Inference of Causal Relationships between Biomarkers and Outcomes in High Dimensions

Felix AGAKOV and Paul MCKEIGUE
Center for Public Health Sciences, University of Edinburgh
Edinburgh EH8 9AG, UK

and

Jon KROHN and Jonathan FLINT
Wellcome Trust Center for Human Genetics
Oxford OX3 7BN, UK

ABSTRACT

We describe a unified computational framework for learning causal dependencies between genotypes, biomarkers, and phenotypic outcomes from large-scale data. In contrast to previous studies, our framework allows for noisy measurements, hidden confounders, missing data, and pleiotropic effects of genotypes on outcomes. The method exploits the use of genotypes as "instrumental variables" to infer causal associations between phenotypic biomarkers and outcomes, without requiring the assumption that genotypic effects are mediated only through the observed biomarkers. The framework builds on sparse linear methods developed in statistics and machine learning and modified here for inferring structures of richer networks with latent variables. Where the biomarkers are gene transcripts, the method can be used for fine mapping of quantitative trait loci (QTLs) detected in genetic linkage studies. To demonstrate our method, we examined effects of gene transcript levels in the liver on plasma HDL cholesterol levels in a sample of 260 mice from a heterogeneous stock.

Key words: sparse linear models, causality, structure learning, Bayesian networks, Mendelian randomization, instrumental variables

1. INTRODUCTION

One of key goals of epidemiology and systems biology is to distinguish causal and non-causal explanations of observed associations between phenotypic biomarkers and disease outcomes. In principle it is possible to exploit genotypic variation that perturbs the biomarkers to learn about these relationships. The classic *Mendelian randomization* method [16] addresses this problem using the framework of *instrumental variable* analysis [2, 3, 30] that perturbs the phenotypic biomarker. The instrumental variable argument assumes that effects of the genetic instrument g on the biomarker x are unconfounded, and that effects of the instrument on the outcome y are mediated only through the biomarker (*no pleiotropy*), e.g. [19]. The assumption of no confounding is guaranteed by the laws of Mendelian genetics, if population stratification has been adequately controlled. However, the requirement to assume no pleiotropy restricts the application of the classic instrumental variable argument to a few biomarkers and genes for which the genotypic effects are well understood. Thus this approach cannot easily be extended to exploit multiple biomarkers and genome-wide genotype data.

A more general approach to exploiting genotypic variation to infer causal relationships between gene transcript levels and quantitative traits of interest, called the *"likelihood-based causality model selection" (LCMS)* by its authors, has been developed by Schadt et. al. [31] and subsequently extended (see e.g. [5]). In contrast with the classical instrumental variable argument, this approach does not require the assumption of no pleiotropy, but instead compares models with and without pleiotropy. After filtering to select a set of gene transcripts $\{x_j\}$ that are associated with the trait y, and loci $\{g_i\}$ at which genotypes have effects on transcript levels x_j, each possible triad of marker locus g_i, transcript x_j and trait y is evaluated to compare three possible models: causal effect of transcript on trait $p(y,x_j|g_i) = p(y|x_j)p(x_j|g_i)$, reverse causation $p(y,x_j|g_i) = p(y|g_i)p(x_j|y)$, and a pleiotropic model $p(y,x_j|g_i) = p(y|x_j,g_i)p(x_j|g_i)$. One of these three models is selected as the most likely explanation of the observed associations according to standard likelihood-based scores penalized by complexity: either Akaike's Information Criterion (AIC) [31], or the Bayesian Information Criterion (BIC) [5].

While the LCMS and related approaches [31, 5] relax the assumption of no pleiotropy of the classic instrumental variable method, they have three key limitations. First, the method is not Bayesian (the BIC score is only a crude approximation to the correct Bayesian procedure for model selection). Thus LCMS may only be heuristically extended to problems where the number of variables exceeds the number of measurements (the so-called large p, small n setting typical for genome-wide studies), and lacks a formal basis for model comparison. A second key limitation is that effects of loci and biomarkers on outcomes are not modeled jointly, so widely varying inferences are possible depending on the choice of the triads $\{g_i,x_j,y\}$. This is demonstrated by Figure 1, which compares differences in the AIC scores for the causal and reverse models for various choices of the genetic instruments and a fixed biomarker-outcome pair. AIC scores shown on Figure 1 have been centered relative to those

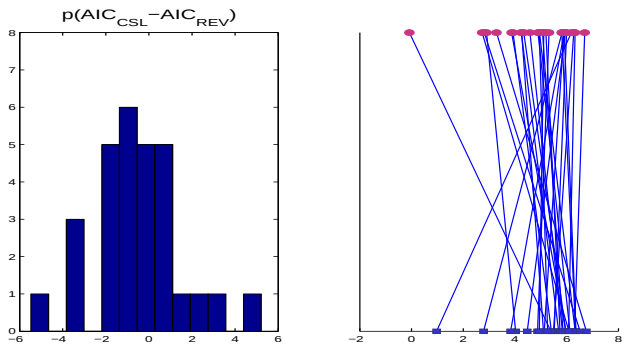

Figure 1: Possible arbitrariness of inference of the likelihood-based causality model selection approach. *Left:* Histogram of the difference of the AIC scores for the causal and reverse hypotheses for liver expressions of *Cyp27b1* and plasma HDL cholesterol in heterogeneous mice, for different choices of genetic instruments g_i. Depending on the choice of g_i, either causal or reverse explanations are favored. No latent confounders are taken into account. *Right:* AIC scores of the causal (top) and reverse (bottom) models for each choice of genetic instrument (the straight lines link the scores corresponding to a fixed choice of g_i).

of the pleiotropic model. It is easy to see that inference of the causal direction of the LCMS may be somewhat arbitrary, depending on the choice of instrument g_i. The LCMS approach may potentially be extended to include multiple instruments, though for high-dimensional datasets with few observations ($p > n$) the Bayesian formulation would be more justified. Most importantly, the LCMS method does not allow for dependencies between multiple biomarkers, measurement noise, or latent variables (such as unobserved confounders of the biomarker-outcome associations). Thus the method of Schadt et. al. [31] can potentially make incorrect conclusions about the direction of causality in situations when an underlying association is best explained by unobserved confounding factors.

Another approach to modeling joint effects of genetic loci and biomarkers (gene expressions) was described by [43]. They modeled the expression measurements as three ordered levels, and used a biased greedy search over model structures from multiple starting points to find models with high BIC scores. Though applicable for large-scale studies, the approach looses information by using categorical measurements, and does not allow for measurement noise or latent variables. Many other recent model selection and structure learning methods from machine learning and systems biology literature are also either not easily extended to include latent confounders, or applicable only for relatively low-dimensional settings with many observations (e.g. [34], [18], [20], [23]).

This paper gives a high-level outline of a unified framework for modeling relations between genotypes, phenotypic biomarkers and outcomes that is both flexible enough to handle realistic models, and computationally tractable enough to handle large datasets. Our *Sparse Instrumental Variables (SPIV)* framework draws on sparse modeling methods developed in engineering, machine learning, and statistics. The approach models joint effects of loci and biomarkers, and may be used for distinguishing causal and non-causal explanations of observed associations between phenotypic biomarkers and outcomes, even when some of the genotypic effects may be pleiotropic. It allows for measurement errors in the phenotypic biomarkers and outcomes, and for latent variables that generate coupling between these biomarkers and confound the biomarker-outcome associations. A somewhat more technical discussion of the method aimed at the machine learning audience is given in [1]. Here we give a general description of the approach in the context of the previous work on causality detection in genetic studies. We discuss some of the more practical issues related to the important problem of feature selection (reducing dimensionality of the genome-wide data) prior to performing the sparse instrumental variables analysis. We also illustrate influence of preprocessing on the results of the SPIV inference.

2. METHODS

Our SPIV approach relies on Bayesian modeling of linear associations between the modeled variables, with the sparseness-inducing prior on links between the variables. The Bayesian framework [11] offers a rigorous approach to model comparison grounded in the rules of probability calculus, which allows formal testing of specific modeling hypotheses if required. It also allows prior biological information to be included if available: for instance cis-acting genotypic effects on transcript levels are likely to be stronger and less pleiotropic than trans-acting effects on transcript levels. The Bayesian framework is valid even when the number of variables exceeds the number of measurements which is typical for genetic studies; the marginal likelihood automatically penalizes models that have more parameters than needed to explain the data.

The SPIV method allows for rich dependencies between genotypes g, biomarkers x, and phenotypic traits y, under the biologically-motivated assumptions that genotypic effects are unconfounded and causal. The prior on the effects favors sparseness, which implies that links specifying dependencies between the variables will tend to be pruned when the model is fitted, unless they are important for explaining the observed data. Even though strong associations may be quite rare, the retained effects are allowed to be large where this is supported by the data. The search over a huge space of dependence hypotheses is replaced by the posterior inference of associations between variables. The posterior distribution of the parameters is approximated at its mode obtained by an efficient optimization algorithm, which iterates to a local maximum of the model best supported by the data. As a result, a heuristic search over parent nodes of a directed graphical model is replaced by a continuous optimization problem, which combines subset selection and regression in the presence of latent variables. The method is motivated by the *automatic relevance determination* approaches (e.g. [21], [26], [38]) and the adaptive shrinkage (e.g. [37], [8], [44]). Here it is adapted for sparse multi-factor instrumental variable analysis in the presence of unobserved confounders, pleiotropy, and noise.

As the number of genetic instruments grows, evidence in favor of the correct model (e.g. causal or pleiotropic) will be less dependent upon the priors on model parameters. For instance, with three genotypic instruments perturbing a single transcript, the pleiotropic model has seven adjustable parameters, while the

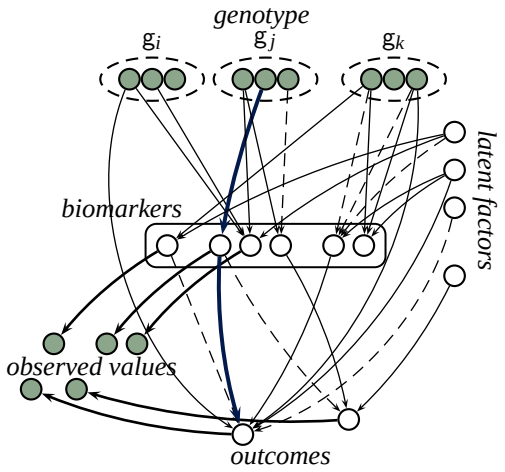

Figure 2: SPIV structure representing dependencies between genotypes, biomarkers, and phenotypes in the presence of noise and latent variables. Genotypic features corresponding to close locations along the genome are grouped together (e.g. g_i, g_j). Dashed lines show weaker links which may shrink to zero in the posterior. Filled and clear nodes correspond to observed and latent variables respectively.

causal model has only four. Where several genotypic variables perturb a single transcript and the causal model fits the data nearly as well as the pleiotropic model, the causal model will tend to be selected, because the slightly better fit of the pleiotropic model will be outweighed by the greater penalty imposed by several extra adjustable parameters.

Model Parameterization

The model is specified with four classes of variables: genotypic and environmental covariates $\mathsf{g} \in \mathbb{R}^{|\mathsf{g}|}$, phenotypic biomarkers $\mathsf{x} \in \mathbb{R}^{|\mathsf{x}|}$, outcomes $\mathsf{y} \in \mathbb{R}^{|\mathsf{y}|}$, and latent factors $z_1, \ldots, z_{|\mathsf{z}|}$. The dimensionality of the latent factors $|\mathsf{z}|$ is fixed at a moderately high value (extraneous dimensions will tend to be pruned under the sparse prior). The latent factors z play two major roles: they represent the shared structure between groups of biomarkers, and confound biomarker-outcome associations. The biomarkers x and outcomes y are specified as hidden variables inferred from noisy observations $\tilde{\mathsf{x}} \in \mathbb{R}^{|\tilde{\mathsf{x}}|}$ and $\tilde{\mathsf{y}} \in \mathbb{R}^{|\tilde{\mathsf{y}}|}$ (note that $|\tilde{\mathsf{x}}| = |\mathsf{x}|$, $|\tilde{\mathsf{y}}| = |\mathsf{y}|$). The effects of genotype on biomarkers and outcome are assumed to be unconfounded. Pleiotropic effects of genotype (effects on outcome that are not mediated through the phenotypic biomarkers) are accounted for by an explicit parameterization of $p(\mathsf{y}|\mathsf{g},\mathsf{x},\mathsf{z})$. The graphical representation of the model is shown on Figure 2, with filled and clear nodes corresponding to the visible and hidden variables respectively.

All the likelihood terms of the corresponding graphical model $p(\mathsf{x},\tilde{\mathsf{x}},\mathsf{y},\tilde{\mathsf{y}},\mathsf{z}|\mathsf{g})$ are linear Gaussians with diagonal covariances, so that

$$\mathsf{y} = \mathsf{W}^T\mathsf{x} + \mathsf{W}_z^T\mathsf{z} + \mathsf{W}_g^T\mathsf{g} + \mathsf{e}_\mathsf{y}, \quad \tilde{\mathsf{y}} = \mathsf{y} + \mathsf{e}_{\tilde{\mathsf{y}}} \tag{1}$$

where $\mathsf{e}_\mathsf{y} \sim \mathcal{N}(0, \Psi_\mathsf{y})$, $\mathsf{e}_{\tilde{\mathsf{y}}} \sim \mathcal{N}(0, \Psi_{\tilde{\mathsf{y}}})$, $\mathsf{z} \sim \mathcal{N}(0, \Psi_\mathsf{z})$, and $\mathsf{W} \in \mathbb{R}^{|\mathsf{x}|\times|\mathsf{y}|}$, $\mathsf{W}_z \in \mathbb{R}^{|\mathsf{z}|\times|\mathsf{y}|}$, $\mathsf{W}_g \in \mathbb{R}^{|\mathsf{g}|\times|\mathsf{y}|}$ are regression coefficients (for clarity, we assume the data is centered). Note that the observed measurements of the outcome variables $\tilde{\mathsf{y}}$ are perturbations of the ground truth outcomes y. The remaining dependencies are expressed analogously.

Prior Distribution

All model parameters are specified as random variables with prior distributions. For computational convenience, the variance components of the diagonal covariances Ψ_y, $\Psi_{\tilde{\mathsf{y}}}$, etc. are specified with inverse Gamma priors $\Gamma^{-1}(a_i, b_i)$, with hyperparameters a_i and b_i fixed at values motivating the prior beliefs about the projection noise (often available to lab technicians collecting trait or biomarker measurements). One way to view the latent confounders z is as missing genotypes or environmental covariates, so that prior variances of the latent factors are peaked at values representative of the empirical variances of the instruments g. Empirically, the choice of priors on the variance components appears to be relatively unimportant as long as such priors are broad, and other choices may be considered [9].

A convenient choice of a sparseness-inducing prior on weight parameters W, W_z, W_g, etc. is a product of zero-mean Laplace and zero-mean normal distributions

$$p(\mathsf{w}) \propto \prod_{i=1}^{|\mathsf{w}|} \mathcal{L}_{w_i}(0,\gamma_1)\mathcal{N}_{w_i}(0,\gamma_2), \tag{2}$$

$\mathcal{L}_{w_i}(0,\gamma_1) \propto \exp\{-\gamma_1|w_i|\}$, and $\mathcal{N}_{w_i}(0,\gamma_2) \propto \exp\{-\gamma_2 w_i^2\}$. Due to the heavy tails of the Laplacian $\mathcal{L}_{w_i}$, the prior $p(\mathsf{w})$ is flexible enough to capture large associations even if they are rare. Higher values of γ_1 give a stronger tendency to shrink irrelevant weights to zero. It is possible to set different γ_1 parameters for different linear weights (e.g. for the cis- and trans-acting effects); however, for clarity of this presentation we shall only use a global parameter γ_1. The isotropic Gaussian component with the inverse variance γ_2 contributes to the grouping effect (see [44], Theorem 1). The considered family of priors (2) induces better consistency properties [42] than the commonly used Laplacians [37, 9, 41, 27, 32]. It has also been shown [15] that important associations between variables may be recovered even for severely under-determined problems ($p \gg n$) common in genetics. The model of Figure 2 with $p(\mathsf{w})$ defined as in (2) generalizes LASSO and elastic net regression [37, 44]. As a special case, it also includes sparse conditional factor analysis. Other sparse priors on linear weights, such as Student-t, "spike-and-slab", or inducing $L_{q<1}$ penalties tend to result in less tractable posteriors even for linear regression [10, 38, 8], which also motivates the choice of (2). Some other intuitions about the influence of the sparse prior (2) on the causal inference is discussed in [1].

Inference

While the choice of prior (2) encourages sparse solutions, it makes the exact inference of the posterior parameters $p(\theta|\mathcal{D})$ analytically intractable. The most efficient approach is based on the maximum-a-posteriori (*MAP*) treatment ([37], [9]), which reduces to solving the optimization problem

$$\theta_{MAP} = \arg\max_\theta \{\log p(\{\tilde{\mathsf{y}}\},\{\tilde{\mathsf{x}}\}|\{\mathsf{g}\},\theta) + \log p(\theta)\} \tag{3}$$

for the joint parameters θ, where the latent variables have been integrated out. Compared to other approximations of inference in sparse linear models based e.g. on sampling or expectation propagation [27, 32], the MAP approximation allows for an efficient handling of very large networks with multiple instruments

and biomarkers, and makes it straightforward to incorporate latent confounders. Depending on the choice of the global sparseness and grouping hyperparameters γ_1, γ_2, the obtained solutions for the weights will tend to be sparse, which is also in contrast to the full inference methods. In high dimensions in particular, the parsimony induced by point-estimates will facilitate structure discovery and interpretations of the findings.

One way to optimize (3) is by an EM-like (expectation-maximization) algorithm, with some tricks for ensuring numerical stability (see [1] for some details). The hyperparameters may be marginalized out for a specific choice of the hyper-prior, set heuristically based on the expected number of links to be retained, or set by cross-validation – in what follows, we use the cross-validation. Once a sparse representation is produced by pruning irrelevant dimensions, other more computationally-intensive inference methods for the full posterior (such as expectation propagation or MCMC) may be used in the resulting lower-dimensional model if required. After fitting SPIV to data, formal hypotheses tests may be performed by comparing the approximate marginal likelihoods of the specific models for the retained instruments, biomarkers, and target outcomes. One way of evaluating these is by the Laplace approximation at θ_{MAP} (e.g. [21]).

Feature Selection

Due to the choice of the sparseness-inducing prior on model parameters (2), our approach may be used for tasks where the number of dimensions p exceeds the number of available observations n. However, for genome-wide studies the dimensionality of data needs to be reduced in order for the inference to be computationally tractable. For example, for a setting with $\sim O(10^5)$ genotypic features and $\sim O(10^4)$ biomarkers corresponding to gene expression profiles, the number of interaction terms between instruments and biomarkers may exceed $O(10^9)$, which is expensive to analyze or even keep in memory. We therefore apply subset selection methods to reduce the number of interactions in the SPIV model to $\sim O(10^5)$.

We note that for the SPIV model with high-dimensional biomarker vectors x (such as gene expressions) and descriptors of genotypic variations g (such as vectors of expected founder haplotypes corresponding to each single nucleotide polymorphism), both the number of biomarkers and the number of instruments may need to be reduced. We combine several common feature selection methods based on filters and forward selection [12], as well as methods based on sparse linear regression [37, 44] in a bootstrapping setting, where feature selection methods are applied multiple times for different samples of the training data.

In order to select features to use in the SPIV model, we applied filters based on several approximations of the unconditional mutual information $I(x_i;y)$ between feature x_i and output y, and the conditional mutual information $I(x_i;y|\mathsf{e})$ given the external covariates e (such as gender and age). While the filter-based methods are computationally efficient and relatively insensitive to the measurement noise, they may produce highly redundant representations and obscure some of the weaker (but nevertheless important) dependencies. The reason for this is that the features are analyzed individually, rather than jointly; thus, filter-based methods are not guaranteed to produce most informative sets of features. We therefore also applied two feature selection approaches based on the forward selection, where the regression accuracy and the approximate mutual information were used as selection criteria to iteratively determine new features to join with the current set in order to explain the residual structure.

The forward selection approach with the information-theoretic criterion is less common than the standard filters or step-wise regression (e.g. [12]), and we describe it in more detail. Here each step of the forward selection procedure was aimed at finding a set of inputs which were *jointly* predictive about the outputs, by maximizing the mutual information $I(\mathsf{x},\mathsf{e};\mathsf{y})$ with respect to the subset of features $\mathsf{x} \stackrel{\text{def}}{=} \{x_i\}$. Note that this is in contrast to filter methods, where the information content between a *single* feature x_i and outcome y is maximized. Since $I(\mathsf{x},\mathsf{e};\mathsf{y}) = I(\mathsf{x};\mathsf{y}|\mathsf{e}) + I(\mathsf{e};\mathsf{y})$, the equivalent optimization task is $\mathsf{x} = \arg\max_{\mathsf{x}} I(\mathsf{x},\mathsf{e};\mathsf{y}) \equiv$

$$\equiv \arg\max_{\mathsf{x}} \left\langle KL\left(p(\mathsf{y}|\mathsf{e},\mathsf{x}) \| p(\mathsf{y}|\mathsf{e})\right)\right\rangle_{p(\mathsf{y},\mathsf{e},\mathsf{x})}, \qquad (4)$$

where $\langle \ldots \rangle_p$ denotes an average over p, and $KL(q\|p)$ is the Kullback-Leibler divergence between distributions q and p (e.g. [6]). By utilizing the chain rule for mutual information, the optimization can be performed sequentially. For Gaussians, it reduces to optimizing the log-determinant of the conditional covariance matrix $\mathrm{cov}(\mathsf{y}|\mathsf{x},\mathsf{e})$. Note that from (4), any new subset of features is deemed to be predictive about the outputs only if the predictions from the new augmented set are significantly different from the predictions from the previously selected features, which helps in finding non-redundant predictors.

In addition to filters and forward-selection approaches, we also applied less greedy methods based on the sparse linear regression, which combine predictive inference with subset selection in a continuous optimization setting. Here we used the LASSO and elastic net methods [37, 44], with the hyperparameters set by cross-validation.

Finally, we note that each feature selection approach was applied in two different contexts:

1. in the *instrumental variable (IV)* setting, we first selected a subset of biomarkers $\tilde{X}$ predictive of the outcome y, and then selected a subset of instruments $\tilde{G}$ predictive of $\tilde{X}$. This selection is useful for identifying target biomarkers most strongly associated with disease outcomes, and genetic instruments explaining regularities in the candidate biomarkers. This setting is important for detecting possible causes of diseases;

2. in the *quantitative trait loci (QTL)* setting, we first selected a subset of genetic instruments $\hat{G}$ predictive of the outcome y, and then selected a subset of biomarkers $\hat{X}$ that was most informative about the chosen instruments $\hat{G}$. This selection is most appropriate for fine-mapping a genotypic region associated with a disease outcome or quantitative trait to a few individual genes.

The IV and QTL features were combined to produce the joint sets $\tilde{X} \cup \hat{X}$ and $\tilde{G} \cup \hat{G}$ for each iteration of the bootstrapping procedure.

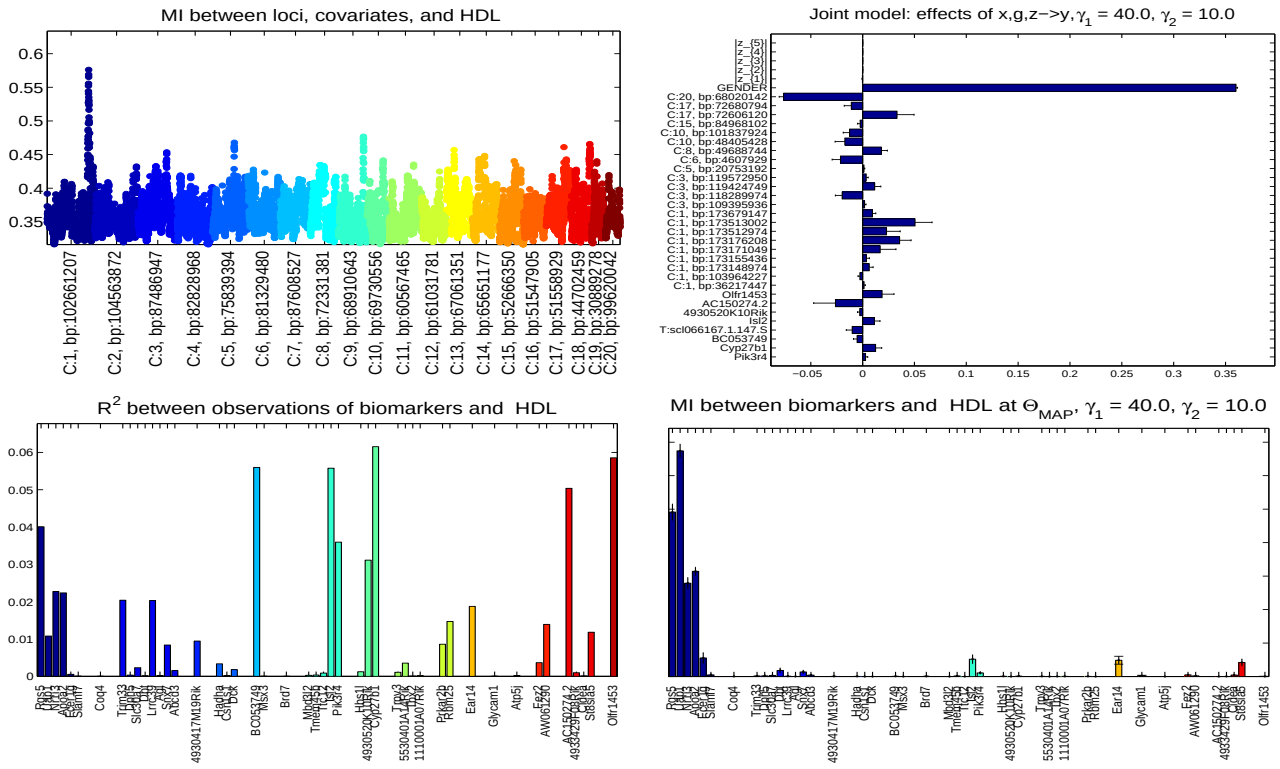

Figure 3: Sparse instrumental variables for the mouse dataset. *Top left:* scatter plot of the mutual information $I(\{g_i, \mathsf{e}\}; \tilde{y})$ between each locus g_i and environmental covariates $\mathsf{e} = \{age, sex\}$, and the observed HDL measurements. *Bottom left:* square of the conditional correlation coefficient $R^2(\tilde{x}_i, \tilde{y}|\mathsf{e})$ between observations of the gene expressions and HDL. Note a cluster of middle-rank correlations at chromosome 1, and several spikes of higher correlations scattered along the genome. *Top right:* maximum a-posteriori weights θ_{MAP}. Concentration parameters $\gamma_1 \approx 40$ and $\gamma_2 \approx 10$ have been obtained by cross-validation. Note a cluster of pleiotropic links on chromosome 1 at about 173 MBP, and nonzero direct effects of 8 biomarkers. *Bottom right:* Mutual information $I(x_i; y|\mathsf{e}, \theta_{MAP})$ between the underlying biomarkers and the unobserved trait expressed from the model at θ_{MAP}, under the joint Gaussian assumption. VSN preprocessing [14] has been used for transforming the RNA transcripts as detailed in [13].

The final set of biomarkers and instruments to use in the SPIV model was then constructed based on the features selected most frequently for multiple bootstrap subsamples.

3. RESULTS

To demonstrate SPIV for causality and fine-mapping studies, we examined effects of gene transcript levels in the liver on plasma high-density lipoprotein (HDL) cholesterol levels for a population of 260 heterogeneous stock mice genotyped and phenotyped for the trait of interest. Details of the data used in the experiments, as well as processing of the gene expressions, is described in Appendix A. At each of the 12500 retained marker loci, genotypes were described by 8-D vectors of expected founder ancestry proportions inferred from the raw marker genotypes by an HMM-based reconstruction method [24]. The genetic variables were augmented by age and sex. The full set of phenotypic biomarkers consisted of levels of 47429 transcripts, appropriately transformed and cleaned (see Appendix A for details). Before applying our method, we decreased the dimensionality of the genetic features and RNA expressions by using a combination of seven feature (subset) selection methods, based on applications of filters, greedy (step-wise) regression, sequential approximations of the mutual information between the retained set and the outcome of interest, and applications of regression methods with LASSO and elastic net shrinkage priors for the genotypes $\mathbf{g}$, observed biomarkers $\tilde{\mathbf{x}}$, and observed HDL measurements $\tilde{\mathbf{y}}$ as discussed in Section 2. For the sparse regression approaches, global hyperparameters were obtained by 10-fold cross-validation. After applying subset selection methods, the data typically contained $\sim O(10^3)$ instruments and $\sim O(10^2)$ biomarkers.

The results of our analysis of this dataset are shown on Figure 3. The *top right* plot shows maximum a-posteriori weights θ_{MAP} computed by running the EM-like optimization procedure to convergence from 20 random initializations, which took approximately 9.5 hours of execution time (unoptimized Matlab code, simple desktop). Note that only a fraction of the variables remains in the posterior. The spikes of the pleiotropic activations in sex chromosome 20 and around chromosome 1 are consistent with the biological knowledge [39]. The biomarker with the strongest effect on HDL (computed as the mean MAP weight $w_i : x_i \to y$ divided by its standard deviation over multiple runs) is the expression of *Cyp27b1* (gene responsible for vitamin D metabolism). Knockout of the *Cyp27b1* gene in mice has been shown to alter body fat stores [25], and this might be expected to affect HDL cholesterol levels. A subsequent comparison of specific reverse, pleiotropic, and causal models for *Cyp27b1*, HDL, and the whole vector of predictive loci indicated a slight prefer-

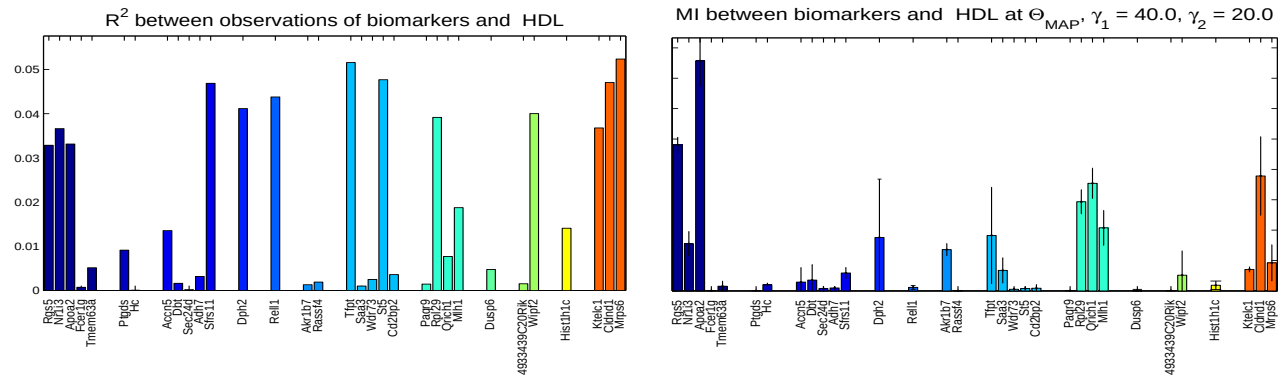

Figure 4: Conditional correlation coefficients $R^2(\tilde{x}_i,\tilde{y}|\mathbf{e})$ and mutual information $I(x_i;y|\mathbf{e},\theta_{MAP})$ for the raw gene expression data. Note significant differences from Figure 3 in both raw correlations and fine-mapping results. Both *Apoa2* and *Rgs5* remain the most significant predictive factors for HDL despite a change in the processing.

ence for the reverse hypothesis (with the ratio of Laplace approximations of the marginal likelihoods of reverse *vs* causal models of $\approx 1.95 \pm 0.27$). This is in contrast to the LCMS which is strongly affected by the choice of an instrument (Figure 1 shows the results for *Cyp27b1*, HDL, and the same choice of instruments). In this case, no hidden confounders appear to have strong effects on the outcome (which is not true in general). Adjusting for sex and age prior to performing feature selection and inference did not significantly change the results.

We can also apply SPIV to map a genotypic region associated with a trait of interest to a few most informative genes (the fine-mapping problem). Figure 3 (*bottom right*) shows the mutual information $I(x_i,y|\mathbf{e}=\{age,sex\})$ between the underlying biomarkers and unobserved HDL levels expressed from the SPIV model at the optimal parameter settings θ_{MAP}. The mutual information takes into account not only the strength of the direct effect of x_i on y (Figure 3, *bottom left*), but also associations with the pleiotropic instruments, strengths of the pleiotropic effects, dependencies between the instruments, and effects of the latent factors confounding the association between the biomarkers and the outcome. The majority of transcripts predictive about HDL are fine-mapped to a small region on chromosome 1 which includes *Uap1*, *Rgs5*, *Apoa2*, and *Nr1i3*. As we note in [1], the informativeness of these genes about the HDL cholesterol cannot be inferred simply from correlations between the measured gene expression and HDL levels; for example, when ranked in accordance to $R^2(\tilde{x}_i,\tilde{y}|age,sex)$, the top 4 genes have the rankings of 838, 961, 6284, and 65 respectively. The findings are also biologically plausible and consistent with high-profile biological literature (with associations between *Apoa2* and HDL described in [39], and strong links of *Rgs5* to metabolic traits discussed in [5], while *Nr1i3* and *Uap1* are their neighbors on chromosome 1 within $\sim 1Mbp$). Note that the couplings are via the links with the pleiotropic genetic markers on chrom 1 at $\sim 173Mbp$.

SPIV results appear to be stable for different choices of feature selection methods, data adjustments, and algorithm runs. We note however that different results could potentially be obtained based on the choice of animal populations or processing of the biomarker (gene expression) measurements. This is demonstrated on Figure 4, which shows the squares of the conditional correlation coefficients $R^2(\tilde{x}_i,\tilde{y}|\mathbf{e})$ and the mutual information $I(x_i;y|\mathbf{e},\theta_{MAP})$ expressed from the SPIV model at Θ_{MAP} for the biomarkers and HDL for a different processing of the gene expression data. Here we used the commonly applied log-transformation of the expressions followed by the centering and scaling (*cf* Huang et. al.'s [13] processing outlined in Appendix A). Despite a significant change in both the data and the mutual information profile, *Apoa2* and *Rgs5* are still selected as the most informative factors for HDL, which is consistent with [39] and [5]. However, a subsequent analysis of causality in this case showed strong effects of the latent confounders. Definitive confirmation of these relationships would require gene knock-out experiments. Significantly extended details of the methodology, experimental setup, and applications to genetic datasets will be published in specialized biology journals.

4. DISCUSSION

Whether or not causation may be inferred from observational data has been a matter of philosophical debate. Pearl [29] argues that causal assumptions cannot be verified without experimental interventions, and that there is nothing in the probability distribution $p(x,y)$ which can tell whether a change in x may have an effect on y. Much of the work of Pearl and his followers focuses on the question of identifiability, i.e. determining sets of graph-theoretic conditions when a post-intervention distribution $p(y|do(x))$ may be uniquely determined from a pre-intervention distribution $p(y,x,z)$, e.g. [28, 4, 33]. If the causal effects are shown to be identifiable, their magnitudes can be obtained by statistical estimation, which for common models often reduces to solving systems of linear equations.

In this paper, we do not explore identifiability conditions of the extended instrumental variables model. Instead, we try to develop a practical approach for determining a set of *candidate causes* of an outcome for a large partially observed under-determined genetic problem. The approach builds on the instrumental variable methods that were historically used in epidemiological studies, and on approximate Bayesian inference in sparse linear latent variable models. Specific modeling hypotheses are tested by comparing approximate marginal likelihoods of the corresponding direct, reverse, and pleiotropic models with and without latent

confounders. The approach is largely motivated by the observation that independent variables do not establish a causal relation, while strong unconfounded direct dependencies retained in the posterior modes even under large sparseness-inducing penalties may indicate potential causality and suggest candidates for further controlled experiments. We note here that from the Bayesian perspective, the problem of inferring the direction of causality may be viewed as that of model selection, where a model $\mathcal{M}_{x\to y}$ is compared with $\mathcal{M}_{y\to x}$. Unless the priors are chosen in such a way that $\mathcal{M}_{x\to y}$ and $\mathcal{M}_{y\to x}$ also have identical posteriors, it may be possible to infer the direction of the causal arrow (see MacKay [22], Section 35). Here we follow [22] in allowing for flexible priors of the models.

Technically, our sparse instrumental variables framework (SPIV) may be viewed as an extension of LASSO and elastic net regression which allows for latent variables and pleiotropic dependencies. While being particularly attractive for genetic studies, SPIV or its modifications may potentially be applied for addressing more general structure learning tasks. For example, when applied iteratively, SPIV may be used to guide search over richer model structures (where a greedy search over parent nodes is replaced by a continuous optimization problem, which combines subset selection and regression in the presence of latent variables). This may be useful, for example, for addressing complex pathway identification studies. In contrast to the vast majority of other recent model selection and structure learning methods from machine learning literature (e.g. [34] and references, [18], [20], [23]), SPIV is applicable in the practically interesting setting with high-dimensional data and latent variables. Other extensions of the framework could involve hybrid (discrete- and real-valued) outcomes with nonlinear/nongaussian likelihoods.

Fundamentally, SPIV extends ideas of the instrumental variable analysis in epidemiological studies by properly addressing situations when the genetic variables may be direct causes of the hypothesized outcomes. It overcomes limitations of the LCMS method by modeling joint effects of genetic loci and biomarkers in the presence of noise and latent variables. This work is based on an efficient MAP treatment of the sparse Bayesian linear modeling framework, which imposes a penalty on rich structures by using a sparsity-inducing prior. In principle this approach may be used for an effective screening of potentially interesting genotype-phenotype and biomarker-phenotype associations in genome-wide studies. It may also be used for identifying specific genes associated with phenotypic outcomes. The approach has wide application in identification of biomarkers as possible targets for intervention, or as proxy endpoints for early-stage clinical trials.

Acknowledgements
The development of the sparse instrumental variables approach was supported by MRC grant G0800604 to Paul McKeigue, and by the European Union's Seventh Framework Programme (FP7/2007-2013) for the Innovative Medicine Initiative under grant agreement No IMI/115006 (the SUMMIT Consortium). Jon Krohn acknowledges support of the Wellcome Trust.

5. REFERENCES

[1] F. V. Agakov, P. McKeigue, J. Krohn, and A. Storkey. Sparse instrumental variables (SPIV) for genome-wide studies. In *Neural Information Processing Systems*, 2010.

[2] J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables (with discussion). *J. of the Am. Stat. Assoc.*, 91:444–455, 1996.

[3] R. J. Bowden and D. A. Turkington. *Instrumental Variables*. Cambridge Uni Press, 1984.

[4] C. Brito and J. Pearl. Generalized instrumental variables. In *UAI*, 2002.

[5] Y. Chen, J. Zhu, and P. Y. Lum et. al. Variations in DNA elucidate molecular networks that cause disease. *Nature*, 452:429–435, 2008.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 1991.

[7] K. Demarest, J. Koyner, Jr. J. McCaughran, L. Cipp, and R. Hitzemann. Further characterization and high-resolution mapping of quantitative trait loci for ethanol-induced locomotor activity. *Behav Genet*, 31:79–91, 2001.

[8] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. of the Am. Stat. Assoc.*, 96(456):1348–1360, 2001.

[9] M. Figueiredo. Adaptive sparseness for supervised learning. *IEEE Trans. on PAMI*, 25(9), 2003.

[10] I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[11] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2003.

[12] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[13] G. J. Huang, S. Shifman, and W. Valdar et. al. High resolution mapping of expression QTLs in heterogeneous stock mice in multiple tissues. *Genome Research*, 19(6):1133–40, 2009.

[14] W. Huber, A. von Heydebreck, and H. Sultmann et. al. Variance stabilisation applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:96–104, 2005.

[15] J. Jia and B. Yu. On model selection consistency of the elastic net when $p \gg n$. Technical Report 756, UC Berkeley, Department of Statistics, 2008.

[16] M. B. Katan. Apolipoprotein E isoforms, serum cholesterol and cancer. *Lancet*, i:507–508, 1986.

[17] W. J. Kent. BLAT the BLAST-like alignment tool. *Genome Research*, 12:656–664, 2002.

[18] S. Kim and E. Xing. Statistical estimation of correlated genome associations to a quantitative trait network. *PLOS Genetics*, 5(8), 2009.

[19] D. A. Lawlor, R. M. Harbord, and J. Sterne et. al. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. in Medicine*, 27:1133–1163, 2008.

[20] M. H. Maathius, M. Kalisch, and P. Buhlmann. Estimating high-dimensional intervention effects from observation data. *The Ann. of Stat.*, 37:3133–3164, 2009.

[21] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4:415–447, 1992.

[22] D. J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge Uni Press, 2003.

[23] J. Mooij, D. Janzing, J. Peters, and B. Schoelkopf. Regression by dependence minimization and its application to causal inference in additive noise models. In *ICML*, 2009.

[24] R. Mott, C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Nat. Acad. Sci. USA*, 97:12649–12654, 2000.

[25] C. J. Narvaez and D. Matthews et. al. Lean phenotype and resistance to diet-induced obesity in vitamin D receptor knockout mice correlates with induction of uncoupling protein-1. *Endocrinology*, 150(2), 2009.

[26] R. M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996.

[27] T. Park and G. Casella. The Bayesian LASSO. *J. of the Am. Stat. Assoc.*, 103(482), 2008.

[28] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge Uni Press, 2000.

[29] J. Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

[30] J. M. Robins and S. Greenland. Identification of causal effects using instrumental variables: comment. *J. of the Am. Stat. Assoc.*, 91:456–458, 1996.

[31] E. E. Schadt, J. Lamb, X. Yang, and J. Zhu et. al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genetics*, 37(7):710–717, 2005.

[32] M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *JMLR*, 9, 2008.

[33] I. Shpitser and J. Pearl. Identification of conditional interventional distributions. In *UAI*, 2006.

[34] R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *JMLR*, 7, 2006.

[35] L. C. Solberg, W. Valdar, D. Gauguier, and G. Nunez et. al. A protocol for high-throughput phenotyping, suitable for quantitative trait analysis in mice. *Mamm Genome*, 17(2):129–146, 2006.

[36] R Development Core Team. A language and environment for statistical computing. R Foundation for Statistical Computing, 2004.

[37] R. Tibshirani. Regression shrinkage and selection via the lasso. *JRSS B*, 58(1):267–288, 1996.

[38] M. E. Tipping. Sparse Bayesian learning and the RVM. *JMLR*, 1:211–244, 2001.

[39] W. Valdar, L. C. Solberg, and S. Burnett et. al. Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics*, 38:879–887, 2006.

[40] W. N. Venables and B. C. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, 2002.

[41] M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using L1-constrained quadratic programmming. *IEEE Trans. on Inf. Theory*, 55:2183 – 2202, 2007.

[42] M. Yuan and Y. Lin. On the nonnegative garrote estimator. *JRSS:B*, 69, 2007.

[43] J. Zhu, M. C. Wiener, and C. Zhang et. al. Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *PLOS Comp. Biol.*, 3(4):692–703, 2007.

[44] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *JRSS:B*, 67(2), 2005.

A. DATA

Original Northport Heterogeneous Stock mice were obtained from Robert Hitzemann of the Oregon Health Sciences Unit (Portland, Oregon). At the time the animals arrived, they had passed 50 generations of pseudorandom breeding [7]. The ancestors of the Heterogeneous Stock mice (i.e., prior to pseudorandom breeding) were eight inbred strains of Mus musculus: A/J, AKR/J, BALBc/J, CBA/J, C3H/HeJ, C57BL/6J, DBA/2J, and LP/J [7]. The animals were bred for phenotyping in a colony established at the University of Oxford. Animals were housed at a maximum of six per cage (mean of four) and maintained on a 12:12 light:dark cycle with ad libitum access to food and water. All of the genotyped Oxford HS mice (n = 1940, including 1000 males) were used for the analyses.

Study design
(From [35]): One thousand nine hundred and forty Heterogeneous Stock mice (1000 male, 940 female) were put through a battery of the tests. Any particular phenotype test was carried out only once for each animal and the test was carried out on approximately the same day after each animals birth. Six experimenters phenotyped 95% of the mice. Six other experimenters phenotyped the other 5%.

Phenotypes
The phenotypes in this study were selected to assess genetic influences on obesity and immunology.

Genotypes
(From [39]): From across the mouse genome, 13459 SNPs were genotyped per animal by Illumina using their BeadArray platform. Where possible, SNPs were selected that were polymorphic in at least some of the eight inbred HS founder strains.

RNA Transcript Collection and Analysis
(From [13]): For 260 of the 1940 genotyped HS mice, liver tissue was frozen in liquid nitrogen and homogenised. RNA was extracted from the tissue and messenger RNA molecules were amplified. Labelled messenger RNA was hybridised to the Illumina Mouse WG-6 v1 BeadArray, which contains 47429 unique RNA probe sequences. Scans of the expression signals on the arrays were imported into Illumina BeadStudio 3.0, allowing background-subtracted signal values to be generated for each of the probe sequences. These data were exported to the R statistical package [36] for normalization via the software package `vsn` [14]. Subsequently, the Box-Cox procedure of the MASS package [40] for R was employed to normalise the residuals of each of the probe sequences when fit to a linear model of relevant predictors, i.e. experimenter and batch. BLAT [17] was used to align the transcripts to physical locations on the mouse genome.