ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 5, *pp.* 1273–1300

A simple new approach to variable selection in regression, with application to genetic fine mapping

Gao Wang, Abhishek Sarkar, Peter Carbonetto and Matthew Stephens

University of Chicago, USA

[Received December 2018. Final revision May 2020]

Summary. We introduce a simple new approach to variable selection in linear regression, with a particular focus on *quantifying uncertainty in which variables should be selected*. The approach is based on a new model—the 'sum of single effects' model, called '*SuSiE*'—which comes from writing the sparse vector of regression coefficients as a sum of 'single-effect' vectors, each with one non-zero element. We also introduce a corresponding new fitting procedure—iterative Bayesian stepwise selection (IBSS)—which is a Bayesian analogue of stepwise selection methods. IBSS shares the computational simplicity and speed of traditional stepwise methods but, instead of selecting a single variable at each step, IBSS computes a *distribution* on variables that captures uncertainty in which variable to select. We provide a formal justification of this intuitive algorithm by showing that it optimizes a variational approximation to the posterior distribution under SuSiE. Further, this approximate posterior distribution naturally yields convenient novel summaries of uncertainty in variable selection, providing a credible set of variables for each selection. Our methods are particularly well suited to settings where variables are highly correlated and detectable effects are sparse, both of which are characteristics of genetic fine mapping applications. We demonstrate through numerical experiments that our methods outperform existing methods for this task, and we illustrate their application to fine mapping genetic variants influencing alternative splicing in human cell lines. We also discuss the potential and challenges for applying these methods to generic variable-selection problems.

Keywords: Genetic fine mapping; Linear regression; Sparsity; Variable selection; Variational inference

1. Introduction

The need to identify, or 'select', relevant variables in regression models arises in a diverse range of applications and has spurred development of a correspondingly diverse range of methods (for example, see O'Hara and Sillanpää (2009), Fan and Lv (2010), Desboulets (2018) and George and McCulloch (1997) for reviews). However, variable selection is a complex problem, and so despite considerable work in this area there remain important issues that existing methods do not fully address. One such issue is *assessing uncertainty in which variables should be selected*, particularly in settings involving *very highly correlated variables*. Here we introduce a simple and computationally scalable approach to variable selection that helps to address this issue.

Highly correlated variables pose an obvious challenge to variable-selection methods, simply because they are difficult to distinguish from one another. Indeed, in an extreme case where two variables (say, x_1 and x_2) are completely correlated, it is impossible to claim, on the basis of a regression analysis, that one variable should be selected as relevant rather than the other. In some applications such ambiguity causes few practical problems. Specifically, in some applications

Address for correspondence: Matthew Stephens, Departments of Statistics and Human Genetics, University of Chicago, Chicago, IL 60637, USA.
E-mail: mstephens@uchicago.edu

variable selection is used only to help *to build an accurate predictor*, in which case it suffices to select one of the two identical variables arbitrarily (or both); prediction accuracy is unaffected by this choice. However, in other scientific applications, variable selection is used as a means to help *to learn something about the world*, and in those applications the ambiguity that is created by highly correlated variables is more problematic because scientific conclusions depend on which variables are selected. In these applications, it is crucial to acknowledge uncertainty in which variables should be selected. This requires methods that can draw conclusions such as 'either x_1 or x_2 is relevant and we cannot decide which' rather than methods that arbitrarily select one of the variables and ignore the other. Although this may seem a simple goal, in practice most existing variable-selection methods do not satisfactorily address this problem (see Section 2 for further discussion). These shortcomings motivate our work here.

One particular application where these issues arise is genetic fine mapping (e.g. Veyrieras *et al.* (2008), Maller *et al.* (2012), Spain and Barrett (2015), Huang *et al.* (2017) and Schaid *et al.* (2018)). The goal of fine mapping is to identify the genetic variants that causally affect some traits of interest (e.g. low density lipoprotein cholesterol in blood and gene expression in cells). In other words, the main goal of fine mapping is to learn something about the world, rather than to build a better predictor. (This is not to say that predicting traits from genetic variants is not important; indeed, there is also a large amount of work on prediction of genetic traits, but this is not the main goal of fine mapping.) The most successful current approaches to fine mapping frame the problem as a *variable-selection problem*, building a regression model in which the regression outcome is the trait of interest, and the candidate predictor variables are the available genetic variants (Sillanpää and Bhattacharjee, 2005). Performing variable selection in a regression model identifies variants that may causally affect the trait. Fine mapping is challenging because the variables (genetic variants) can be *very* highly correlated, owing to a phenomenon called *linkage disequilibrium* (Ott, 1999). Indeed, typical studies contain many pairs of genetic variants with sample correlations exceeding 0.99, or even equalling 1.

Our approach builds on previous work on Bayesian variable selection in regression (BVSR) (Mitchell and Beauchamp, 1988; George and McCulloch, 1997), which has already been widely applied to genetic fine mapping and related applications (e.g. Meuwissen *et al.* (2001), Sillanpää and Bhattacharjee (2005), Servin and Stephens (2007), Hoggart *et al.* (2008), Stephens and Balding (2009), Logsdon *et al.* (2010), Guan and Stephens (2011), Bottolo *et al.* (2011), Maller *et al.* (2012), Carbonetto and Stephens (2012), Zhou *et al.* (2013), Hormozdiari *et al.* (2014), Chen *et al.* (2015), Wallace *et al.* (2015), Moser *et al.* (2015), Wen *et al.* (2016) and Lee *et al.* (2018)). BVSR is an attractive approach to these problems because it can, in principle, assess uncertainty in which variables to select, even when the variables are highly correlated. However, applying BVSR in practice remains difficult for at least two reasons. First, BVSR is computationally challenging, often requiring implementation of sophisticated Markov chain Monte Carlo or stochastic search algorithms (e.g. Bottolo and Richardson (2010), Bottolo *et al.* (2011), Guan and Stephens (2011), Wallace *et al.* (2015), Benner *et al.* (2016), Wen *et al.* (2016) and Lee *et al.* (2018)). Second, and perhaps more importantly, the output from BVSR methods is typically a complex posterior distribution—or samples approximating the posterior distribution—and this can be difficult to distil into results that are easily interpretable.

Our work addresses these shortcomings of BVSR through several innovations. We introduce a new formulation of BVSR, which we call the 'sum of single effects' model *SuSiE*. This model, although similar to existing BVSR models, has a different structure that naturally leads to a simple, intuitive and fast procedure for model fitting—iterative Bayesian stepwise selection (IBSS)—which is a Bayesian analogue of traditional stepwise selection methods (and which enjoys important advantages over these traditional selection methods, as we explain below).

We provide a principled justification for this intuitive algorithm by showing that it optimizes a variational approximation to the posterior distribution under SuSiE. Although variational approaches to BVSR already exist (Logsdon *et al.*, 2010; Carbonetto and Stephens, 2012), our new approach introduces a different family of approximating distributions that provides much more accurate inferences in settings with highly correlated variables.

A key feature of our method, which distinguishes it from most existing BVSR methods, is that it produces 'credible sets' of variables that quantify uncertainty in which variable should be selected when multiple, highly correlated variables compete with one another. These credible sets are designed to be as small as possible while still each capturing a relevant variable. Arguably, this is exactly the kind of posterior summary that we would like to obtain from Markov chain Monte Carlo based or stochastic search BVSR methods, but doing so would require non-trivial post-processing of their output. In contrast, our method provides this posterior summary directly, and with little extra computational effort.

The structure of this paper is as follows. Section 2 provides further motivation for our work, and a brief background on BVSR. Section 3 describes the new SuSiE model and fitting procedure. Section 4 uses simulations, designed to mimic realistic genetic fine mapping studies, to demonstrate the effectiveness of our approach compared with existing methods. Section 5 illustrates the application of our methods to fine mapping of genetic variants affecting splicing, and Section 6 briefly highlights the promise (and limitations) of our methods for other applications such as change point problems. We end with a discussion highlighting avenues for further work.

2. Background

2.1. A motivating toy example

Suppose that the relationship between an n-vector $\mathbf{y}$ and an $n \times p$ matrix $\mathbf{X} = (x_1, \ldots, x_p)$ is modelled as a multiple regression:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \qquad \mathbf{e} \sim N_n(0, \sigma^2 I_n), \tag{2.1}$$

where $\mathbf{b}$ is a p-vector of regression coefficients, $\mathbf{e}$ is an n-vector of error terms, $\sigma^2 > 0$ is the residual variance, I_n is the $n \times n$ identity matrix and $N_r(\mu, \Sigma)$ denotes the r-variate normal distribution with mean μ and variance Σ. For brevity, we shall refer to variables j with non-zero effects ($b_j \neq 0$) as 'effect variables'.

Assume now that exactly two variables are effect variables—variables 1 and 4, say—and that these two effect variables are each completely correlated with another non-effect variable, say $x_1 = x_2$ and $x_3 = x_4$. Further suppose that no other pairs of variables are correlated. Here, because the effect variables are completely correlated with other variables, it is impossible to select the correct variables confidently, even when n is very large. However, given sufficient data it should be possible to conclude that there are (at least) two effect variables, and that

$$(b_1 \neq 0 \text{ or } b_2 \neq 0) \text{ and } (b_3 \neq 0 \text{ or } b_4 \neq 0). \tag{2.2}$$

Our goal, in short, is to *provide methods that directly produce this kind of inferential statement.* Although this example is simplistic, it mimics the kind of structure that occurs in, for example, genetic fine mapping applications, where it often happens that an association can be narrowed down to a small set of highly correlated genetic variants, but not down to an individual variant.

Most existing approaches to sparse regression do not provide statements like expression (2.2); nor do they attempt to do so. For example, methods that maximize a penalized likelihood, such as the lasso (Tibshirani, 1996) or elastic net (EN) (Zou and Hastie, 2005), select a single 'best'

combination of variables and make no attempt to assess whether other combinations are also plausible. In our toy example, the EN selects all four variables (1–4), implying that $b_1 \neq 0$, $b_2 \neq 0$, $b_3 \neq 0$ and $b_4 \neq 0$, which is quite different from expression (2.2). Recently developed selective inference approaches (Taylor and Tibshirani, 2015) do not solve this problem, because they do not assess uncertainty in *which* variables should be selected; instead they assess uncertainty in the coefficients of the selected variables within the selected model. In our toy motivating example, selective inference methods sometimes select the wrong variables (inevitably, because of the complete correlation with other variables) and then assign them highly significant p-values (see Wang *et al.* (2020a) for an explicit example accompanied by code). The p-values are significant because, even though the wrong variables are selected, their coefficients—within the (wrong) selected model—can be estimated precisely. An alternative approach, which does address uncertainty in variable selection, is to control the false discovery rate (FDR) among selected variables—e.g. by using stability selection (Meinshausen and Bühlmann, 2010) or the knockoff filter (Barber and Candès, 2015). However, in examples with very highly correlated variables no individual variable can be confidently declared an effect variable, and so controlling the FDR among selected variables results in no discoveries, and not inferences like expression (2.2).

One approach to producing inferences like expression (2.2) is to reframe the problem, and to focus on selecting *groups* of variables, rather than individual variables. A simple version of this idea might first cluster the variables into groups of highly correlated variables, and then to perform some kind of 'group selection' (Huang *et al.*, 2012). However, whereas this could work in our toy example, in general this approach requires *ad hoc* decisions about which variables to group, and how many groups to create—an unattractive feature that we seek to avoid. A more sophisticated version of this idea is to use hierarchical testing (Meinshausen, 2008; Yekutieli, 2008; Mandozzi and Bühlmann, 2016; Renaux *et al.*, 2018), which requires specification of a hierarchy on the variables, but avoids an *a priori* decision on where to draw group boundaries. However, in applications where variables are not precisely arranged in a known hierarchy—which includes genetic fine mapping—this approach is also not entirely satisfactory. In numerical assessments that are shown later (Section 4), we find that this approach can considerably overstate the uncertainty in which variables should be selected.

Another approach that could yield statements like expression (2.2), at least in principle, is the Bayesian approach to variable selection (BVSR; see Section 1 for references). BVSR methods introduce a prior distribution on $\mathbf{b}$ that favours sparse models (few effect variables) and then compute a posterior distribution assessing relative support for each combination of variables. In our toy example, the posterior distribution would roughly have equal mass (approximately 0.25) on each of the four equivalent combinations $\{1,3\}$, $\{1,4\}$, $\{2,3\}$ and $\{2,4\}$. This posterior distribution contains exactly the information that is necessary to infer statement (2.2). Likewise, in more complex settings, the posterior distribution contains information that could, in principle, be translated to simple statements analogous to expression (2.2). This translation is, however, highly non-trivial in general. Consequently, most implementations of BVSR do not provide statements like expression (2.2), but rather summarize the posterior distribution with a simpler but less informative quantity: the marginal posterior inclusion probability (PIP) of each variable:

$$\mathrm{PIP}_j := \Pr(b_j \neq 0 | \mathbf{X}, \mathbf{y}). \tag{2.3}$$

In our example, $\mathrm{PIP}_1 = \mathrm{PIP}_2 = \mathrm{PIP}_3 = \mathrm{PIP}_4 \approx 0.5$. Although not inaccurate, the PIPs do not contain the information in expression (2.2). In Wang *et al.* (2020a), we illustrate inference of credible sets in two additional toy examples in which the variables are correlated in more complicated ways.

2.2. Credible sets

To define our main goal more formally, we introduce the concept of a *credible set* of variables.

Definition 1. In the context of a multiple-regression model, *a level ρ credible set* is defined to be a subset of variables that has probability ρ or greater of containing at least one effect variable (i.e. a variable with non-zero regression coefficient). Equivalently, the probability that all variables in the credible set have zero regression coefficients is $1 - \rho$ or less.

Our use of the term credible set here indicates that we have in mind a Bayesian inference approach, in which the probability statements in the definition are statements about uncertainty in which variables are selected given the available data and modelling assumptions. One could analogously define a *confidence set* by interpreting the probability statements as referring to the set, considered random.

Although the term credible set has been used in fine mapping applications before, most previous uses either assumed that there was a single-effect variable (Maller *et al.*, 2012) or defined a credible set as a set that contains *all* effect variables (Hormozdiari *et al.*, 2014), which is a very different definition (and, we argue, both less informative and less attainable; see further discussion below). Our definition here is closer to the 'signal clusters' from Lee *et al.* (2018) and is related to the idea of 'minimal true detection' in Mandozzi and Bühlmann (2016).

With definition 1 in place, our primary aim can be restated: we wish to report as many credible sets as the data support, each with as few variables as possible. For example, to convey statement (2.2) we would report two credible sets: $\{1, 2\}$ and $\{3, 4\}$. As a secondary goal, we would also like to prioritize the variables within each credible set, assigning each a probability that reflects the strength of the evidence for that variable being an effect variable. Our methods achieve both of these goals.

It is important to note that, if a variable is *not* included in any credible set produced by our method, this does not imply that it is *not* an effect variable. This is analogous to the fact that, in hypothesis testing applications, a non-significant p-value does not imply that the null hypothesis is true. In practice no variable-selection method can guarantee identifying *every* effect variable unless it simply selects all variables, because finite data cannot rule out that every variable has a (possibly tiny) effect. This is why the credible set definition of Hormozdiari *et al.* (2014) is unattainable, at least without strong assumptions on sparsity. It also explains why attempting to form confidence or credible sets for identifying the *true model* (i.e. the exact combination of effect variables) leads to very large sets of models; see Ferrari and Yang (2015) for example.

2.3. The single-effect regression model

We now describe the building block for our approach, the 'single-effect regression' (SER) model, which we define as a multiple-regression model in which *exactly one of the p explanatory variables has a non-zero regression coefficient.* This idea was introduced in Servin and Stephens (2007) to fine-map genetic associations, and consequently it has been adopted and extended by others, including Veyrieras *et al.* (2008) and Pickrell (2014). Although of very narrow applicability, the SER model is trivial to fit. Furthermore, when its assumptions hold, SER provides exactly the inferences that we desire, including credible sets. For example, if we simplify our motivating example (Section 2.1) to have a single-effect variable—variable 1, for example—then the SER model would, given sufficient data, infer a 95% credible set containing both of the correlated variables, 1 and 2, with PIPs of approximately 0.5 each. This credible set tells us that we can be confident that one of the two variables has a non-zero coefficient, but we do not know which.

Specifically, we consider the following SER model, with hyperparameters for the residual variance, σ^2, the prior variance of the non-zero effect, σ_0^2, and the prior inclusion probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_p)$, in which π_j gives the prior probability that variable j is the effect variable:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{2.4}$$

$$\mathbf{e} \sim N_n(0, \sigma^2 I_n), \tag{2.5}$$

$$\mathbf{b} = b\boldsymbol{\gamma}, \tag{2.6}$$

$$\boldsymbol{\gamma} \sim \text{Mult}(1, \boldsymbol{\pi}), \tag{2.7}$$

$$b \sim N_1(0, \sigma_0^2). \tag{2.8}$$

Here, $\mathbf{y}$ is the n-vector of response data, $\mathbf{X} = (x_1, \ldots, x_p)$ is an $n \times p$ matrix containing n observations of p explanatory variables, b is a scalar representing the 'single effect', $\boldsymbol{\gamma} \in \{0, 1\}^p$ is a p-vector of indicator variables, $\mathbf{b}$ is the p-vector of regression coefficients, $\mathbf{e}$ is an n-vector of independent error terms and $\text{Mult}(m, \boldsymbol{\pi})$ denotes the multinomial distribution on class counts that is obtained when m samples are drawn with class probabilities given by $\boldsymbol{\pi}$. We assume that $\mathbf{y}$ and the columns of $\mathbf{X}$ have been centred to have mean 0, which avoids the need for an intercept term (Chipman *et al.*, 2001).

Under the SER model (2.4)–(2.8), the effect vector $\mathbf{b}$ has exactly one non-zero element (equal to b), so we refer to $\mathbf{b}$ as a 'single-effect vector'. The element of $\mathbf{b}$ that is non-zero is determined by the binary vector $\boldsymbol{\gamma}$, which also has exactly one non-zero entry. The probability vector $\boldsymbol{\pi}$ determines the prior probability distribution on which of the p variables is the effect variable. In the simplest case, $\boldsymbol{\pi} = (1/p, \ldots, 1/p)$; we assume this uniform prior here for simplicity, but our methods require only that $\boldsymbol{\pi}$ is fixed and known (so in fine mapping one could incorporate different priors based on genetic annotations; e.g. Veyrieras *et al.* (2008)). To lighten the notation, we henceforth make conditioning on $\boldsymbol{\pi}$ implicit.

2.3.1. *Posterior under single-effect regression model*

Given σ^2 and σ_0^2, the posterior distribution on $\mathbf{b} = \boldsymbol{\gamma} b$ is easily computed:

$$\boldsymbol{\gamma} | \mathbf{X}, \mathbf{y}, \sigma^2, \sigma_0^2 \sim \text{Mult}(1, \boldsymbol{\alpha}), \tag{2.9}$$

$$b | \mathbf{X}, \mathbf{y}, \sigma^2, \sigma_0^2, \gamma_j = 1 \sim N(\mu_{1j}, \sigma_{1j}^2), \tag{2.10}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)$ is the vector of PIPs, with $\alpha_j := \Pr(\gamma_j = 1 | \mathbf{X}, \mathbf{y}, \sigma^2, \sigma_0^2)$, and μ_{1j} and σ_{1j}^2 are the posterior mean and variance of b given $\gamma_j = 1$. Calculating these quantities simply involves performing the p univariate regressions of $\mathbf{y}$ on columns x_j of $\mathbf{X}$, for $j = 1, \ldots, p$, as shown in the on-line appendix A. From $\boldsymbol{\alpha}$, it is also straightforward to compute a level ρ credible set (definition 1), $\text{CS}(\boldsymbol{\alpha}; \rho)$, as described in Maller *et al.* (2012), and detailed in appendix A. In brief, this involves sorting variables by decreasing α_j and then including variables in the credible set until their cumulative probability exceeds ρ.

For later convenience, we introduce a function, SER, that returns the posterior distribution for $\mathbf{b}$ under the SER model. Since this posterior distribution is uniquely determined by the values of $\boldsymbol{\alpha}$, $\boldsymbol{\mu}_1 := (\mu_{11}, \ldots, \mu_{1p})$ and $\boldsymbol{\sigma}_1^2 := (\sigma_{11}^2, \ldots, \sigma_{1p}^2)$ in distributions (2.9)–(2.10), we can write

$$\text{SER}(\mathbf{X}, \mathbf{y}; \sigma^2, \sigma_0^2) := (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2). \tag{2.11}$$

2.3.2. Empirical Bayes approach for single-effect regression model

Although most previous treatments of the SER model assume that σ_0^2 and σ^2 are fixed and known, we note here the possibility of estimating σ_0^2 and/or σ^2 by maximum likelihood before computing the posterior distribution of $\mathbf{b}$. This is effectively an empirical Bayes approach. The log-likelihood for σ_0^2 and σ^2 under the SER,

$$l_{\mathrm{SER}}(\mathbf{y}; \sigma_0^2, \sigma^2) := \log\{p(\mathbf{y}|\mathbf{X}, \sigma_0^2, \sigma^2)\}, \tag{2.12}$$

is available in closed form, and can be maximized over one or both parameters (on-line appendix A).

3. The sum of single-effects regression model

We now introduce a new approach to variable selection in multiple regression. Our approach is motivated by the observation that the SER model provides simple inference if there is indeed exactly one effect variable; it is thus desirable to extend SER to allow for multiple variables. The conventional approach to doing this in BVSR is to introduce a prior on $\mathbf{b}$ that allows for multiple non-zero entries (e.g. using a 'spike-and-slab' prior; Mitchell and Beauchamp (1988). However, this approach no longer enjoys the convenient analytic properties of the SER model; posterior distributions become difficult to compute accurately, and computing credible sets is even more difficult.

Here we introduce a different approach which better preserves the desirable features of the SER model. The key idea is simple: introduce multiple single-effect vectors $\mathbf{b}_1, \ldots, \mathbf{b}_L$ and construct the overall effect vector $\mathbf{b}$ as the sum of these single effects. We call this the 'sum of single effects' regression model SuSiE:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \tag{3.1}$$

$$\mathbf{e} \sim N_n(0, \sigma^2 I_n), \tag{3.2}$$

$$\mathbf{b} = \sum_{l=1}^{L} \mathbf{b}_l, \tag{3.3}$$

$$\mathbf{b}_l = \gamma_l b_l, \tag{3.4}$$

$$\gamma_l \sim \mathrm{Mult}(1, \boldsymbol{\pi}), \tag{3.5}$$

$$b_l \sim N_1(0, \sigma_{0l}^2). \tag{3.6}$$

For generality, we have allowed the variance of each effect, σ_{0l}^2, to vary among the components, $l = 1, \ldots, L$. The special case in which $L = 1$ recovers the SER model. For simplicity, we initially assume that σ^2 and $\boldsymbol{\sigma}_0^2 = (\sigma_{01}^2, \ldots, \sigma_{0L}^2)$ are given, and defer estimation of these hyperparameters to Section 3.1.3.

If $L \ll p$ then SuSiE is approximately equivalent to a standard BVSR model in which L randomly chosen variables have non-zero coefficients (see proposition A2 in the on-line appendix C for a formal statement). The main difference is that with some (small) probability some of the single effects $\mathbf{b}_l$ in SuSiE have the same non-zero co-ordinates, and so the number of non-zero elements in $\mathbf{b}$ has some (small) probability of being less than L. Thus, at most L variables have non-zero coefficients in this model. We discuss the choice of L in Section 3.3.

Although SuSiE is approximately equivalent to a standard BVSR model, its novel structure has two major advantages. First, it leads to a simple, iterative and deterministic algorithm for computing approximate posterior distributions. Second, it yields a simple way to calculate the

credible sets. In essence, because each $\mathbf{b}_l$ captures only one effect, the posterior distribution on each γ_l can be used to compute a credible set that has a high probability of containing an effect variable. The remainder of this section describes both these advantages, and other issues that may arise in fitting the model.

3.1. Fitting SuSiE: iterative Bayesian stepwise selection

A key motivation for the SuSiE model (3.1)–(3.6) is that, given $\mathbf{b}_1, \ldots, \mathbf{b}_{L-1}$, estimating $\mathbf{b}_L$ involves simply fitting an SER model, which is analytically tractable. This immediately suggests an iterative approach to fitting this model: at each iteration use the SER model to estimate $\mathbf{b}_l$ given current estimates of $\mathbf{b}_{l'}$, for $l' \neq l$; see algorithm 1 in Table 1. This algorithm is simple and computationally scalable, with computational complexity $O(npL)$ per outer loop iteration.

We call algorithm 1 IBSS because it can be viewed as a Bayesian version of stepwise selection approaches. For example, we can compare it with an approach that was referred to as 'forward stagewise' (FS) selection in Hastie *et al.* (2009), section 3.3.3 (although subsequent literature often uses this term to mean something slightly different), also known as 'matching pursuit' (Mallat and Zhang, 1993). In brief, FS selection first selects the single 'best' variable among p candidates by comparing the results of the p univariate regressions. It then computes the residuals from the univariate regression on this selected variable, and then selects the next best variable by comparing the results of univariate regression of the residuals on each variable. This process repeats, selecting one variable each iteration, until some stopping criterion has been reached.

IBSS is similar in structure to FS selection, but, instead of selecting a single best variable at each step, it computes a *distribution* on which variable to select by fitting the Bayesian SER model. Similarly to FS selection, this distribution is based on the results of the p univariate regressions; consequently each selection step in IBSS has the same computational complexity as in FS selection: $O(np)$. However, by computing a distribution on variables—rather than choosing a single best variable—IBSS captures uncertainty about which variable should be selected at each step. This uncertainty is taken into account when computing residuals by using a *model-averaged* (posterior mean) estimate for the regression coefficients. In IBSS, we use an iterative procedure, whereby early selections are re-evaluated in light of the later selections (as in 'backfitting'; Friedman and Stuetzle (1981)). The final output of IBSS is L distributions on variables, parameterized by $(\boldsymbol{\alpha}_l, \boldsymbol{\mu}_{1l}, \boldsymbol{\sigma}_{1l})$, for $l = 1, \ldots, L$, in place of the L variables that are

Table 1. Algorithm 1: IBSS

Require data $\mathbf{X}, \mathbf{y}$
Require number of effects, L, and hyperparameters $\sigma^2, \boldsymbol{\sigma}_0^2$
Require a function $\text{SER}(\mathbf{X}, \mathbf{y}; \sigma^2, \sigma_0^2) \to (\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \boldsymbol{\sigma}_1)$ that computes the posterior distribution for $\mathbf{b}_l$ under the SER model; see expression (2.11)

1, initialize posterior means $\bar{\mathbf{b}}_l = 0$, for $l = 1, \ldots, L$ ▷ other initializations are possible (see algorithm 3 in the on-line appendix B)
2, *repeat*
3, *for* l in $1, \ldots, L$ *do*
4, $\bar{\mathbf{r}}_l \leftarrow \mathbf{y} - \mathbf{X}\Sigma_{l' \neq l} \bar{\mathbf{b}}_{l'}$. ▷ expected residuals without lth single effect
5, $(\boldsymbol{\alpha}_l, \boldsymbol{\mu}_{1l}, \boldsymbol{\sigma}_{1l}) \leftarrow \text{SER}(\mathbf{X}, \bar{\mathbf{r}}_l; \sigma^2, \sigma_{0l}^2)$ ▷ fit SER to residuals
6, $\bar{\mathbf{b}}_l \leftarrow \boldsymbol{\alpha}_l \circ \boldsymbol{\mu}_{1l}$ ▷ '$\circ$' denotes elementwise multiplication
7, *until* convergence criterion satisfied
return $\boldsymbol{\alpha}_1, \boldsymbol{\mu}_{11}, \boldsymbol{\sigma}_{11}, \ldots, \boldsymbol{\alpha}_L, \boldsymbol{\mu}_{1L}, \boldsymbol{\sigma}_{1L}$

selected by FS selection. Each distribution is easily summarized, for example, by a 95% credible set for each selection.

To illustrate, consider our motivating example (Section 2.1) with $x_1 = x_2$ and $x_3 = x_4$, and with variables 1 and 4 having non-zero effects. To simplify the example, suppose that the effect of variable 1 is substantially larger than the effect of variable 4. Then FS selection would first (arbitrarily) select either variable 1 or 2, and then select (again arbitrarily) variable 3 or 4. In contrast, given enough data, the first IBSS update would select variables 1 and 2, i.e. it would assign approximately equal weights of 0.5 to variables 1 and 2, and small weights to other variables. The second IBSS update would similarly select variables 3 and 4 (again, with equal weights of approximately 0.5). Summarizing these results would yield two credible sets, $\{1, 2\}$ and $\{3, 4\}$, and the inference (2.2) is achieved. This simple example is intended only to sharpen intuition; later numerical experiments demonstrate that IBSS also works well in more realistic settings.

3.1.1. Iterative Bayesian stepwise selection computes a variational approximation to the SuSiE posterior distribution

The analogy between the IBSS algorithm and the simple FS procedure emphasizes the intuitive and computational simplicity of IBSS, but of course does not give it any formal support. We now provide a formal justification for IBSS. Specifically, we show that it is a co-ordinate ascent algorithm for optimizing a *variational approximation (VA) to the posterior distribution* for $\mathbf{b}_1, \ldots, \mathbf{b}_L$ under the SuSiE model (3.1)–(3.6). This result also suggests a method for estimating the hyperparameters σ^2 and $\boldsymbol{\sigma}_0^2$.

The idea behind VA methods for Bayesian models (e.g. Jordan *et al.* (1999) and Blei *et al.* (2017)) is to find an approximation $q(\mathbf{b}_1, \ldots, \mathbf{b}_L)$ to the posterior distribution $p_{\text{post}} := p(\mathbf{b}_1, \ldots, \mathbf{b}_L | \mathbf{y})$ by minimizing the Kullback–Leibler (KL) divergence from q to p_{post}, written as $D_{\text{KL}}(q, p_{\text{post}})$, subject to constraints on q that make the problem tractable. Although $D_{\text{KL}}(q, p_{\text{post}})$ itself is difficult to compute, it can be formulated in terms of an easier-to-compute function F, known as the 'evidence lower bound' (ELBO):

$$D_{\text{KL}}(q, p_{\text{post}}) = \log\{p(\mathbf{y}|\sigma^2, \boldsymbol{\sigma}_0^2)\} - F(q; \sigma^2, \boldsymbol{\sigma}_0^2).$$

Because $\log\{p(\mathbf{y}|\sigma^2, \boldsymbol{\sigma}_0^2)\}$ does not depend on q, minimizing D_{KL} over q is equivalent to maximizing F; and, since F is easier to compute, this is how the problem is usually framed. See the on-line appendix B.1 for further details. (Note that the ELBO also depends on the data, $\mathbf{X}$ and $\mathbf{y}$, but we make this dependence implicit to lighten the notation.)

We seek an approximate posterior q that factorizes as

$$q(\mathbf{b}_1, \ldots, \mathbf{b}_L) = \prod_{l=1}^{L} q_l(\mathbf{b}_l). \tag{3.7}$$

Under this approximation, $\mathbf{b}_1, \ldots, \mathbf{b}_L$ are independent *a posteriori*. We make no assumptions on the form of q_l; in particular, we do *not* require that each q_l factorizes over the p elements of $\mathbf{b}_l$. This is a crucial difference from previous VA approaches for BVSR (e.g. Logsdon *et al.* (2010) and Carbonetto and Stephens (2012)), and it means that q_l can accurately capture strong dependences between the elements of $\mathbf{b}_l$ under the assumption that exactly one element of $\mathbf{b}_l$ is non-zero. Intuitively, each factor q_l captures one effect variable and provides inferences of the form that 'we need one of variables $\{A, B, C\}$, and we are unsure about which one to select'. By extension, approximation (3.7) provides inferences of the form 'we need to select one variable among the set $\{A, B, C\}$, one variable among the set $\{D, E, F, G\}$, and so on'.

Under the assumption that the VA factorizes as equation (3.7), finding the optimal q reduces to the following problem:

$$\underset{q_1,\ldots,q_L}{\text{maximize}}\ F(q_1,\ldots,q_L;\sigma^2,\boldsymbol{\sigma}_0^2). \tag{3.8}$$

Although jointly optimizing F over $q_1,\ldots,q_L$ is difficult, optimizing an individual factor q_l is straightforward and in fact reduces to fitting an SER model, as formalized in the following proposition.

Proposition 1.

$$\arg\max_{q_l}\ F(q_1,\ldots,q_L;\sigma^2,\boldsymbol{\sigma}_0^2) = \mathrm{SER}(\mathbf{X},\mathbf{r}_l;\sigma^2,\sigma_{0l}^2), \tag{3.9}$$

where $\bar{\mathbf{r}}_l$ denotes the expected value of the residuals obtained by removing the estimated effects other than l,

$$\bar{\mathbf{r}}_l := \mathbf{y} - \mathbf{X}\sum_{l'\neq l}\bar{\mathbf{b}}_{l'}, \tag{3.10}$$

and where $\bar{\mathbf{b}}_l$ denotes the expected value of $\mathbf{b}_l$ with respect to the distribution q_l.

For intuition, note that computing the posterior distribution for $\mathbf{b}_l$ under model (3.1)–(3.6), given the other effects $\mathbf{b}_{l'}$ for $l'\neq l$, involves fitting an SER to the residuals $\mathbf{y}-\mathbf{X}\Sigma_{l'\neq l}\mathbf{b}_{l'}$. Now consider computing an (approximate) posterior distribution for $\mathbf{b}_l$ when $\mathbf{b}_{l'}$ are not known, and we have approximations $q_{l'}$ to their posterior distributions. Proposition 1 states that we can solve for $\arg\max_{q_l} F(q_1,\ldots,q_L)$ by using a similar procedure, except that each $\mathbf{b}_{l'}$ is replaced with the (approximate) posterior mean $\bar{\mathbf{b}}_{l'}$.

The following corollary is an immediate consequence of proposition 1.

Corollary 1. IBSS (algorithm 1) is a co-ordinate ascent algorithm for maximizing the ELBO F over q satisfying approximation (3.7). Equivalently, it is a co-ordinate ascent algorithm for minimizing the KL divergence $D_{\mathrm{KL}}(q, p_{\mathrm{post}})$ over q satisfying approximation (3.7), where p_{post} is the true posterior distribution under SuSiE.

Further, as a consequence of being a co-ordinate ascent algorithm, IBSS converges to a stationary point of F under conditions that are easily satisfied.

Proposition 2. Provided that $0<\sigma,\sigma_0<\infty$ and $\pi_j>0$ for all $j=1,\ldots,p$, the sequence of iterates q that are generated by the IBSS method (parameterized by $\boldsymbol{\alpha}_1,\boldsymbol{\mu}_{11},\boldsymbol{\sigma}_{11},\ldots,\boldsymbol{\alpha}_L,\boldsymbol{\mu}_{1L},\boldsymbol{\sigma}_{1L}$) converges to a limit point that is a stationary point of F.

The proofs of propositions 1 and 2 and corollary 1 are given in the on-line appendix B.

3.1.2. *Contrast with previous variational approximations*

A critical point is that the VA being computed by IBSS is different from previous 'fully factorized' VAs for BVSR (e.g. Logsdon *et al.* (2010) and Carbonetto and Stephens (2012)). In settings with highly correlated variables, the new VA produces results that are not only *quantitatively* different, but also *qualitatively* different from the fully factorized VA. For example, in our motivating example (Section 2.1), the new VA provides statements like expression (2.2), whereas the fully factorized VAs do not. Rather, a fully factorized VA often selects at most one of two identical variables without adequately capturing uncertainty in which variable should be selected (Carbonetto and Stephens, 2012). This feature makes the fully factorized VA unsuitable for applications where it is important to assess uncertainty in *which variables are selected*.

More generally, the new VA computed by IBSS satisfies the following intuitive condition: when two variables are identical, inferences that are drawn about their coefficients are identical (assuming that the priors on their coefficients are the same). Despite the simplicity of this condition, it is not satisfied by existing VAs, nor by point estimates from penalized likelihood approaches with L_0- or L_1-penalty terms. (In fact, Zou and Hastie (2005) used this condition as motivation for the EN method, which does ensure that point estimates for coefficients of identical variables are equal.) This property is formalized in the following proposition.

Proposition 3. Consider applying the IBSS algorithm (algorithm 1) to a data set in which two columns of $\mathbf{X}$ are identical, i.e. $x_j = x_k$ for some $j \neq k$. Further suppose that the prior distributions on selecting these two variables are equal ($\pi_j = \pi_k$). Then the approximate posterior computed by IBSS will be exchangeable in j and k, i.e. if $\omega_{jk} : \mathbb{R}^p \to \mathbb{R}^p$ denotes the function that permutes elements j and k of a p-vector, and q denotes the approximate posterior obtained from the IBSS algorithm, then

$$q\{\omega_{jk}(\mathbf{b}_1), \ldots, \omega_{jk}(\mathbf{b}_L)\} = q(\mathbf{b}_1, \ldots, \mathbf{b}_L). \tag{3.11}$$

Proof. Since $q(\mathbf{b}_1, \ldots, \mathbf{b}_L) = \prod_{l=1}^{L} q_l(\mathbf{b}_l)$, it suffices to show that each q_l is exchangeable in j and k, i.e. $q_l\{\omega_{jk}(\mathbf{b}_l)\} = q_l(\mathbf{b}_l)$ for all $l = 1, \ldots, L$. This exchangeability is satisfied after every iteration of the IBSS algorithm because the algorithm computes q_l (parameterized by $\boldsymbol{\alpha}_l$, $\boldsymbol{\mu}_{1l}$ and $\boldsymbol{\sigma}_{1l}$) as the exact posterior distribution under an SER model (step 5 of algorithm 1), and this posterior is exchangeable in j and k because both the prior and the likelihood are exchangeable.

Because the exchangeability is satisfied after every iteration of IBSS, and not just at convergence, the result is not sensitive to stopping criteria. By contrast, the corresponding EN property (Zou and Hastie, 2005) holds only at convergence—for example, in numerical implementations of the EN method (e.g. the `glmnet` R package), the coefficient estimates for identical variables can differ substantially. Similarly, Markov chain Monte Carlo based implementations of BVSR may satisfy this exchangeability property only asymptotically.

3.1.3. Estimating σ^2 and $\boldsymbol{\sigma}_0^2$

Algorithm 1 can be extended to estimate the hyperparameters σ^2 and $\boldsymbol{\sigma}_0^2$ by adding steps to maximize $F(q_1, \ldots, q_L; \sigma^2, \boldsymbol{\sigma}_0^2)$ over σ^2 and/or $\boldsymbol{\sigma}_0^2$. Estimating the hyperparameters by maximizing the ELBO can be viewed as an expectation–maximization algorithm (Dempster *et al.*, 1977) in which the expectation step is approximate (Heskes *et al.*, 2004; Neal and Hinton, 1998).

Optimizing F over σ^2 involves computing the expected residual sum of squares under the VA, which is straightforward; see the on-line appendix B for details.

Optimizing F over $\boldsymbol{\sigma}_0^2 = (\sigma_{01}^2, \ldots, \sigma_{0L}^2)$ can be achieved by modifying the step that computes the posterior distribution for $\mathbf{b}_l$ under the SER model to estimate first the hyperparameter σ_{0l}^2 in the SER model by maximum likelihood, i.e. by maximizing the SER likelihood (2.12) over σ_0^2, keeping σ^2 fixed (step 5 of algorithm 3 in on-line appendix B). This is a one-dimensional optimization which is easily performed numerically (we used the R function `optim`).

Algorithm 3 in appendix B extends algorithm 1 to include both these steps.

3.2. Posterior inference: posterior inclusion probabilities and credible sets

Algorithm 1 provides an approximation to the posterior distribution of $\mathbf{b}$ under SuSiE, parameterized by $(\boldsymbol{\alpha}_1, \boldsymbol{\mu}_{11}, \boldsymbol{\sigma}_{11}), \ldots, (\boldsymbol{\alpha}_L, \boldsymbol{\mu}_{1L}, \boldsymbol{\sigma}_{1L})$. From these results it is straightforward to compute approximations to various posterior quantities of interest, including PIPs and credible sets.

3.2.1. Posterior inclusion probabilities

Under SuSiE, the effect of explanatory variable j is $b^{(j)} := \sum_{l=1}^{L} b_{lj}$, which is 0 if and only if $b_{lj} = 0$ for all $l = 1, \ldots, L$. Under our VA the b_{lj} are independent across l, and therefore

$$\text{PIP}_j := \Pr(b^{(j)} \neq 0 | \mathbf{X}, \mathbf{y}) \approx 1 - \prod_{l \in \mathcal{L}} (1 - \boldsymbol{\alpha}_{lj}). \tag{3.12}$$

Here, we set $\mathcal{L} := \{l : \sigma_{0l}^2 > 0\}$ to treat the case where some σ_{0l}^2 are 0, which can happen if $\boldsymbol{\sigma}_0^2$ is estimated.

3.2.2. Credible sets

Computing the sets $\text{CS}(\boldsymbol{\alpha}_l; \rho)$ (A.4) in the on-line appendix A for $l = 1, \ldots, L$ immediately yields L credible sets that satisfy definition 1 under the VA to the posterior.

If L exceeds the number of detectable effects in the data, then in practice many of the L credible sets are large, often containing the majority of variables. The intuition is that, once all the detectable signals have been accounted for, the IBSS algorithm becomes very uncertain about which variable to include at each step, and so the distributions $\boldsymbol{\alpha}$ become very diffuse. Credible sets that contain very many uncorrelated variables are of essentially no inferential value—whether or not they contain an effect variable—and so in practice it makes sense to ignore them. To automate this, in this paper we discard credible sets with 'purity' less than 0.5, where we define purity as the smallest absolute correlation between all pairs of variables within the credible set. (To reduce computation for credible sets containing over 100 variables, we sampled 100 variables at random to estimate the purity.) The purity threshold of 0.5 was chosen primarily for comparing with Lee *et al.* (2018), who used a similar threshold in a related context. Although any choice of threshold is somewhat arbitrary, in practice we observed that most credible sets are either very pure (greater than 0.95) or very impure (less than 0.05), with intermediate cases being rare (Fig. S2 in the on-line appendix), so most results are robust to this choice of threshold.

3.3. Choice of L

It may seem that SuSiE would be sensitive to the choice of L. In practice, however, key inferences are often robust to overstating L; for example, in our simulations below, the simulated number of effects was between 1 and 5, whereas we still obtain good results with $L = 10$. This is because, when L is larger than necessary, the method is very uncertain about where to place the extra effects—consequently, it distributes them broadly among many variables, and therefore they are too diffuse to impact key inferences. For example, setting L to be larger than necessary inflates the PIPs of many variables, but only slightly, and the extra components result in credible sets with low purity.

Although inferences are generally robust to overstating L, we also note that the empirical Bayes version of our method, which estimates $\boldsymbol{\sigma}_0^2$, also effectively estimates the number of effects: when L is greater than the number of signals in the data, the maximum likelihood estimate of σ_{0l}^2 will be 0 or close to 0 for many l, which in turn forces b_l to 0. This is closely related to the idea behind 'automatic relevance determination' (Neal, 1996).

3.4. Identifiability and label switching

The parameter vectors $\mathbf{b}_1, \ldots, \mathbf{b}_L$ that were introduced in SuSiE are technically non-identifiable, in that the likelihood $p(\mathbf{y} | \mathbf{b}_1, \ldots, \mathbf{b}_L)$ is unchanged by permutation of the labels $1, \ldots, L$. As a result, the posterior distribution p_{post} is symmetric with respect to permutations of the labels (as-

suming that the prior is also symmetric)—i.e., for any permutation $\nu : \{1, \ldots, L\} \to \{1, \ldots, L\}$, we have $p(\mathbf{b}_1, \ldots, \mathbf{b}_L | \mathbf{y}) = p(\mathbf{b}_{\nu(1)}, \ldots, \mathbf{b}_{\nu(L)} | \mathbf{y})$. A similar non-identifiability also occurs in mixture models, where it is known as the 'label switching problem' (Stephens, 2000).

In principle, non-identifiability due to label switching does not complicate Bayesian inference; the posterior distribution is well defined and correctly reflects uncertainty in the parameters. In practice, however, complications can arise. Specifically, label switching typically causes the posterior distribution to be multimodal, with $L!$ symmetric modes corresponding to the $L!$ different labellings (ν above). Care is then needed when summarizing this posterior distribution. For example, the posterior mean will not be a sensible estimate for $\mathbf{b}_1, \ldots, \mathbf{b}_L$ because it averages over the $L!$ modes (Stephens, 2000).

Fortunately, our variational approximation (Section 3.1.1) avoids these potential complications of label switching. This is due to the way that variational approximations behave when approximating the posterior distribution of a mixture model; they typically produce a good approximation to one of the permutations, effectively ignoring the others (Wang and Titterington, 2006; Blei *et al.*, 2017; Pati *et al.*, 2018). See also the discussion of 'spontaneous symmetry breaking' in Wainwright and Jordan (2007). Consequently, our posterior approximation $q(\mathbf{b}_1, \ldots, \mathbf{b}_L)$ approximates just one of the $L!$ symmetric modes of the true posterior, avoiding the issues with label switching that can occur when summarizing the true posterior distribution.

Formally, this non-identifiability causes the objective F that is optimized by the IBSS algorithm to be invariant to relabelling—i.e.

$$F(q_{\nu(1)}, \ldots, q_{\nu(L)}; \sigma^2, \sigma^2_{0\nu(1)}, \ldots, \sigma^2_{0\nu(L)})$$

is the same for all permutations ν—and therefore every solution $\hat{q}$, $\hat{\sigma}$ and $\hat{\sigma}_0^2$ that is returned by our IBSS algorithm has $L!$ equivalent solutions that achieve the same value of the ELBO F, each corresponding to a different labelling (and a different mode of the true posterior). These $L!$ solutions are inferentially equivalent; they all imply the same distribution for the unordered set $\{\mathbf{b}_1, \ldots, \mathbf{b}_L\}$ and the same distribution for the sum $\mathbf{b} = \sum_{l=1}^{L} \mathbf{b}_l$ (which does not depend on the labelling), and they all produce the same PIPs and the same credible sets. Thus, it does not matter which mode is used.

4. Numerical comparisons

We performed numerical comparisons on data generated to mimic closely our main motivating application: genetic fine mapping. Specifically, we generated data for fine mapping of expression quantitative trait loci (QTLs), which are genetic variants associated with gene expression. We used these simulations to assess our methods and compare with state of the art BVSR methods that were specifically developed for this problem. We also compared against a (frequentist) hierarchical testing method (Mandozzi and Bühlmann, 2016; Renaux *et al.*, 2018).

In genetic fine mapping, $\mathbf{X}$ is a matrix of genotype data, in which each row corresponds to an individual, and each column corresponds to a genetic variant, typically a single-nucleotide polymorphism (SNP). In our simulations, we used the real human genotype data from $n = 574$ genotype samples collected as part of the genotype–tissue expression project (GTEx Consortium, 2017). To simulate fine mapping of *cis* effects on gene expression, we randomly selected 150 genes out of the more than 20000 genes on chromosomes 1–22 and then assigned $\mathbf{X}$ to be the genotypes for genetic variants near the transcribed region of the selected gene. For a given gene, between $p = 1000$ and $p = 12000$ SNPs were included in the fine mapping analysis; for more details on how SNPs were selected, see the on-line appendix D.

These real genotype matrices $\mathbf{X}$ exhibit complex patterns of correlations; see Fig. A1 for

example. Furthermore, many variables are strongly correlated with other variables: for a randomly chosen variable, the median number of other variables with which its correlation exceeds 0.9 is 8, and the median number of other variables with which its correlation exceeds 0.99 is 1. Corresponding means are even larger—26 and eight other variables respectively—because some variables are strongly correlated with hundreds of other variables. Thus these genotype matrices lead to challenging, but realistic, variable-selection problems.

We generated synthetic outcomes $\mathbf{y}$ under the multiple-regression model (2.1), with assumptions on $\mathbf{b}$ specified by two parameters: S, the number of effect variables, and ϕ, the proportion of variance in $\mathbf{y}$ explained by $\mathbf{X}$. Given S and ϕ, we simulated $\mathbf{b}$ and $\mathbf{y}$ as follows.

(a) Sample the indices of the S effect variables, $\mathcal{S}$, uniformly at random from $\{1, \ldots, p\}$.
(b) For each $j \in \mathcal{S}$, independently draw $b_j \sim N(0, 0.6^2)$ and, for all $j \notin \mathcal{S}$, set $b_j = 0$.
(c) Set σ^2 to achieve the desired proportion of variance explained ϕ; specifically, we solve for σ^2 in

$$\phi = \frac{\text{var}(\mathbf{Xb})}{\sigma^2 + \text{var}(\mathbf{Xb})},$$

where var$(\cdot)$ denotes sample variance.
(d) For each $i = 1, \ldots, n$, draw $y_i \sim N(x_{i1}b_1 + \ldots + x_{ip}b_p, \sigma^2)$.

We generated data sets under two simulation scenarios. In the first scenario, each data set has $p = 1000$ SNPs. We generated data sets by using all pairwise combinations of $S \in \{1, \ldots, 5\}$ and $\phi \in \{0.05, 0.1, 0.2, 0.4\}$. These settings were chosen to span typical expected values for expression QTL studies. We simulated two replicates for each gene and for each combination of S and ϕ. Therefore, in total we generated $2 \times 150 \times 5 \times 4 = 6000$ data sets for the first simulation scenario.

In the second simulation scenario, we generated data sets with more SNPs, ranging from 3000 to 12000 SNPs, and, to generate the outcomes $\mathbf{y}$, we set $S = 10$ and $\phi = 0.3$. We generated two replicates for each gene, resulting in a total of $2 \times 150 = 300$ data sets in the second simulation scenario.

4.1. Illustrative example

We begin with an example to illustrate that the IBSS algorithm (algorithm 1) can perform well in a challenging fine mapping setting. This example is summarized in Fig. 1.

We draw this example from one of our simulations in which the variable with the strongest marginal association (SMA) with $\mathbf{y}$ is not one of the actual effect variables (in this example, there are two effect variables). This situation occurs because the SMA variable has moderate correlation with both effect variables, and these effects combine to make its marginal association stronger than the marginal associations of the individual effect variables. Standard forward selection in this case would select the wrong (SMA) variable in the first step; indeed, after one iteration, IBSS also yields a credible set that includes the SMA variable (Fig. 1(b)). However, as the IBSS algorithm proceeds, it recognizes that, once other variables have been accounted for, the SMA variable is no longer required. After 10 iterations (at which point the IBSS solution is close to convergence) IBSS yields two high purity credible sets, neither containing the SMA, and each containing one of the effect variables (Fig. 1(c)). Our manuscript resource repository includes an animation showing the iteration-by-iteration progress of the IBSS algorithm (Wang *et al.*, 2020b).

This example, where the SMA variable does not appear in a credible set, also illustrates that multiple regression can sometimes result in conclusions that are very different from those of a marginal association analysis.

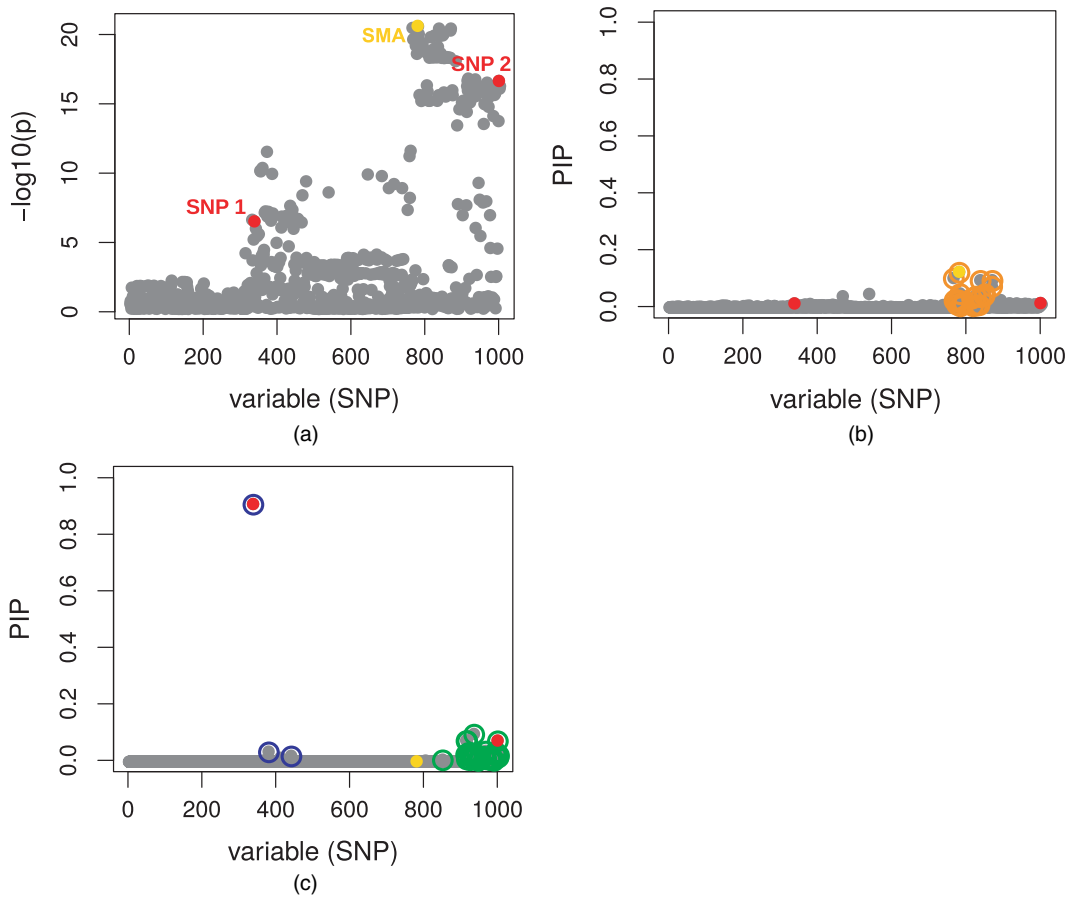

Fig. 1. Fine mapping example to illustrate that the IBSS algorithm can deal with a challenging case: the results are from a simulated data set with $p = 1000$ variables (SNPs); some of these variables are very strongly correlated (Fig. A1 in the on-line appendix); two out of the 1000 variables are effect variables (●, labelled 'SNP 1' and 'SNP 2' in (a)); we chose this example from our simulations because the strongest marginal association is a non-effect variable (●, labelled 'SMA' in (a)); after one iteration ((b)), IBSS incorrectly identifies a credible set containing the strongest marginal association and no effect variable (●); however, after 10 iterations (and also at convergence) the IBSS algorithm has corrected itself ((c)), finding two 95% credible sets (○, ○), each containing a true effect variable; additionally, neither credible set contains the strongest marginal association variable; one credible set (●) contains only three SNPs (purity 0.85), whereas the other credible set (●) contains 37 very highly correlated variables (purity 0.97); in the latter credible set, the individual PIPs are small, but the inclusion of the 37 variables in this credible set indicates, correctly, high confidence in at least one effect variable among them

4.2. Posterior inclusion probabilities

Next, we seek to assess the effectiveness of our methods more quantitatively. We focus initially on one of the simpler tasks in BVSR: computing PIPs. Most implementations of BVSR compute PIPs, making it possible to compare results across several implementations. Here we compare our methods (henceforth SuSiE, implemented in R package `susieR`, version 0.4.29) with three other software implementations that were specifically developed for genetic fine mapping applications: CAVIAR (Hormozdiari *et al.* (2014), version 2.2), FINEMAP (Benner *et al.* (2016), version 1.1) and DAP-G (Wen *et al.* (2016) and Lee *et al.* (2018), installed by using source code from the git repository, commit `ef11b26`). These methods are all implemented as C++

programs. They implement similar BVSR models and differ in the algorithms that are used to fit these models and the priors on the effect sizes. CAVIAR exhaustively evaluates all possible combinations of up to L non-zero effects among the p variables. FINEMAP and DAP-G approximate this exhaustive approach by heuristics that target the best combinations. Another important difference between the methods is that FINEMAP and CAVIAR perform inference by using summary statistics that are computed from each data set—specifically, the marginal association Z-scores and the $p \times p$ correlation matrix for all variables—whereas, as we apply them here, DAP-G and SuSiE use the full data. The summary statistic approach can be viewed as approximating inferences from the full data; see Lee *et al.* (2018) for discussion.

For SuSiE, we set $L = 10$ for all the data sets that were generated in the first simulation scenario, and $L = 20$ for the second scenario. We assessed performance both estimating the hyperparameters σ^2 and $\boldsymbol{\sigma}_0^2$, and fixing one or both of these hyperparameters. The overall performances of these different approaches were similar, and here we show results when σ^2 was estimated, and σ_{0l}^2 was fixed to $0.1\,\mathrm{var}(\mathbf{y})$ (consistent with data applications in Section 5); other results are in the on-line supplementary data (Fig. S4 and Fig. S5). Parameter settings for other methods are given in appendix D. We ran CAVIAR and FINEMAP only on simulations with $S \leqslant 3$ since these methods are computationally more intensive than the others (particularly for larger S).

Since these methods differ in their modelling assumptions, we should not expect their PIPs to be equal. Nonetheless, we found generally reasonably good agreement (Figs 2(a)–2(c)). For $S = 1$, the PIPs from all four methods agree closely. For $S > 1$, the PIPs from the various methods are also highly correlated; correlations between PIPs from SuSiE and the other methods vary from 0.94 to 1 across individual data sets, and the number of PIPs differing by more than 0.1 is always small—the proportions vary from 0.013% to 0.2%. In the scatter plots, this agreement appears less strong because the eye is drawn to the small proportion of points that lie away from the diagonal, but the vast majority of points lie on or near the origin. In addition, all four methods produce reasonably well-calibrated PIPs (Fig. S1 in the on-line appendix).

The general agreement of PIPs from the four methods suggests that

(a) all four methods are mostly accurate for computing PIPs for the data set sizes that were explored in our numerical comparisons and
(b) the PIPs themselves are usually robust to details of the modelling assumptions.

Nonetheless, some non-trivial differences in PIPs are clearly visible from Figs 2(a)–2(c). Visual inspection of these differences suggests that the SuSiE PIPs may better distinguish effect variables from non-effect variables, in that there appears a higher ratio of red–grey points below the diagonal than above the diagonal. This is confirmed in our analysis of power *versus* FDR, obtained by varying the PIP threshold independently for each method; at a given FDR, the SuSiE PIPs always yield higher power (Figs 2(d) anf 2(e)).

Notably, even though SuSiE is implemented in R, its computations are much faster than the other methods implemented in C++: for example, in the data sets simulated with $S = 3$, SuSiE is, on average, roughly four times faster than DAP-G, 30 times faster than FINEMAP and 4000 times faster than CAVIAR (Table 2).

Because SuSiE computations scale linearly with data size (computational complexity $O(npL)$ per iteration) it can easily handle data sets that are much larger than those in these simulations. To illustrate, running SuSiE ($L = 10$) on two larger simulated data sets—one with $n = 100000$, $p = 500$; another with $n = 1000$, $p = 50000$; each with four effect variables—took 25 s and 43 s on a modern Linux workstation (see the on-line appendix D.2 for details). This is competitive with the lasso, implemented in the `glmnet` R package, version 2.0.18, which

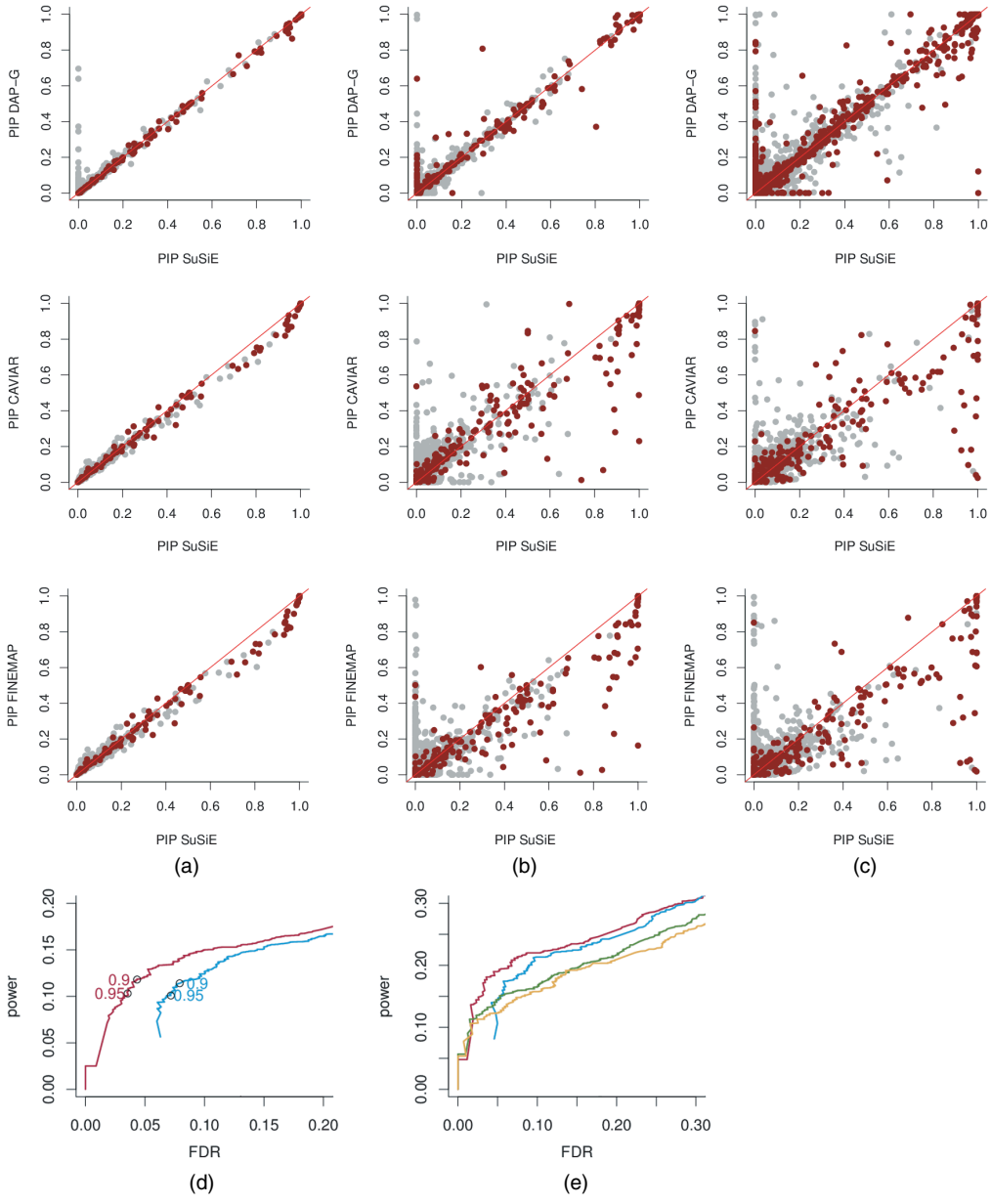

Fig. 2. Evaluation of PIPs (scatter plots in (a)–(c) compare PIPs computed by SuSiE against PIPs computed by using the other methods (DAP-G, CAVIAR and FINEMAP) (each point depicts a single variable in one of the simulations; ●, true effect variables; ●, variables with no effect; the scatter plots in (d) and (e) combine results across the first set of simulations; (d) and (e) summarize power *versus* FDR from the first simulation scenario; these curves are obtained by independently varying the PIP threshold for each method (○, PIP thresholds of 0.9 and 0.95); these quantities are calculated as FDR := FP/(TP + FP) (also known as the 'false discovery proportion') and power := TP/(TP + FN), where FP, TP, FN and TN denote the number of false positive, true positive, false negative and true negative results respectively (this plot is the same as a *precision–recall curve* after reversing the x-axis, because precision = TP/(TP + FP) = 1 − FDR, and recall = power) (note that CAVIAR and FINEMAP were run only on data sets with 1–3 effect variables): (a) one effect variable; (b) two effect variables; (c) 3–5 effect variables; (d) 1–5 effect variables (——, SuSiE; ——, DAP-G); (e) 1–3 effect variables (——, SuSiE; ——, DAP-G; ——, CAVIAR; ——, FINEMAP)

Table 2. Run times from data sets simulated with $S = 3$

Method	*Mean (s)*	*Minimum (s)*	*Maximum (s)*
SuSiE	0.64	0.34	2.28
DAP-G	2.87	2.23	8.87
FINEMAP	23.01	10.99	48.16
CAVIAR	2907.51	2637.34	3018.52

with tenfold cross-validation (and other parameters at their defaults) took 82 s for each data set.

In summary, in the settings that are considered here, SuSiE produces PIPs that are as reliable as or more reliable than existing BVSR methods and does so at a fraction of the computational effort.

4.3. Credible sets

4.3.1. Comparison with DAP-G

A key feature of SuSiE is that it yields multiple credible sets, each aimed at capturing an effect variable (definition 1). The only other BVSR method that attempts something similar, as far as we are aware, is DAP-G, which outputs 'signal clusters' defined by heuristic rules (Lee *et al.*, 2018). Although Lee *et al.* (2018) did not refer to their signal clusters as credible sets, and they did not give a formal definition of a signal cluster, the intent of these signal clusters is similar to our credible sets, and so for brevity we henceforth refer to them as credible sets.

We compared the level 95% credible sets produced by SuSiE and DAP-G in several ways. First we assessed their empirical (frequentist) coverage levels, i.e. the proportion of credible sets that contain an effect variable. Since our credible sets are Bayesian credible sets, 95% credible sets are not designed, or guaranteed, to have a frequentist coverage of 0.95 (Fraser, 2011). Indeed, coverage will inevitably depend on the simulation scenario; for example, in completely null simulations, in which the data are simulated with $\mathbf{b} = \mathbf{0}$, *every* credible set would necessarily contain no effect variable, and so the coverage would be zero. Nonetheless, under reasonable circumstances that include effect variables, one might hope that the Bayesian credible sets would have coverage that is near the nominal levels. And, indeed, we confirmed this was so: in the simulations, credible sets from both methods typically had coverage slightly below 0.95, and in most cases above 0.90 (Fig. 3; see Fig. S3 in the on-line appendix for additional results).

Having established that the methods produce credible sets with similar coverage, we compared them by three other criteria:

(a) power (the overall proportion of simulated effect variables included in a credible set),
(b) average size (the median number of variables included in a credible set) and
(c) purity (here, measured as the average squared correlation of variables in a credible set since this statistic is provided by DAP-G).

By all three metrics, the credible sets from SuSiE are consistently an improvement over DAP-G—they achieve higher power, smaller size and higher purity (Fig. 3).

Although the way that we construct credible sets in SuSiE does not require that they be disjoint, we note that the credible sets rarely overlapped (after filtering out low purity credible

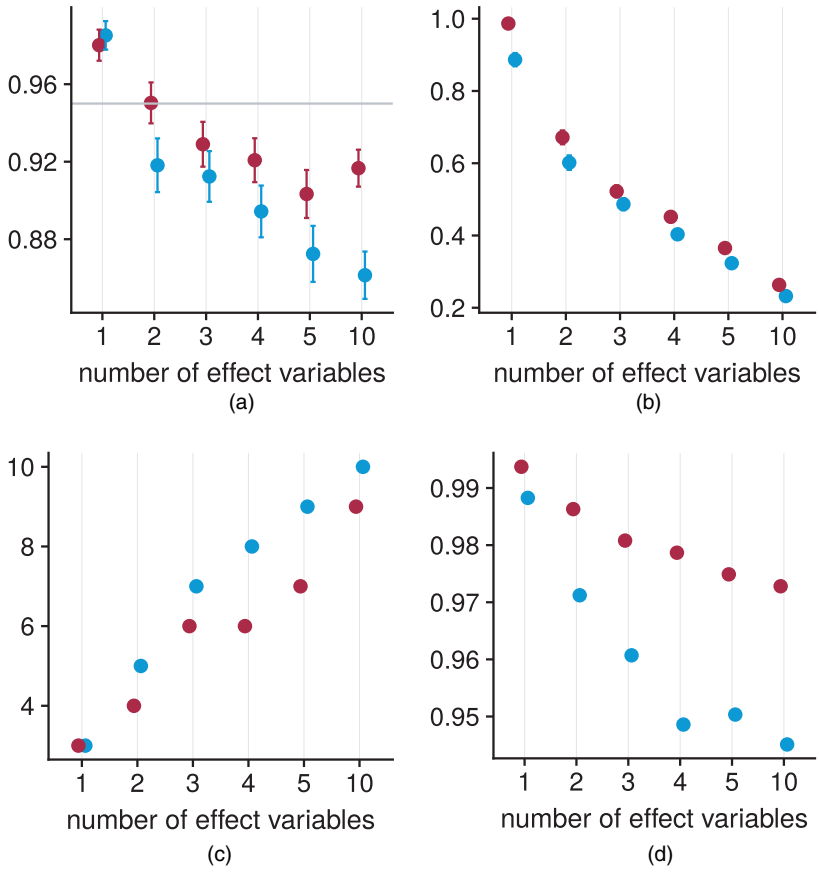

Fig. 3. Comparison of 95% credible sets from SuSiE (●) and DAP-G (●): (a) coverage, (b) power, (c) median size and (d) average squared correlation of the variables in each credible set (these statistics are taken as the mean over all credible sets computed in all data sets; error bars in (a) show 2 × standard error; simulations with 1–5 effect variables are from the first simulation scenario, and simulations with 10 effect variables are from the second scenario)

sets; see Section 3.2.2). Indeed, across the thousands of simulations, there was only one example of two credible sets overlapping.

4.3.2. Comparison with hierarchical testing

Finally, we compared our credible sets with results produced by the R package `hierinf` (Renaux *et al.*, 2018) (version 1.3.1), which implements a frequentist approach to identifying significant clusters of variables based on hierarchical testing (Meinshausen, 2008; Mandozzi and Bühlmann, 2016). In brief, this approach starts by assuming that the variables are organized in a given hierarchy. Then, starting from the top of the hierarchy, it proceeds to test whether groups of variables (clades in the hierarchy) contain at least one non-zero effect. Each time that a group is deemed significant, the method proceeds to test clades in the next level of the hierarchy. The procedure ultimately reports the smallest significant clades detected, where the significance criteria are designed to control the overall familywise error rate FWER at a prespecified level α. We note that FWER-control is not guaranteed when $p > n$ and variables are highly correlated (Mandozzi and Bühlmann, 2016), which is the situation in our simulations.

Although the theory for controlling FWER in hierarchical testing is elegant, genetic variants do not come in a natural hierarchy, and so for fine mapping the need to specify a hierarchy is a drawback. Here we use the `cluster_var` function from `hierinf`, which infers a hierarchical clustering. There is no simple correspondence between the level α and (frequentist) coverage rates of the significant clusters, so selecting a suitable α is non-trivial; in our simulations, we found that empirical coverage was typically close to 0.95 when $\alpha = 0.1$, so we report results for $\alpha = 0.1$.

The results (Table 3) show that the `hierinf` clusters are substantially larger and have lower purity than the credible sets from SuSiE, as well as from DAP-G. For example, in simulations with five effect variables, the SuSiE credible sets have a median size of seven variables with an average r^2 of 0.97, whereas the `hierinf` clusters have a median size of 54 variables with an average r^2 of 0.56. Further, the number of credible sets that were reported by SuSiE and DAP-G is higher than the number of significant clusters from `hierinf`.

We believe that the much larger number of variables that are included in the `hierinf` clusters partly reflects a fundamental limitation of the hierarchical approach to this problem. Specifically, by assuming a hierarchy that does not truly exist, the method artificially limits the clusters of variables that it can report. This will sometimes force it to report clusters that are larger than necessary. For example, with three variables, if variables 2 and 3 are grouped at the bottom of the hierarchy, then the method could never report a cluster $\{1, 2\}$, representing the statement 'either variable 1 or 2 is an effect variable, but we cannot tell which', even if the data support such an inference. Instead, it would have to report the larger cluster, $\{1, 2, 3\}$.

While our work here was under peer review and available as a preprint (Wang *et al.*, 2019), we became aware of new related work in Sesia *et al.* (2020). Similarly to `hierinf` this new method tests groups of variables at multiple resolutions in a hierarchy; but it improves on `hierinf` by controlling the FDR of selected groups (rather than type I error), and with statistical guarantees that hold even in the presence of highly correlated variables. Comparisons with our method find that their significant groups are typically larger than ours (Sesia *et al.* (2020), Fig. 4), presumably in part because of the fundamental limitation with the hierarchical approach (discussed above).

5. Application to fine mapping splice quantitative trait loci

To illustrate SuSiE for a real fine mapping problem, we analysed data from Li *et al.* (2016) aimed at detecting genetic variants (SNPs) that influence splicing (known as 'splice QTLs'). Li *et al.* (2016) quantified alternative splicing by estimating, at each intron in each sample, a ratio capturing how often the intron is used relatively to other introns in the same 'cluster' (roughly, gene). The data involve 77345 intron ratios measured on lymphoblastoid cell lines from 87 Yoruban individuals, together with genotypes of these individuals. Following Li *et al.* (2016), we preprocessed the intron ratios by regressing out the first three principal components of the matrix of intron ratios; the intent is to control for unmeasured confounders (Leek and Storey, 2007). For each intron ratio, we fine-mapped SNPs within 100 kilobases of the intron, which is approximately 600 SNPs on average. In short, we ran SuSiE on 77345 data sets with $n = 87$ and $p \approx 600$.

To specify the prior variance σ_{0l}^2, we first estimated typical effect sizes from the data on all introns. Specifically, we performed univariate (SNP-by-SNP) regression analysis at every intron and estimated the proportion of variance explained of the top (strongest associated) SNP. The mean proportion of variance explained of the top SNP across all introns was 0.096, so we applied SuSiE with $\sigma_{0l}^2 = 0.096\,\mathrm{var}(\mathbf{y})$, and with the columns of $\mathbf{X}$ standardized to have unit variance. The residual variance parameter σ^2 was estimated by IBSS.

We then ran SuSiE to fine-map splice QTLs at all 77345 introns. After filtering for purity, this

Table 3. Comparison of credible sets from SuSiE and DAP-G with significant clusters from hierarchical inference (hierinf software, with FWER-level $\alpha = 0.1$)†

Number of effects	*Power*			*Coverage*			*Median size*			*Average r^2*		
	SuSiE	*DAP-G*	*hierinf*	*SuSiE*	*DAP-G*	*hierinf*	*SuSiE*	*DAP-G*	*hierinf*	*SuSiE*	*DAP-G*	*hierinf*
1	0.99	0.89	0.97	0.98	0.99	0.94	3	3	8	0.99	0.99	0.82
2	0.67	0.60	0.55	0.95	0.92	0.96	4	5	20	0.99	0.97	0.71
3	0.52	0.49	0.39	0.93	0.91	0.95	6	7	34	0.98	0.96	0.64
4	0.45	0.40	0.29	0.92	0.89	0.95	6	8	37	0.98	0.95	0.60
5	0.37	0.32	0.24	0.90	0.87	0.98	7	9	54	0.97	0.95	0.56

†Results are averages across all data sets in the first simulation scenario.

yielded a total of 2652 credible sets (level 0.95) spread across 2496 intron units. These numbers are broadly in line with the original study, which reported 2893 significant introns at 10% FDR. Of the 2652 credible sets that were identified, 457 contain exactly one SNP, representing strong candidates for being the causal variants that affect splicing. Another 239 credible sets contain exactly two SNPs. The median size of a credible set was 7, and the median purity was 0.94.

The vast majority of intron units with a credible set had exactly one credible set (2357 of 2496). Thus, SuSiE could detect at most one splice QTL for most introns. Of the remainder, 129 introns yielded two credible sets, five introns yielded three credible sets, three introns yielded four credible sets and two introns yielded five credible sets. This represents a total of $129+10+9+8=156$ additional ('secondary') signals that would be missed in conventional analyses that report only one signal per intron. Both primary and secondary signals were enriched in regulatory regions (on-line appendix E), lending some independent support that SuSiE is detecting real signals. Although these data show relatively few secondary signals, this is a small study ($n=87$); in larger studies, the ability of SuSiE to detect secondary signals will probably be greater.

6. An example beyond fine mapping: change point detection

Although our methods were motivated by genetic fine mapping, they are also applicable to other sparse regression problems. Here we apply SuSiE to an example that is quite different from fine mapping: change point detection. This application also demonstrates that the IBSS algorithm can sometimes produce a poor fit—due to becoming stuck in a local optimum—which was seldom observed in our fine mapping simulations. We believe that examples where algorithms fail are just as important as examples where they succeed—perhaps more so—and that this example could motivate improvements.

We consider the simple change point model

$$y_t = \mu_t + e_t, \qquad t = 1, \ldots, T, \tag{6.1}$$

where t indexes a dimension such as space or time, and the errors e_t are independently normal with zero mean and variance σ^2. The mean vector $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_T)$ is assumed to be piecewise constant; the indices t where changes to $\boldsymbol{\mu}$ occur, $\mu_t \neq \mu_{t+1}$, are called the 'change points'.

To capture change points being rare, we formulate the change point model as a sparse multiple regression (2.1) in which $\mathbf{X}$ has $T-1$ columns, and the tth column is a step function with a step at location t, i.e. $x_{st}=0$ for $s \leqslant t$, and $x_{st}=1$ for all $s>t$). The tth element of $\mathbf{b}$ then determines the change in the mean at position t, $\mu_{t+1}-\mu_t$. Therefore, the non-zero regression coefficients in this multiple-regression model correspond to change points in $\boldsymbol{\mu}$.

The design matrix $\mathbf{X}$ in this setting has a very special structure, and quite different from fine mapping applications; the $(T-1)\times(T-1)$ correlation matrix decays systematically and very slowly away from the diagonal. By exploiting this special structure of $\mathbf{X}$, SuSiE computations can be made $O(TL)$ rather than the $O(T^2L)$ of a naive implementation; for example, the matrix–vector product $\mathbf{X}^{\mathrm{T}}\mathbf{y}$, naively an $O(T^2)$ computation, can be computed as the cumulative sum of the elements of the reverse of $\mathbf{y}$, which is an $O(T)$ computation.

Change point detection has a wide range of potential applications, such as segmentation of genomes into regions with different numbers of copies of the genome. Software packages in R that can be used for detecting change points include `changepoint` (Killick and Eckley, 2014), `DNAcopy` (Seshan and Olshen, 2018; Olshen *et al.*, 2004), `bcp` (Erdman and Emerson, 2007) and `genlasso` (Tibshirani, 2014; Arnold and Tibshirani, 2016); see Killick and Eckley (2014) for a longer list. Of these, only `bcp`, which implements a Bayesian method, quantifies uncertainty in estimated change point locations, and `bcp` provides only PIPs, not credible sets

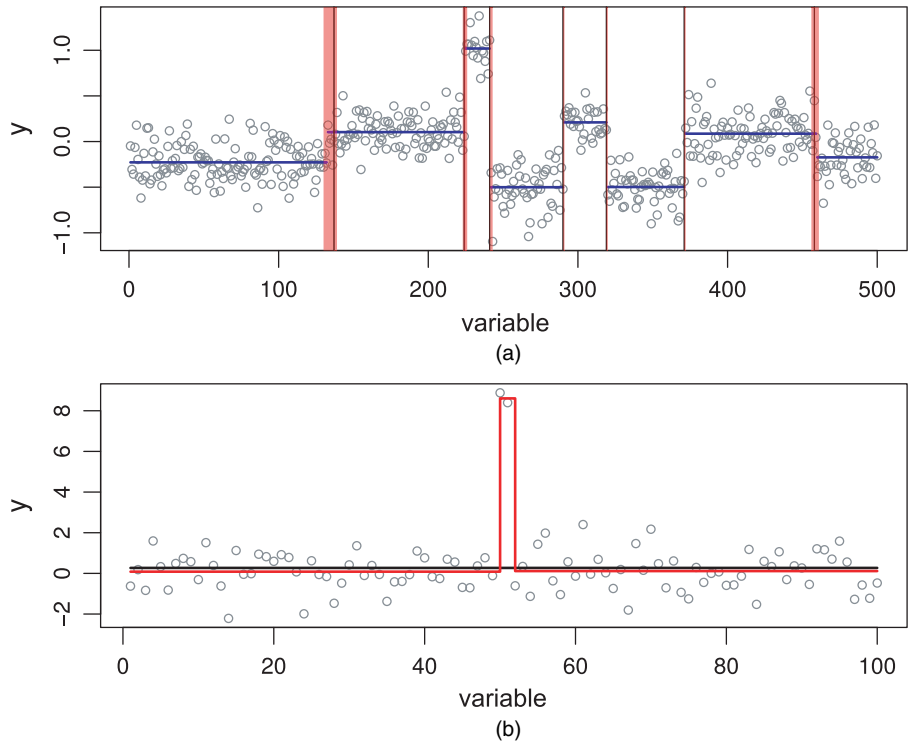

Fig. 4. Illustration of SuSiE applied to two change point problems: (a) a simulated example with seven change points (|) (———, mean function inferred by the segment method from the DNAcopy R package (version 1.56.0); the inference is reasonably accurate—all change points except the leftmost are nearly exactly recovered—but provides no indication of uncertainty in the locations of the change points; ———, 95% credible sets for change point locations inferred by SuSiE; each of these credible sets contains a true change point); (b) a simulated example with two change points in quick succession (this example is intended to illustrate convergence of the IBSS algorithm to a (poor) local optimum) (———, fit from the IBSS algorithm when it is initialized to a null model in which there are no change points; this fit results in no change points being detected; ———, result also of running IBSS, but this time the fitting algorithm is initialized to the true model with two change points; the latter accurately recovers both change points and attains a higher value of the objective function (−148.2 *versus* −181.8))

for change point locations. Therefore, the ability of SuSiE to provide credible sets is unusual, and perhaps unique, among existing change-point-detection methods.

To illustrate its potential for change point estimation, we applied SuSiE to a simulated example that is included with the `DNAcopy` R package. In this example, all settings for running SuSiE remain unchanged from the fine mapping simulations (Section 4). Fig. 4(a) shows results of applying SuSiE and `DNAcopy` to the data set. Both methods provide accurate estimates of the change points; indeed all change point locations except the leftmost are recovered nearly exactly. However, only SuSiE provides 95% credible sets for each estimate of a change point location. And, indeed, SuSiE is most uncertain about the leftmost change point. All the true change points in this example are contained in a SuSiE credible set, and every credible set contains a true change point. This occurs even though we set $L = 10$ to be greater than the number of true change points (7); the three extra credible sets were filtered out because they contained variables that were very uncorrelated. (To be precise, SuSiE reported eight credible sets after filtering, but two of the credible sets overlapped and contained the same change point; this observation

of overlapping of credible sets contrasts with the fine mapping simulations in Section 4 where overlapping credible sets occurred very rarely.)

To demonstrate that IBSS can converge to a poor local optimum, consider the simulated example that is shown in Fig. 4(b), which consists of two change points in quick succession that cancel each other out (the means before and after the change points are the same). This example was created specifically to illustrate a limitation of the IBSS procedure: IBSS can introduce or update only one change point at a time, and every update is guaranteed to increase the objective, whereas in this example introducing one change point will make the fit worse. Consequently, when SuSiE is run from a null initialization, IBSS finds no change points and reports no credible sets.

This poor outcome represents a limitation of the IBSS algorithm, not a limitation of SuSiE or the variational approximation. To show this, we reran the IBSS algorithm, but initializing at a solution that contained the two true change points. This yielded a fit with two credible sets, each containing one of the correct change points. This also resulted in a much improved value of the objective function (-148.2 *versus* -181.8). Better algorithms for fitting SuSiE, or more careful initializations of IBSS, will be needed to address this shortcoming,

7. Discussion

We have presented a simple new approach to variable selection in regression. Compared with existing methods, the main benefits of our approach are its computational efficiency, and its ability to provide credible sets summarizing uncertainty in which variables should be selected. Our numerical comparisons demonstrate that for genetic fine mapping our methods outperform existing methods at a fraction of the computational cost.

Although our methods apply generally to variable selection in linear regression, further work may be required to improve performance in difficult settings. In particular, whereas the IBSS algorithm worked well in our fine mapping experiments, for change point problems we showed that IBSS may converge to poor local optima. We have also seen convergence problems in experiments with many effect variables (e.g. 200 effect variables out of 1000). Such problems may be alleviated by better initialization, e.g. by using fits from convex objective functions (e.g. the lasso) or from more sophisticated algorithms for non-convex problems (Bertsimas *et al.*, 2016; Hazimeh and Mazumder, 2018). More ambitiously, one could attempt to develop better algorithms to optimize the SuSiE variational objective function reliably in difficult cases. For example, taking smaller steps each iteration, rather than full co-ordinate ascent, may help.

At its core, SuSiE is based on adding up simple models (SERs) to create more flexible models (sparse multiple regression). This additive structure is the key to our variational approximations, and indeed our methods apply generally to adding up any simple models for which exact Bayesian calculations are tractable, not only SER models (on-line appendix B; algorithm 2). These observations suggest connections with both additive models and boosting (e.g. Friedman *et al.* (2000) and Freund *et al.* (2017)). However, our methods differ from most work on boosting in that each 'weak learner' (here, SER model) itself yields a model-averaged predictor. Other differences include our use of backfitting, the potential to estimate hyperparameters by maximizing an objective function rather than cross-validation, and the interpretation of our algorithm as a variational approximation to a Bayesian posterior. Although we did not focus on prediction accuracy here, the generally good predictive performance of methods based on model averaging and boosting suggest that SuSiE should work well for prediction as well as variable selection.

It would be natural to extend our methods to generalized linear models, particularly logistic

regression. In genetic studies with small effects, Gaussian models are often adequate to model binary outcomes (e.g. Pirinen *et al.* (2013) and Zhou *et al.* (2013)). However, in other settings this extension may be more important. One strategy would be to modify the IBSS algorithm directly, replacing the SER fitting procedure with a logistic or generalized linear model equivalent. This strategy is appealing in its simplicity, although it is not obvious what objective function the resulting algorithm is optimizing. Alternatively, for logistic regression one could use the variational approximations that were developed by Jaakkola and Jordon (2000).

For genetic fine mapping, it would also be useful to modify our methods to deal with settings where only summary data are available (e.g. the p univariate regression results). Many recent fine mapping methods deal with this (e.g. Chen *et al.* (2015), Benner *et al.* (2016) and Newcombe *et al.* (2016)) and ideas that are used by these methods can also be applied to SuSiE. Indeed, our software already includes preliminary implementations for this problem.

Beyond genetic fine mapping, one could consider applying SuSiE to related tasks, such as genetic prediction of complex traits and heritability estimation (Yang *et al.*, 2011). However, we do not expect SuSiE to provide substantial improvements over existing methods for these tasks. This is because, in general, the best existing approaches to these problems do not make strict sparsity assumptions on the effect variables; they allow for models in which many (or all) genetic variants affect the outcome (Meuwissen *et al.*, 2001; Moser *et al.*, 2015; Speed and Balding, 2014; Vilhjálmsson *et al.*, 2015; Zhou *et al.*, 2013). Nonetheless, it is possible that the ideas that were introduced here for sparse modelling could be combined with existing methods allowing non-sparse effects to improve prediction and heritability estimation, similarly to Zhou *et al.* (2013).

Finally, we are particularly interested in extending these methods to select variables simultaneously for multiple outcomes (*multivariate regression* and *multitask learning*). Joint analysis of multiple outcomes should greatly enhance power and precision to identify relevant variables (e.g. Stephens (2013)). The computational simplicity of our approach makes it particularly appealing for this complex task, and we are currently pursuing this direction by combining our methods with those from Urbut *et al.* (2018).

8. Data and resources

SuSiE is implemented in the R package `susieR` that is available from `https://github.com/stephenslab/susieR`. Source code and a website detailing the analysis steps for numerical comparisons and data applications are available from our manuscript resource repository (Wang *et al.*, 2020a), and also available from `https://github.com/stephenslab/susie-paper`.

Acknowledgements

We thank Kaiqian Zhang and Yuxin Zou for their substantial contributions to the development and testing of the `susieR` package. Computing resources were provided by the University of Chicago Research Computing Center. This work was supported by National Institutes of Health grant HG002585 and by a grant from the Gordon and Betty Moore Foundation (grant GBMF #4559).

References

Arnold, T. and Tibshirani, R. (2016) Efficient implementations of the generalized lasso dual path algorithm. *J. Computnl Graph. Statist.*, **25**, 1–27.

Barber, R. F. and Candès, E. J. (2015) Controlling the false discovery rate via knockoffs. *Ann. Statist.*, **43**, 2055–2085.
Benner, C., Spencer, C. C., Havulinna, A. S., Salomaa, V., Ripatti, S. and Pirinen, M. (2016) FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*, **32**, 1493–1501.
Bertsimas, D., King, A. and Mazumder, R. (2016) Best subset selection via a modern optimization lens. *Ann. Statist.*, **44**, 813–852.
Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017) Variational inference: a review for statisticians. *J. Am. Statist. Ass.*, **112**, 859–877.
Bottolo, L., Petretto, E., Blankenberg, S., Cambien, F., Cook, S. A., Tiret, L. and Richardson, S. (2011) Bayesian detection of expression quantitative trait loci hot spots. *Genetics*, **189**, 1449–1459.
Bottolo, L. and Richardson, S. (2010) Evolutionary stochastic search for Bayesian model exploration. *Baysn Anal.*, **5**, 583–618.
Carbonetto, P. and Stephens, M. (2012) Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Baysn Anal.*, **7**, 73–108.
Chen, W., Larrabee, B. R., Ovsyannikova, I. G., Kennedy, R. B., Haralambieva, I. H., Poland, G. A. and Schaid, D. J. (2015) Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics*, **200**, 719–736.
Chipman, H., George, E. I. and McCulloch, R. E. (2001) The practical implementation of Bayesian model selection. In *Model Selection* (ed. P. Lahiri), pp. 65–116. Beachwood: Institute of Mathematical Statistics.
Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.* B, **39**, 1–22.
Desboulets, L. D. D. (2018) A review on variable selection in regression analysis. *Econometrics*, **6**, no. 4, article 45.
Erdman, C. and Emerson, J. W. (2007) bcp: an R package for performing a Bayesian analysis of change point problems. *J. Statist. Softwr.*, **23**, 1–13.
Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. *Statist. Sin.*, **20**, 101–148.
Ferrari, D. and Yang, Y. (2015) Confidence sets for model selection by F-testing. *Statist. Sin.*, **25**, 1637–1658.
Fraser, D. A. S. (2011) Is Bayes posterior just quick and dirty confidence? *Statist. Sci.*, **26**, 299–316.
Freund, R. M., Grigas, P. and Mazumder, R. (2017) A new perspective on boosting in linear regression via subgradient optimization and relatives. *Ann. Statist.*, **45**, 2328–2364.
Friedman, J., Hastie, T. and Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. *Ann. Statist.*, **28**, 337–407.
Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Am. Statist. Ass.*, **76**, 817–823.
George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statist. Sin.*, **7**, 339–373.
GTEx Consortium (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.
Guan, Y. and Stephens, M. (2011) Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Ann. Appl. Statist.*, **5**, 1780–1815.
Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*, 2nd edn. New York: Springer.
Hazimeh, H. and Mazumder, R. (2018) Fast best subset selection: coordinate descent and local combinatorial optimization algorithms. *Preprint arXiv1803.01454*.
Heskes, T., Zoeter, O. and Wiegerinck, W. (2004) Approximate expectation maximization. In *Advances in Neural Information Processing Systems 16* (eds S. Thrun, L. K. Saul and B. Schölkopf), pp. 353–360. Cambridge: MIT Press.
Hoggart, C. J., Whittaker, J. C., De Iorio, M. and Balding, D. J. (2008) Simultaneous analysis of all SNPS in genome-wide and re-sequencing association studies. *PLOS Genet.*, **7**, article e1000130.
Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. and Eskin, E. (2014) Identifying causal variants at loci with multiple signals of association. *Genetics*, **198**, 497–508.
Huang, H., Fang, M., Jostins, L., Umićević Mirkov, M., Boucher, G., Anderson, C. A., Andersen, V., Cleynen, I., Cortes, A., Crins, F., D'Amato, M., Deffontaine, V., Dmitrieva, J., Docampo, E., Elansary, M., Farh, K. K.-H., Franke, A., Gori, A.-S., Goyette, P., Halfvarson, J., Haritunians, T., Knight, J., Lawrance, I. C., Lees, C. W., Louis, E., Mariman, R., Meuwissen, T., Mni, M., Momozawa, Y., Parkes, M., Spain, S. L., Théâtre, E., Trynka, G., Satsangi, J., van Sommeren, S., Vermeire, S., Xavier, R. J., Weersma, R. K., Duerr, R. H., Mathew, C. G., Rioux, J. D., McGovern, D. P. B., Cho, J. H., Georges, M., Daly, M. J., Barrett, J. C. and Barrett, J. C. (2017) Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, **547**, 173–178.
Huang, J., Breheny, P. and Ma, S. (2012) A selective review of group selection in high-dimensional models. *Statist. Sci.*, **27**, 481–499.
Jaakkola, T. S. and Jordan, M. I. (2000) Bayesian parameter estimation via variational methods. *Statist. Comput.*, **10**, 25–37.
Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999) An introduction to variational methods for graphical models. *Mach. Learn.*, **37**, 183–233.
Killick, R. and Eckley, I. (2014) changepoint: an R package for changepoint analysis. *J. Statist. Softwr.*, **58**, 1–19.

Lee, Y., Francesca, L., Pique-Regi, R. and Wen, X. (2018) Bayesian multi-SNP genetic association analysis: control of FDR and use of summary statistics. *Preprint bioRxiv 10.1101/316471*.
Leek, J. T. and Storey, J. D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genet.*, **3**, article e161.
Li, Y. I., van de Geijn, B., Raj, A., Knowles, D. A., Petti, A. A., Golan, D., Gilad, Y. and Pritchard, J. K. (2016) RNA splicing is a primary link between genetic variation and disease. *Science*, **352**, 600–604.
Logsdon, B. A., Hoffman, G. E. and Mezey, J. G. (2010) A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinform.*, **11**, article 58.
Mallat, S. and Zhang, Z. (1993) Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signl Process.*, **41**, 3397–3415.
Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M. M., Auton, A., Myers, S., Morris, A., Pirinen, M., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Hall, A. S., Hattersley, A. T., Hill, A. V. S., Mathew, C. G., Pembrey, M., Satsangi, J., Stratton, M. R., Worthington, J., Craddock, N., Hurles, M., Ouwehand, W., Parkes, M., Rahman, N., Duncanson, A., Todd, J. A., Kwiatkowski, D. P., Samani, N. J., Gough, S. C. L., McCarthy, M. I., Deloukas, P., Donnelly, P. and Donnelly, P. (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.*, **44**, 1294–1301.
Mandozzi, J. and Bühlmann, P. (2016) Hierarchical testing in the high-dimensional setting with correlated variables. *J. Am. Statist. Ass.*, **111**, 331–343.
Meinshausen, N. (2008) Hierarchical testing of variable importance. *Biometrika*, **95**, 265–278.
Meinshausen, N. and Bühlmann, P. (2010) Stability selection (with discussion). *J. R. Statist. Soc.* B, **72**, 417–473.
Meuwissen, T. H., Hayes, B. J. and Goddard, M. E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.
Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression. *J. Am. Statist. Ass.*, **83**, 1023–1032.
Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R. and Visscher, P. M. (2015) Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLOS Genet.*, **11**, article e1004969.
Neal, R. M. (1996) *Bayesian Learning for Neural Networks*. New York: Springer.
Neal, R. M. and Hinton, G. E. (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models* (ed. M. I. Jordan), pp. 355–368. New York: Springer.
Newcombe, P. J., Conti, D. V. and Richardson, S. (2016) JAM: a scalable Bayesian framework for joint analysis of marginal SNP effects. *Genet. Epidem.*, **40**, 188–201.
O'Hara, R. B. and Sillanpää, M. J. (2009) A review of Bayesian variable selection methods: what, how and which. *Baysn Anal.*, **4**, 85–117.
Olshen, A. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, **5**, 557–572.
Ott, J. (1999) *Analysis of Human Genetic Linkage*, 3rd edn. Baltimore: Johns Hopkins University Press.
Pati, D., Bhattacharya, A. and Yang, Y. (2018) On statistical optimality of variational Bayes. In *Proc. 21st Int. Conf. Artificial Intelligence and Statistics* (eds A. Storkey and F. Perez-Cruz), pp. 1579–1588. American Association for Artificial Intelligence.
Pickrell, J. K. (2014) Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am. J. Hum. Genet.*, **94**, 559–573.
Pirinen, M., Donnelly, P. and Spencer, C. C. A. (2013) Efficient computation with a linear mixed model on large-scale data sets with applications to genetic studies. *Ann. Appl. Statist.*, **7**, 369–390.
Renaux, C., Buzdugan, L., Kalisch, M. and Bühlmann, P. (2018) Hierarchical inference for genome-wide association studies: a view on methodology with software. *Computnl Statist.*, to be published.
Schaid, D. J., Chen, W. and Larson, N. B. (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.*, **19**, 491–504.
Servin, B. and Stephens, M. (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLOS Genet.*, **3**, 1296–1308.
Seshan, V. E. and Olshen, A. (2018) DNA copy: DNA copy number data analysis. *R Package Version 1.56.0*.
Sesia, M., Katsevich, E., Bates, S., Candès, E. and Sabatti, C. (2020) Multi-resolution localization of causal variants across the genome. *Nat. Communs*, **11**, article 1093.
Sillanpää, M. J. and Bhattacharjee, M. (2005) Bayesian association-based fine mapping in small chromosomal segments. *Genetics*, **169**, 427–439.
Spain, S. L. and Barrett, J. C. (2015) Strategies for fine-mapping complex traits. *Hum. Molec. Genet.*, **24**, R111–R119.
Speed, D. and Balding, D. J. (2014) MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.*, **24**, 1550–1557.
Stephens, M. (2000) Dealing with label switching in mixture models. *J. R. Statist. Soc.* B, **62**, 795–809.
Stephens, M. (2013) A unified framework for association analysis with multiple related phenotypes. *PLOS One*, **8**, article e65245.

Stephens, M. and Balding, D. J. (2009) Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.*, **10**, 681–690.
Taylor, J. and Tibshirani, R. J. (2015) Statistical learning and selective inference. *Proc. Natn. Acad. Sci. USA*, **112**, 7629–7634.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.
Tibshirani, R. J. (2014) Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.*, **42**, 285–323.
Urbut, S., Wang, G., Carbonetto, P. and Stephens, M. (2018) Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.*, to be published.
Veyrieras, J.-B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M. and Pritchard, J. K. (2008) High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLOS Genet.*, **4**, article e1000214.
Vilhjálmsson, B. J., Yang, J., Finucane, H. K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., Hayeck, T., Won, H.-H. and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015) Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.*, **97**, 576–592.
Wainwright, M. J. and Jordan, M. I. (2007) *Graphical Models, Exponential Families, and Variational Inference*. Boston: Now.
Wallace, C., Cutler, A. J., Pontikos, N., Pekalski, M. L., Burren, O. S., Cooper, J. D., García, A. R., Ferreira, R. C., Guo, H., Walker, N. M., Smyth, D. J., Rich, S. S., Onengut-Gumuscu, S., Sawcer, S. J., Ban, M., Richardson, S., Todd, J. A. and Wicker, L. S. (2015) Dissection of a complex disease susceptibility region using a Bayesian stochastic search approach to fine mapping. *PLOS Genet.*, **11**, article e1005272.
Wang, B. and Titterington, D. M. (2006) Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Baysn Anal.*, **1**, 625–650.
Wang, G., Sarkar, A., Carbonetto, P. and Stephens, M. (2019) A simple new approach to variable selection in regression, with application to genetic fine-mapping. *Preprint bioRxiv 10.1101/501114*.
Wang, G., Sarkar, A., Carbonetto, P. and Stephens, M. (2020a) Code and data accompanying this manuscript. (Available from `https://doi.org/10.5281/zenodo.2368676`.)
Wang, G., Sarkar, A., Carbonetto, P. and Stephens, M. (2020b) An animation illustrating the IBSS algorithm. (Available from `https://doi.org/10.6084/m9.figshare.11819997`.)
Wen, X., Lee, Y., Luca, F. and Pique-Regi, R. (2016) Efficient integrative multi-SNP association analysis via deterministic approximation of posteriors. *Am. J. Hum. Genet.*, **98**, 1114–1129.
Yang, J., Lee, S. H., Goddard, M. E. and Visscher, P. M. (2011) GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, **88**, 76–82.
Yekutieli, D. (2008) Hierarchical false discovery rate-controlling methodology. *J. Am. Statist. Ass.*, **103**, 309–316.
Zhou, X., Carbonetto, P. and Stephens, M. (2013) Polygenic modeling with Bayesian sparse linear mixed models. *PLOS Genet.*, **9**, article e1003264.
Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'A simple new approach to variable selection in regression, with application to genetic fine-mapping'.