JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2024, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2024.2347667

Controlled Discovery and Localization of Signals via Bayesian Linear Programming

Asher Spector[a] and Lucas Janson[b]

[a]Department of Statistics, Stanford University, Stanford, CA; [b]Department of Statistics, Harvard University, Cambridge, MA

ABSTRACT

Scientists often must simultaneously localize and discover signals. For instance, in genetic fine-mapping, high correlations between nearby genetic variants make it hard to identify the exact locations of causal variants. So the statistical task is to output as many disjoint regions containing a signal as possible, each as small as possible, while controlling false positives. Similar problems arise, for example, when locating stars in astronomical surveys and in changepoint detection. Common Bayesian approaches to these problems involve computing a posterior distribution over signal locations. However, existing procedures to translate these posteriors into credible regions for the signals fail to capture all the information in the posterior, leading to lower power and (sometimes) inflated false discoveries. We introduce Bayesian Linear Programming (BLiP), which can efficiently convert any posterior distribution over signals into credible regions for signals. BLiP overcomes an extremely high-dimensional and nonconvex problem to verifiably nearly maximize expected power while controlling false positives. Applying BLiP to existing state-of-the-art analyses of UK Biobank data (for genetic fine-mapping) and the Sloan Digital Sky Survey (for astronomical point source detection) increased power by 30%–120% in just a few minutes of additional computation. BLiP is implemented in `pyblip` (Python) and `blipr` (R). Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY
Received January 2023
Accepted April 2024

KEYWORDS
Astronomical source detection; Genetic fine-mapping; Signal detection

1. Introduction

1.1. Motivation

Localizing signals is an important statistical task across disciplines. For example, consider the problem of controlled variable selection: given variables $X_1, \ldots, X_p$, analysts seek to identify a few key variables which impact an outcome Y. Here, $X_1, \ldots, X_p$ could represent genetic mutations or demographic data, and Y could represent a disease status or economic outcome. However, when variables are highly correlated, it can be very challenging to certify that any individual variable is important. For example, if X_1 and X_2 are nearly perfectly correlated, analysts may not be able to distinguish between them even if the data make clear that $\{X_1, X_2\}$ contains at least one important variable. Yet selecting different variables may lead to qualitatively different scientific conclusions, such as in genetic studies, where intervening on a causal variant (e.g., X_1) could help cure a disease, whereas intervening on its highly correlated neighbor (e.g., X_2) could have no impact at all. As a result, analysts often must contend with significant uncertainty in which variables ought to be selected. Unfortunately, as observed by Wang et al. (2020), many modern variable selection methods cannot accomplish this task.

More generally, analysts in many settings may be able to tell that a signal exists without perfectly localizing it. Indeed, astronomers often can guarantee that a light source exists somewhere in a region of space without knowing its exact location. Similarly, economists may know that a time series has changed without knowing precisely when it did so. Lastly, this problem is particularly relevant in *genetic fine-mapping*, where researchers attempt to identify genetic variants which cause outcomes such as cardiovascular disease. Indeed, genetic variants are highly locally correlated—for example, this article analyzes a UK Biobank dataset where over 40% of genetic variants are at least 99% correlated with a nearby variant. In these settings, analysts often want to localize signals as precisely as possible, that is, to make statements like "there is a signal in this region, even if we do not know exactly where it is." These discovered regions should be as small as possible to yield precise scientific insights, all while controlling the false discovery rate (FDR) to ensure findings are replicable, interpretable, and do not waste future resources. Namely, we classify a discovered region as a false discovery if it contains no signals, and we require that the expected proportion of false discoveries is at most $q \in (0, 1)$ (see Section 2 for a mathematical definition).

With this motivation, our article introduces a novel procedure, called *Bayesian Linear Programming* (BLiP), which takes a posterior distribution over signals as an input and uses it to localize signals as precisely as possible while controlling false positives. Before describing our contribution, however, we pause to survey related literature.

1.2. Related Literature

1. Frequentist methods. There is an enormous literature on frequentist analysis of spatially distributed signals. However, most methods are either tailored to a specific application or solve a

CONTACT Asher Spector aspector@stanford.edu Department of Statistics, Stanford University, Stanford, CA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

different problem than we do. For example, there is a large literature on hierarchical testing (e.g., Meinshausen 2008; Goeman and Solari 2012; Mandozzi and Bühlmann 2016; Renaux et al. 2018; Bogomolov et al. 2020), but only two methods (Yekutieli 2008; Katsevich, Sabatti, and Bogomolov 2021) control the FDR as defined in Section 2, which we focus on because it is a standard error rate in our real applications. These two methods can be powerful, but they require computing a very large number of p-values, making them computationally expensive in some cases. Furthermore, they either assume independence of the p-values (Yekutieli 2008) or make conservative assumptions (Katsevich, Sabatti, and Bogomolov 2021); in our simulations, this leads to FDR control violations and power loss, respectively.

Similarly, there is an important literature on post-hoc simultaneous inference (e.g., Goeman and Solari 2011; Katsevich and Ramdas 2020; Rosenblatt et al. 2018; Goeman et al. 2019; Blanchard, Neuvial, and Roquain 2020), which can be used to output a set of disjoint regions which each contain at least one signal with high probability. Yet to the best of our knowledge, all computationally tractable methods in this literature apply on top of p-values for individual locations, that is, p-values testing "is there a signal at location ℓ?" Unfortunately, this causes a very large loss of power in our setting. For example, consider a linear regression of $\{X_1, X_2\}$ on Y where X_1, X_2 and Y are all nearly perfectly correlated. In this case, due to collinearity, the individual p-values for X_1 and X_2 will not be significant, and existing post-hoc approaches will be powerless; however, an F-test testing whether *either* X_1 or X_2 influences Y may be highly significant (see Appendix A for a concrete example). While some closed testing methods can leverage (e.g.) F-test p-values, these methods require computing $O(2^{|\mathcal{L}|})$ p-values, making them impractical at scale (see Appendix B for discussion). Furthermore, to our knowledge, no existing post-hoc methods control the FDR as defined in Section 2. That said, post-hoc bounds can also be applied to find regions where (e.g.) 95% of the region is a signal. This is useful when signals are clustered together, for example, in neuroscience. However, in our setting (e.g., genetics), the object of inference is fundamentally different: signals are sparse and not necessarily clustered together, so analysts do not aim to discover "clusters" of signals; rather, they seek to isolate individual signals as precisely as possible. Thus, this literature solves a different problem than the one we consider. See Appendix B for further discussion.

There are several other existing frequentist methods that solve related, but distinct, problems from the one in this article. For brevity, we give a further review in Appendix B.

2. Bayesian approaches. Many Bayesian works discuss how to compute or approximate the posterior distribution over signal locations (see Brooks et al. 2011; Blei, Kucukelbir, and McAuliffe 2017 for review). For example, our work builds on methods for approximating the posterior law of the regression coefficients in sparse Bayesian regression (Mitchell and Beauchamp 1988; Albert and Chib 1993; George and McCulloch 1997; Wang et al. 2020; Shin and Liu 2021), many of which are commonly used in genetic fine-mapping (e.g., Guan and Stephens 2011; Carbonetto and Stephens 2012; Benner et al. 2016; Lee et al. 2018; Weissbrod et al. 2020). However, a high-dimensional posterior distribution is not directly interpretable. Indeed, even after computing the posterior distribution, localizing signals can still be difficult because there are combinatorially many regions which could contain signals, making it hard to identify the smallest regions which each contain at least one signal with high probability. Thus, our work asks the question: given a posterior distribution over signal locations, how can we output a set of disjoint regions which (a) each contain a signal with high probability, (b) are as small as possible, and (c) are as numerous as possible?

To our knowledge, only a small number of prior works have addressed this question. Our work is perhaps closest in spirit to that of Wang et al. (2020), who introduced "SuSiE," a method for sparse Bayesian linear regression. SuSiE localizes signals via an iterative Bayesian stepwise selection (IBSS) algorithm, which, roughly speaking, sequentially creates a credible region for the signal with the largest signal size and then proceeds to the next largest signal, and so on (see Appendix E.3 for a more detailed review). This procedure is equivalent to (i) using a novel variational approximation (which is accurate when the number of signals is small) to approximate the posterior distribution of the signals and (ii) then greedily processing that posterior to localize signals. In contrast, our method BLiP performs only the latter task, but it can do so on any posterior. For example, it is not clear how to apply SuSiE to astronomical point source detection problems, but BLiP can easily wrap around other Bayesian methods in this field (see Section 5.2). Additionally, when appropriate, BLiP can apply directly to the posterior obtained from SuSiE, as we show in Section 4, where BLiP uniformly improves SuSiE's power. Indeed, even when SuSiE's variational approximation is inaccurate, BLiP can often partially correct this issue, leading to improved power and sometimes improved FDR control (see Section 4). (Note that BLiP can also apply on top of the refined SuSiE procedure suggested in Zou et al. 2021). To our knowledge, the only other comparable Bayesian method is DAP-G (Lee et al. 2018), which also requires a specific approximation to the posterior to localize signals. In principle, BLiP can also wrap on top of DAP-G to improve its power, although we did not explore this possibility because SuSiE outperformed DAP-G. Alternatively, BLiP can be combined with any other method to approximate the posterior, for example, MCMC, and thus BLiP offers an attractive alternative to perform resolution-adaptive inference without (necessarily) making any variational approximation.

1.3. Contribution

Our key contribution is to introduce *Bayesian Linear Programming* (BLiP), a method for performing resolution-adaptive signal detection. As an input, BLiP takes a posterior distribution over the location of the signals (defined formally in Section 2). For example, if $Y \mid X$ follows a generalized linear model (GLM) where nonzero coefficients are signals, one can use a Markov chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution of the model coefficients and use the MCMC samples as the input for BLiP. Thus, BLiP can accurately be described as a type of post-processing on the posterior (albeit with appealing statistical guarantees). As we shall see in Section 5, this "post-processing" can dramatically improve the

power and calibration of applied analyses in settings ranging from GWAS to astronomical point-source detection. Indeed, to quote Wang et al. (2020), "the output from Bayesian Variable Selection methods is typically a complex posterior distribution, and this can be difficult to distill into results that are easily interpretable." BLiP is designed to solve exactly this problem (in the more general signal detection setting).

Given this input, BLiP will output a set of disjoint regions, each containing a signal, which maximizes a natural measure of power (defined in Section 2) while controlling false positives. For example, in variable selection problems, BLiP will return a set of disjoint groups of variables so that (a) nearly all groups contain at least one signal variable, (b) we discover as many groups as possible, and (c) the groups are as small as possible.

We now highlight a few attractive features of BLiP. First, BLiP is often much more powerful than other methods (where competitors exist), as we demonstrate in simulations and two real data analyses. Indeed, BLiP is verifiably nearly optimal in the sense that one can compare its power to an upper bound on the achievable expected power, and these quantities were indistinguishable in all our analyses. Second, given a correct posterior, BLiP can provably control one of several error rates, including the FDR, familywise error rate (FWER), and local FDR. Third, since BLiP acts directly on a posterior distribution, it can be applied on top of any Bayesian model or algorithm, allowing analysts to leverage arbitrary advances in Bayesian MCMC or variational inference. Finally, although computing the posterior over signal locations may be expensive, BLiP itself is extremely computationally efficient, allowing it to search over billions of candidate regions to find a near-optimal set of discoveries.

2. Problem Statement

We now introduce the problem of resolution-adaptive signal detection. To start, let $\mathcal{L}$ denote a set of locations at which there may be signals and let $\mathcal{S} \subset \mathcal{L}$ denote the true (unknown) set of signals. It may be helpful to keep the following two examples in mind.

Example *1* (*Variable selection in regression*). Suppose we observe variables $X \in \mathbb{R}^p$ and a response $Y \in \mathbb{R}$, and we seek to discover "important" variables. Here, the locations $\mathcal{L} = \{1, \dots, p\}$ represent $X_1, \dots, X_p$, and the signals $\mathcal{S}$ are the set of "important" variables. For example, if Y depends on X through linear coefficients $\beta \in \mathbb{R}^p$, we set $\mathcal{S} = \{\ell \in [p] : \beta_\ell \neq 0\}$. Note when $(X_1, \dots, X_p)$ are highly correlated, we may not be able to discover individual signal variables with confidence. However, for a group $G \subset \mathcal{L}$ of highly correlated variables, we may have power to discover that at least one variable in G is important.

Example *2* (*Point source detection*). Astronomers often seek to locate point sources (e.g., stars) in the night sky. Here, the locations $\mathcal{L}$ represent a region of the sky, so $\mathcal{L} \subset \mathbb{R}^2$ is a continuous (infinite) set, and $\mathcal{S} \subset \mathcal{L}$ denotes the true set of sources. Since most images have blur, it is difficult to identify the exact location of a source. This motivates a resolution-adaptive approach, where we output regions $G_1, \dots, G_R$ which each contain a source with high confidence and are as small as possible.

We take a Bayesian approach and assume that the analyst has a prior on $\mathcal{S}$ and a model for the data $\mathcal{D}$; however, our method applies to any choice of model and prior with a well-defined set of signals. We do require that the model and prior are sufficiently tractable such that the analyst can compute or well-approximate the posterior law of $\mathcal{S} \mid \mathcal{D}$.

Requirement 2.1. The analyst can compute the posterior distribution of $\mathcal{S} \mid \mathcal{D}$.

Computing the law of $\mathcal{S} \mid \mathcal{D}$ is not easy, but an immense amount of literature has studied this problem (see Section 1.2). For example, in sparse regression problems (Example 1) following, for example, a two-groups model (Efron 2008), one can sample from $\mathcal{S} \mid \mathcal{D}$ by sampling from the posterior of the coefficients β, which is a well-studied task (Brooks et al. 2011). Choosing a good method to compute $\mathcal{S} \mid \mathcal{D}$ is very important, although it is a domain-specific problem; Sections 3.3, 4 and Appendix E review general guidelines which help ensure robustness to misspecification and convergence issues. However, BLiP can wrap around any such method. Thus, this choice is orthogonal to our contribution.

Based on the posterior $\mathcal{S} \mid \mathcal{D}$, we aim to output a disjoint set of regions $G_1, \dots, G_R \subset \mathcal{L}$, where any group $G \subset \mathcal{L}$ is a true discovery if it contains at least one signal, that is, $G \cap \mathcal{S} \neq \emptyset$. Our goal is to maximize true discoveries subject to false positive control. However, it is not obvious how to count the number of true discoveries, because discovering a large region G only asserts that at least one signal exists in G; as a result, large discovered regions are less valuable than small discovered regions. For example, in genetic fine-mapping, discovering $G_0 = \{\ell_1\}$ identifies ℓ_1 as a causal variant, whereas discovering $G_1 = \{\ell_1, \ell_2\}$ only asserts that at least one of ℓ_1, ℓ_2 is a causal genetic variant, which provides strictly less information. Of course, if ℓ_1 and ℓ_2 are highly correlated, it is much easier to discover G_1 than G_0 because it is hard to determine which of $\{\ell_1, \ell_2\}$ is the causal effect. (Note this logic holds even if ℓ_1 is a signal and ℓ_2 is not—see Appendix A for a concrete example.)

To resolve this ambiguity, we suggest weighting discoveries to prioritize discovering smaller groups, so that, for example, discovering a group of size 3 counts as "fewer" discoveries than discovering a group of size 2. For example, one proposal from Mandozzi and Bühlmann (2016) is to assign a discovered region of size m weight $\frac{1}{m}$, so that discovering a region of size 1 and a region of size 2 would count as 1.5 discoveries total. Formally, let $2^{\mathcal{L}}$ denote the set of all subsets of $\mathcal{L}$. For any weighting function $w : 2^{\mathcal{L}} \to \mathbb{R}$ and discoveries $G_1, \dots, G_R$, we define the *Resolution-Adjusted number of True Positives* ($\mathrm{TP_{RA}}$) as the sum of the weights associated with each true discovery:

$$\mathrm{TP_{RA}}(G_1, \dots, G_R) \triangleq \sum_{r=1}^{R} w(G_r)\mathbb{I}(G_r \cap \mathcal{S} \neq \emptyset), \tag{1}$$

where we remind the reader that $\mathbb{I}(G_r \cap \mathcal{S} \neq \emptyset)$ is the indicator that G_r is a true discovery. Colloquially, when we say that a method is "powerful," we mean that it has high expected $\mathrm{TP_{RA}}$. Following Mandozzi and Bühlmann (2016), we will argue that $w(G) = \frac{1}{|G|}$ is a good default choice, but first we give a formal problem statement.

Definition 2.1 (*Resolution-adaptive signal detection*). Suppose we seek to discover signals among locations $\mathcal{L}$. Let $R \geq 0$ be the number of discoveries and let $G_1, \ldots, G_R \subset \mathcal{L}$ denote the discovered regions, so R and $G_1, \ldots, G_R$ are our optimization variables. For a weighting function $w : 2^{\mathcal{L}} \to \mathbb{R}$, we seek to maximize expected TP_{RA} subject to FDR control:

$$\max_{R \geq 0, G_1, \ldots, G_R} \quad \mathbb{E}[\text{TP}_{\text{RA}}(G_1, \ldots, G_R) \mid \mathcal{D}] \tag{2}$$

$$\text{s.t.} \quad \text{FDR} \triangleq \mathbb{E}\left[\frac{\#\{1 \leq r \leq R : G_r \cap \mathcal{S} = \emptyset\}}{\max(1, R)} \mid \mathcal{D}\right] \leq q, \tag{3}$$

$$G_1, \ldots, G_R \subset \mathcal{L} \text{ are disjoint.} \tag{4}$$

Note that above, all expectations are taken over the posterior law of $\mathcal{S} \mid \mathcal{D}$.

Remark 1 (*Disjointness*). We constrain $G_1, \ldots, G_R$ to be disjoint to improve interpretability and prevent double-counting. For example, discovering $\{\ell_1\} \subset \mathcal{L}$ makes discovering $\{\ell_1, \ell_2\} \subset \mathcal{L}$ logically redundant, so these should not count as two separate discoveries.

Remark 2 (*Terminology*). We use the words "region" and "group" interchangeably. Both refer to an arbitrary, possibly non-contiguous subset $G \subset \mathcal{L}$ of the locations.

Remark 3 (*Error rate*). We focus on the FDR because it is a popular and appealing error rate, but this problem is still well-defined if we replace the FDR with another error rate. Indeed, BLiP can also control (e.g.) the FWER or local FDR (see Appendix C.2).

Remark 4 (*Default Weight Function*). BLiP can optimize for *any* weight function, but by default, we suggest choosing $w(G) = |G|^{-1}$, because this choice is simple, interpretable, and it reflects the intuition that discovering a region of size m gives roughly m times less information than discovering an individual signal. Indeed, this choice has been used in recent papers (Mandozzi and Bühlmann 2016; Buzdugan et al. 2016; Renaux et al. 2018; Guo et al. 2021). Of course, in variable selection problems, one may wonder if $w(G)$ should account for the correlation structure of X_G. For example, if X_1 and X_2 are perfectly correlated, does it still make sense to count $G = \{X_1, X_2\}$ as only half a discovery? By default, we argue that the answer is yes when X_1 and X_2 represent distinct scientific hypotheses. For example, in genetic fine-mapping, if X_1 is causal and X_2 is not, then discovering X_1 could help develop a drug, whereas discovering X_2 is a false positive, no matter the correlation between X_1 and X_2. Lastly, we note that BLiP is not too sensitive to the precise definition of $w(G)$; for example, Appendix F.8 shows empirically that using (e.g.) $w(G) = \frac{1}{\log_2(|G|)+1}$ yields similar results. That said, there are settings where this default choice is not ideal. For example, in our astronomical application, we apply BLiP using two different choices of w to optimize for two different scientific objectives.

Remark 5 (*Why adaptivity is important*). A simplification of this problem would be to fix a prespecified partition $G_1, \ldots, G_m \subset \mathcal{L}$ and test whether a signal exists in each of $G_1, \ldots, G_m$. However, this nonadaptive approach will not optimally localize signals, because the best choice of partition $G_1, \ldots, G_m$ depends on the unknown data-generating process. Informally, when the "signal size" is large, we may be able to perfectly localize individual signals, whereas we may only be able to detect that a weak signal exists somewhere in a relatively large region. Indeed, in Appendix A, we give a concrete example of a regression problem where the best partition depends on the unknown relationship between Y and X. In contrast, resolution-adaptive methods can use the data to discover regions $G_1, \ldots, G_R$ which are as small as possible. Of course, for computational reasons, it is not possible to consider *every* region $G \subset \mathcal{L}$ as a potential discovery. However, we expect that methods which are more adaptive, meaning they can use the data to choose from among a larger set of candidate regions, will perform better.

3. Bayesian Linear Programming

3.1. Bayesian Linear Programming for FDR Control

We now introduce BLiP. We focus on controlling the FDR, although Appendix C.2 also considers the FWER, local FDR, and per-family error rate (PFER).

While Problem 2.1 may seem intractable, it turns out that high quality solutions can be found via a convex relaxation. That said, it is still too computationally challenging to search over all regions $G \subset \mathcal{L}$, since there are combinatorially many subsets of $\mathcal{L}$. To narrow the search space, we require that the discovered regions are members of a set of *candidate regions*, $\mathcal{G}$. This requirement is not particularly restrictive, since one can make $\mathcal{G}$ as large as is computationally feasible, and our algorithm can handle billions of candidate regions. For example, in variable selection problems, one can cluster the variables $X_1, \ldots, X_p$ using (literally) a thousand different clustering algorithms and let $\mathcal{G}$ equal the union of the clusters created by the algorithms. We offer more suggestions for constructing $\mathcal{G}$ in Section 3.2.

Given candidate regions $\mathcal{G}$, the first key observation is that maximizing expected TP_{RA} corresponds to maximizing a linear function. To see this, let $p_G = \mathbb{P}(G \cap \mathcal{S} \neq \emptyset \mid \mathcal{D})$ be the posterior probability that there is a signal in region G, also known as a posterior inclusion probability (PIP). PIPs are similar to the "local true discovery rate" in Efron's two-group model (Efron 2008), and they are easily computable if we have access to the posterior distribution of $\mathcal{S} \mid \mathcal{D}$ as per Requirement 2.1. For example, if $\mathcal{S}^{(1)}, \ldots, \mathcal{S}^{(N)}$ denote N ergodic samples from the law of $\mathcal{S} \mid \mathcal{D}$ from an appropriate MCMC algorithm, then

$$p_G \approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\mathcal{S}^{(i)} \cap G \neq \emptyset), \tag{5}$$

where this approximation becomes exact as $N \to \infty$. The choice of MCMC algorithm is a domain-specific question which we will discuss further in Sections 3.3–5, although this choice is orthogonal to our contribution. Thus, for now, we assume that we know the PIPs—later, Section 4 will study whether BLiP is robust to PIPs which are only approximately correct.

With this notation, let $x_G \in \{0, 1\}$ be the indicator of whether our procedure discovers region G. Here, $\{x_G\}_{G \in \mathcal{G}}$ are our optimization variables which determine our discovery set

$\mathcal{G}_{\text{disc}} \triangleq \{G \in \mathcal{G} : x_G = 1\} \subset \mathcal{G}$. This notation plus the definition of TP_{RA} yields

$$\mathbb{E}[\text{TP}_{\text{RA}}(\mathcal{G}_{\text{disc}}) \mid \mathcal{D}] \triangleq \mathbb{E}\left[\sum_{G \in \mathcal{G}_{\text{disc}}} w(G)\mathbb{I}(G \cap \mathcal{S} \neq \emptyset) \mid \mathcal{D}\right] = \sum_{G \in \mathcal{G}} p_G w(G) x_G, \tag{6}$$

where the last equality sums over $\mathcal{G}$ because $x_G = 1$ if and only if G is discovered. In other words, the objective is a linear function of $\{x_G\}_{G \in \mathcal{G}}$. Notably, the FDR constraint can also be formulated as a linear constraint. In particular, let $V = \sum_{G \in \mathcal{G}_{\text{disc}}} \mathbb{I}(G \cap \mathcal{S} = \emptyset)$ be the number of false discoveries and let $R = |\mathcal{G}_{\text{disc}}|$ denote the number of discoveries. Controlling the FDR at level q requires that

$$\text{FDR} \triangleq \mathbb{E}\left[\frac{V}{R} \mid \mathcal{D}\right] = \frac{\mathbb{E}[V \mid \mathcal{D}]}{R} = \frac{\sum_{G \in \mathcal{G}}(1 - p_G)x_G}{\sum_{G \in \mathcal{G}} x_G} \leq q, \tag{7}$$

where in the above equation we use the convention that $0/0 = 0$. Multiplying by $\sum_{G \in \mathcal{G}} x_G$ on both sides yields the linear constraint $\sum_{G \in \mathcal{G}}(1 - p_G - q)x_G \leq 0$. Thus, as stated below, the resolution-adaptive signal detection problem can be formulated as an integer linear program (LP). See Appendix C.1 for a proof. Naturally, the same result holds for any error rate which can be expressed as linear constraints on $\{x_G\}_{G \in \mathcal{G}}$ (see Appendix C.2).

Proposition 3.1. The solution to the resolution-adaptive signal detection problem in Definition 2.1 is the same as the solution to the following integer LP:

$$\max_{\{x_G\}_{G \in \mathcal{G}}} \quad \sum_{G \in \mathcal{G}} p_G w(G) x_G \tag{8}$$

$$\text{s.t.} \quad \sum_{G \in \mathcal{G}} (1 - p_G - q)x_G \leq 0, \tag{9}$$

$$\sum_{G \in \mathcal{G}: \ell \in G} x_G \leq 1 \quad \forall \ell \in \mathcal{L}, \tag{10}$$

$$x_G \in \{0, 1\} \quad \forall G \in \mathcal{G}. \tag{11}$$

For simplicity, we now assume $\mathcal{L}$ and $\mathcal{G}$ are finite sets (as in genetics), and thus (8)–(11) is finite-dimensional. That said, when $\mathcal{G}$ and $\mathcal{L}$ are infinite sets (e.g., Example 2), one can efficiently reduce (8)–(11) to an equivalent finite-dimensional problem assuming (i) the expected number of signals is finite and (ii) a mild regularity condition on $\mathcal{G}$. Intuitively, this is because when there are only finitely many signals, only finitely many regions have non-negligible PIPs. For brevity, we discuss this in Appendix D.

Integer LPs are NP complete but well studied (Jünger et al. 2010), so when $|\mathcal{L}|$ and $|\mathcal{G}|$ are small, it may be possible to directly solve the problem in Proposition 3.1. However, this naive approach is usually too expensive. Thus, we suggest two strategies to improve efficiency.

Strategy 1: Adaptive preprocessing. After observing the data, we can often tell that many locations and candidate regions almost certainly do not contain a signal, and thus we can discard them. Formally, let $\mathcal{L}_0 = \{\ell \in \mathcal{L} : p_{\{\ell\}} > 0.01\}$ denote the set of locations with at least a 1% chance of being a signal, and let $\mathcal{G}_0 = \{G \subset \mathcal{G} : p_G > 0.01, G \subset \mathcal{L}_0\}$ denote the set of regions with at least a 1% chance of containing a signal; we recommend replacing $\mathcal{L}$ with $\mathcal{L}_0$ and $\mathcal{G}$ with $\mathcal{G}_0$ in (8)–(11). This approach can improve computation by multiple orders of magnitude, since when the signals are sparse, most regions have a low posterior probability of containing a signal and can be discarded. This common sense heuristic should have almost no impact on the final discovery set, since (e.g.) if a region G has only a 0.001% chance of containing a signal, we almost certainly would not have discovered it anyway. Indeed, unlike methods which restrict $\mathcal{G}$ a priori, adaptive preprocessing does not sacrifice adaptability, since it uses the full data to prune $\mathcal{G}$.

Strategy 2: LP relaxation. After adaptive preprocessing, we recommend approximately solving the integer LP (8)-(11) by first solving the *relaxed problem* which replaces the integer constraint $x_G \in \{0, 1\}$ with the relaxed constraint $x_G \in [0, 1]$. The relaxed problem is a simple LP with sparse constraints, so it can be solved efficiently using standard software even when $|\mathcal{G}_0|$ and $|\mathcal{L}_0|$ have millions of elements (Boyd and Vandenberghe 2004). For example, our python and R implementations use the "CBC" solver in the packages `cvxpy` and `cvxr`, respectively. Although the relaxed LP may return a non-integer solution set $\{x^\star_G\}_{G \in \mathcal{G}_0}$, empirically, the solutions $\{x^\star_G\}_{G \in \mathcal{G}_0}$ are usually composed almost entirely of integers, making it easy to "post-process" $\{x^\star_G\}_{G \in \mathcal{G}_0}$ to obtain a fully integer solution. While we cannot prove that this will always happen, we now give some intuition to explain this phenomenon based on the properties of *knapsack problems*.

Definition 3.1. A knapsack problem with variables $z_1, \ldots, z_m$ is an integer LP of the form

$$\max_{z_1, \ldots, z_m \in \{0,1\}} \sum_{i=1}^m a_i z_i \text{ such that } \sum_{i=1}^m b_i z_i \leq c \tag{12}$$

for $a_1, \ldots, a_m, b_1, \ldots, b_m, c \in \mathbb{R}$.

The integer LP in Proposition 3.1 is a knapsack problem with additional *sparse* constraints, where $\{x_G\}_{G \in \mathcal{G}}$ correspond to $\{z_i\}_{i \in [m]}$, $\{w(G)p_G\}_{G \in \mathcal{G}}$ correspond to $\{a_i\}_{i \in [m]}$, and $\{(1 - p_G - q)\}_{G \in \mathcal{G}}$ correspond to $\{b_i\}_{i \in [m]}$. The only difference is that Proposition 3.1 enforces the extra constraints in (10). However, relaxed knapsack problems with added sparse constraints are known to admit solutions which are largely composed of integers (see Yang and Bulfin 2009; Scatamacchia 2017; also Appendix C.3 for further intuition). We are not aware of any theory for the specific problem in Proposition 3.1, but this partially explains why the relaxed LP admits a mostly integer solution. Indeed, Figure 1 shows empirically that the relaxed LP typically outputs a single-digit number of non-integer values even when $|\mathcal{G}| > 50,000$.

Introducing BLiP. Algorithm 1 defines BLiP for FDR control. After adaptive preprocessing, BLiP solves the relaxed LP of Proposition 3.1. As discussed above, this typically yields a solution $\{x^\star_G\}_{G \in \mathcal{G}_0}$ with only a few non-integer values, denoted $\mathcal{H} = \{G \in \mathcal{G}_0 : x^\star_G \notin \{0, 1\}\}$. To obtain an integer solution, BLiP solves the integer LP (8)-(11) while holding the values of $\{x^\star_G : G \in \mathcal{G}_0 \setminus \mathcal{H}\}$ constant and only optimizing over $\{x_G : G \in \mathcal{H}\}$. Typically, this integer LP has only a few variables, so it runs

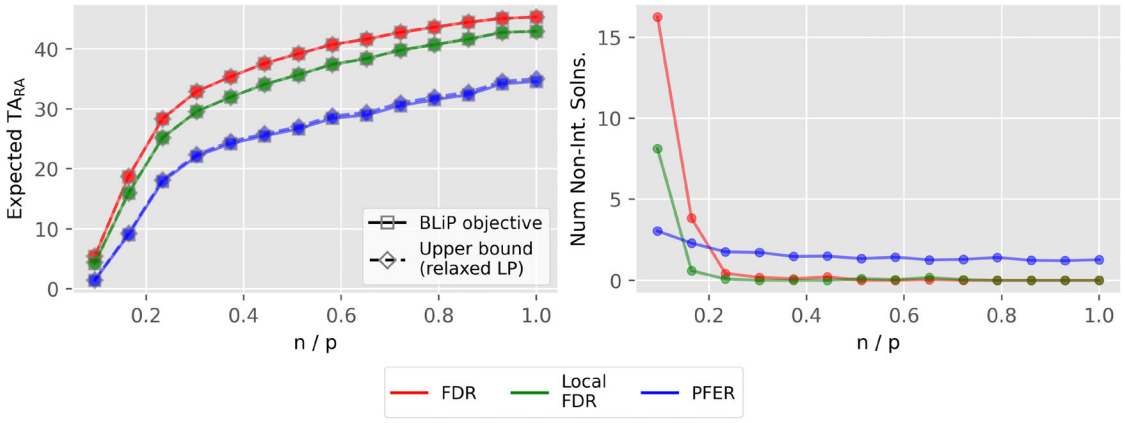

Figure 1. In a regression problem with $p = 1000$ features, 50 signals, and $>$ 50,000 candidate regions, the left plot shows the objective function (expected TP_{RA}) achieved by BLiP and the upper bound from the relaxed LP, which are almost indistinguishable. The right plot shows the number of non-integer solutions to the relaxed LP. We applied BLiP on top of PIPs from a standard Gibbs sampler for sparse regression problems as detailed in Section 4. See Appendices F.1 and C.3 for simulation details and an analogous plot for the FWER.

Algorithm 1 BLiP for FDR control.

Input: Candidate regions $\mathcal{G}$, PIPs $\{p_G\}_{G\in\mathcal{G}}$, a weighting function w, a nominal level q.

1: *Adaptive preprocessing*: Replace $\mathcal{L}$ with $\mathcal{L}_0 \triangleq \{\ell \in \mathcal{L} : p_{\{\ell\}} > \epsilon\}$ and $\mathcal{G}$ with $\mathcal{G}_0 \triangleq \{G \in \mathcal{G} : \mathcal{G} \subset \mathcal{L}_0, p_G > \epsilon\}$ for some small ϵ, for example, $\epsilon = 0.01$.

2: *LP relaxation*: Solve the relaxed variant of (8)–(11) to obtain a solution set $\{x^\star_G\}_{G\in\mathcal{G}}$. Let $\mathcal{H} = \{G \in \mathcal{G} : x^\star_G \notin \{0, 1\}\}$ denote the non-integer solutions.

3: *Convert to integers*: Fix $x_G = x^\star_G$ for $G \in \mathcal{G} \setminus \mathcal{H}$ and run the integer LP (8)–(11) on the remaining variables $\{x_G : G \in \mathcal{H}\}$, yielding (integer) solutions $\{x^{\star\star}_G : G \in \mathcal{G}\}$.

4: *Ensure feasibility*: If the integer LP in Step 3 is feasible, detect signals in $\mathcal{G}_{\text{disc}} \triangleq \{G : x^{\star\star}_G = 1\}$ and terminate.
Else, define $G_{\min} \triangleq \arg\min_{\{G:x^\star_G=1\}} p_G$, set $\mathcal{H} = \mathcal{H} \cup G_{\min}$, and return to Step 3.

practically instantly, yielding integer solutions $\{x^{\star\star}_G : G \in \mathcal{G}_0\}$. Then, BLiP outputs $\mathcal{G}_{\text{disc}} \triangleq \{G \in \mathcal{G}_0 : x^{\star\star}_G = 1\}$.

Technically, up to two things could go wrong with this algorithm. First, if the final integer LP is large, it may be challenging to solve efficiently. However, the user will know this in advance and can use polynomial-time heuristic methods instead (see Appendix C.4). Second, it is technically possible for the integer LP to be infeasible, in which case we propose a backtracking algorithm that iteratively finds the group $G_{\min}$ in $\{G : x^\star_G = 1\}$ with the smallest PIP and adds $G_{\min}$ to $\mathcal{H}$. This guarantees that BLiP can find a feasible solution, because after $|\{G : x^\star_G = 1\}|$ steps, it is possible for the integer LP to set $x_G = 0$ for all G, which is always feasible. That said, neither of these two phenomena ever occurred in any of our analyses, despite our applying BLiP thousands of times in large-scale settings. This suggests that these modifications are only required in pathological examples. Even in pathological cases, however, the output of BLiP is always a feasible solution to (8)–(11) and thus provably controls the FDR by Proposition 3.1.

To aid intuition, note that when backtracking is not required (as in all of our simulations and applications), BLiP outputs $\{G \in \mathcal{G}_0 : x^\star_G = 1\}$, the regions selected by the relaxed LP, plus a few regions from $\mathcal{H}$, the non-integer solutions from the relaxed LP.

The overall runtime of BLiP is dominated by a single large sparse linear program, whose computational complexity is at most $O(|\mathcal{G}_0|^2|\mathcal{L}_0|)$ (Boyd and Vandenberghe 2004). In practice, LP solvers may be much faster than their worst-case performance; usually, it is possible to solve LPs with millions of variables (Boyd and Vandenberghe 2004). In Sections 4–5, we find that BLiP is always less expensive than computing its input PIPs.

We now discuss the claim that BLiP finds "nearly" the optimal set of discoveries among $\mathcal{G}_0$. This is because the relaxed LP solution $\{x^\star_G\}_{G\in\mathcal{G}_0}$ is usually almost entirely integral. Thus, the expected TP_{RA} achieved by BLiP is very close to the expected TP_{RA} obtained by the relaxed LP, which is an upper-bound on the maximum achievable expected TP_{RA} of any valid method whose discoveries are elements of $\mathcal{G}_0$. Of course, even when this is not true, one can compute and compare the expected TP_{RA} achieved by the relaxed LP and BLiP, so BLiP also comes with "warning lights" which signal when it is not optimal. However, Figure 1 confirms empirically that in a high-dimensional regression problem with $>$ 50,000 candidate regions, the nominal expected TP_{RA} achieved by BLiP is indistinguishable from the upper bound provided by the relaxed LP. This suggests BLiP is effectively optimal.

3.2. Choosing the Candidate Groups

We now discuss the choice of candidate groups $\mathcal{G}$. Up to computational limits, we suggest adding every conceivably useful region to $\mathcal{G}$, since adding more regions should increase expected TP_{RA} without affecting validity, and BLiP is usually efficient enough to handle millions of candidate regions. That said, we recommend two general approaches.

First, we can include all *contiguous* groups below some maximum size m. This approach makes sense when the locations

have spatial or temporal structure. For example, suppose $\mathcal{L} = \{\ell_1, \ldots, \ell_p\}$ consists of p ordered locations of genetic variants on the genome. In this context, a contiguous group is of the form $\{\ell_i, \ell_{i+1}, \ldots, \ell_{i+k}\}$ for some $i, k \in \mathbb{N}$. Since genetic variants exhibit local biological similarities and mostly local correlations, considering contiguous groups is often more interpretable and useful than considering groups of far-flung genetic variants. Of course, correlations among genetic variants are not perfectly explained by spatial structure, which is why the next paragraph recommends including some non-contiguous groups as well—however, including contiguous groups is a good first step. This option is also attractive in change point detection (Appendix F.10), where the set of locations $\mathcal{L} = \{1, \ldots, T\}$ is a set of ordered times. Notably, there are roughly $m \cdot p$ contiguous groups of length m or less when considering p locations, so the number of candidate groups scales linearly with p. Lastly, when there are more spatial dimensions, we use spherical subsets of $\mathcal{L}$ as candidate regions. For example, in our astronomical application in Section 5.2, $\mathcal{L} = [0,1]^2$, and we let $\mathcal{G}$ include the set of circles of radius ϵ centered at one of a few million equidistant lattice points in $[0,1]^2$, for many values of ϵ. We review efficient algorithms for this in Appendix D.

The second main approach we recommend is tailored to regression problems, where we seek to discover important variables among $X = (X_1, \ldots, X_p)$. Here, we recommend applying many clustering algorithms to X and letting $\mathcal{G}$ denote the union of the clusters. For example, one could generate candidate regions by hierarchically clustering X based on its correlation matrix, or its partial correlation matrix (Bühlmann et al. 2012), or any combination thereof. Furthermore, we suggest running these algorithms many times using different tuning parameters to add more candidate regions to $\mathcal{G}$. Finally, there is no need to choose between multiple approaches: when $(X_1, \ldots, X_p)$ exhibit spatial structure, we can combine this approach with that of the previous paragraph.

3.3. Robustly Computing the PIPs

BLiP's theoretical guarantees assume that its input PIPs are accurate. That said, estimating PIPs can be challenging in large-scale problems where (a) the prior may be misspecified and (b) standard MCMC algorithms may not converge. Below, we describe heuristics to improve robustness to these issues. Using these heuristics, our simulations in Section 4 show that BLiP is highly robust to misspecification and convergence issues.

To address (a), we recommend using *hierarchical priors*. For example, many sparse Bayesian models require some knowledge of s, the proportion of signals. We suggest picking fairly uninformative priors for such parameters, for example, letting $s \sim \text{Unif}(0, s_{\max})$ for some $s_{\max} \leq 1$. To determine the hyperparameters (e.g., $s_{\max}$), we suggest using a conservative choice or taking an empirical Bayesian approach. (Here, "conservative" choices are choices that may yield an error rate below the nominal level.) Section 4 shows empirically that fairly conservative choices, such as when $s_{\max} = \frac{s}{2} \ll s$, do not lose much power and reliably control the FDR even when the hyperprior is quite different from the true sparsity.

To address (b), we recommend sampling from *multiple MCMC chains with random initialization*. Even if each chain does not converge, we expect that aggregating results across chains will usually overestimate the uncertainty in the location of a signal. This allows BLiP to empirically control the error rate, even if it is conservative. For example, in a bivariate regression problem, suppose $\{X_1, X_2\}$ clearly contains a signal variable, but X_1 and X_2 are highly correlated, so it is not clear which one is the signal variable. Consider a worst-case scenario where the MCMC algorithm randomly initializes (e.g.) X_1 to be a signal variable, but then keeps X_1 as a signal variable at every iteration. If we compute $p_{\{1\}}$ just using this chain, we will falsely conclude that $p_{\{1\}} \approx 1$. However, if we run 10 MCMC chains which each have a 50% chance of initializing X_1 or X_2 as a signal variable, then we will conclude $p_{\{1\}}, p_{\{2\}} \approx 50\%$, or equivalently, that we are maximally uncertain about which of $\{X_1, X_2\}$ is a signal. This will yield (conservative) error rate control in this toy example, even though the MCMC algorithm did not converge whatsoever. Note this intuition also extends to variational approaches which use random initialization.

To empirically demonstrate the effect of using multiple MCMC chains, in Section 4, we rerun our core simulations with only 200 MCMC samples per chain (our default is 5000 samples per chain). We describe these simulations completely in Section 4—for now, we note this is a high-dimensional setting with $p = 1000$ highly correlated covariates, so we should not expect the first 200 MCMC samples to converge. Indeed, Figure 5 shows that using only one chain leads to substantial FDR control violations. Using 10 chains, however, yields FDR control without reducing power. See Section 4 for details.

4. Simulations

We now show that BLiP is powerful, robust, and efficient compared to its competitors. We focus on variable selection in Gaussian linear models. However, the appendix contains more simulations, including a concrete example demonstrating the strengths of different methods (Appendix F.4), comparisons to more competitors (Appendix F.2), sensitivity analyses for the weight function (Appendix F.8) and prior (Appendix F.6), analysis of the correlations among discovered groups (Appendix F.5), simulations for binary regression (Appendix F.9), change point detection (Appendix F.10), and simulations using real genotype data (Appendix G.2). All code is publicly available at *https://github.com/amspector100/blip_sims*.

We simulate $Y \mid X \sim \mathcal{N}(X\beta, \sigma^2)$ where β has $\lceil sp \rceil$ randomly chosen nonzero coefficients, for sparsity $s \in (0,1)$. The nonzero coefficients are iid $\mathcal{N}(0, \tau^2)$ random variables; this is often called a "Linear Spike and Slab" (LSS) model (Mitchell and Beauchamp 1988). The locations $\mathcal{L} = [p]$ represent $(X_1, \ldots, X_p)$ and the signals are $\mathcal{S} = \{\ell \in [p] : \beta_\ell \neq 0\}$. To capture a challenging setting, we sample $\boldsymbol{X}$ from a nonstationary AR(k) model, meaning $\boldsymbol{X}$ exhibits high local correlations; for example, the average correlation between two adjacent variables is $\approx 90\%$ but can be as high as 99.99% (see Appendix F.1 for details and a picture of the covariance matrix). Unless otherwise specified, we set $k = 3$, $p = 1000$, and the FDR level is $q = 0.1$. We compare the performance of four classes of methods:

1. BLiP: We apply BLiP on top of a standard Gibbs sampler for the LSS model (George and McCulloch 1997). We consider a well-specified case, where the sampler uses the true values of s, τ^2, and σ^2, and a misspecified case, where the sampler uses standard choices of uninformative conjugate priors (see Appendix E.1).[1] We use the default settings in `pyblip`.

2. SuSiE: SuSiE approximates the posterior law of the signals $\mathcal{S} \mid \mathcal{D}$ using an efficient variational approximation which is accurate when $|\mathcal{S}|$ is small. The form of SuSiE's posterior automatically yields one set of regions $G_1^{\text{susie}}, \ldots, G_R^{\text{susie}} \subset [p]$ which localize signals and control the FDR as per Definition 2.1. However, we can also apply BLiP directly to the posterior from SuSiE; since BLiP explicitly maximizes expected TP_{RA} subject to FDR control, we should expect SuSiE + BLiP to have weakly higher expected TP_{RA} than SuSiE alone. For brevity, see Appendix E.3 for further review of SuSiE and SuSiE + BLiP. We apply SuSiE with the default settings in `susieR`, except we input the true number of signals.

3. FBH: Katsevich, Sabatti, and Bogomolov (2021) introduced the Focused Benjamini-Hochberg (FBH), a method for localizing signals based on a set of frequentist p-values $\{p_G^{\text{freq}} : G \in \mathcal{G}_{\text{tree}}\}$, where $\mathcal{G}_{\text{tree}}$ is a set of groups from a hierarchical clustering of the variables X, and p_G^{freq} tests the null hypothesis $H_G : G \cap \mathcal{S} = \emptyset$ that there is no signal in G. Our simulations apply the FBH on top of p-values from a lasso-based distilled conditional randomization test (dCRT) (Candès et al. 2018; Liu et al. 2021). We used the dCRT because it was powerful empirically and it can produce frequentist p-values in high dimensions. We implement these methods in the python package `blip_sims`.

4. Baselines: We also apply the standard Benjamini-Hochberg (BH) method for FDR control on top of individual p-values $p_{\{j\}}^{\text{freq}}$ from the dCRT. Second, the "LSS (indiv. only)" method discovers as many individual signals as possible based on individual PIPs from the well-specified LSS sampler. That is, after sorting the PIPs, this method rejects as many signals as possible such that the average rejected PIP is $\geq 1 - q$. We compare to these methods (which can only discover individual signals) to assess the benefit of resolution-adaptivity.

See Appendix F.2 for a rigorous review of each method and implementation details. We also compared to methods from Yekutieli (2008) and Lee et al. (2018), but the FBH and SuSiE uniformly outperformed these methods, so we defer this analysis to Appendix F.2.

We compare these methods using three main metrics. First, we compute the realized FDR of each method. Note that given correct PIPs, BLiP provably controls the FDR conditional on the data, and thus it should control the FDR in our plots, which average over the randomness in both the data and the data-generating parameters, for example, β. Of course, we cannot perfectly compute the PIPs due to prior misspecification or other error in approximating the posterior. Thus, these simulations also assess BLiP's robustness to (somewhat) inaccurate PIPs. Second, Figure 3 plots the distribution of the sizes of true discoveries from each method. Third, we measure *resolution-adjusted power* (Power_{RA}), defined as the expected TP_{RA} using the default weight function from Section 2 divided by the total number of signals. Formally, if $G_1, \ldots, G_R \subset \mathcal{L}$ are the discoveries from a method:

$$\text{Power}_{\text{RA}} \triangleq \frac{\mathbb{E}\left[\text{TP}_{\text{RA}}(G_1, \ldots, G_R)\right]}{|\mathcal{S}|}, \tag{13}$$

where the expectation is taken over both the data and the data-generating parameters (note that the number of signals $|\mathcal{S}|$ is nonrandom in our simulations). Power_{RA} is proportional to expected TP_{RA}, but it is more interpretable because it takes values in $[0, 1]$. For example, $\text{Power}_{\text{RA}} = 1$ indicates that we perfectly localized all signals and $\text{Power}_{\text{RA}} = \frac{1}{4}$ is consistent with perfectly localizing 25% of the signals or localizing all the signals in regions of size 4.

Figure 2 shows the Power_{RA} and FDR for each method while varying the number of data-points n and the sparsity s. It shows that LSS + BLiP uniformly has the highest Power_{RA} by wide margins and reliably controls the FDR, even when $p \gg n$. Remarkably, LSS + BLiP (misspec.) achieves essentially the same performance as the well-specified LSS model with oracle knowledge of the hyperparameters. We emphasize that the prior and posterior of LSS (misspec.) can be quite misspecified. For example, the prior mean of τ^2 is five times larger than its true value, and when $s = 0.2$ and $\kappa = 2$, the true value of τ^2 is on average the $\approx 5 \times 10^{-4}$ quantile of its estimated posterior law. That said, the posteriors of s and σ^2 are more accurate (despite the prior misspecification—see Appendix F.6 for precise details). Appendix F.6 contains additional experiments which make the prior on s even more misspecified and anti-conservative. It shows that BLiP controls the FDR even when the prior mean of s is 9 times larger than and 8 prior standard deviations above its true value (causing the posterior mean of s to be 4 times larger than and 5 posterior standard deviations above its true value). These results show BLiP's robustness in challenging estimation settings.

Besides LSS + BLiP, only FBH and the baselines (BH and LSS (indiv. only)) reliably control the FDR. Furthermore, SuSiE + BLiP has uniformly higher Power_{RA} than SuSiE and sometimes simultaneously improves upon SuSiE's FDR control. Finally, note that these results are not sensitive to the definition of Power_{RA}; Figure 3 shows that for almost every integer k, LSS + BLiP makes more true discoveries of size k or less than every other method. Similarly, SuSiE+BLiP makes either as many or more true discoveries of size k or less than SuSiE (up to MCMC error). Thus, BLiP improves "power" by nearly any metric. The next two paragraphs give intuition to explain this result.

First, we discuss SuSiE. Recall that SuSiE makes a variational approximation which is accurate when the number of signals is small. When $s = 0.01$ and there are 10 signals, the approximation is accurate and SuSiE almost matches the performance of LSS + BLiP. Yet for $s \geq 0.05$, SuSiE has much lower Power_{RA} than LSS + BLiP and violates FDR control. Indeed, SuSiE's approximation is inaccurate when the absolute number of signals is large, so SuSiE may perform poorly even in very sparse problems with large p (see also the simulations in Appendix F.7). However, applying BLiP on top of SuSiE can partially remedy this problem. Indeed, any disjoint output from

[1] The "well-specified" case is not perfectly well-specified because the prior assumes $|\mathcal{S}| \sim \text{Bin}(p, s)$ signals, whereas in our simulations $|\mathcal{S}| = \lceil sp \rceil$ deterministically. Nonetheless, it is almost perfectly well-specified.

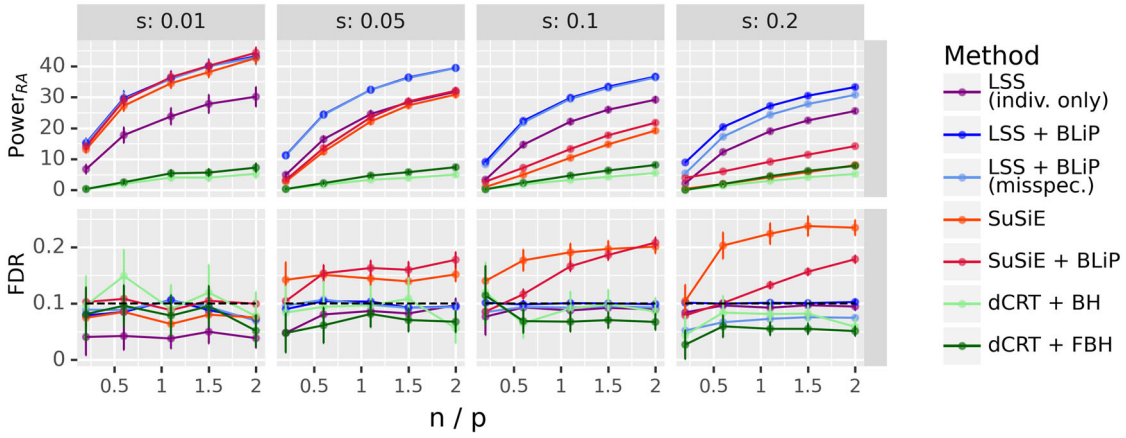

Figure 2. Resolution-adaptive variable selection for Gaussian linear models as described in Section 4 with $p = 1000$ and $\lceil sp \rceil$ signals. Note Power_{RA} is defined in (13). See Appendix F.1 for further simulation details.

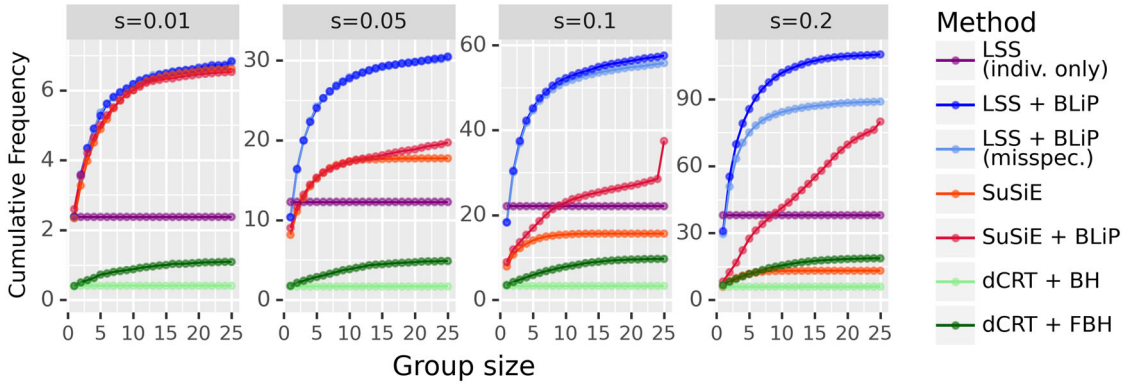

Figure 3. This figure plots the cumulative frequency of the discovered group sizes in the same setting as Figure 2 with $n = 1100, p = 1000$, and $\lceil sp \rceil$ signals. That is, the point with x-value k on the blue curve counts the expected number of true discoveries of size k or less made by LSS + BLiP.

SuSiE is a feasible output for BLiP, so we expect BLiP to have uniformly higher Power_{RA} than SuSiE, which is supported by all of our simulations. Furthermore, by increasing the number of true discoveries, BLiP can simultaneously improve FDR control, as shown in Figure 2 for $s \geq 0.1$. See Appendix E.3 for details on how SuSiE's approximation breaks down and intuition explaining how BLiP can partially correct this problem.

LSS + BLiP also has uniformly higher Power_{RA} than the FBH procedure, we suspect because BLiP can search over hundreds of times more candidate regions than the FBH, which is restricted to search over a single hierarchical tree. Furthermore, BLiP explicitly maximizes Power_{RA} when searching over candidate regions, whereas the FBH only searches heuristically over its input p-values and may not find a rejection set which maximizes Power_{RA} (or any measure of power). For example, Figure 3 confirms that FBH makes over twice as many true discoveries as the baseline BH procedure; however, it makes many of these discoveries at very coarse resolutions, presumably because it cannot search over many candidate regions. That said, there are other possible explanations. For example, perhaps there are more powerful p-values that could be used with the FBH, although we are not aware of p-values more powerful than the ones we used. Either way, it is not clear how one could apply the FBH in a way that is more powerful than LSS + BLiP.

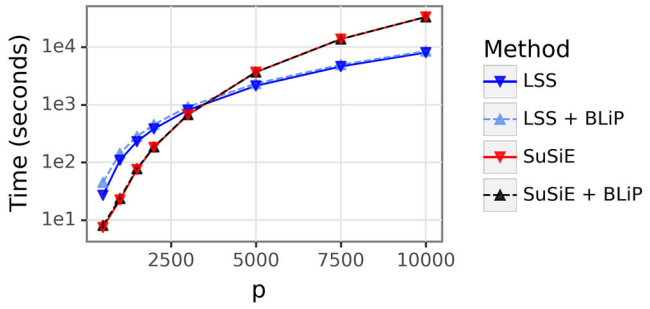

Figure 4. In sparse linear regression, this figure shows the computation time required to fit the underlying model and the total time to both fit the model and run BLiP. In this setting, $n = 0.5p$ and there are $\lceil 0.05p \rceil$ signals. Note all methods controlled the FDR (see Appendix F.7).

Next, Figure 4 analyzes the runtime of BLiP in large-scale settings with p varied from 500 to 10,000. It shows that the cost of applying BLiP is trivial compared to the cost of running SuSiE or the LSS sampler: when $p = 10{,}000$ and $|\mathcal{S}| = 500$, BLiP runs in a few minutes, whereas fitting LSS and SuSiE requires 2 and 9 hours, respectively. All other methods (e.g., dCRT + FBH) were too expensive to fit with $p > 1000$.

Lastly, Figure 5 shows that the heuristics in Section 3.3 make BLiP quite robust to convergence issues for MCMC algorithms. See Section 3.3 and the figure caption for details.

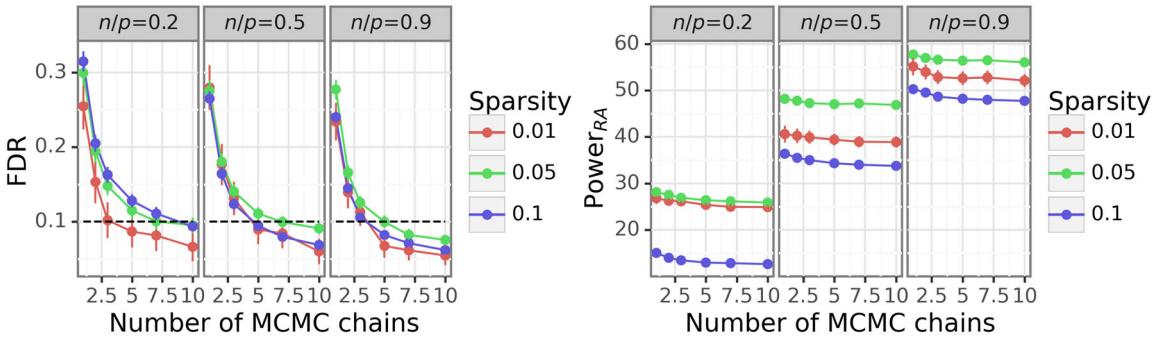

Figure 5. This figure replicates the "LSS + BLiP (misspec.)" method from Figure 2 with sparsity $s = 0.05$ but uses only 200 samples per MCMC chain (Figure 2 uses 5000). The individual chains do not converge, since the realized FDR is up to three times the nominal level when using just one chain. Despite this, aggregating results from 5 to 10 chains leads to FDR control without losing much Power_{RA}. The simulation details are otherwise identical to Figure 2 (see Appendix F.1).

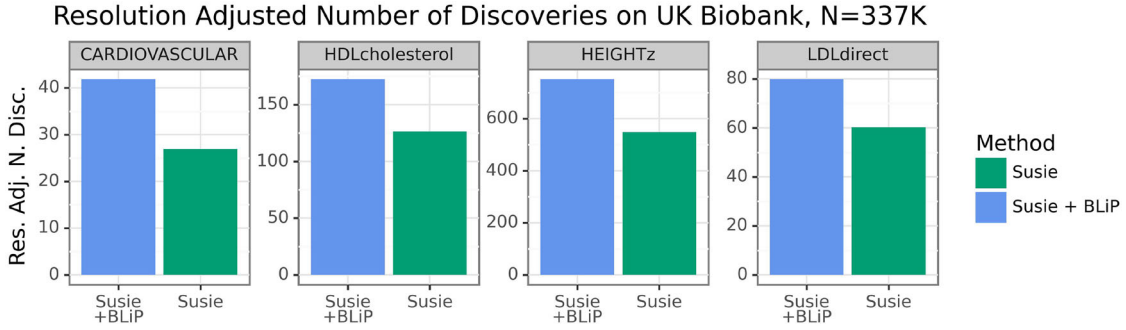

Figure 6. This figure shows that SuSiE + BLiP made 30%–50% more resolution-adjusted discoveries than SuSiE alone in our application to UK Biobank data.

5. Real Data Applications

5.1. Application to Genetic Fine-Mapping

As discussed in Section 1, resolution-adaptive methods are particularly attractive in fine-mapping problems, where correlations among genetic variants are very strong. Thus, it can be very challenging to detect individually important genetic variants. Resolution-adaptive methods instead allow the analyst to localize causal variants as precisely as possible given the data at hand, and for this reason, a few recent works (Weissbrod et al. 2020; Wang et al. 2020; Wallace 2021) have used resolution-adaptive methods in fine-mapping problems. Furthermore, Bayesian variable selection methods are commonly used in genetic fine-mapping (Guan and Stephens 2011; Carbonetto and Stephens 2012; Benner et al. 2016; Lee et al. 2018; Weissbrod et al. 2020). All this suggests that BLiP can help solve an important problem in the domain of fine-mapping.

To test BLiP's effectiveness, we apply BLiP to a dataset of $n \approx$ 337,000 individuals from the UK Biobank with $p \approx$ 19,000,000 genetic variants. We seek to identify causal genetic variants for four traits of interest: cardiovascular disease, height, low-density lipoprotein (LDL) cholesterol, and high-density lipoprotein (HDL) cholesterol. This dataset was previously analyzed by Weissbrod et al. (2020), and indeed, our work explicitly builds upon theirs. SuSiE is an attractive model in this setting because we expect that each genetic locus has a small number of causal variants, and our simulations suggest that SuSiE performs almost as well as full Bayesian inference when the number of signals is small. Thus, we run BLiP directly on top of the SuSiE model that Weissbrod et al. (2020) fit on this dataset. For each method's discoveries $G_1, \ldots, G_R \subset \mathcal{L}$, we calculate the resolution-adjusted number of discoveries, defined as $\sum_{r=1}^{R} \frac{1}{|G_r|}$. (This is identical to TP_{RA} except we do not include the indicators that G_r are true discoveries since we do not know the ground truth.)

Running BLiP requires less than 1 min of computation per trait, but as shown by Figure 6, SuSiE + BLiP makes 30%–50% more resolution-adjusted discoveries than SuSiE alone. Crucially, this result is not sensitive to the metric of power: Figure 7 shows that for every k, SuSiE + BLiP discovers more groups of size k or less than SuSiE alone, and thus SuSiE + BLiP makes more discoveries at finer resolutions by nearly any metric. Indeed, for every region G discovered by SuSiE, SuSiE + BLiP discovers a group G' which overlaps with G. This suggests that BLiP is successfully optimally localizing signals based on the information available in the SuSiE model—see Appendix E.3 for more intuition on why SuSiE + BLiP can outperform SuSiE alone. Notably, SuSiE + BLiP makes more singleton discoveries than SuSiE alone, in part because SuSiE + BLiP uses PIPs which are provably more powerful than the default SuSiE algorithm. However, we caution that discovering singleton groups is not the primary purpose of BLiP and the interpretation of this result is subtle, so we discuss this further in Appendix G.4. Lastly, note that each group we discover is roughly (but not perfectly) contiguous, meaning that each group only contains nearby genetic variants. Thus, BLiP's outputs are interpretable: each discovery asserts that one of k nearby genetic variants has a causal effect on a trait. See Appendix G.1 for methodological details.

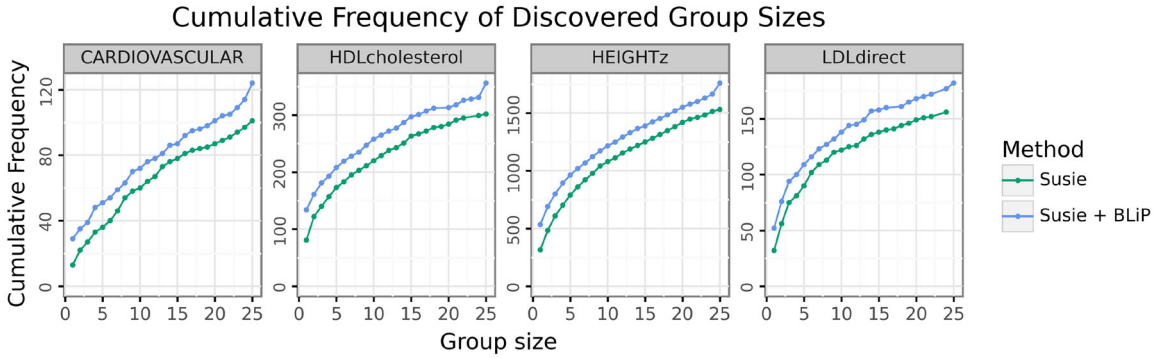

Figure 7. This figure plots the cumulative frequency of the discovered group sizes. That is, the point with x-value *k* on the green curve (resp. blue curve) counts the number of groups of size *k* or less discovered by SuSiE (resp. SuSiE + BLiP).

To validate our findings, we first confirm that SuSiE + BLiP controls the FDR in simulations using the real genotype data (shown in Appendix G.2). Furthermore, we compare our findings to those of previous work. To start, as a sanity check, we compare the discoveries from SuSiE + BLiP with those of the SuSiE model from Weissbrod et al. (2020) (i.e., the model represented by the green bars in Figure 6).[2] Appendix G.3 shows that SuSiE + BLiP replicates every finding from the SuSiE model but makes roughly 15%–20% more discoveries (note this number is not resolution-adjusted). Since SuSiE + BLiP makes 30%–50% more resolution-adjusted discoveries than SuSiE, this shows that the power gain comes both from more precisely localizing existing discoveries and from making entirely new discoveries. Crucially, of the new discoveries made by SuSiE + BLiP, we found that 45%–65% are corroborated by a separate study in the NHGRI-EBI GWAS Catalog (Buniello et al. 2018), which is comparable to the corroboration rate of the initial analysis from Weissbrod et al. (2020). This is arguably a remarkable (positive) result, since one might expect that any novel discoveries would informally be "harder to discover" and thus corroborate, since the initial model did not discover them. Nonetheless, the *additional* discoveries from SuSiE + BLiP were corroborated at a similar rate to the original discoveries. See Appendix G.3 for details. Overall, these results suggest that BLiP enhanced SuSiE's power to find real causal variants, and they give no indication that the increased resolution-adjusted power of SuSiE + BLiP results from false discoveries. All code and data are publicly available at *https://github.com/amspector100/ukbb_blip*.

Lastly, we emphasize that BLiP can be applied on top of any Bayesian model, yielding more discoveries at finer resolutions with little additional computational cost. For example, in this section, SuSiE uses an uninformative prior for simplicity, but several recent works have used priors based on (e.g.) functional annotations, other complex traits, and prior knowledge about genetic effect sizes (Weissbrod et al. 2020; O'Connor 2021; Trippe, Finucane, and Broderick 2021). Similarly, the fine-mapping literature contains many inferential algorithms besides SuSiE (Carbonetto and Stephens 2012; Hormozdiari et al. 2014; Benner et al. 2016; Kichaev et al. 2016). BLiP can wrap on top of any of these methods, and, we hope, enhance their power to make meaningful scientific discoveries.

[2]Note that this is not a replication analysis, since both analyses use the same dataset and model.

5.2. Application to Astronomical Point Source Detection

Appendix H performs a similarly detailed application to detect and localize astronomical point sources (e.g., stars). We apply BLiP on top of pretrained Bayesian models from the literature and show that BLiP dramatically increases Power_{RA} compared to the state-of-the-art. Furthermore, BLiP achieves remarkably good FDR calibration as verified by comparison to the ground truth, since the true positions of the stars in our dataset were later observed using a much more powerful telescope. We also demonstrate the flexibility of BLiP by using two weight functions to accommodate two scientific objectives.

6. Discussion

This article introduces BLiP, a method for performing resolution-adaptive signal detection. Our simulations and two applications show that BLiP is computationally efficient, robust, and powerful while providing provable error control. That said, BLiP does have a few limitations. First, BLiP's provable guarantees assume that its input PIPs are correct. In practice, this condition will not hold exactly due to (e.g.) misspecification or MCMC error. Although we have devised methods which empirically make BLiP robust to this issue, there is certainly room for improvement. Indeed, it may be worthwhile to design BLiP-like methods which have provable guarantees under misspecification. Second, although BLiP was verifiably nearly optimal (as described in Section 3.1) in all of our analyses, we can only give a heuristic explanation of this, and we cannot fully rigorously explain this result. Future theoretical work is needed to better understand this phenomenon.

A last notable benefit of BLiP is that it is a highly flexible method for a very general problem. As a result, there are many possible extensions of BLiP which may be of interest, including optimizing for different objective functions (see Appendix C.5 for an example of an objective function which flexibly balances resolution-adjusted power against false positives), applying BLiP on different Bayesian models, and using BLiP in different application areas.

Supplementary Materials

Supplementary Material: The supplement contains further discussion, simulations, an additional real application, and a few proofs (.pdf file).

Data Availability Statement

Code and Data are publicly available at *https://github.com/amspector100/blip_sims*.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

A.S. was partially supported by the Two Sigma Graduate Fellowship Fund and a Graduate Research Fellowship from the National Science Foundation. L.J. was partially supported by the William F. Milton Fund and a CAREER grant from the National Science Foundation (grant #DMS2045981).

References

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679. [2]

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. [2,10,11]

Blanchard, G., Neuvial, P., and Roquain, E. (2020), "Post Hoc Confidence Bounds on False Positives Using Reference Families," *The Annals of Statistics*, 48, 1281–1303. [2]

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877. [2]

Bogomolov, M., Peterson, C. B., Benjamini, Y., and Sabatti, C. (2020), "Hypotheses On A Tree: New Error Rates and Testing Strategies," *Biometrika*, 108, 575–590. [2]

Boyd, S., and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge: Cambridge University Press. [5,6]

Brooks, S., Gelman, A., Jones, G., and Meng, X. (2011), *Handbook of Markov Chain Monte Carlo*, Boca Raton, FL: CRC Press. [2,3]

Buniello, A., MacArthur, J. A., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F., and Parkinson, H. (2018), "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019," *Nucleic Acids Research*, 47, D1005–D1012. [11]

Buzdugan, L., Kalisch, M., Navarro, A., Schunk, D., Fehr, E., and Bühlmann, P. (2016), "Assessing Statistical Significance in Multivariable Genome Wide Association Analysis," *Bioinformatics*, 32, 1990–2000. [4]

Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C.-H. (2012), "Correlated Variables in Regression: Clustering and Sparse Estimation," *Journal of Statistical Planning and Inference 2013*, 143, 1835–1858. [7]

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for Gold: Model-X Knockoffs for High-Dimensional Controlled Variable Selection," *Journal of the Royal Statistical Society*, Series B, 80, 551–577. [8]

Carbonetto, P., and Stephens, M. (2012), "Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies," *Bayesian Analysis*, 7, 73–108. [2,10,11]

Efron, B. (2008), "Microarrays, Empirical Bayes and the Two-Groups Model," *Statistical Science*, 23, 1–22. [3,4]

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [2,8]

Goeman, J. J., Meijer, R. J., Krebs, T. J. P., and Solari, A. (2019), "Simultaneous Control of All False Discovery Proportions in Large-Scale Multiple Hypothesis Testing," *Biometrika*, 106, 841–856. [2]

Goeman, J. J., and Solari, A. (2011), "Multiple Testing for Exploratory Research," *Statistical Science*, 26, 584–597. [2]

——— (2012), "The Sequential Rejection Principle of Familywise Error Control," *The Annals of Statistics*. [2]

Guan, Y., and Stephens, M. (2011), "Bayesian Variable Selection Regression for Genome-Wide Association Studies and Other Large-Scale Problems," *The Annals of Applied Statistics*, 5, 1780–1815. [2,10]

Guo, Z., Renaux, C., Bühlmann, P., and Cai, T. (2021), "Group Inference in High Dimensions with Applications to Hierarchical Testing," *Electronic Journal of Statistics*, 15, 6633–6676. [4]

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. [11]

Jünger, M., Liebling, T., Naddef, D., Nemhauser, G., Pulleyblank, W., Reinelt, G., Rinaldi, G., and Wosley, L. (2010), *50 Years of Integer Programming 1958-2008*, Berlin: Springer. [5]

Katsevich, E., and Ramdas, A. (2020), "Simultaneous High-Probability Bounds on the False Discovery Proportion in Structured, Regression and Online Settings," *The Annals of Statistics*, 48, 3465–3487. [2]

Katsevich, E., Sabatti, C., and Bogomolov, M. (2021), "Filtering the Rejection Set While Preserving False Discovery Rate Control," *Journal of the American Statistical Association*, 118, 165–176. [2,8]

Kichaev, G., Roytman, M., Johnson, R., Eskin, E., Lindström, S., Kraft, P., and Pasaniuc, B. (2016), "Improved Methods for Multi-Trait Fine Mapping of Pleiotropic Risk Loci," *Bioinformatics*, 33, 248–255. [11]

Lee, Y., Luca, F., Pique-Regi, R., and Wen, X. (2018), "Bayesian Multi-SNP Genetic Association Analysis: Control of fdr and Use of Summary Statistics," bioRxiv. [2,8,10]

Liu, M., Katsevich, E., Janson, L., and Ramdas, A. (2021), "Fast and Powerful Conditional Randomization Testing via Distillation," *Biometrika*, 109, 277–293. [8]

Mandozzi, J., and Bühlmann, P. (2016), "Hierarchical Testing in the High-Dimensional Setting with Correlated Variables," *Journal of the American Statistical Association*, 111, 331–343. [2,3,4]

Meinshausen, N. (2008), "Hierarchical Testing of Variable Importance," *Biometrika*, 95, 265–278. [2]

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [2,7]

O'Connor, L. J. (2021), "The Distribution of Common-Variant Effect Sizes," *Nature Genetics*, 53, 1243–1249. [11]

Renaux, C., Buzdugan, L., Kalisch, M., and Bühlmann, P. (2018), "Hierarchical Inference for Genome-Wide Association Studies: A View on Methodology with Software," *Computational Statistics*, 35, 1–40. [2,4]

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., and Goeman, J. J. (2018), "All-Resolutions Inference for Brain Imaging," *NeuroImage*, 181, 786–796. [2]

Scatamacchia, R. (2017), "Knapsack Problems with Side Constraints." DOI:10.6092/polito/porto/2667802. [5]

Shin, M., and Liu, J. S. (2021), "Neuronized Priors for Bayesian Sparse Linear Regression," *Journal of the American Statistical Association*, 117, 1695–1710. [2]

Trippe, B. L., Finucane, H. K., and Broderick, T. (2021), "For High-Dimensional Hierarchical Models, Consider Exchangeability of Effects Across Covariates Instead of Across Datasets," in *Advances in Neural Information Processing Systems*, eds. A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan. *https://openreview.net/forum?id=28NikxkK6kJ* [11]

Wallace, C. (2021), "A More Accurate Method for Colocalisation Analysis Allowing for Multiple Causal Variants," *PLOS Genetics*, 17, 1–11. [10]

Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020), "A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping," *Journal of the Royal Statistical Society*, Series B, 82, 1273–1300. [1,2,3,10]

Weissbrod, O., Hormozdiari, F., Benner, C., Cui, R., Ulirsch, J., Gazal, S., Schoech, A., van de Geijn, B., Reshef, Y., Márquez-Luna, C., O'Connor, L., Pirinen, M., Finucane, H. K., and Price, A. L. (2020), "Functionally-Informed Fine-Mapping and Polygenic Localization of Complex Trait Heritability," *Nature Genetics*, 52, 1355–1363. [2,10,11]

Yang, Y., and Bulfin, R. L. (2009), "An Exact Algorithm for the knapsack Problem with Setup," *International Journal of Operational Research*, 5, 280–291. [5]

Yekutieli, D. (2008), "Hierarchical False Discovery Rate-Controlling Methodology," *Journal of the American Statistical Association*, 103, 309–316. [2,8]

Zou, Y., Carbonetto, P., Wang, G., and Stephens, M. (2021), "Fine-Mapping from Summary Data with the "Sum of Single Effects" Model," bioRxiv. [2]