean misclassification errors for the three-parameter beta prior, the Gauss hypergeometric prior and the Kiefer–Wolfowitz nonparametric maximum likelihood estimator*

Sparsity	0·3	0·27	0·25	0·23	0·21	0·18	0·16	0·14	0·12	0·1
HS/TPB	7·1	7·1	7·0	6·5	6·5	6·3	6·2	6·3	5·8	5·6
GH	5·9	5·4	4·7	4·4	4·0	3·5	3·0	2·7	2·1	1·7
KW	13·7	12·7	11·7	10·5	9·0	7·6	6·1	4·8	3·4	2·8

HS, horseshoe; TPB, three-parameter beta; GH, Gauss hypergeometric; KW, Kiefer–Wolfowitz.

when the θ_i are drawn from a mixture of gamma distributions with the null hypothesis favouring higher concentration near zero.

5. DETECTING RARE MUTATIONAL HOTSPOTS

In this section, we apply our methods to count data arising from a massive sequencing study called the Exome Aggregation Consortium. This database reports the total number of mutated alleles or variants along the whole exome for 60 076 individuals, and provides information about genetic variation in the human population. It is widely recognized in the scientific community that these rare changes are responsible for both common and rare diseases (Pritchard, 2001). The frequency of mutated alleles is very low or zero at a vast majority of locations across the genome but is substantially higher in certain functionally relevant genomic locations, such as promoters and insulators (Lewin et al., 2014). An important problem in the study of rare mutations is identification of such mutational hotspots where the mutation rate significantly exceeds the background rate. This is important since these mutational hotspots might be enriched with disease-causing rare variants (Ionita-Laza et al., 2012).

Our goal is to use the method developed in this article to identify potential hotspots harbouring rare variants in a genomic region. We first filter out the common variants with minor allele frequency greater than 0·05% on a gene, PIK3CA, known to be responsible for ovarian and cervical cancers (Shayesteh et al., 1999). The mutation dataset contains the number of mutated alleles along the gene PIK3CA for 240 amino acid positions ranging from 0 to 1066. The flexible shrinkage property of the Gauss hypergeometric prior should yield better detection of the true hotspots by better shrinking low counts. The number of mutated alleles at the ith position, denoted by y_i, ranges from 0 to 58. We model $y_i \sim \text{Po}(N_i\theta_i)$ independently, where θ_i is the mutation rate and N_i is the number of alleles sequenced at location i. We make the simplifying assumption of uniform sequencing depth across the gene such that $N_i = N$ for all i, but in general the sequencing depth is dependent on location. Since each individual carries two copies of each allele that could harbour a rare variant, N_i is also equal to twice the number of individuals sequenced at that position.

We compare the shrinkage profile of the Gauss hypergeometric prior with those of the three-parameter beta/horseshoe prior and the Kiefer–Wolfowitz estimator in terms of the number of mutational hotspots identified. We apply the multiple testing rule proposed in § 4·2 to identify the variants for which the number of mutated alleles is substantially higher than the background. The number of variants identified as nonnull using the Gauss hypergeometric prior, three-parameter beta/horseshoe prior and Kiefer–Wolfowitz method are 7, 81 and 56, respectively.

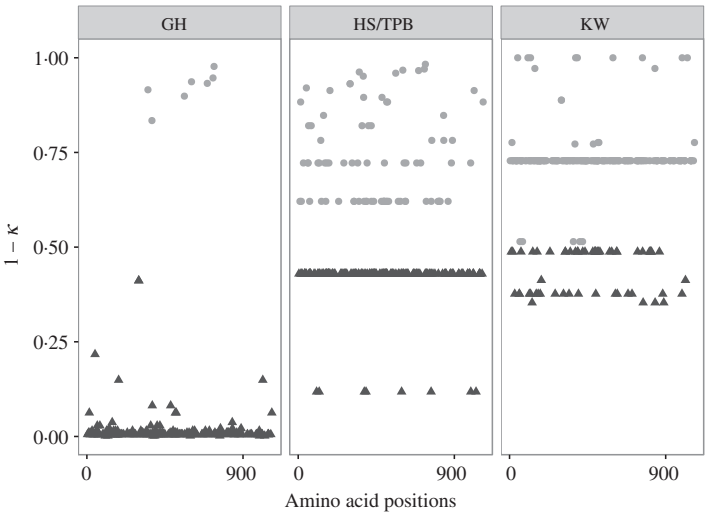

Fig. 5. Comparison of the shrinkage profiles for the Gauss hypergeometric (GH) prior, three-parameter beta/horseshoe (TPB/HS) prior, and Kiefer–Wolfowitz (KW) nonparametric maximum likelihood estimator via scatterplots of the corresponding shrinkage weights.

Figure 5 shows the posterior probabilities ω_i of the competing methods; the Gauss hypergeometric prior has a sharpened ability to segregate the substantially higher signals from the background noise.

6. Identifying north american cities with the most terrorist attacks

We consider an application to a database containing details of all terrorist attacks in the world since 1970, including the location and type of each attack. The global terrorism database defines a terrorist attack as 'the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation'. We focus on terrorist attacks in North America, aggregated over the years of observation at the level of cities, and our goal is to identify the cities that have been worst hit. As expected, there are many cities with zero or small counts, and a few cities with large counts, such as New York City, Mexico City and Miami. We apply our method to identify the cities with high attack rates. Figure 6 shows the observed counts of terrorist attacks along with the posterior mean estimates of the rates for North American cities obtained using three different methods; we show only nonzero observations. As expected, the Gauss hypergeometric method selects the fewest cities, while the Kiefer–Wolfowitz method selects every city that experienced at least one attack. This illustrates the robustness of the Gauss hypergeometric estimator with respect to very small counts, which may correspond to either very sporadic events or errors in recording; for example, an attack may be mislabelled as terrorism-related. Methods that naïvely select all locations with nonzero counts are not very useful in practice.

Acknowledgement

The authors thank Dr Sandeep Dave for invaluable help in interpreting the exome sequencing data. This research was funded by the U.S. National Science Foundation through the Statistical

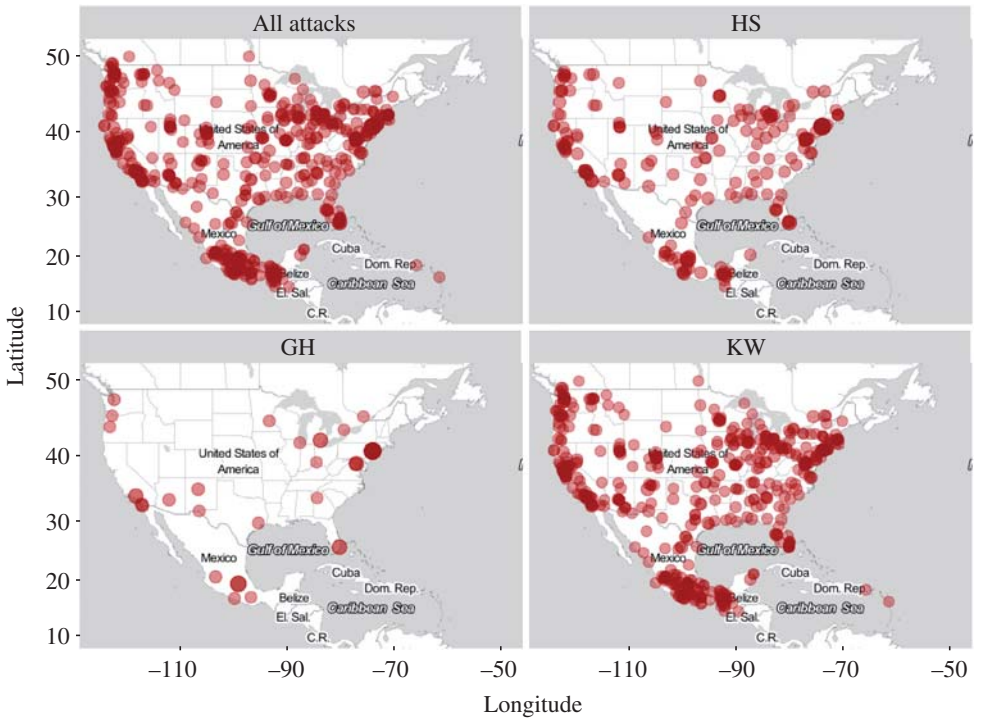

Fig. 6. The total number of terror attacks in North American cities since 1970, together with the posterior mean estimates of the rate of terror attacks under the horseshoe (HS) prior, the Gauss hypergeometric (GH) prior, and the Kiefer–Wolfowitz (KW) estimator. The size and opacity of the points on the maps increase with the value of the observation.

and Applied Mathematical Sciences Institute and by the U.S. National Institutes of Health. We also thank the reviewers for comments that have led to substantial improvements to the article.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes proofs of the theoretical results, a further simulation study, more details about the data analysed in § 5, and discussion of the performance of the Gauss hypergeometric prior for zero-inflated data without covariates.

REFERENCES

ARMAGAN, A., CLYDE, M. & DUNSON, D. B. (2011). Generalized beta mixtures of Gaussians. *Adv. Neural. Info. Proces. Syst.* **24**, 523–31.

ARMAGAN, A., DUNSON, D. B. & LEE, J. (2013). Generalized double Pareto shrinkage. *Statist. Sinica* **23**, 119–43.

ARMERO, C. & BAYARRI, M. J. (1994). Prior assessments for prediction in queues. *Statistician* **43**, 139–53.

BHATTACHARYA, A., PATI, D., PILLAI, N. S. & DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *J. Am. Statist. Assoc.* **110**, 1479–90.

BROWN, L. D., GREENSHTEIN, E. & RITOV, Y. (2013). The Poisson compound decision problem revisited. *J. Am. Statist. Assoc.* **108**, 741–9.

CARVALHO, C. M., POLSON, N. G. & SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–80.

DONOHO, D. L. & JOHNSTONE, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–55.

Ionita-Laza, I., Makarov, V., Buxbaum, J. D. & ARRA Autism Sequencing Consortium (2012). Scan-statistic approach identifies clusters of rare disease variants in LRP2, a gene linked and associated with autism spectrum disorders, in three datasets. *Am. J. Hum. Genet.* **90**, 1002–13.

Kiefer, J. & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27**, 887–906.

Kimber, A. C. (1983). A note on Poisson maxima. *Z. Wahr. verw. Geb.* **63**, 551–2.

Koenker, R. & Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Am. Statist. Assoc.* **109**, 674–85.

Lewin, B., Krebs, J., Kilpatrick, S. T. & Goldstein, E. S. (2014). *Lewin's Genes XI*, vol. 11. Burlington, Massachusetts: Jones & Bartlett Learning.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–37.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proc. 3rd Berkeley Symp. Math. Statist. Prob., 1954–1955*, J. Neyman, ed., vol. I. Berkeley: University of California Press.

Shayesteh, L., Lu, Y., Kuo, W.-L., Baldocchi, R., Godfrey, T., Collins, C., Pinkel, D., Powell, B., Mills, G. B. & Gray, J. W. (1999). PIK3CA is implicated as an oncogene in ovarian cancer. *Nature Genet.* **21**, 99–102.

van der Pas, S., Kleijn, B. & van der Vaart, A. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Statist.* **8**, 2585–618.

Yang, Z., Hardin, J. W. & Addy, C. L. (2009). Testing overdispersion in the zero-inflated Poisson model. *J. Statist. Plan. Infer.* **139**, 3340–53.

[*Received June* 2015. *Revised October* 2016]