Large-scale Bayesian Inference for Genetic Association with Multiple Outcomes

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the service académique.

Thèse n. 9139
présentée le 11 décembre 2018
à la Faculté des Sciences de Base
Chaire de Statistique programme doctoral en Mathématique
École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Hélène Ruffieux

acceptée sur proposition du jury:

Prof K. Hess Bellwald, président du jury
Prof A. C. Davison, directeur de thèse
Dr J. Hager, co-directeur de thèse
Prof C. C. Holmes, rapporteur
Prof S. Morgenthaler, rapporteur
Prof S. Richardson, rapporteur

Lausanne, EPFL, 2018

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Support statement

This work was supported by Nestlé Research and by the Chair of Statistics of École Polytechnique Fédérale de Lausanne. I would also like to express my gratitude to Leonardo Bottolo and Sylvia Richardson for welcoming me at the MRC Biostatistics Unit in Cambridge from September to December 2017, as well as for their invaluable guidance then and since. I am also grateful to Benjamin Fairfax for enlightening discussions and for providing us data.

Lausanne, 6 November 2018 H. R.

Abstract

Genetic association studies have become increasingly important in understanding the molecular bases of complex human traits. The specific analysis of intermediate molecular traits, via *quantitative trait locus* (QTL) studies, has recently received much attention, prompted by the advance of high-throughput technologies for quantifying gene, protein and metabolite levels. Of major interest is the detection of weak *trans*-regulatory effects between a genetic variant and a distal gene product. In particular, *hotspot* genetic variants, which remotely control the expression of many molecular outcomes, may initiate decisive functional mechanisms underlying disease endpoints.

This thesis proposes a Bayesian hierarchical approach for joint analysis of QTL data on a genome-wide scale. We consider a series of parallel sparse regressions combined in a hierarchical manner to flexibly accommodate many responses (molecular expression outcomes) and predictors (genetic variants), and we present new methods for large-scale inference.

Existing approaches have limitations. Conventional marginal screening does not account for local dependencies and association patterns common to multiple outcomes and genetic variants, whereas joint modelling approaches are restricted to relatively small datasets by computational constraints. Our novel framework allows information-sharing across outcomes and variants, thereby enhancing the detection of weak hotspot effects, and implements tailored variational inference procedures that allow simultaneous analysis of data for an entire QTL study, comprising hundreds of thousands of predictors, and thousands of responses and samples.

The present work also describes extensions to leverage spatial and functional information on the genetic variants, for example, using predictor-level covariates such as epigenomic marks. Moreover, we augment variational inference with simulated annealing and parallel expectation-maximization schemes in order to enhance exploration of highly multimodal spaces and allow efficient empirical Bayes estimation.

Our methods, publicly available as packages implemented in R and C++, are extensively assessed in realistic simulations. Their advantages are illustrated in several QTL applications, including a large-scale proteomic QTL study on two clinical cohorts that highlights novel candidate biomarkers for metabolic disorders.

Keywords: Bayesian sparse regression; Hierarchical model; High-dimensional data; Molecular quantitative trait locus analysis; Pleiotropy; Statistical genetics; Variable selection; Variational inference.

Résumé

Les études d'association génétique sont aujourd'hui largement utilisées pour tenter de comprendre les bases moléculaires de traits humains complexes. L'analyse de phénotypes moléculaires intermédiaires, via des études de *locus à caractères quantitatifs* (QTL), a récemment fait l'objet d'une attention particulière, eu égard aux progrès techniques réalisés en matière de quantification à haut rendement de niveaux de gènes, de protéines et de métabolites. La détection de faibles effets *trans* entre un variant génétique et le produit d'un gène distant est d'un intérêt majeur. En particulier, les variants génétiques *pléiotropiques*, qui contrôlent à distance l'expression de nombreux phénotypes moléculaires, pourraient être à l'origine de mécanismes fonctionnels déterminants pour le développement de maladies.

Cette thèse propose une approche hiérarchique bayésienne pour l'analyse multivariée de données QTL à l'échelle du génome. Nous considérons une série de régressions parallèles éparses, combinées de manière hiérarchique, pour modéliser de manière flexible un grand nombre de réponses (phénotypes moléculaires) et de prédicteurs (variants génétiques), et nous présentons de nouvelles méthodes pour l'inférence à grande échelle.

Les approches existantes ont des limitations. Les méthodes marginales conventionnelles ne tiennent pas compte des dépendances locales et des motifs d'association communs à plusieurs phénotypes moléculaires et variants génétiques, alors que les approches de modélisation conjointe sont limitées à des jeux de données relativement petits en raison de contraintes computationelles. Notre nouvelle approche permet un transfert d'information entre les phénotypes et variants, améliorant ainsi la détection des effets pléiotropiques faibles, et implémente des procédures d'inférence variationnelle permettant l'analyse simultanée de données pour une étude QTL complète, comprenant des centaines de milliers de prédicteurs, et des milliers de réponses et d'échantillons.

La thèse décrit également des extensions permettant de tirer parti d'informations spatiales et fonctionnelles sur les variants génétiques, par exemple à l'aide de covariables sur les prédicteurs telles que les marques épigénomiques. De plus, nous couplons nos algorithmes variationnels à des schémas de recuit simulé et d'espérance-maximisation parallèles afin de faciliter l'exploration d'espaces hautement multi-modaux et de permettre une estimation bayésienne empirique efficace.

Nos méthodes, en libre accès sous forme de packages implémentés en R et C++, sont rigoureusement évaluées par des simulations réalistes. Leurs avantages sont illustrés par plusieurs applications QTL, notamment par une étude QTL protéomique à grande échelle sur deux cohortes cliniques mettant en évidence de nouveaux biomarqueurs candidats pour les troubles métaboliques.

Mots clés : Analyse de locus à caractères quantitatifs ; Données en hautes dimensions ; Génétique statistique ; Inférence variationnelle ; Modèle hiérarchique ; Pléiotropie ; Régression bayésienne éparse ; Sélection de variables.

Contents

1 Introduction

The past decades have seen a multiplication of large-scale statistical applications, prompted by the proliferation of devices capable of measuring large volumes of information, whether on buying habits, health and life-style parameters, molecular entities, or even galaxies. As well as growing in size, the datasets collected are also growing in complexity, which calls for elaborate and flexible modelling strategies. Bayesian hierarchical modelling is a powerful framework for describing intricate dependencies across multiple sources of information, while conveying uncertainty in a coherent fashion.

Many applications entail many samples for a comparatively modest number of variables; this can pose computational challenges, but is relatively unproblematic from a statistical viewpoint, and a rich statistical and machine learning literature tackles capitalizing on the amount of data to effectively answer questions of interest. The large-scale paradigm faced by genomic researchers is of a completely different nature. The data produced by high-throughput technologies have unprecedented numbers of variables, many more than samples. Such complex *large p, small n* setups have triggered intense research efforts since the early 2000s, directed towards reconsidering traditional asymptotics, assessing finite-sample accuracy and characterizing computational complexity. But despite important progress, full awareness that statistical and computational approaches should be developed jointly has long been lacking, with inference algorithms often employed as mere secondary tools on complicated high-dimensional models. This is particularly true for Bayesian procedures, whose statistical advantages are often hindered by computational disadvantages that prevent their adoption in applications. Addressing this conflict in the context of genetic association with many outcomes is a central ambition of this thesis.

1.1 Motivation

This thesis is concerned with variable selection from high-throughput genetic data. Understanding the genetic architecture of complex human traits is an important step towards predicting health risks and developing effective therapies. Dramatic technological developments leading to the sequencing of the human genome at the end of the last millennium gave hopes for rapid breakthroughs in medical research. Several thousand studies have been designed with the twin purposes of giving deeper insights into the molecular processes underlying certain phenotypes (e.g., hypertension, obesity or types of cancer), and of the detection of reliable biomarkers for them. Some of these studies have led to major discoveries, such as obesity-associated risk alleles, whose encoded enzymes have shed light on

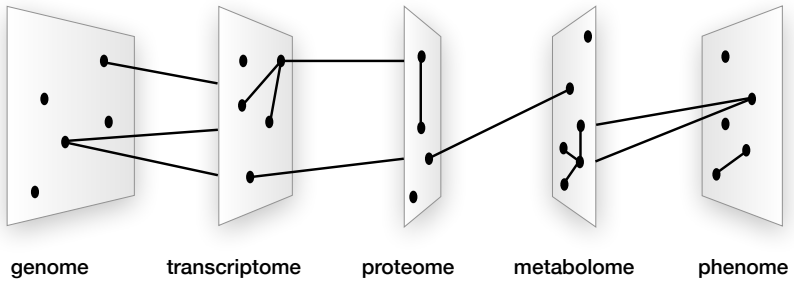

Figure 1.1 – Genotype-to-phenotype path (reproduced from Lusis, 2006).

the functional mechanisms behind body weight regulation (Loos and Yeo, 2014). But the expected transformation of medicine has barely begun, because of the complexity of biological mechanisms, the size, heterogeneity and noisy nature of the collected data, and the slow development of solid inference procedures.

The challenge of bridging this gap is currently being widely addressed, and a first step towards this involves a proper understanding of the molecular data available for analyses. The biomarkers sought in genetic association studies are genetic variants that influence the susceptibility to a disease or clinical *phenotype* of interest. The term *genetic variant* refers to a change at a specific location on the genome (*locus*), and the different versions of this change are called *alleles*. All the genetic variants contribute to defining the genetic makeup or *genotype* of an individual, but the most common variants in a population are *single nucleotide polymorphisms* (SNPs). A SNP is a variation in the nucleotide, A, T, G, or C, that is present to some appreciable extent in a population, i.e., the *minor* allele of this variation has frequency > 0.01 (Lewin et al., 2011, Chap. 2 and 5).

Most human conditions and diseases are *complex traits*. Complex traits are controlled by variants from many genetic loci, each of which may contribute a relatively small effect, yet whose cumulative contribution can be substantial (Lander and Schork, 1994). A variety of molecular mechanisms mediate the action of the genotype on the phenotype, via different types of entities (Figure 1.1). Genetic variants can regulate the expression of genes (the transcriptome), which may have downstream consequences at the protein (the proteome) or metabolite (the metabolome) level. These mechanisms can involve complex interactions that may subtly perturb pathways or networks underpinning the phenotype. They can also be tissue- and cell-type-specific, as well as retroactive, meaning that all molecular layers, except the genotype, can react to the phenotype and to environmental factors.

At the advent of high-throughput technologies, the editorial entitled “Talkin’ Omics” of the journal *Disease Markers* (issue of September 2001) thus observes

> “We are no longer satisfied to study a gene or a gene product in isolation, but rather we strive to view each gene within the complex circuitry of a cell. Understanding how genes and their products interact will open many exciting avenues.”

The concurrent surge of interest for such holistic approaches has led to using the generic designation of *systems biology* for them (Kitano, 2002). In the context of association studies, gene, protein and metabolite levels are often called *endophenotypes*, as they may be regarded as intermediate molecular proxies for clinical endpoints of interest. The former should have clearer connections with genetic variants than the latter, as they are less subject to environmental and behavioural effects (Gottesman

and Gould, 2003). This stimulated much of the current focus in statistical genetics on *molecular quantitative trait locus (QTL) analyses*, which assess how genetic variants control molecular levels at a genome-wide scale; this thesis tackles efficient modelling and inference for such molecular QTL studies.

Molecular QTL studies differ from classical genome-wide association studies in the types of analyses and questions addressed. We discuss this for *expression quantitative trait locus* (eQTL) analyses, which study the effects of genetic variants on the expression of transcripts or genes, but the same considerations apply to protein or metabolite outcomes, involved in so-called pQTL or mQTL analyses. The data used for eQTL studies usually involve several hundreds thousand SNPs and thousand transcript expression outcomes. Genetic variants can act locally, affecting the expression of a nearby gene (*cis*-eQTL) or they can alter expression of remote transcripts (*trans*-eQTL). Understanding by which mechanisms *trans*-regulation can take place, via a local *cis* gene that acts on a whole network or via other means, is a subject of active debate (Westra et al., 2013; Solovieff et al., 2013; Brynedal et al., 2017; Yao et al., 2017). In particular, the detection of *pleiotropic* variants, regulating the expression of tens or possibly hundreds of transcripts, is of great interest: such "*trans*-hotspot" genetic variants may provide insight into the regulatory landscape of the transcriptome, and hence into the mechanisms shaping the evolution of the human genome. They may also shed light on important functional processes underlying clinical traits and diseases. Hence, the task of identifying *trans* effects and hotspot variants is a central endeavour of molecular QTL studies, which is absent from genome-wide association studies that involve a single or a handful of clinical outcomes.

The locations and abundance of hotspots on the genome are largely unknown; as we next illustrate, detecting *trans* effects is difficult, and conventional approaches to association analyses have several drawbacks. We consider eQTL data comprising $> 24,400$ transcripts from $CD14^+$ monocytes and $> 380,000$ SNPs determined using Illumina arrays, for $n = 432$ samples from healthy European individuals; details are in Section 5.5, which describes a more extensive analysis using our work. Here we perform a classical marginal screening on all the transcripts and the $p = 29,607$ SNPs of chromosome one, that is, we regress each expression outcome on each genetic variant, one by one. This leads to the following observations (Table 1.1 and Figure 1.2): first, the estimated effect sizes of *trans* associations uncovered at Benjamini–Hochberg false discovery rate of 20% are substantially smaller than those of the *cis* effects. Second, although the screening uncovers about 2.5 times more *cis* associations than *trans* associations, about one-third of the former are essentially redundant: because of the local correlation structure on the genome (*linkage disequilibrium*), a single transcript is often assessed as under control by several genetic variants at a same locus, yet these variants are likely to be proxies for a single causal variant. Such scenarios are much less represented among the uncovered *trans* associations, as they concern only about 2% of them. Hence the large number of false positive *cis* associations reported by the marginal screening is likely to have hampered the detection of, weaker, *trans* effects.

It is easy to formalize this as a model misspecification issue in ordinary least squares regression. For an $n \times 1$ centered expression outcome vector $\boldsymbol{y}$, screening approaches assume a series of marginal models

$$\boldsymbol{y} = \boldsymbol{X}_s \beta_s + \boldsymbol{\varepsilon}, \qquad s = 1, \ldots, p, \tag{1.1}$$

where $\boldsymbol{X}_s$ an $n \times 1$ centered SNP vector, β_s is its regression coefficient, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ error term. Suppose that the true model is simply

$$\boldsymbol{y} = \boldsymbol{X}_1 \beta_1 + \boldsymbol{\varepsilon}, \qquad \beta_1 \neq 0$$

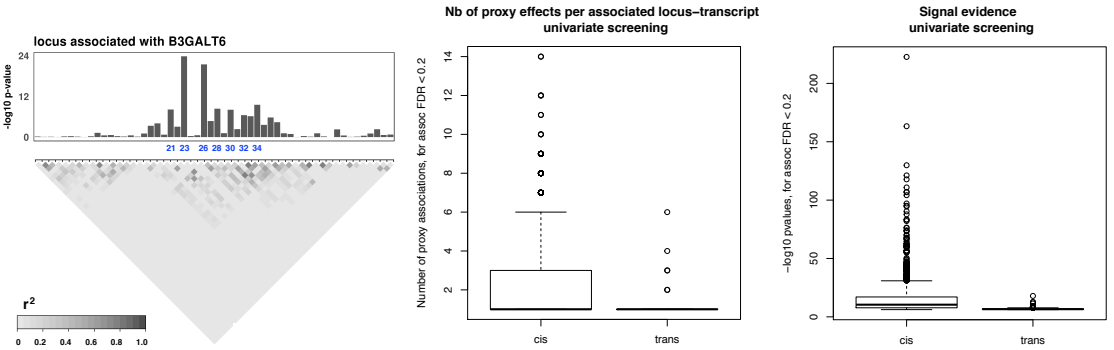

Figure 1.2 – Detection of *cis* and *trans* associations by a univariate screening approach using a Benjamini–Hochberg FDR of 20%. Effects between a gene and a SNP located less than 2 megabases (Mb) to it were defined as *cis* effects; the remaining effects were defined as *trans* effects. Left: example of linkage disequilibrium plot and Manhattan plot, here for associations with transcript *B3GALT6*. The blue labels indicate seven SNPs *cis*-acting on *B3GALT6* at FDR 20%; these effects are likely to be proxies for a single signal in the locus and arise because of the failure of univariate approaches to handle local correlation structures. Middle: number of such "proxy" associations for *cis* and *trans* effects, based on a linkage disequilibrium threshold of 0.5 (r^2 correlation) and window size 2 Mb. Right: $-\log_{10} p$-values for the declared effects.

	Number	Number after LD pruning	Magnitude of estimated effects
Cis effects	1,611	1,049	0.11 (0.10)
Trans effects	655	641	0.04 (0.03)

Table 1.1 – Detection of *cis* and *trans* associations by univariate screening using a Benjamini–Hochberg false discovery rate threshold of 0.2. Left: number of detected pairwise associations. Middle: number of detected pairwise associations after grouping those between a given transcript and several SNPs in linkage disequilibrium (LD) using r^2 correlation > 0.5 and window size 2 Mb. Right: average magnitude of regression estimates, and standard deviation in parentheses.

(the argument easily generalizes to an additive contribution of multiple SNPs). Based on (1.1), the ordinary least squares estimate $\hat{\beta}_s$ has mean

$$\mathrm{E}(\hat{\beta}_s) = \begin{cases} \beta_1, & s = 1, \\ \beta_1 (\boldsymbol{X}_s^T \boldsymbol{X}_s)^{-1} \boldsymbol{X}_s^T \boldsymbol{X}_1, & \text{otherwise}, \end{cases}$$

so, omitting the variable $\boldsymbol{X}_1$ yields a biased estimate for $\beta_s = 0$ ($s \neq 1$) if $\boldsymbol{X}_s$ is correlated with $\boldsymbol{X}_1$, which explains the redundant spurious effects declared by the marginal screening in regions of high linkage disequilibrium. This also suggests that the biases for SNPs in a locus with a *cis* effect (say, β_1 large) may be larger than many estimated *trans* effects. In Section 2.3.2, we will briefly review some general properties and recent modifications of marginal screening approaches.

There is a broad consensus in the biological community about the generality of the above remarks (Gilad et al., 2008; Mackay et al., 2009; Nica and Dermitzakis, 2013). It may be tempting to view them as consequences of the multiplicity burden entailed by molecular QTL problems. To date, marginal approaches have focused on detecting proximal *cis* associations, either to limit this burden or because the distal *trans* associations uncovered would fail to replicate. False discovery rate techniques with different corrections for *cis* and *trans* effects have been proposed (Peterson et al., 2016) and may

alleviate the issue. Rather than pursue this approach, we anticipate and tackle the question upfront, at the modelling stage, by building a hierarchical sparse regression model that can directly borrow information across genes.

Compelling evidence indicates that joint approaches outperform univariate screening approaches for variable selection in genetic association studies. Collectively accounting for the SNPs is needed to avoid the above omitted variable misspecification (Stephens and Balding, 2009; Guan and Stephens, 2011; Yang et al., 2012; Goddard et al., 2016), while leveraging information across multiple correlated transcripts is important to uncover pleiotropy (Jia and Xu, 2007; Richardson et al., 2010; Bottolo et al., 2011; Scott-Boyer et al., 2012). The reason why marginal screening approaches prevail relates to a combination of firmly-established practices and computational concerns. Indeed, none of the existing approaches allow joint analysis at the scale required by current molecular QTL studies, and they often necessitate drastic preliminary dimension reduction.

In addition to the *large p, small n* paradigm, whereby the number of genetic variants p greatly exceeds the number of samples n, molecular QTL studies also have a *large q* characteristic, entailed by the large number of expression levels q. The present thesis develops methodologies for this setup at both modelling and inference levels, thereby enabling expressive joint inference for realistic molecular QTL problem sizes. It builds on the flexible hierarchical model of Richardson et al. (2010), which it tailors to accommodate specific biological structures and heterogeneous sources of information, and it develops efficient variational inference procedures applicable to hundreds of thousands of SNPs and thousands of molecular expression outcomes.

1.2 Thesis outline

This thesis is structured as follows. Chapter 2 surveys basic concepts of high-dimensional Bayesian statistics and recent developments therein. Its first part reviews sparse modelling, from discrete mixture priors to continuous shrinkage priors, also touching on multiplicity control. Its second part discusses inference approaches, focusing on scalability issues, and presents variational inference.

Chapter 3 introduces our hierarchical sparse regression model for molecular QTL data and describes our variational procedure for it. The candidate predictors are SNPs and the responses are molecular expression outcomes. The use of variational approaches is relatively "non-standard" for Bayesian inference, for which sampling algorithms still prevail; the chapter strives to show the adequacy our algorithm for the model and for molecular QTL data.

The next three chapters propose enhancements and variants of the model and variational algorithm. They revolve around the concepts of linkage disequilibrium and pleiotropy: Chapters 4 and 6 propose using SNP data structures and external information to improve selection from loci with marked linkage disequilibrium, and Chapter 5 enhances the borrowing of strength from related expression outcomes to further adapt the model to the detection of hotspots.

More precisely, Chapter 4 is concerned with better handling dependence structures. It evaluates the performance of two model variants that encode linkage disequilibrium, and augments the inference procedure with a simulated annealing scheme that enhances exploration of multimodal spaces, without resorting to any structural information.

Chapter 5 describes a fully Bayesian second-stage model for hotspots, which bypasses sensitivity issues affecting the estimated propensity of SNPs to be hotspots. This proposal involves a flexible global-local representation for hotspots, which significantly improves their detection.

Chapter 6 considers involving a third data source: it extends the model with a second-stage regression to accommodate predictor-level covariates that inform the probability of candidate predictors to be involved in associations. In genetics, such covariates may be epigenomic marks that annotate the SNPs on their function, location or other features that may relate to their regulatory potential. This potential is inferred by the model, which selects the relevant marks using a dedicated spike-and-slab prior.

All chapters illustrate the performances of the approaches in numerical experiments that involve real molecular QTL data (transcriptomic, proteomic or metabolomic) or simulated data emulating real ones. Chapter 7 details an application on pQTL datasets from two obesity clinical cohorts. It supports the relevance of the replicated hits by assessing colocalization with epigenomic marks and known eQTL effects, and takes advantage of comprehensive clinical data to study links with dyslipidemia and insulin sensitivity.

We conclude with a general discussion and outline possible extensions in Chapter 8.

2 Hierarchical modelling and inference for high-dimensional data

In this chapter we review modelling approaches to high-dimensional regression in the Bayesian framework, and we evaluate possibilities for reliable and scalable inference. We restrict our discussion to the linear regression setting, from which essential properties of Bayesian variable selection procedures can be learnt. We focus on presenting an overview of general methodological aspects and recent developments, and defer the review of approaches tailored to the genetic context to subsequent chapters.

The chapter is organized as follows. Section 2.1 reviews sparse Bayesian modelling, its relations to frequentist penalized regression, and two different perspectives on it. Section 2.2 discusses multiplicity in variable selection tasks. Section 2.3 presents the main approaches to scaling Bayesian inference for high-dimensional problems. Section 2.4 focuses on the inference approach used in this thesis, namely, variational inference; it gives a general presentation of the approach and briefly outlines new developments.

2.1 Bayesian sparse regression modelling

2.1.1 Frequentist and Bayesian approaches to regularization

Consider inference about a p-variate parameter $\boldsymbol{\beta}$ in the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_n\right), \tag{2.1}$$

where $\boldsymbol{y}$ is an n-variate response, $\boldsymbol{X}$, an $n \times p$ design matrix and σ_ε^2, the error variance. For simplicity, the intercept is set to zero and the columns $\boldsymbol{X}_s$ ($s = 1, \ldots, p$) of $\boldsymbol{X}$ are measured on the same scale; this is without loss of generality, as it can always be achieved by centering $\boldsymbol{y}$ and standardizing each $\boldsymbol{X}_s$ to have mean zero and variance one.

We discuss model (2.1) in presence of high-dimensional predictors, i.e., for $p \gg n$. When $p > n$, the design matrix is singular, so inference requires structural assumptions on $\boldsymbol{\beta}$ in order to be well-posed. A natural approach is to enforce sparsity by assuming that only a handful of predictors may be associated with the response, constraining most regression coefficients to be (nearly) zero. It is questionable whether the sparsity assumption is always reasonable in practice; it is however not unreasonable in

genome-wide applications, as one typically expects a small fraction of the analyzed molecular entities to control an outcome of interest.

In the frequentist setup, Bühlmann and van de Geer (2011, Chap. 1) explain that optimal estimation properties may be obtained by imposing sparsity conditions of the form

$$\|\boldsymbol{\beta}\|_q^q \times \log p \ll n,$$

where $\|\cdot\|_q$ is the ℓ_q norm with $0 \leq q < \infty$ chosen depending on the context, and writing $\|\boldsymbol{\beta}\|_0^0 = \|\boldsymbol{\beta}\|_0 = \#\left\{1 \leq s \leq p : \beta_s \neq 0\right\}$. To achieve such regularization, high-dimensional regression is often framed as a minimization problem with objective function

$$\mathcal{L}(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{s=1}^{p} \rho\left(\beta_s\right), \tag{2.2}$$

for some penalty $\rho(\cdot)$ and parameter $\lambda > 0$. For instance, the choice of $\rho\left(\beta_s\right) = \left|\beta_s\right|$ corresponds to the popular least absolute shrinkage and selection operator (LASSO, Tibshirani, 1996), but there are many more possibilities; the related literature is reviewed by Bühlmann and van de Geer (2011). Frequentist uncertainty estimates are often difficult to construct because asymptotic normality arguments do not transfer trivially from the $p < n$ to the $p \gg n$ regime (Bhattacharya et al., 2015). Bootstrapping may instead be used but can produce unstable or poor variance estimates, see Kyung et al. (2010) for an extensive discussion in the context of the LASSO.

Most penalized estimators can be interpreted as the mode of a posterior distribution under an independent and identically distributed prior (Polson and Scott, 2010)

$$\mathrm{pr}\left(\beta_s\right) \propto \exp\left\{-\lambda\rho\left(\beta_s\right)\right\}, \qquad s = 1, \ldots, p. \tag{2.3}$$

For instance, the LASSO finds a maximum a posteriori (MAP) estimate for a Laplace shrinkage prior on β_s. This suggests that the entire posterior distribution may be used for inference, rather than just its posterior mode, and this is what Bayesian sparse regression enables. Hence, in contrast to frequentist penalized regression, Bayesian regression implicitly conveys uncertainty. Fully Bayesian versions of (2.2) also place a prior on the regularization parameter λ and infer it simultaneously with the model parameters, thereby avoiding plug-in solutions based on cross-validation or marginal maximum likelihood estimation.

The scope of Bayesian sparse regression goes beyond the relation with frequentist penalized regression, as sparsity is induced via the prior placed on the regression coefficients and MAP estimation under this prior may or may not coincide with an existing penalized method. A wide range of sparsity priors can be expressed using scale mixtures of normal densities:

$$\mathrm{pr}(\beta_s) = \int \mathcal{N}\left(\beta_s \mid 0, \omega_s\right) \mathrm{d}G\left(\omega_s\right), \qquad s = 1, \ldots, p, \tag{2.4}$$

for some distribution function G (Griffin and Brown, 2017). Bayesian optimality properties, however, usually do rely on frequentist perspectives. For posterior consistency, for example, one assumes the existence of an underlying true parameter $\boldsymbol{\beta}_0$, and evaluates convergence, in a suitable sense, of the posterior distribution to the Dirac measure of $\boldsymbol{\beta}_0$ as the amount of data grows indefinitely. Another characterization of posterior consistency requires the posterior probability assigned to any neighborhood of $\boldsymbol{\beta}_0$ to converge to unity (Ghosal and van der Vaart, 2017, Chap. 6). On top of assessing

consistency, it is often of interest to derive concentration rates, that is, to study at what rate $\boldsymbol{\beta}_0$ can be learnt, as this can be informative about the number of samples n needed to reach a desired accuracy, up to constants. This corresponds to finding the smallest shrinking ball that still contains all the posterior mass as $n \to \infty$. The first general results on consistency may be attributed to Doob (1949) and Schwartz (1965). The former stated that if the true parameter $\boldsymbol{\beta}_0$ is drawn from the prior, then the posterior of $\boldsymbol{\beta}$ is consistent almost everywhere. The latter obtained guarantees for consistency when the prior places positive mass on Kullback–Leibler neighborhoods of the true density (assuming that densities exist). Barron et al. (1999) and Ghosal et al. (1999) derived further similar theoretical results. Posterior concentration rates have been studied by Ghosal et al. (2000), Genovese and Wasserman (2000) and Shen and Wasserman (2001), among others, for general density estimation or estimation with mixture of normals. Ghosal and van der Vaart (2007) derived extensions for observations that are not necessarily independent nor identically distributed.

Asymptotic normality is another central ingredient of large-sample theory. The Bernstein–von Mises theorem is the Bayesian analog of the central limit theorem: it states that the posterior converges in distribution to a Gaussian distribution centered at the maximum likelihood estimator, under some conditions, including that the prior is strictly positive in a neighborhood of $\boldsymbol{\beta}_0$, and that the problem dimension is fixed and finite; the full conditions can be found in van der Vaart (2000, Chap. 10). Hence, when applicable, the Bernstein–von Mises theorem provides frequentist justifications for Bayesian inference; for instance, it ensures robustness of inferences to the choice of priors, and an asymptotic agreement between credible intervals and classical Wald intervals. Unfortunately the vanilla Bernstein–von Mises theorem doesn't hold in high dimensions: it requires fixed p, and the condition on the prior being strictly positive in a neighborhood of $\boldsymbol{\beta}_0$ is very strong for large parameter spaces. To date, attempts to obtain Bernstein–von Mises-type of theorems when $p \gg n$ were essentially unsuccessful: Ghosal (1999) obtained sufficient conditions for the asymptotic normality of the posterior of $\boldsymbol{\beta}$ the under model (2.1), but under the assumption that p grows much slower than n. Bontemps (2011) derived a Bernstein–von Mises theorem for semiparametric and non-parametric regression models, with a faster growth rate than Ghosal (1999), but assuming $p < n$. Moreover, for both results, the prior must be sufficiently flat in the vicinity of $\boldsymbol{\beta}_0$, which essentially rules out sparsity priors.

In the $p \gg n$ regime, the prior typically remains influential; in Chapter 5, we will illustrate how its choice may severely distort posterior inferences in high dimensions, and propose a solution for our model. As general asymptotic statements are difficult to obtain, guarantees for high-dimensional asymptotics are typically established for given likelihoods, priors and data-generating truths. For instance, common lines of research seek consistency and optimal concentration rates for the posterior distribution under a prior (2.4) with a specific choice G. In particular, strong results have been obtained for two important classes of priors (2.4), which we now discuss. To ease the presentation and unless stated otherwise, we describe these priors in the context of the normal means problem (Stein, 1981),

$$y_i = \beta_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0,1), \qquad i = 1, \dots, n, \tag{2.5}$$

where n is the dimensionality and is large. Theory for model (2.5) is frequently examined under a *nearly-black* sparsity assumption, i.e., assuming that the unknown true parameter $\boldsymbol{\beta}_0$ belongs to (Donoho et al., 1992; Johnstone, 1994)

$$l_0[p_n; n] = \left\{\boldsymbol{\beta} \in \mathbb{R}^n : \|\boldsymbol{\beta}\|_0 \leq p_n\right\}, \quad p_n = o(n), \quad n \to \infty. \tag{2.6}$$

2.1.2 Two-group shrinkage priors

We first discuss the *two-group* shrinkage priors, which model the signal and the noise with two dedicated components (Polson and Scott, 2010). They use a discrete mixture distribution for G in (2.4), which yields

$$\beta_i \mid \pi \sim \pi g_\beta + (1-\pi)\delta_0, \qquad i = 1, \dots, n, \tag{2.7}$$

where π is a mixing weight, g_β is an absolutely continuous density on $\mathbb{R}$ (centered normal in (2.4)), and δ_0 is the Dirac distribution. These priors are better known as *spike-and-slab* priors, the null effects being attributed to the "spike" degenerate distribution δ_0 at zero, and the non-null effects to the "slab" density g_β. They also allow handling separately the typical size of the nonzero coefficients, via g_β, and the sparsity level, via $\pi = \mathrm{pr}\left(\beta_i \neq 0\right)$, with prior average model size $n\pi$.

Spike-and-slab priors were proposed by Mitchell and Beauchamp (1988), George and McCulloch (1993), and further studied by Clyde et al. (1996) and Chipman (1996); they have become very popular in recent decades with the increasing availability of high-dimensional data. In the original formulation of Mitchell and Beauchamp (1988), g_β was a uniform distribution, but other unimodal symmetric distributions are today preferred, such as the centered normal distribution, which is used in many applications. The two-group class also comprises continuous relaxations of (2.7), whereby the Dirac mass is replaced with a peaked continuous density, although such specifications are less common in practice; the case of two centered normal distributions with variances $\sigma_0^2 \ll \sigma_\beta^2$ is often employed to discuss theoretical properties of two-group priors, see, e.g., George and McCulloch (1993), Ishwaran and Rao (2003), Ishwaran and Rao (2005) and Narisetty and He (2014).

The following lemma characterizes the shrinkage enforced by two-group priors on the posterior mean of $\boldsymbol{\beta}$.

Lemma 2.1.1 (Bhadra et al., 2017a)**.** *Assume the normal means model (2.5) and prior (2.7) for β_i, where g_β is a centered Gaussian distribution with variance σ_β^2. Then the posterior mean of β_i is*

$$\mathrm{E}\left(\beta_i \mid y_i\right) = \pi(y_i)\frac{\sigma_\beta^2}{1+\sigma_\beta^2} y_i, \tag{2.8}$$

where $\pi(y_i) = \mathrm{pr}\left(\beta_i \neq 0 \mid y_i\right)$, so that

$$\mathrm{E}\left(\beta_i \mid y_i\right) = \{1 + o(1)\}\,\pi(y_i)\, y_i, \qquad \sigma_\beta^2 \to \infty. \tag{2.9}$$

Equality (2.8) indicates that σ_β^2 enforces global shrinkage, while $\pi(y_i)$ adapts to the individual effects. If the mixing weight π in (2.7) is treated as unknown, then $\pi(y_i)$ adjusts to the overall sparsity in the data through its shared dependence upon π. The approximation $\mathrm{E}\left(\beta_i \mid y_i\right) \approx \pi(y_i)\, y_i$ holds for appropriately heavy tailed densities g_β; indeed,

$$\mathrm{E}\left(\beta_i \mid y_i\right) = \pi(y_i)\mathrm{E}_{g_\beta}\left(\beta_i \mid y_i\right),$$

where $\mathrm{E}_{g_\beta}(\cdot)$ is the expectation taken with respect to g_β (Carvalho et al., 2010).

The asymptotic behaviour of two-group priors is generally well understood. Specific optimality properties hold for the posterior of $\boldsymbol{\beta}$, depending on whether the mixing weight is estimated via empirical Bayes (Johnstone and Silverman, 2004), or assigned suitable Beta prior (Castillo and van der Vaart,

2012). The density g_β must satisfy appropriate tail conditions; in particular, Castillo and van der Vaart (2012) obtained optimal contraction rates when g_β is a Laplace or a Cauchy density. Ishwaran and Rao (2011) established an oracle property of the posterior mean for non-orthogonal and low-dimensional setups, and Narisetty and He (2014) obtained model selection consistency in high dimensions for suitable data-driven hyperparameters; both studies concern the continuous spike-and-slab model.

A reparametrization of spike-and-slab models consists in adding one level of hierarchy, by introducing a coefficient-specific latent variable γ_i,

$$\beta_i \mid \gamma_i \sim \gamma_i g_\beta + (1-\gamma_i)\delta_0, \qquad \gamma_i \mid \pi \sim \text{Bern}(\pi), \qquad i = 1, \ldots, n. \tag{2.10}$$

As we will see in Section 2.2, this formulation is prevalent in regression models as it yields useful posterior quantities for variable selection: replacing the index i in (2.10) by the predictor index s, the posterior mean of γ_s corresponds to the marginal posterior probability of inclusion of variable $\boldsymbol{X}_s$ in model (2.1), $\mathrm{E}(\gamma_s \mid \boldsymbol{y}) = \mathrm{pr}(\gamma_s = 1 \mid \boldsymbol{y})$, and hence is a direct measure of support for the presence or absence of individual associations. The binary variables γ_s are independent conditional on the prior probability of inclusion π but dependent marginally. If it is not justified in practice, this assumption could be relaxed by using a variable-specific inclusion probability π_s; we will see in Chapter 3 how such predictor-specific probability parameter will serve us to model "hotspot" SNPs in multiple-response contexts.

We end this section by discussing two important priors which, although they do not strictly fit within the classes (2.4) or (2.7), are two-group priors in essence, since they also frame inference about a sparse vector as a classification problem between the null and non-null hypotheses, $\beta_s = 0$ and $\beta_s \neq 0$. The g-prior (Zellner, 1986) for variable selection is a non-exchangeable prior having, for the alternative hypothesis,

$$\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, g \sim \mathcal{N}_{p_\gamma}\left(0, g\sigma_\varepsilon^2 \left(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma\right)^{-1}\right),$$

where $\boldsymbol{\beta}_\gamma$ is the vector gathering the $p_\gamma = \sum_{s=1}^{p} \gamma_s$ nonzero β_s, $\boldsymbol{X}_\gamma$ is the corresponding $n \times p_\gamma$ design matrix, $g > 0$ is a parameter controlling the expected sizes of effects, and σ_ε^2 is the response error variance of model (2.1). This prior is often employed for its conjugacy properties; the analytical expression of the marginal likelihood $\mathrm{pr}(\boldsymbol{y} \mid \boldsymbol{\gamma})$ involves a determinant term under the independent spike-and-slab prior (2.10), which is absent under the g-prior, allowing cheaper computations. The g-prior is data-dependent as it involves the design matrix $\boldsymbol{X}$, and should be used cautiously in presence of highly collinear predictors, as nearly singular $\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma$ may give rise to instabilities and poor inferences. The choice of the hyperparameter g has been the subject of much discussion, resulting in a variety of proposals of hyperpriors for it; some are reviewed in Liang et al. (2008) and more recent work towards more robust priors in terms of tail behaviour includes Maruyama and George (2011), and Bayarri et al. (2012).

Finally, Johnson and Rossell (2010) proposed a *nonlocal* prior whose signal component has negligible mass in a small neighborhood around zero, possibly allowing better distinction between the signal and the noise. This is in contrast with most sparse priors, which have highest probability mass at zero. To date, theoretical guarantees for nonlocal priors are few: in variable selection settings, Johnson and Rossell (2012) showed strong model selection consistency, but the proof is for regimes $p < n$.

2.1.3 One-group shrinkage priors

There is growing interest in priors termed *one-group* shrinkage priors, which, instead of classifying coefficients using a mixture of signal/noise components, infer them in a unified fashion by enforcing shrinkage with a single component.

The James–Stein shrinkage estimator (Stein, 1956; James and Stein, 1961), based on the prior

$$\beta_i \mid \sigma_0^2 \sim \mathcal{N}\left(0, \sigma_0^2\right), \tag{2.11}$$

may be seen as an early one-group prior. Stein (1956) made the counterintuitive observation that the maximum likelihood estimator is inadmissible when estimating three or more independent parameters. In fact, its quadratic loss always exceeds that of the James–Stein estimator. This revealed that shared dependence upon a global parameter, such as σ_0^2 in (2.11), can induce beneficial shrinkage and can yield estimates with substantially lower risk than conventional estimators. This finding has had a considerable impact since the 1960s, and laid the ground for intense developments in shrinkage estimation later, this time focusing on sparse problems.

Modern one-group priors correspond to choosing G in (2.4) to be an absolutely continuous distribution. For instance, the Laplace prior discussed in Section 2.1.1 is a one-group prior: it is obtained by taking G to be the exponential distribution with parameter $\left(2\sigma_0^2\right)^{-1}$, for σ_0, the Laplace scale parameter. Polson and Scott (2010) expressed one-group priors as *global-local* scale mixture priors,

$$\beta_i \mid \sigma_0^2, \lambda_i^2 \sim \mathcal{N}\left(0, \sigma_0^2 \lambda_i^2\right), \qquad \sigma_0 \sim f, \qquad \lambda_i \sim g, \tag{2.12}$$

where f and g are densities on $\mathbb{R}^+$. The global scale σ_0 controls overall shrinkage toward the origin, while the local scales λ_i allow coefficient-specific deviations in the level of shrinkage. These parameters are learnt from the data through their hyperpriors f and g. This interplay of local and global shrinkages gives rise to non-normal marginal densities for β_i that can model both sparse and heavy-tailed signals. This improves upon James–Stein-type estimators, which lose their optimal risk properties in sparse settings (Polson and Scott, 2009).

The emergence of one-group priors also relates to the idea that sparsity may arguably be better induced in a *weak sense*, i.e., assuming that most true coefficients are small, yet not exactly zero. For instance, despite their relation, the Laplace prior-based and LASSO estimations yield answers of different nature, namely weakly and strongly sparse estimates respectively, the former being based on posterior means and the latter, on posterior modes. There is much discussion as to whether *weak* or *strong* sparsity should be preferred. The question is linked to that of the pertinence of hypothesis testing in the Bayesian paradigm, which we will touch on briefly in Section 2.2.

Gelman (2006) and Polson and Scott (2010) discuss the adequacy of several prior choices for variances, f or g in (2.12), in terms of their degree of (un)informativeness and robustness. Gelman argues that the standard inverse-Gamma priors InvGamma$(\varepsilon, \varepsilon)$ with small $\varepsilon > 0$ proposed by Spiegelhalter et al. (1996) are often inappropriate for variance components. He illustrates that, on data for which the variance can take low values, inferences can be sensitive to the choice of ε, and InvGamma$(\varepsilon, \varepsilon)$ is not truly uninformative despite its reputation. Apart from the Laplace and Student-t marginal priors for β_i, with exponential and inverse-Gamma local variances respectively, two important examples of (2.12)

are the Strawderman–Berger prior (Strawderman, 1971; Berger, 1980),

$$\beta_i \mid \kappa_i \sim \mathcal{N}\left(0, \frac{1}{\kappa_i} - 1\right), \qquad \kappa_i \sim \text{Beta}\left(\frac{1}{2}, 1\right),$$

and the normal/inverted-beta prior with local variances following an inverted-Beta density, for $\alpha > 0$ and $\beta > 0$,

$$\text{pr}\left(\lambda_i^2\right) = \frac{\left(\lambda_i^2\right)^{\alpha-1}\left(1+\lambda_i^2\right)^{-\alpha-\beta}}{\text{B}(\alpha,\beta)}, \tag{2.13}$$

where $\text{B}(\cdot,\cdot)$ is the beta function. A popular special case is the horseshoe prior whose half-Cauchy local scales correspond to taking $\alpha = \beta = 1/2$ in (2.13),

$$\beta_i \mid \sigma_0^2, \lambda_i^2 \sim \mathcal{N}\left(0, \sigma_0^2\lambda_i^2\right), \qquad \lambda_i \sim \text{C}^+(0,1). \tag{2.14}$$

The horseshoe+ prior (Bhadra et al., 2017a) adds a level in the hierarchy,

$$\beta_i \mid \sigma_0^2, \lambda_i^2 \sim \mathcal{N}\left(0, \sigma_0^2\lambda_i^2\right), \qquad \lambda_i \mid \eta_i \sim \text{C}^+\left(0, \eta_i\right), \qquad \eta_i \sim \text{C}^+(0,1).$$

This leads to heavier marginal tails compared to the original horseshoe prior, which permits a better separation of the signals, or so the authors claim.

The differences between these proposals in handling shrinkage is best understood by inspecting the *shrinkage profiles* linked with their posterior expectations, as we next explain.

Lemma 2.1.2 (Adapted from Carvalho et al., 2009)**.** *Assume the normal means model (2.5) and prior (2.12) for β_i. Let*

$$\kappa_i = \frac{1}{1+\sigma_0^2\lambda_i^2}, \qquad \kappa_i \in (0,1). \tag{2.15}$$

Then the conditional posterior mean of β_i can be expressed as

$$\text{E}\left(\beta_i \mid y_i, \sigma_0^2, \lambda_i^2\right) = (1-\kappa_i) \times y_i + \kappa_i \times 0,$$

so

$$\text{E}\left(\beta_i \mid y_i, \sigma_0^2\right) = \left\{1 - \text{E}\left(\kappa_i \mid y_i, \sigma_0^2\right)\right\} y_i.$$

The parameter κ_i is called the *shrinkage factor*, as it represents the amount of weight placed on zero by the posterior mean of β_i. Lemma 2.1.2 is the global-local analog of Lemma 2.1.1, with $1 - \text{E}\left(\kappa_i \mid y_i, \sigma_0^2\right)$ mimicking the posterior probability of inclusion $\pi(y_i)$. The definition of κ_i in (2.15) also suggests that σ_0^2 must have substantial mass near zero to promote a strong shrinkage of unimportant coefficients (κ_i close to 1), while λ_i must have sufficiently fat tails to ensure that large coefficients are only minimally shrunk (κ_i close to 0). In Chapter 5, we will use these shrinkage profiles to define a multiplicity penalty on the response dimensionality.

Inspecting the marginal prior density of κ_i allows one to appreciate the specificities of the different global-local scale priors in terms of their behaviour near zero and in the tails; Figure 2.1 displays this density for the above-cited global-local scale priors, conditional on $\sigma_0 = 1$, as in Carvalho et al. (2009).

In particular, the horseshoe shape $\kappa_i \sim \text{Beta}(1/2, 1/2)$, obtained under the horseshoe half-Cauchy scales $\lambda_i \sim \text{C}^+(0,1)$, encodes the assumption that signals are *a priori* either large or nearly zero. The Student-t and Strawderman–Berger shrinkage profiles both have a pole at zero, reflecting the fat tails

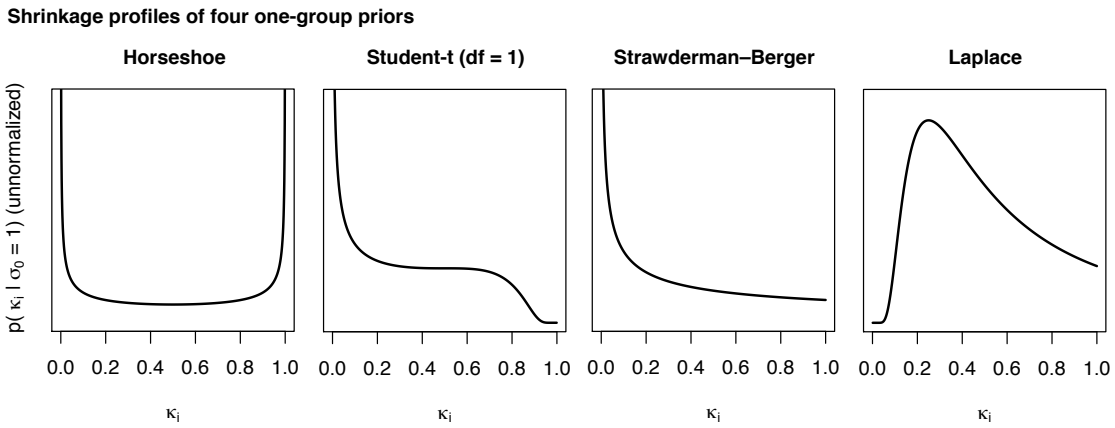

Figure 2.1 – (Unnormalized) prior densities of the shrinkage factor κ_s conditional on $\sigma_0 = 1$ for the sparse priors: horseshoe, Student-t (with one degrees of freedom, i.e., Cauchy), Strawderman–Berger and Laplace. Mass close to 1 induces shrinkage, while mass close to 0 leaves effects unshrunk. The difference between the horseshoe and the horseshoe + (not shown) shrinkage factors is imperceptible; the latter places slightly more mass near 0 and 1 than the former. This figure is reproduced from Carvalho et al. (2009).

of these priors, but are also bounded near $\kappa_i = 1$, indicating that irrelevant coefficients may be only partially shrunk to zero. The Laplace shrinkage profile has very little mass near zero because of the lighter tails of the exponential local variances. Hence, the Laplace prior can't prevent shrinkage of genuine coefficients. The concept of tail heaviness is often linked with that of the robustness of the prior distribution. However, it is not clear if heavy tails must always be preferred, especially in problems with flat likelihoods. This is discussed by Gelman et al. (2008) and Polson and Scott (2012), who recommend Student-t priors with degrees of freedom $\nu > 1$ over priors with Cauchy-like tails for local and global scales in weakly informative data situations, e.g., in case of separation with logistic regression. We will discuss and compare these choices in the context of our model in Chapter 5.

Theory for one-group shrinkage priors is at a relatively early stage. For the reasons outlined above, optimality properties for a given prior hinge on suitable tail conditions and probability mass near zero. Such properties are being intensively researched for the horseshoe prior, and yielded convincing results on concentration rates, see van der Pas et al. (2014), van der Pas et al. (2016) and van der Pas et al. (2017). van der Pas et al. (2016) and Ghosh and Chakrabarti (2017) extended the study to a general class of global-local shrinkage priors.

2.2 Multiplicity in Bayesian variable selection

The previous section focused on describing important approaches to modelling sparse vectors. This section builds on this overview to discuss Bayesian variable selection in high-dimensional regression. We thus leave the normal means setting and return to the linear regression model (2.1), using indices $s = 1,\ldots,p$ for variables and $i = 1,\ldots,n$ for samples.

In principle, deciding whether each parameter β_s inferred from (2.1) should be classified as signal or noise depends on the type of sparsity prior used for inference. In two-group models, the binary latent

variables γ_s are useful for selection. It is natural to perceive the highest posterior probability model

$$\arg \max_{\boldsymbol{\gamma} \in \{0,1\}^p} \text{pr}\left(\mathcal{M}_\gamma \mid \boldsymbol{y}\right), \tag{2.16}$$

where $\mathcal{M}_\gamma$ is the model that includes or excludes each of the p candidate predictors according to $\boldsymbol{\gamma}$, as the best model for selection purposes (Clyde, 1999). However in high dimensions, posterior probabilities for individual models are usually very small and their interpretation is less appealing. More importantly, identifying (2.16) is beyond reach for moderate-to-large p, as this would involve exploring 2^p possible models. In practice, it is therefore customary to measure the importance of each variable $\boldsymbol{X}_s$ separately, using its marginal posterior inclusion probability, $\text{pr}(\gamma_s = 1 \mid \boldsymbol{y})$. Barbieri and Berger (2004) prove that retaining the variables with marginal posterior inclusion probability greater than 0.5 yields optimal predictions in the orthogonal design case. Because it is also inexpensive computationally, this *median probability model* approach is a popular alternative to the highest posterior probability model approach.

In one-group models, there is no such binary variable to provide direct measures of support, and exact zeros for β_s also are unavailable. Assessing whether the posterior credible interval for β_s covers zero or not may result in poor selection, as reliable parameter uncertainty estimates can be difficult to obtain in high dimensions (Li and Pati, 2017). In practice, selection is done by hard-thresholding, and there is no natural rationale for the choice of threshold. Carvalho et al. (2010) apply heuristics to transfer the median probability model rule to one-group priors: based on the interpretation of shrinkage factors in terms of posterior inclusion probabilities (Lemmas 2.1.1 and 2.1.2), they suggest to threshold $1 - \text{E}\left(\kappa_s \mid \boldsymbol{y}\right)$ at 0.5. However they caution that g_β should be sufficiently heavy-tailed for the relation with posterior probabilities of inclusion to hold. Datta and Ghosh (2013) established that selection under this decision rule has certain optimality properties for the horseshoe prior, and Ghosh et al. (2016) extended their results to a rich class of one-group priors.

Although these arguments make intuitive sense, summaries from marginal posterior probabilities must be interpreted with care when using data with complex dependencies. For instance, Ghosh and Ghattas (2015) gave examples where summaries from marginal posteriors strongly contradict those of joint posteriors in assessing the importance of highly-correlated predictors.

Moreover, variable selection is inherently related to the question of the multiplicity of comparisons, and of a correction for this multiplicity by controlling the number of false discoveries as the number of candidate variables increases. There is much debate as to whether the number of comparisons should be corrected for in the Bayesian paradigm. In their paper "Why We (Usually) Don't Have to Worry About Multiple Comparisons", Gelman et al. (2012) argue that

> "the problem is not multiple testing but rather insufficient modeling of the relationship between the parameters of the model"

suggesting that adjustment should be solely achieved from suitable shrinkage, rather than being imposed post hoc. They explain that shrinkage built in the model through the prior in principle induces an appropriately conservative selection, by pooling and shrinking parameter estimates. The earliest observation along this line dates back to 1939 and is due to Jeffreys, who examined ways of assigning probabilities across various types of model spaces and referred to this as "correcting for selection" (see Sections 1.6, 5.0 and 6.0 of Jeffreys, 1961, 3rd edition).

In order to enforce the right degree of shrinkage, it is crucial to adequately learn parameters controlling the overall sparsity. In two-group priors, the choice of mixing parameter (or prior inclusion probability) π governs the level of sparsity being induced. Its treatment is the focus of the influential paper by Scott and Berger (2010) on Bayesian multiplicity adjustment. Clearly, $\pi = 1/2$ yields a uniform prior on the model space, with each variable having equal probability of being included in the model; no multiplicity control is obtained in this case. The paper discusses inferring π using a fully Bayes approach with

$$\pi \sim \text{Beta}(\alpha, \beta),$$

for certain $\alpha, \beta > 0$, or using an empirical Bayes approach, and demonstrates that both can provide suitable multiplicity adjustment, in the sense that the variable marginal posterior probabilities of inclusion tend to decrease as the number of candidate "noise" variables increases. The paper also provides theoretical comparisons of the full and empirical Bayes approaches via an asymptotic analysis of their respective estimates.

In one-group models, the role of π in exerting multiplicity control is played by the global scale parameter σ_0, and similar investigations as in Scott and Berger (2010) have been undertaken. In the horseshoe model case, Carvalho et al. (2010) advise a fully Bayesian treatment, using a half-Cauchy hyperprior for σ_0. Datta and Ghosh (2013) provide an empirical assessment of the relationship between σ_0 under this prior and the underlying sparsity level, and conclude that the posterior of σ_0 tends to put mass on smaller values in sparser conditions, as expected. Piironen and Vehtari (2016) further propose a procedure to set the scale hyperparameter of the half-Cauchy prior, based on a prior expected number of signals. Plug-in strategies based on naive estimates of σ_0, e.g., from cross-validation, have been implemented but, in *nearly-black* settings (2.6), caution is warranted as σ_0 can collapse to zero, resulting in possible degeneracy in inference (Bhadra et al., 2017b). van der Pas et al. (2014) studied and showed optimality conditions for concentration rates for both the fully Bayes and marginal maximum likelihood approaches for σ_0.

Another reason why Gelman et al. (2012), following others (Greenland and Robins, 1991; Poole, 1991; Krantz, 1999), tend to disregard multiplicity correction, relates to their belief that true effects are unlikely to be exactly zero, in line with the *weak sparsity* perspective. They therefore consider pointless the attempts to control the type I error, which assumes a sharp point null hypothesis of zero effect. This idea is not new; for instance Tukey (1991) claims that

> "All we know about the world teaches us that the effects of A and B are always different – in some decimal place – for any A and B. Thus asking 'Are the effects different?' is foolish".

and before him, Cox et al. (1977) had made a similar point. However, Gelman et al. (2012) seem to acknowledge that the assumption of true zeros is reasonable in some contexts. For instance, they admit:

> "Methods that control for the false discovery rate may make particular sense in fields like genetics where one would expect to see a number of real effects amidst a vast quantity of zero effects".

In high-throughput genetic analyses, the number of tests being performed is indeed usually very large, and practitioners can't afford many false discoveries, as each reported signal may lead to extensive subsequent investigation. Efron et al. (2001) were among the first to highlight this need in the context of

gene expression microarray experiments, and to present a Bayesian reinterpretation of the frequentist false discovery rate of Benjamini and Hochberg (1995); Efron later gathered important ideas on large-scale empirical Bayes inference in a now-seminal book (Efron, 2010).

Let $H_1, \ldots, H_p$ be p null hypotheses considered simultaneously, and let $\hat{z}_1, \ldots, \hat{z}_p$ be corresponding summary statistics. Efron et al. (2001) use the two-group view to propose a mixture model for the test statistics z,

$$z \sim f = (1 - \pi_0) f_1 + \pi_0 f_0, \tag{2.17}$$

with mixing weight π_0, signal density f_1 and noise density f_0. They define the *local false discovery rate* (fdr) as the posterior probability of the null H_0,

$$\text{fdr}(z) = \text{pr}(H_0 \mid z) = \frac{\pi_0 f_0(z)}{f(z)}.$$

A fully Bayesian treatment would require priors for π_0, f_0 and f_1. Instead, they propose nonparametric empirical Bayes estimation of f and f_0, or posit a standard Gaussian density for a *theoretical null* f_0. The null prior probability π_0 should be large to promote sparsity, and may also be estimated under (strong) parametric assumptions; alternatively, Efron et al. (2001) suggest $\pi_0 = 1$ as a suitable, yet conservative, option.

The classical Benjamini–Hochberg false discovery rate is based on tail areas as opposed to densities. It corresponds to the expected proportion of type I errors using the rejection rule $\{Z \leq z\}$ (the rationale is the same for the rule $\{Z \geq z\}$). Efron and Tibshirani (2002) relate the fdr to the tail-area false discovery rate through the *Bayesian false discovery rate* (Fdr) for $\{Z \leq z\}$ which they define as the posterior probability of the null under model (2.17) given $\{Z \leq z\}$,

$$\text{Fdr}(z) = \text{pr}(H_0 \mid Z \leq z) = \frac{\pi_0 F_0(z)}{F(z)},$$

where $F = (1 - \pi_0) F_1 + \pi_0 F_0$, with F_1 and F_0 the cumulative distribution functions of f_1 and f_0, respectively. Efron and Tibshirani (2002) further highlight that the local and tail-area false discovery rates are linked through

$$\text{Fdr}(z) = \text{E}_f \{\text{fdr}(z) \mid Z \leq z\},$$

where $\text{E}_f(\cdot)$ is the expectation with respect to f. Hence, the former is greater than the latter in the general case where $\text{fdr}(z)$ decreases as the magnitude of z increases. Moreover, the local false discovery rate is more specific than the tail-area false discovery rate as it provides a measure of support for a given value z, rather than for a set of values containing the value z; a set which may cover a wide range of fdr values. The *positive false discovery rate* (pFDR) proposed by Storey (2002) also has a Bayesian posterior probability formulation. From the pFDR, Storey (2002) obtains a *q-value*, which he terms as the "Bayesian posterior p-value"; for more on pFDR and q-values, see Storey (2002), Storey (2003) and Storey and Tibshirani (2003).

The above Bayesian approaches to false discovery use summary statistics that may be obtained from frequentist or Bayesian analyses. For instance, they may be t-statistics or marginal posterior means of β_s. Because they already relate to noise-signal mixtures, two-group posterior quantities don't need further modelling, they have de facto connections with hypothesis testing problems: the posterior quantity $\text{pr}(\gamma_s = 0 \mid \boldsymbol{y})$ corresponds to the probability of making a false discovery when selecting variable s. Following this idea, Newton et al. (2004) define another Bayesian false discovery rate as the

fraction of false discoveries relative to the total number of discoveries under a threshold rule τ:

$$\mathrm{FDR}(\tau) = \frac{\sum_{s=1}^{p} (1 - \mathrm{PPI}_s)\, \mathbb{1}(\mathrm{PPI}_s > \tau)}{\sum_{s=1}^{p} \mathbb{1}(\mathrm{PPI}_s > \tau)}, \qquad 0 < \tau < 1, \tag{2.18}$$

writing $\mathrm{PPI}_s = \mathrm{pr}(\gamma_s = 1 \mid \boldsymbol{y})$.

More recently, Stephens (2016) proposed an empirical Bayes false discovery rate approach, called "ASH" for "adaptive shrinkage", based on the one-group perspective. ASH uses both the estimated effect sizes, $\hat{\beta}_s$, and their standard errors, $\hat{s}_s$, assuming a unimodal distribution of the true effects. Specifically, ASH links the observations $(\hat{\beta}_s, \hat{s}_s)$ with the effects β_s using the normal means model,

$$\hat{\beta}_s \mid \beta_s, \hat{s}_s \sim \mathcal{N}(\beta_s, \hat{s}_s^2), \qquad \beta_s \overset{\text{iid}}{\sim} G, \qquad s = 1, \ldots, p,$$

where G belongs to the space $\mathcal{G}$ of unimodal distributions with mode at zero. Then, it maximizes the marginal likelihood,

$$\hat{G} = \arg\max_{G \in \mathcal{G}} \mathrm{pr}(\hat{\boldsymbol{\beta}} \mid G, \hat{\boldsymbol{s}}) = \arg\max_{G \in \mathcal{G}} \prod_{s=1}^{p} \int \mathcal{N}(\beta_s, \hat{s}_s^2)\, \mathrm{d}G(\beta_s), \tag{2.19}$$

and computes summaries from the posterior distributions $\mathrm{pr}(\beta_s \mid \hat{G}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{s}})$, $s = 1, \ldots, p$; the task (2.19) is cast as a convex optimization problem by approximating G by a finite mixture of uniform distributions. The posterior summaries formed after estimating G are Efron's local false discovery rates and novel *local false sign rates* (lfsr) which report measures of support for the signs of effects, see Stephens (2016) for details. A strength of Stephens's procedure is that it allows incorporating variable precision measures $\hat{s}_s$, unlike conventional approaches to FDR.

In practice, all the above approaches may suffer from correlation among tests: correlated summary statistics can produce strong deviations from theoretical null distributions, often resulting in failure to control false discovery rates, with too many rejections of the null (Efron, 2007; Leek and Storey, 2007). The issue reaches beyond these examples: in frequentist settings, only a handful of procedures allow for some form of correlation among tests, the best-known of which was proposed by Benjamini and Yekutieli (2001), and even then, the assumptions underlying these procedures are often violated in real scenarios. Practitioners tend to circumvent the problem by resorting to permutation analyses, where the permuted data maintain the original correlation structure, but the computational cost of this is prohibitive in many large-scale applications. Deriving more formal procedures with relaxed assumptions regarding correlation is an open research area. To obtain more robust estimates in our applications to real data, we will use a permutation-based variant of the FDR estimate (2.18).

2.3 Scaling high-dimensional Bayesian inference

2.3.1 Preliminaries

The Bayesian paradigm essentially involves integration tasks. Except in very specific cases, however, integrals are not amenable in closed form and posterior inference requires approximation procedures. Monte Carlo techniques, and in particular, Markov Chain Monte Carlo (MCMC) algorithms, have been the workhorse for this task since the 1990s (Gelfand and Smith, 1990). For reasonably small datasets, MCMC methods can yield accurate inference in a timely manner. For large-scale problems however,

obtaining samples from the target distribution involves substantial, if not prohibitive, computational expenses.

Whether via MCMC approximations or via other approaches, scaling Bayesian inference is a challenge. This has triggered considerable work in the statistical and machine learning communities to devise new strategies aiming for accuracy, robustness, tractability, and asymptotic guarantees. The nature of the difficulties faced in this endeavour differ depending on the specific "large-scale" regime under consideration: the *large n, small p* case is the most favourable; with more samples, the posterior should concentrate to a point mass (following the Bernstein–von Mises theorem), so point estimation may seamlessly replace the full Bayesian machinery. This is no longer warranted in settings with more variables, e.g., p growing with the number of samples n.

Most of the literature on scaling Bayesian inference is devoted to designing effective solutions for such *moderate-to-large p, large n* cases. There, procedures requiring evaluations of all data points for each sampling step or parameter update can be excessively expensive. Common remedies implement data subsampling (e.g., Quiroz et al., 2018; Bardenet et al., 2014; Hoffman et al., 2013) or data partitioning (e.g., Mcdonald et al., 2009; Wang and Dunson, 2013; Wang et al., 2015; Scott et al., 2016). The former evaluate likelihoods of a random subset of samples at each iteration. The latter divide the samples and fit the original model on all variables to each subset in order to restrict to cheaper-to-manipulate batch quantities; they are also called *divide-and-conquer* approaches, as they can leverage parallel and distributed computing resources, leading to dramatic computational savings in some instances. An extensive review on subsampling and partitioning approaches for MCMC inference in the *large n* regime is that of Bardenet et al. (2017).

Although prominent in applications, the case that interests us, whereby p is large and n is small, is understudied. It is also arguably more difficult to tackle: dividing the sample space may no longer be beneficial or recommended, and naively transposing this to the variable space is not without risks, since important dependencies may be omitted by treating groups of variables separately. The next two sections clarify such tensions between modelling assumptions and inference tractability in this high-dimensional regime, and discuss the two dominant strategies to achieve scalability in the Bayesian paradigm. The first approach sees the scalability requirement as a prerequisite which should precede the inference procedure and shape the modelling strategy itself. The second approach strives to leave the modelling approach free of any practical consideration, and pursue scalability solely by adapting the inference procedure. In some cases, the boundaries between the two strategies may be fuzzy, as it may be desirable to design methods combining bits of both approaches, yet we here focus on representative examples of each kind in order to convey general themes of research.

2.3.2 Rethinking the modelling approach

Many recent approaches to variable selection seek to develop methodologies that lend themselves to efficient inference in the *large p, small n* paradigm. These approaches not only acknowledge the need for sparsity assumptions, but also anticipate, at the modelling level, the computational burden attached to the exploration of large parameter spaces. They often take the form of modelling procedures with two or more consecutive stages, aiming to partition, prescreen or reduce the data to a lower-dimensional subspace.

The variable partitioning idea mentioned in Section 2.3.1 has been adopted in several independent works. For instance, Song and Liang (2015) proposed a *split-and-merge* (SAM) method which divides

the data into lower dimensional subsets, screens out variables uncorrelated with the response by performing Bayesian variable selection within each subset, and then refines the selection by modelling the response and the surviving variables from the aggregated dataset. Their authors prove selection consistency under conditions that they describe as mild. However, for finite sample sizes, the size of the subsets can affect the effectiveness of the prescreening: the smaller the subsets, the greater the chance to select false discoveries because of spurious correlation with the response through a variable from another subset. To address such issues, Wang et al. (2016) developed a procedure called DECO, which proceeds with a "decorrelation" step before the partitioning stage, to ensure that variables handled in distinct subsets are uncorrelated. DECO differs from the SAM method in two other respects: first, it doesn't involve any merging stage to refine the set of selected variables, and second, it is a frequentist procedure that performs penalized regression on each subset, though its rationale is paradigm-free.

Taken to its extreme, the partitioning strategy reduces to a purely marginal screening (such as illustrated in the motivation of Section 1.1), where the importance of each variable is assessed separately. By giving up on modelling complex dependencies in the data, this screening strategy permits embarrassingly parallel computing schemes that often lead to massive speedups. For this reason, it is widespread: most published genome-wide association studies result from screening hundreds of thousands of genetic variants, one by one. The performance of marginal screening has been the object of theoretical investigations; notably Fan and Lv (2008) wrote an influential paper on their frequentist *sure independence screening* (SIS) strategy. The authors prove that SIS has the *sure screening property*, which they define as the property of retaining all the relevant variables with probability tending to one after screening. These guarantees holds for situations with "fairly uncorrelated" variables, to quote P. Bühlmann in his discussion of the paper. In its contribution to the discussion, S. Richardson points out that this is

> "a favourable case for asymptotics, but an unlikely situation in most applications, in particular in genetics and genomics".

E. Levina and J. Zhu further report a degraded performance for low signal-to-noise ratios when using equicorrelated predictors. Aware of this weakness, Fan and Lv (2008) propose an *iterative sure independence screening* (ISIS) algorithm, which iteratively re-screens based on residuals from previous screenings. They demonstrate that their procedure alleviate issues attributable to correlations in practice, but provide no theoretical foundation for it.

Many other techniques may be used to map high-dimensional data to lower-dimensional objects in a tractable manner. Among them, factor models are appealing because they confine inference to a relatively small number of parameters compared to classical sparse regression, see, e.g., Bhattacharya and Dunson (2011) and Murray et al. (2013). However variable selection is not the original goal of factor modelling, and this task appears somewhat add-hoc, as it requires additional interpretation of the extracted factors.

In summary, addressing the practical concerns associated with high dimensionality at the modelling stage is a sensible approach on scaling Bayesian methods, provided that the modelling assumptions used are carefully reviewed and lend themselves to variable selection. The resulting methodology should also be compared to more canonical methods, on both theoretical and empirical grounds, even if this exercise implies considering smaller problems. Conversely, evaluating the scalability of a given modelling strategy is always necessary. For instance, in a regression setting, modelling the correlation of tens of thousands of response variables without further structural assumption is unrealistic, from

both a memory and CPU-time standpoints, and this, regardless of the inference approach employed; we will discuss this further in Chapter 3.

2.3.3 Scaling inference algorithms

The strategy summarized in the previous section focuses on attaining scalability from a methodological perspective, thereby acknowledging that some modelling approaches are more amenable to fast computation than others. The present section describes a different perspective to scaling Bayesian methods, which is independent of the chosen modelling approach and therefore avoids possible tradeoffs between enforcing structural constraints for scalability reasons and modelling data complexity purely for statistical reasons. This perspective is concerned with designing inference algorithms that can efficiently search through high-dimensional parameter spaces.

There are two main paths to approximate Bayesian inference. The first relies on Monte Carlo sampling methods, developed by Stanislas Ulam, Nicholas Metropolis and John von Neumann (Metropolis and Ulam, 1949; Von Neumann and Ulam, 1951). Monte Carlo sampling methods frame integrations as expectations to be estimated by the sample mean of (a function of) simulated random variables. Because direct simulation from the desired ("target") distribution is often not possible, a special class of Monte Carlo algorithms was developed: Markov Chain Monte Carlo (MCMC) algorithms devise Markov chains whose stationary distributions, i.e., reached after convergence, correspond to the target distribution (Robert and Casella, 2013). A natural characterization of Markov chains is through their transition kernels, which specify conditional distributions defining moves to the next states.

MCMC inference is difficult in high dimensions. First, likelihoods and sometimes gradients need to be evaluated at each iteration, and the computational cost associated to these evaluations increases with the number of parameters. Second, the mixing tends to deteriorate in high dimensions, so extensive explorations of the posterior space require a large number of iterations. Moreover, assessing convergence is hard in practice: indeed, all the parameters need to have converged, including the possible nuisance parameters (Gill, 2008), and storing and checking diagnostics for millions of parameters in some applications is essentially impossible.

Approaches to accelerate MCMC inference again depend on the large-scale setting under consideration, and the *large n* data case entails more active research compared to the high-dimensional, *large p small n*, case. A common strategy for this latter regime involves replacing the Markov transition kernel with an approximation that is cheaper to sample from. One example is to substitute point estimates in some sampling steps (Guhaniyogi et al., 2018), and another example is to approximate full conditional distributions by simpler distributions (Bhattacharya and Dunson, 2010; O'Brien and Dunson, 2004). While such strategies have long been ad-hoc, Johndrow et al. (2015) recently provided bounds on approximation errors that can be tolerated to achieve the best statistical performance under a given loss function and computational budget.

While approximating transition kernels aims to reduce the cost per iteration, it is also sensible to try reducing the number of iterations through improved exploration of the parameter space. For instance, simulated tempering techniques (Marinari and Parisi, 1992; Geyer and Thompson, 1995) aim to better handle posterior multimodality, which often exacerbates in high dimensions. They embed the target distribution $\mathrm{pr}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^p$, into an augmented space by introducing auxiliary distributions indexed by a so-called "temperature" parameter $T \geq 1$, namely, $\mathrm{pr}_T(\boldsymbol{\theta}) \propto \mathrm{pr}(\boldsymbol{\theta})^{1/T}$. Sufficiently large temperatures

T flatten out the density allowing the chain to search through wider regions in the parameter space without being trapped in local modes.

Another natural approach to more efficient exploration is to perform "automatic" adaptations as the algorithm proceeds, to find better parameter values. This is often attempted in Metropolis–Hastings algorithms, a canonical class of MCMC algorithms, which draw samples from a *proposal distribution*, and accept them with a certain probability (Metropolis et al., 1953; Hastings, 1970). Adaptive Metropolis–Hastings algorithms carefully tune the proposal distribution during execution, using information from previous samples (Liang et al., 2011, Chap. 8). A variety of adaptive MCMC methods have been designed for different contexts (Haario et al., 2001; Rosenthal and Roberts, 2007; Andrieu and Moulines, 2006), including genetics (Turro et al., 2007); in particular, the evolutionary stochastic search algorithm proposed by Bottolo and Richardson (2010) combines both tempering ideas and adaptive moves.

The second way to perform Bayesian inference is via deterministic methods, which turn inference onto an optimization problem. Vanilla deterministic inference doesn't rely on sampling. It doesn't necessarily target the exact posterior distribution either, but seeks for a good surrogate that may be obtained in a more tractable fashion. Variational inference is an instance of a deterministic inference approach. In order to allow cheaper computations, it uses to a class of approximating distributions and seeks the distribution of this class that is the "closest" (in Kullback–Leibler divergence) to the target distribution. Because it is the object of much of the present work, we defer the full description of variational inference to Section 2.4, and we only provide here general considerations.

Much discussion opposing Monte Carlo and deterministic inference concerns tradeoffs between asymptotic unbiasedness and computational efficiency in large-scale applications. MCMC inference is often perceived as the gold standard because it is asymptotically unbiased; the approximation becomes arbitrarily good as iterative sampling proceeds. But if stationarity can't be attained within an acceptable timeframe, then one should consider deterministic alternatives; yet the lack of theoretical understanding of deterministic procedures is an objection that often arises. Scaling MCMC inference and obtaining guarantees for deterministic inference are two endeavours that would permit exploiting the full potential of large-scale data.

As noted by Angelino et al. (2016), an important step in this direction could be to acknowledge the need for balancing bias and variance. Indeed, variational classes of approximating distributions rarely cover the target distribution, so approximations may suffer from large bias when obtained from an overly restrictive variational class. As we will see in Section 2.4, there is a substantial amount of work on using more expressive families of distributions to lower this bias. There are also attempts to quantify and control the bias both in asymptotic and finite regimes, even if these are still mostly limited to specific contexts. Unlike variational approximations, MCMC approximations are asymptotically unbiased. However, Angelino et al. (2016) observe that

> "Insisting on zero asymptotic bias from Monte Carlo estimates of expectations may leave us swamped in errors from high variance or transient bias".

Indeed, in practice, MCMC estimates are obtained using a finite number of iterations, after which the error can still be significant; this error can be decomposed in a transient bias, linked to a residual dependence on the burn-in, and the Monte Carlo standard error, related to the collected samples being too few or too correlated. Welling and Teh (2011), Korattikara et al. (2014) and Angelino et al. (2016), among others, therefore point out that exact inference may have inferior statistical properties under a

finite computational budget. This may be put in parallel with the discussion of Section 2.1.3 on the improved risk properties of the biased James–Stein estimator over unbiased estimators. Although there seems to be a long way to go towards full acceptance and leveraging of this rationale, recent work mentioned above has already targeted relaxing asymptotic exactness in order to reduce the transient bias or the Monte Carlo variance for given computational resources: approximate transition kernels (e.g., Korattikara et al., 2014; Johndrow et al., 2015) speed up execution of the chain at the cost of introducing some asymptotic bias.

Finally, there have been attempts to pursue both scalability and asymptotic unbiasedness by marrying MCMC and variational inference. The best-known example is probably that of de Freitas et al. (2001), who use the variational approximation in the proposal distribution of a Metropolis–Hastings algorithm. Salimans et al. (2015) exploit the latest developments of variational inference to approximate an MCMC chain. It is, however, not clear if these attempts have yielded the expected practical payoff. Designing hybrid algorithms with sound theoretical and empirical properties in a principled manner is an interesting avenue for future research.

2.4 Variational inference

2.4.1 Origins and main ideas

The roots of variational inference date back to the late 1980s. They emerged from statistical physics, and in particular statistical mechanics for spin glasses with the work of Mézard et al. (1987) applying variational principles for fitting Ising-type models. In the same period, Anderson and Peterson (1987) published a now-seminal paper, where they employed variational methods to study neural networks. These methods were then adopted by the computer science community and started to be an appealing alternative to sampling-based methods in the 1990s: Jaakkola et al. (1996) Saul et al. (1996), Jordan et al. (1999) and Opper and Saad (2001) generalized their applicability to many probabilistic models while, in parallel and probably independently, Hinton and van Camp (1993) contributed to further developing the original ideas in the context of neural networks. With the availability of large datasets, recent years have seen renewed interest in variational inference, which has now been extended in various fashions and applied in many different contexts. We next present the main ideas of variational inference, and draw connections with other types of inference procedures, deterministic or sampling-based. The material of this chapter is mostly based on Blei et al. (2017) and Zhang et al. (2017).

Variational inference approximates the posterior density, $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, for a parameter vector $\boldsymbol{\theta}$ and data $\boldsymbol{y}$, with a simpler density, $q(\boldsymbol{\theta})$, obtained as the solution of an optimization problem. This problem corresponds to minimizing of a measure of "closeness" to $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, namely the Kullback–Leibler (KL) divergence

$$\mathrm{KL}\big(q \,\|\, p\big) = \int q\,(\boldsymbol{\theta}) \log\left\{\frac{q\,(\boldsymbol{\theta})}{p\,(\boldsymbol{\theta} \mid \boldsymbol{y})}\right\} \mathrm{d}\boldsymbol{\theta}. \tag{2.20}$$

The reasoning behind variational inference is thus very natural, yet two aspects need to be clarified: the meaning of a "simple" distribution $q(\boldsymbol{\theta})$ and the rationale for the choice of divergence.

2.4.2 The Kullback–Leibler divergence and other α-divergences

The KL divergence is probably the most prominent divergence measure used in statistics, machine learning and information theory. It was introduced by Kullback and Leibler (1951) who described it as a "directed divergence", referring to its asymmetry, i.e., $\mathrm{KL}(q\|p) \neq \mathrm{KL}(p\|q)$. The divergence from p to q, $\mathrm{KL}(p\|q)$, is sometimes called *forward* KL divergence and that from q to p, $\mathrm{KL}(q\|p)$, is called *reverse* KL divergence. These divergences are special cases of *α-divergences*, indexed by $\alpha \in \mathbb{R}$, which have several formulations, successively introduced by Rényi (1961), Amari (1985) and Tsallis (1988), among others. Rényi's α-divergence is arguably the most studied, while Amari α-divergence has a long history in differential geometry (Cichocki and Amari, 2010). The former is defined as

$$\mathrm{D}_\alpha^R(p\|q) = \frac{1}{\alpha-1}\log\int p(\boldsymbol{\theta}\mid\boldsymbol{y})^\alpha\, q(\boldsymbol{\theta})^{1-\alpha}\,\mathrm{d}\boldsymbol{\theta}, \tag{2.21}$$

for $\alpha \in \mathbb{R}_+ \setminus \{1\}$ such that $\mathrm{D}_\alpha^R(p\|q) < +\infty$, and the latter, as

$$\mathrm{D}_\alpha^A(p\|q) = \frac{4}{1-\alpha^2}\left(1 - \int p(\boldsymbol{\theta}\mid\boldsymbol{y})^{\frac{1+\alpha}{2}}\, q(\boldsymbol{\theta})^{\frac{1-\alpha}{2}}\,\mathrm{d}\boldsymbol{\theta}\right), \tag{2.22}$$

for $\alpha \in \mathbb{R} \setminus \{\pm 1\}$ such that $\mathrm{D}_\alpha^A(p\|q) < +\infty$. The forward KL divergence is obtained by extending (2.21) and (2.22) by continuity to $\alpha = 1$, namely,

$$\mathrm{D}_1^R(p\|q) = \lim_{\alpha\to 1}\mathrm{D}_\alpha^R(p\|q) = \mathrm{KL}(p\|q), \qquad \mathrm{D}_1^A(p\|q) = \lim_{\alpha\to 1}\mathrm{D}_\alpha^A(p\|q) = \mathrm{KL}(p\|q),$$

and, similarly, the reverse KL divergence is obtained by extending (2.22) to $\alpha = -1$,

$$\mathrm{D}_{-1}^A(p\|q) = \lim_{\alpha\to -1}\mathrm{D}_\alpha^A(p\|q) = \mathrm{KL}(q\|p).$$

When used as an optimization criterion, the choice of α-divergence results in approximations with different behaviours. For instance, examining the integrand of (2.20) reveals that the reverse KL divergence tends to prevent q from putting mass in regions where p has little mass, and penalizes less q placing little mass where p is has positive mass. Hence, inference under $\mathrm{KL}(q\|p)$ tends to produce underdispersed distributions. Conversely, the forward KL divergence prevents q from having low probability mass in areas where p has positive mass, and penalizes less q putting large mass where p has negligible mass; approximations under $\mathrm{KL}(p\|q)$ tend to overestimate the support of p. Figure 2.2 illustrates such behaviours under Amari α-divergence minimization. It describes the approximation of a mixture of normal p with a density q restricted to be Gaussian. The forward and reverse KL divergences correspond to $\alpha = 1$ and $\alpha = -1$, respectively. More generally, the approximation tends to cover the entire distribution p when α takes large positive values, and it tends to concentrate on the mode with largest probability mass when α takes large negative values. Neither extreme seems ideal. If the target distribution p has many modes, then focusing on one of its modes misrepresents the complexity of p, while covering all of them in a mode averaging fashion may assign high probability in regions where p has negligible mass.

But the decision of which α to pick is mainly driven by more practical considerations: evaluating the α-divergence, whether in its form (2.21) or (2.22), would require computing the marginal likelihood $p(\boldsymbol{y})$, whose intractability motivated the use of inference algorithms in the first place. Hence, optimization can't be carried out by directly using divergences as objective functions. Variational inference bypasses

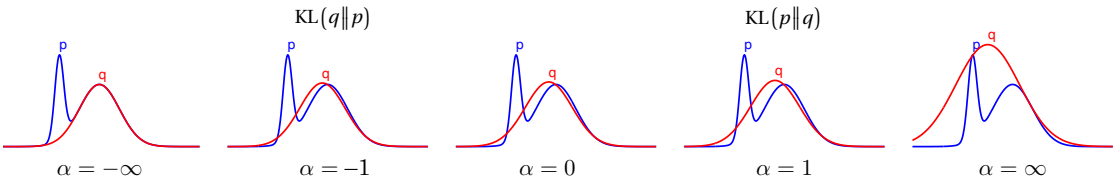

Figure 2.2 – (Unnormalized) Gaussian approximation q to a mixture of Gaussian densities p under Amari α-divergences for various α. Assuming finite divergences, for $\alpha \to -\infty$, q focuses on one mode, while for $\alpha \to \infty$, it entirely covers p. This figure is adapted from Minka (2005).

this issue by exploiting the following decomposition of the reverse KL divergence:

$$\mathrm{KL}\left(q \middle\| p\right) = \log p(\boldsymbol{y}) - \mathcal{L}(q), \tag{2.23}$$

where

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \log\left\{\frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\} \mathrm{d}\boldsymbol{\theta}. \tag{2.24}$$

To paraphrase (2.23), the KL divergence is the difference between the marginal log likelihood, which has no effect on the optimization procedure, and a functional, $\mathcal{L}(q)$. Hence, minimizing the KL divergence amounts to maximizing $\mathcal{L}(q)$, which does not involve the marginal likelihood. The former problem is intractable, but the latter has a closed form under suitably-chosen variational distribution families and models. In such cases, $\mathcal{L}(q)$ can be used as a surrogate objective function.

A wide class of models enabling analytical $\mathcal{L}(q)$ is formed by conditionally-conjugate models, which Gelman et al. (2013, Sect. 2.4) define as follows. Let $\mathcal{F}$ be a class of sampling distributions $p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{\lambda})$, and $\mathcal{P}$, a class of prior distributions for $\boldsymbol{\theta}$ conditional on $\boldsymbol{\lambda}$, then the class $\mathcal{P}$ is *conditionally conjugate for* $\mathcal{F}$ if $p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{\lambda}) \in \mathcal{P}$ for all $p(\cdot \mid \boldsymbol{\theta}, \boldsymbol{\lambda}) \in \mathcal{F}$ and $p(\cdot \mid \boldsymbol{\lambda}) \in \mathcal{P}$.

Because $\mathrm{KL}\left(q \middle\| p\right) \geq 0$, $\mathcal{L}(q)$ is also a lower bound on the marginal likelihood; it is therefore often referred to as a *variational lower bound* or *evidence lower bound* (ELBO). Equality (2.23) results from a simple rearrangement of terms, but the lower bound relation can also be obtained without explicitly invoking the KL divergence,

$$\begin{aligned}
\log p(\boldsymbol{y}) &= \log \int p(\boldsymbol{y},\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&= \log \int \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} q(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \\
&\geq \int q(\boldsymbol{\theta}) \log\left\{\frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\} \mathrm{d}\boldsymbol{\theta} \\
&= \mathcal{L}(q),
\end{aligned}$$

by Jensen's inequality.

The variational lower bound also provides useful insights on the type of inference obtained under the reverse KL divergence. We can rewrite it as

$$\mathcal{L}(q) = \mathrm{E}_q \log p(\boldsymbol{y},\boldsymbol{\theta}) - \mathrm{E}_q \log q(\boldsymbol{\theta}) \tag{2.25}$$

where $\mathrm{E}_q(\cdot)$ is the expectation with respect to q, that is, $\mathcal{L}(q)$ is the sum of two terms, the expected log joint distribution of the observations and the parameter vector, and an entropy term. The entropy term acts as a regularization; without it, the optimization would correspond to a MAP estimation.

The technical advantage of reverse KL divergence's decomposition (2.23) can be reproduced for the more general (reverse) Rényi α-divergences. We have,

$$\mathrm{D}_\alpha^R\left(q\middle\|p\right) = \log p(\boldsymbol{y}) - \mathcal{L}_\alpha(q)\,,$$

where

$$\mathcal{L}_\alpha(q) = \frac{1}{\alpha-1}\log\int q(\boldsymbol{\theta})\left\{\frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\}^{1-\alpha}\mathrm{d}\boldsymbol{\theta}, \qquad \alpha\in\mathbb{R}_+\setminus\{1\}, \tag{2.26}$$

so $\mathcal{L}_\alpha(q)$ is also a lower bound on $p(\boldsymbol{y})$ whose evaluation doesn't require computing $p(\boldsymbol{y})$. The Rényi α-divergences are the only α-divergences whose formulation leads to such a bound; the bound is continuous and non-increasing on $\left\{\alpha\geq 0 : |\mathcal{L}_\alpha| < +\infty\right\}$, in particular, for all $\alpha\in(0,1)$,

$$\mathcal{L}(q) = \lim_{\alpha\to 1}\mathcal{L}_\alpha(q) \leq \mathcal{L}_\alpha(q) \leq \mathcal{L}_0(q),$$

where the left hand-side is the variational lower bound, obtained by reverse KL divergence, and the right hand-side equals $\log p(\boldsymbol{y})$ if and only if the support of $p(\boldsymbol{\theta}\mid\boldsymbol{y})$ is included in that of $q(\boldsymbol{\theta})$ (Li and Turner, 2016). Hence, optimization under Rényi divergence with $\alpha<1$ may yield tighter bounds on the marginal log-likelihood compared to variational inference. This improved accuracy comes at a price, however: no closed-form expression can usually be obtained for (2.26), even by using restricted families. We shall return to this question in Section 2.4.4.

We saw how variational inference minimizes the reverse KL divergence via an auxiliary decomposition, and how this can be extended to reverse Rényi divergences, albeit losing some tractability. Other approaches to minimizing divergences exist. Expectation propagation (Minka, 2001) is one of them; in a nutshell, expectation propagation optimizes the forward KL divergence by leveraging the factorization structure of the posterior. The algorithm also imposes restricted families of candidate distributions and can result in exact inferences in simple cases. From a practical perspective, expectation propagation tends to converge in fewer iterations than variational inference, but often has a higher cost per iteration. An important drawback is that convergence is not guaranteed. Extensions of expectation propagation to α-divergences for general α exist and are often termed *power expectation propagation*. More about expectation propagation can be found in Bishop (2006, Section 10.7).

2.4.3 Gaussian and mean-field variational families

Returning to variational inference, we mentioned that the expectation (2.24) can be evaluated analytically by forcing the approximation to belong to a restricted family of distributions. If the family contains the target distribution p, the variational approximation will be exact, that is, $q(\cdot)=p(\cdot\mid\boldsymbol{y})$ and $\mathrm{KL}\left(q\middle\|p\right)=0$. This rarely happens; for instance, in the example displayed in Figure 2.2, the family is restricted to Gaussian distributions and can't capture the bimodal target posterior. Such Gaussian approximations form an important branch of variational algorithms. While they are poor proxies for the posterior when the latter is multimodal or complicated, they are particularly interesting in the *large n* setting, where the true posterior is expected to behave like a Gaussian.

Algorithm 1: Mean-field variational inference

Input: Data $\boldsymbol{y}$, model $p(\boldsymbol{y}, \boldsymbol{\theta})$, tolerance tol
initialize: For $s = 1, \ldots, p$, variational parameter α_s indexing factor $q(\theta_s) = q(\theta_s; \alpha_s)$ in (2.27),
$\mathcal{L}(q) \leftarrow -\infty$
repeat
  for $s = 1, \ldots, p$ **do**
    Set the variational parameters α_s according to (2.28)
  end
  $\mathcal{L}^{\text{old}}(q) \leftarrow \mathcal{L}(q)$
  $\mathcal{L}(q) \leftarrow \mathrm{E}_q \log p(\boldsymbol{y}, \boldsymbol{\theta}) - \mathrm{E}_q \log q(\boldsymbol{\theta})$
until $\left|\mathcal{L}(q) - \mathcal{L}^{\text{old}}(q)\right| < \text{tol}$;
return Variational approximation $q(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \prod_{s=1}^{p} q(\theta_s; \alpha_s)$

The strengths and limitations of Gaussian variational approximation can be confronted to those of Laplace approximation (Laplace, 1986). Laplace approximation uses the maximum of the posterior and the inverse of its Hessian (provided that the log posterior is twice-differentiable) as mean and covariance for a Gaussian posterior approximation. It is purely local, as it depends only on the curvature of the posterior in the vicinity of the optimum, whereas the Gaussian variational approximation, by optimizing the KL divergence, typically captures the overall posterior shape more accurately. A second drawback of the Laplace approximation is that it requires the inversion of a potentially large Hessian, which makes it intractable in high dimensions. Finally, the Laplace method is restricted to Gaussian approximations; variational inference often posits families of distributions without this parametric restriction, as we now explain.

The mean-field formulation gives a natural approach to constructing variational families of candidate distributions. The approach appeared in the context of the mean-field theory of physics (Mézard et al., 1987) and was a central tool in the first applications of variational inference in the 1980s. Mean-field variational inference assumes that the approximation factorizes over the components of the parameter vector,

$$q(\boldsymbol{\theta}) = \prod_{s=1}^{p} q(\theta_s), \tag{2.27}$$

without imposing any constraint on the functional forms of each $q(\theta_s)$. This assumption allows simple parameter updates based on Lemma 2.4.1.

Lemma 2.4.1 (Blei et al., 2017)**.** *Let $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ be the target posterior distribution for observations $\boldsymbol{y}$ and parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, and consider the mean-field formulation (2.27). For $s \in \{1, \ldots, p\}$, and fixing all variational factors $q(\theta_{s'})$, $s' \neq s$, the optimal variational factor $q(\theta_s)$ is*

$$\log q(\theta_s) = \mathrm{E}_{-s}\{\log p(\boldsymbol{\theta}, \boldsymbol{y})\} + \text{cst}, \tag{2.28}$$

where cst *is constant with respect to θ_s.*

Proof. The variational lower bound can be written

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_{s'=1}^{p} q(\theta_{s'}) \left\{ \log p(\boldsymbol{\theta}, \boldsymbol{y}) - \sum_{s'=1}^{p} \log q(\theta_{s'}) \right\} \mathrm{d}\theta_1 \cdots \mathrm{d}\theta_p \\ &= \int q(\theta_s) \left\{ \int \log p(\boldsymbol{\theta}, \boldsymbol{y}) \prod_{s' \neq s} q(\theta_{s'}) \mathrm{d}\theta_{s'} - \log q(\theta_s) \right\} \mathrm{d}\theta_s + \text{cst} \\ &= \int q(\theta_s) \log \left\{ \frac{p_{-s}(\theta_s; \boldsymbol{y})}{q(\theta_s)} \right\} \mathrm{d}\theta_s + \text{cst}, \end{aligned} \tag{2.29}$$

where cst is constant with respect to θ_s and where we introduced the distribution

$$p_{-s}(\theta_s; \boldsymbol{y}) = \text{cst} \times \exp\left[\mathrm{E}_{-s} \left\{ \log p(\boldsymbol{\theta}, \boldsymbol{y}) \right\} \right],$$

with $\mathrm{E}_{-s}(\cdot)$ denoting the expectation with respect to the distributions $q(\theta_{s'})$ over all variables $\theta_{s'}$, $s' \neq s$. The right-hand side of (2.29) corresponds to the negative KL divergence between $q(\theta_s)$ and $p_{-s}(\theta_s; \boldsymbol{y})$, plus a constant. So, fixing the factors $q(\theta_{s'})$, $s' \neq s$, the distribution $q(\theta_s)$ which maximizes $\mathcal{L}(q)$ is $q(\theta_s) = p_{-s}(\theta_s; \boldsymbol{y})$; the result follows. □

If the posterior is in the conditionally-conjugate exponential family, then expressions (2.28) can be obtained in closed form using to the fully factorized form of q (2.27). The relations (2.28) give rise to cyclic dependencies among the distributions $q(\theta_s)$ and form a coordinate ascent algorithm. Convergence is ensured by the concavity of $\mathcal{L}(q)$ in each of the variational parameter α_s indexing the factors $q(\theta_s) = q(\theta_s; \alpha_s)$ (Boyd and Vandenberghe, 2004, Sections 3.1.5, 3.2.4, 3.2.5). Moreover, $\mathcal{L}(q)$ is guaranteed to increase monotonically at every iteration, which provides a useful check against mistakes in the computations or the implementation. A sketch of the procedure is outlined in Algorithm 1 and an example of computations of variational updates and objective function is given in Appendix A.2.

The coordinate updates of the mean-field algorithm have close connections with Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990), which draws successively from each variable's distribution, conditional on the current realisations of the other variables. Equation (2.28) relates to the same conditional distribution since $p(\boldsymbol{\theta}, \boldsymbol{y}) \propto p(\theta_s \mid \boldsymbol{\theta}_{-s}, \boldsymbol{y})$, where the proportionality constant doesn't depend on θ_s. Similar ideas underlie the expectation-maximization (EM) algorithm (Dempster et al., 1977): the EM algorithm alternates between taking expectations (E-step) of the log joint distribution of the observations and the parameters, $\log p(\boldsymbol{\theta} \mid \boldsymbol{y})$, and maximizing this expectation (M-step). In fact, the parallel with variational inference is not limited to the alternating form of the mean-field algorithm (2.28). Indeed, the expected log joint distribution corresponds to the first term of the variational objective function (2.25) with the expectation taken with respect to $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, and the EM algorithm relies on the fact that it equals the log marginal likelihood when $q(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \boldsymbol{y})$. Unlike in variational inference, the EM algorithm estimates fixed model parameters $\boldsymbol{\theta}$ and takes the expectation of $\log p(\boldsymbol{y}, \boldsymbol{\theta})$ with respect to $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ and not $q(\boldsymbol{\theta})$, assuming it can be computed.

2.4.4 Recent trends

A quantity of new work in variational inference addresses novel directions such as enhancing accuracy, scalability and making it applicable to a wider range of complicated models. This work is often primarily designed for *large n* settings; in this section, we outline these developments and briefly discuss their relevance in the *large p, small n* setting.

We saw how fully factorized approximations (2.27) improve tractability and make the mathematics simpler. They introduce strong independence assumptions however, leading to a posterior which is less expressive than when maintaining the dependencies, and to underestimated posterior variances. To improve the accuracy of mean-field approximations, it is worth trying to restore some structure by grouping factors in such a way that the coordinate updates are still obtained analytically. Richer variational families along these lines have been studied, e.g., by Saul and Jordan (1996) and Barber and Wiegerinck (1999), and the resulting inference was referred to as *structured variational inference*. More recent advances on this can be found in Tran et al. (2015), Guo et al. (2016) and Ranganath et al. (2016); many of these approaches give up closed-form formulations and introduce model-specific or generic approximations, which may be costly. In general, choices about which dependencies to retain should be made in a customized fashion for the model considered, as some dependencies may impact inference more crucially than others.

The workhorse of variational algorithms in the *large n* regime is *stochastic variational inference* (Hoffman et al., 2013), designed to further scale computation to the massive datasets routinely encountered in machine learning. It is applied on models with local parameters λ_i ($i = 1,\dots,n$), and a global parameter $\boldsymbol{\theta}$, and typically assumes a mean-field formulation

$$q(\boldsymbol{\lambda},\boldsymbol{\theta}) = q(\boldsymbol{\theta}) \prod_{i=1}^{n} q(\lambda_i)\,,$$

yielding a variational lower bound that involves a sum over contributions from the n data points, e.g.,

$$\mathcal{L}(q) = \mathrm{E}_q\left\{\log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})\right\} + \sum_{i=1}^{n} \mathrm{E}_q\left\{\log p(y_i \mid \lambda_i,\boldsymbol{\theta}) + \log p(\lambda_i \mid \boldsymbol{\theta}) - \log q(\lambda_i)\right\}. \tag{2.30}$$

For conditionally-conjugate exponential families, expectations in (2.30) are available in closed form, but evaluating the sum may be expensive when n is large. To alleviate this burden, stochastic variational inference substitutes the coordinate ascent optimization based on (2.28) with a stochastic gradient optimization (Robbins and Monro, 1951), which uses noisy yet easily-computed gradients and is guaranteed to converge under certain conditions on the step size sequence. In the context of variational inference, stochastic optimization approximates (2.30) at each iteration by employing a single sample or a batch of samples selected randomly, that is,

$$\widehat{\mathcal{L}}(q) = \mathrm{E}_q\left\{\log p(\boldsymbol{\theta}) - \log q(\boldsymbol{\theta})\right\} + \frac{n}{B}\sum_{b=1}^{B} \mathrm{E}_q\left\{\log p(y_{i_b} \mid \lambda_{i_b},\boldsymbol{\theta}) + \log p(\lambda_{i_b} \mid \boldsymbol{\theta}) - \log q(\lambda_{i_b})\right\},$$

where i_b is the variable index from the batch and $B \ll n$ is the batch size. The choice of B is dictated by a tradeoff between the computational benefits of taking smaller B, and the gradient noise, which is reduced by taking larger B thanks to the law of large numbers. The choice of step size ρ_t at iteration t also requires tuning: the *Robbins and Monro* conditions,

$$\sum_{t=1}^{\infty} \rho_t = \infty, \qquad \sum_{t=1}^{\infty} \rho_t^2 < \infty,$$

ensure that the entire parameter space can be explored and that the gradient noise decreases sufficiently quickly to guarantee convergence.

Step sizes can be larger if the gradient variance is small, which leads to faster convergence. They may be learnt using adaptive rules based on the current empirical gradient variances (Duchi et al., 2011; Zeiler,

2012; Kingma and Ba, 2014). Alternatively, the batch size B can be set adaptively while keeping the step size sequence fixed (Balles et al., 2016; Byrd et al., 2012; De et al., 2017). More elaborate variance reduction techniques may also be needed, such as resorting to control variates (Paisley et al., 2012; Wang et al., 2013; Johnson and Zhang, 2013) or non-uniform sampling (Perekrestenko et al., 2017; Zhao and Zhang, 2015).

For conditionally conjugate models, the stochastic variational inference updates entail so-called *natural gradients* (Amari, 1985, 1998), which enjoy interesting properties: these gradients allow optimization to take place in Riemann space, where "closeness" is measured by KL divergence, rather than in Euclidean space. This allows the optimization process to exploit the geometry of the parameter space; details can be found in Sato (2001) and Honkela et al. (2008).

Stochastic variational inference has been applied to various domains and is still the object of much research. But its primary, purely computational, motivation has a limited impact in high-dimensional settings where the number of samples is not particularly large. In such contexts, closed-form updates tend to be more effective, as they avoid questions on gradient variance and optimized step size schedules.

The genericity of stochastic optimization has also triggered recent research to broaden the applicability of variational inference. For instance, it may be of interest to relax the requirement of conditionally-conjugate models or to optimize α-divergences other than the reverse KL divergence, as discussed in Section 2.4.2. Because analytical objective functions are typically unavailable, this research implements further levels of approximations, relying on Monte Carlo approximations of intractable expectations, and uses them in stochastic gradient optimizations. More precisely, this can be achieved by computing the variational lower bound gradient with respect to the variational parameter vector $\boldsymbol{\alpha}$ to be optimized, expressing it as an expectation with respect to q,

$$\nabla_\alpha \mathcal{L}(q) = \mathrm{E}_q \left[\nabla_\alpha \log q(\boldsymbol{\theta}; \boldsymbol{\alpha}) \left\{ \log p(\boldsymbol{\theta}, \boldsymbol{y}) - \log q(\boldsymbol{\theta}; \boldsymbol{\alpha}) \right\} \right], \tag{2.31}$$

and obtaining unbiased estimates by Monte Carlo techniques. Gradients of the Rényi α-divergence lower bound can also be estimated by Monte Carlo sums using a simple reparametrization rule; see Li and Turner (2016). As it bypasses the tedious model-specific analytical derivations (2.24) and (2.28) entailed by standard variational algorithms, this approach is often sold as capable of generic inference that can be applied in a "black-box" fashion, see, e.g, Ranganath et al. (2014). The requirements for black-box variational inference are typically very low, whether on the approximating family or on the model itself. For instance, to form Monte Carlo estimates of (2.31), one need only be able to evaluate the model joint distribution, sample from q, and compute the gradient (2.31) efficiently. The main difficulty of black-box variational inference concerns reducing the gradient variance. Owing to their closed-form updates, standard variational methods require no tuning and their convergence can be easily monitored through the changes in the variational objective function. In contrast, the stochastic optimization procedure behind black-box and stochastic variational inferences require careful batch and step-size adaptation, and need to be assessed on a problem-specific basis.

2.4.5 Asymptotic guarantees and finite-sample diagnostics

Theory for variational inference has long seemed understudied, especially when contrasted with the abundant work on Monte Carlo inference. It has also mainly been investigated assuming specific models and variational families. For instance, mean-field variational inference was studied for exponential

family models with missing values (Wang and Titterington, 2004), mixture models with conjugate priors (Wang and Titterington, 2005), the classical Bayesian linear model (You et al., 2014), latent Gaussian models (Sheth and Khardon, 2017), and latent Dirichlet allocation (Ghorbani et al., 2018). Most results concern assessing whether the posterior mean of the variational approximation has standard frequentist asymptotic properties. In particular, for the regression setting that interests us, You et al. (2014) obtain consistency using a normal prior for the regression coefficients and an inverse gamma prior for the error variance. They later extend their work to spike-and-slab priors (Ormerod et al., 2014), also providing selection consistency guarantees. Their results hold for a fixed dimension p, but, based on empirical grounds, they hypothesize that they remain be valid for high-dimensional cases. Huang et al. (2016) establish selection consistency for spike-and-slab regression, allowing p to grow exponentially fast with n. Both Ormerod et al. (2014) and Huang et al. (2016) perform selection using the median probability model of Barbieri and Berger (2004). Recent research tackles more general settings, e.g., Wang and Blei (2018) obtain a Bernstein–von Mises-type of theorem for general parametric models, and Zhang and Gao (2017) study convergence rates of variational posterior distributions for nonparametric and high-dimensional inferences.

Approaches for evaluating the quality of variational approximations in a finite-sample setting are also important, and more so as the variational lower bound can't be used to measure accuracy because of its uninterpretable scale. A natural yet heuristic way to evaluate a variational approximation is to regard MCMC inference as a reference, leaning on its well-established theoretical properties, and benchmark the obtained variational posterior summaries against those of an MCMC approximation for the same model. Carbonetto and Stephens (2012) employ this viewpoint to provide a detailed empirical evaluation of their variational algorithm in the context of genome-wide association studies; we will follow this approach in some numerical experiments of Chapter 3. A constraint however is that the problem sizes used in such comparisons are limited to those for which the MCMC algorithm converges in a reasonable timeframe. MCMC inference is asymptotically exact but the properties of approximation after a finite number of iterations are less well understood, so, for larger problems, the relative accuracy of variational inference and MCMC inference can be difficult to interpret.

Another approach to measuring the quality of variational inference is to rely on diagnostic tools. A recent paper by Yao et al. (2018b) discusses two diagnostics with different purposes. The first employs *Pareto smoothed importance sampling* (PSIS) to assess the quality of the full variational posterior, using the shape parameter of the Pareto distribution to do diagnostics on the variational approximation. Pareto smoothing allows controlling the potentially large variance of the importance sampling weights, by fitting a generalized Pareto distribution to the importance ratios and by replacing the largest weights using the inverse cumulative density function of the fitted distribution. The second diagnostic focuses on assessing the quality of variational point estimates. It is referred to as *variational simulation-based calibration diagnostic* (VSBC) because it obtains calibration probabilities by running variational inference multiple times on simulated datasets, which it then uses to test for asymmetry in the distribution of calibration probabilities. When applied to high-dimensional problems, these diagnostics may require substantial computational resources, possibly more than those needed for the original approximation. Putting aside questions of tractability, importance sampling is often deemed too difficult in high dimensions, so it is not clear how informative the PSIS diagnostic can be in such settings. Still, the practical importance of performance evaluations for variational inference is incontestable, especially as the asymptotic properties of variational inference are less appealing than those of MCMC inference and as it is not clear how this affects the inferred variational posterior distributions. Reliable diagnostic tools should also make us better armed to understand, compare and improve variational algorithms.

2.5 Summary

In this chapter, we reviewed some basic material related to large-scale Bayesian modelling and inference, with Sections 2.1 and 2.2 focusing on the former task and Sections 2.3 and 2.4, on the latter task. The subsequent chapters will attempt tailoring these tools to hierarchical regression for large predictor and response spaces, in molecular quantitative trait locus settings. Our work will heavily rely on two-group mixture priors, which lend themselves to variable selection, but we will also resort to the one-group horseshoe prior (in Chapter 5), whose global-local specification is particularly useful for modelling hotspot predictors, that associate with many responses. We will also capitalize on the flexibility of hierarchical modelling to leverage structural information, via a tailored modelling of the spike-and-slab mixing probability π. In settings where the predictor and response vectors are both high-dimensional, it is important to study how inference behaves with their dimensionality; this is done in Chapters 3 and 5, either by suitably setting hyperparameters or by embedding a multiplicity penalty within a fully Bayesian modelling framework. Both approaches allow enforcing appropriate shrinkage at modelling level, as advocated by Gelman and co-authors; recall Section 2.2.

Variational inference, upon which we will rely throughout this thesis, is a non-standard tool for Bayesian inference, so we need to evaluate its validity on our model, as a replacement of more standard sampling approaches. A lot of emphasis will be put on this in Chapter 3, but all other chapters will complement this support with further algorithmic enhancements and numerical experiments. In particular, in Chapter 4, we will undertake making variational mean-field inference more robust to multimodality by coupling it with the ancestor of simulated tempering discussed in Section 2.3.3, namely, simulated annealing.

3 Variational inference for multiple-response hierarchical regression

In this chapter we present our approach to variable selection in molecular quantitative trait locus (QTL) studies. The marginal screening analysis presented in Section 1.1 illustrates the need for joint modelling of the molecular outcomes and genetic variants; our proposal addresses this. We saw in Chapter 2 that Bayesian variable selection for single-response regression is the subject of a vast research area, by comparison with which, variable selection in multiple-response linear models has received little attention. While most concepts developed for the former setting naturally extend to the latter when a handful of response variables is considered, we will see that further issues arise when the number of responses is large.

Our endeavour has two parts. First, we aim to improve the modelling of complex molecular QTL data in order to better uncover weak associations, and in particular distal pleiotropic effects. The hierarchical sparse regression approach of Richardson et al. (2010) and Bottolo et al. (2011) provides an appealing modelling basis for this (for brevity, we hereafter only cite the former reference). Second, we make joint inference feasible on the hundreds of thousands of genetic variants and thousands of expression outcomes entailed by molecular QTL studies. Most of this chapter is devoted to this latter task: we propose a novel variational procedure for inference on a model adapted from that of Richardson et al. (2010).

A legitimate concern is whether fast deterministic inference is an adequate alternative to MCMC inference for variable selection in molecular QTL studies. Attempting an answer to this is a central objective of this chapter and we follow two complementary lines to this end: first, we evaluate the quality of our variational procedure in small problems, where exact computations and comparisons with MCMC inference are workable. Second, we assess variable selection performance on simulated datasets of realistic sizes, compare it to several existing methods and illustrate it on a real QTL dataset.

The chapter is organized as follows. Section 3.1 presents the model, contrasts it with the general multiple-response regression literature, and discusses its relations to earlier formulations by Jia and Xu (2007), Richardson et al. (2010) and Scott-Boyer et al. (2012). It also proposes a procedure for adjusting for the dimensionality of the predictor space. Section 3.2 recalls some useful variational inference principles from Section 2.4 and explains our choice of approximation. Section 3.3 compares variational and MCMC inferences on our model, also using direct approximations of posterior quantities. Section 3.4 describes numerical experiments for large problems: it compares the method with several predictor selection methods, including the single-response variational approach varbvs (Carbonetto and Stephens, 2012), and with methods performing combined predictor and response

selection, namely HESS (Richardson et al., 2010) and iBMQ (Scott-Boyer et al., 2012). Section 3.5 evaluates the computational efficiency of our proposal by extensive runtime profiling. Section 3.6 reports a permutation-based comparison of our method with varbvs in a metabolite QTL problem. Finally, Section 3.7 extends modelling and inference to handle other data scenarios, including binary and mixed response data.

Most of the work presented in this chapter has been published in Ruffieux et al. (2017). The method is freely available as a package implemented in R, with C++ subroutines (https://github.com/hruffieux/locus).

3.1 Hierarchical sparse regression for multiple responses

3.1.1 Model

Consider a series of parallel regressions

$$\boldsymbol{y}_t = \boldsymbol{X}\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \qquad \boldsymbol{\varepsilon}_t \sim \mathcal{N}_n\left(\boldsymbol{0}, \tau_t^{-1}\boldsymbol{I}_n\right), \qquad t = 1, \ldots, q, \tag{3.1}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q)$ is an $n \times q$ matrix of q centered responses, and $\boldsymbol{X}$, an $n \times p$ matrix of p centered candidate predictors, for each of n samples. Each response, $\boldsymbol{y}_t$, is related linearly to the predictors and has a specific residual precision, τ_t, to which we further assign a Gamma prior, $\tau_t \sim \text{Gamma}(\eta_t, \kappa_t)$. The regressions (3.1) are intended to accommodate any type of molecular QTL data; in this context, the candidate predictors are genetic variants, typically single nucleotide polymorphisms (SNPs), and the responses might represent gene, protein or metabolite levels, depending on whether an eQTL, pQTL or mQTL problem is considered. Here and throughout this thesis, the candidate predictors will be indexed by s for "SNPs", and the responses, by t, for "traits".

As both p and q can be very large compared to n, we enforce sparsity on the $p \times 1$ regression parameters $\boldsymbol{\beta}_t$ by placing a spike-and-slab prior on each of their components, namely, for $s = 1, \ldots, p$,

$$\beta_{st} \mid \gamma_{st}, \sigma^2, \tau_t \sim \gamma_{st}\mathcal{N}\left(0, \sigma^2\tau_t^{-1}\right) + (1 - \gamma_{st})\,\delta_0, \qquad \gamma_{st} \mid \omega_s \sim \text{Bernoulli}\left(\omega_s\right), \tag{3.2}$$

where δ_0 is the Dirac distribution. Hence, to each regression parameter β_{st} corresponds a binary latent parameter γ_{st}, which acts as a "predictor-response association indicator": the predictor $\boldsymbol{X}_s$ is associated with the response $\boldsymbol{y}_t$ if and only if $\gamma_{st} = 1$. The parameter σ represents the typical size of nonzero effects and is modulated by the residual scale, $\tau_t^{-1/2}$, of the response concerned by the effect; we infer σ from the data using a Gamma prior specification, $\sigma^{-2} \sim \text{Gamma}\left(\lambda, \nu\right)$. Finally, we let the probability parameter ω_s have a classical Beta distribution,

$$\omega_s \sim \text{Beta}\left(a_s, b_s\right), \qquad a_s, b_s > 0. \tag{3.3}$$

As it is involved in the Bernoulli prior specification of all $\gamma_{s1}, \ldots, \gamma_{sq}$, the parameter ω_s controls the proportion of responses associated with the predictor $\boldsymbol{X}_s$, and hence directly represents the propensity of predictors to be "hotspots". A graphical representation of the model is provided in Figure 3.1. Our proposal is a variant of the models proposed by Jia and Xu (2007), Richardson et al. (2010) and Scott-Boyer et al. (2012); we discuss their differences in Section 3.1.2.

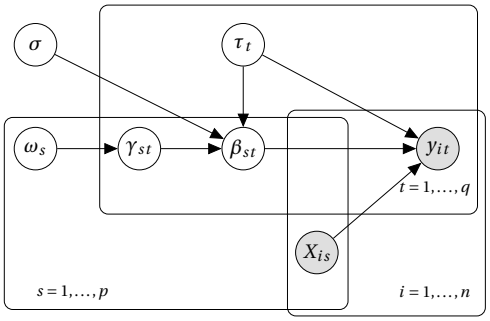

Figure 3.1 – Graphical representation of model (3.1)–(3.2)–(3.3). The shaded nodes are observed, the others are inferred.

Two keywords for model (3.1)–(3.2)–(3.3) are *flexibility* and *interpretability*. We next explain why, by contrasting our model with classical multiple-response approaches. Consider the canonical sparse multivariate regression model,

$$\boldsymbol{Y} = \boldsymbol{X}_\gamma \boldsymbol{B}_\gamma + \boldsymbol{E}, \tag{3.4}$$

where $\boldsymbol{Y}$ is the $n \times q$ response matrix and $\boldsymbol{E}$, the $n \times q$ error term (capital letters are used to emphasize matrix forms). The parameter $\boldsymbol{\gamma}$ is now a $p \times 1$ indicator vector, and, writing $p_\gamma = \sum_{s=1}^{p} \gamma_s$, $\boldsymbol{X}_\gamma$ is the $n \times p_\gamma$ design matrix restricted to variables with $\gamma_s = 1$ and $\boldsymbol{B}_\gamma$ is the corresponding $p_\gamma \times q$ matrix of regression coefficients. The error $\boldsymbol{E}$ is often assigned a matrix-variate normal distribution, $\boldsymbol{E} \sim \mathcal{MN}_{n\times q}(\boldsymbol{0}, \boldsymbol{I}_n, \boldsymbol{\Sigma})$, as defined by Dawid (1981), meaning that the rows $\boldsymbol{E}_{i\cdot} = (\varepsilon_{i1}, \ldots, \varepsilon_{iq})$ are independent identically distributed as $\mathcal{N}_q(\boldsymbol{0}, \boldsymbol{\Sigma})$.

When $q > n$, estimating the precision matrix $\boldsymbol{\Sigma}^{-1}$ is numerically unstable, in addition to being difficult from both computational and memory requirement viewpoints. Hence, although proposals based on such multivariate response modelling for genome-wide association problems exist, they don't accommodate joint modelling of molecular QTL data in full generality. Most of them limit the number of outcomes to just a few, as in Petretto et al. (2010) and Lewin et al. (2015), who extend the original model of Richardson et al. (2010) by applying (3.4) to a small number of different conditions, such as tissues, cell types or time points. Some approaches further simplify the task by considering one candidate SNP predictor at a time, thereby avoiding sparse modelling, see, e.g., Flutre et al. (2013) and Zhou and Stephens (2014) for a mixed modelling approach. Structural assumptions on the precision matrix may alleviate the burden to some extent; for instance, Bhadra and Mallick (2013) enforce sparsity on $\boldsymbol{\Sigma}^{-1}$ and apply their model to problems with a few hundred predictors and responses and around one hundred samples. However, they confess that

> "inferring the covariance graph for a given sample size n scales as $O(q^2)$, where q is the number of correlated traits. Therefore, our approach is useful if one is interested in inferring the interaction among a modest number of traits."

Envisioning reliable joint estimation of both regression coefficient and precision matrices for QTL problems, having q of order $10^3 - 10^4$, seems unreasonable, even with fast determinist inference approaches.

Our model (3.1)–(3.2)–(3.3) takes a different approach. Central to its formulation is the borrowing of information through the model hierarchy. This idea is also at the heart of Richardson et al. (2010) and circumvents modelling the covariance of the residuals, thereby allowing *flexible* inference in $q \gg n$

setups. More precisely, although responses are conditionally independent across the regressions, some dependence is captured via the prior on the regression coefficients, namely, via parameters ω_s and σ, which are common to all the responses. This naturally serves variable selection by leveraging strength across responses associated with the same predictors, and we will see in the experiments of Section 3.4 that this suffices to greatly improve on separate single-response regressions, even when the residual correlations are substantial; some of these experiments involve joint inference for $q = 20,000$ responses or are performed on a genome-wide scale.

The second aspect concerns *interpretable* inference for variable selection, which permits unified selection of outcomes and associated predictors. This is overlooked by most existing Bayesian or frequentist methods, whose focus is on selecting either predictors or outcomes. For instance, O'Reilly et al. (2012) reverse the classical regression setup and fit a succession of models, where each genetic variant is regressed on several outcomes. Because it entails a marginal treatment of the SNPs, their approach is best suited to selecting outcomes for a few candidate SNPs; the authors provide no strategy for handling the massive multiplicity burden that would arise from performing genome-wide screens. Moreover, the method does not penalize model complexity, which may cause instabilities when many outcomes are modelled. More classical proposals for handling multiple outcomes in genome-wide association studies involve direct extensions of the sparse single-response model, whereby to a given SNP $\boldsymbol{X}_s$ corresponds a scalar regression coefficient β_s that measures the overall association of $\boldsymbol{X}_s$ with any of the outcomes (Simon et al., 2013). When there is an effect β_s, the model doesn't provide any information on which outcome(s) are associated with $\boldsymbol{X}_s$. Setting aside computational questions, formulation (3.4) somewhat relaxes this assumption by modelling associations between each pair SNP-outcome via the $p \times q$ regression coefficient matrix $\boldsymbol{B}$. However, selection is performed on the rows of $\boldsymbol{B}$ using the p-variate vector $\boldsymbol{\gamma}$, that is, inclusion or exclusion of a candidate predictor is based on all response variables. Such an assumption can make sense in settings with a few related outcomes, see, e.g., the work of Petretto et al. (2010) and Lewin et al. (2015) mentioned above. But, in molecular QTL problems, each row of $\boldsymbol{B}_\gamma$ is believed to be sparse, as a given SNP may control a small subset of molecular entities, and this subset may vary depending on the SNP considered. One may attempt to enforce *within-row* sparsity by reformulating (3.4) in a Bayesian *seemingly unrelated regression* (SUR) framework (Holmes et al., 2002; Bhadra and Mallick, 2013). SUR models generalize (3.4) by allowing the modelling of different covariates for each response, and therefore open possibilities for response-specific model selection, yet they still entail the modelling of the $q \times q$ response covariance matrix.

In contrast to the above approaches, our proposal (3.1)–(3.2)–(3.3) naturally lends itself to the selection of predictors or responses, and the detection of pairwise associations. Both the posterior means of ω_s and γ_{st} offer direct *interpretable* measures of support for associations; the former quantify the importance of each SNP across outcomes and can serve to select hotspot SNPs in pleiotropic contexts, and the latter correspond to marginal posterior probabilities of inclusion of individual effects, $\mathrm{pr}(\gamma_{st} = 1 \mid \boldsymbol{y})$, and can serve to select SNP-outcome pairs. Support for the inclusion of each SNP or each outcome can also be assessed by summing the marginal posterior probabilities across responses or predictors, respectively.

3.1.2 Relations to earlier proposals

Our model differs from that of Richardson et al. (2010) in two respects. One concerns the treatment of the regression coefficient parameters β_{st}: we use independent priors, whereas Richardson et al. rely on

g-priors (Zellner, 1986). The g-prior specification assumes that the correlation structure in the prior (i.e., that of the regression coefficients) matches that in the likelihood (i.e., that of the predictors). A motivation for considering independent priors instead is that, in genome-wide association problems, effects can take place at locations of the genome that are far apart, and their correlation structure need not reflect the spatial correlation of the SNPs (see, e.g., Guan and Stephens, 2011); we will further discuss encoding predictor dependence structures in prior specifications in Chapter 4. Jia and Xu (2007) also rely on independent priors for regression coefficients, but they model the latter with a mixture of two normal distributions rather than a spike-and-slab prior and impose a residual variance parameter common to all responses. This stringent assumption may represent a weakness of their proposal.

The second difference concerns the third level of the model. Richardson et al. (2010) opt for a multiplicative specification, in which

$$\omega_{st} = \rho_s \omega_t\,, \qquad \omega_t \sim \text{Beta}(a_t, b_t)\,, \qquad \rho_s \sim \text{Gamma}(c_s, d_s)\,, \qquad 0 \leq \omega_{st} \leq 1\,. \tag{3.5}$$

In their case, the predictor inclusion probability is modelled via ω_t; it is specific to each response $\boldsymbol{y}_t$ but modulated by the parameter ρ_s shared by all responses. Jia and Xu (2007) and Scott-Boyer et al. (2012) propose other variants for this prior. The former choose a treatment similar to ours, with $\omega_{st} \equiv \omega_s \sim \text{Dirichlet}(1,1)$, and the latter consider the mixture prior

$$\omega_{st} \mid a_s, b_s, \pi_s \sim \pi_s \delta_0 + (1-\pi_s)\text{Beta}(a_s, b_s)\,, \qquad \pi_s \sim \text{Beta}(a_0, b_0)\,, \tag{3.6}$$

with $a_s \sim \text{Exp}(\lambda_a)$ and $b_s \sim \text{Exp}(\lambda_b)$. This encodes the belief that most SNPs have no association with the responses and hence may be selected using (3.6); this idea was originally described in Lucas et al. (2006). Our choice $\omega_{st} \equiv \omega_s \sim \text{Beta}(a_s, b_s)$ is partly driven by our wish to design a simpler model and partly by practical considerations, since it ensures a closed form for our variational algorithm, unlike with (3.5). While such a formulation was mentioned by Richardson et al. (2010) and by Scott-Boyer et al. (2012), they did not pursue it because of concerns regarding its ability to control for multiplicity. These authors adjust for multiplicity by specifying an expected number of predictors associated with each response and a variance for this number; we instead consider the number of predictors entering the model, as we next explain.

3.1.3 Predictor multiplicity control

The role of the prior specification in inducing sparsity and controlling the association pattern is largely taken by the hotspot propensity parameter ω_s. Thus, it is important to evaluate how different choices of hyperparameters, a_s and b_s, for ω_s may affect inference as the predictor and response dimensions grow. To assess this we consider the *prior odds ratio* representing the support for a model to have an additional response associated with a given predictor $\boldsymbol{X}_s$,

$$\text{POR}(q_s - 1 : q_s) = \frac{\text{pr}\left(\mathcal{M}_{q_s-1}\right)}{\text{pr}\left(\mathcal{M}_{q_s}\right)} = \frac{b_s + q - q_s}{a_s + q_s - 1}\,, \tag{3.7}$$

where $\mathcal{M}_{q_s}$ is a model in which $\boldsymbol{X}_s$ is associated with $1 \leq q_s \leq q$ responses. Such prior odds ratios were first employed for other multiplicity adjustment considerations by Scott and Berger (2010). Clearly, a penalty arises and increases with the total number of responses in the model, q. But no inherent adjustment exists when the total number of candidate predictors, p, increases. We propose to fill this

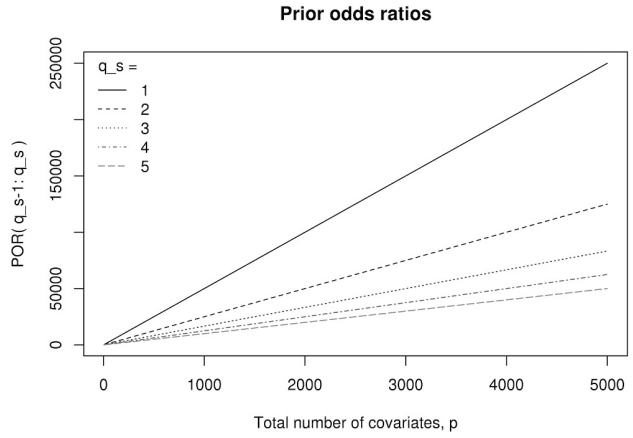

Figure 3.2 – Prior odds ratios, POR($q_s - 1 : q_s$), for $q_s = 1,\ldots,5$, a_s and b_s as in (3.9), $q = 100$, $p^* = 2$, and for a total number of predictors ranging from $p = 5$ to 5,000; see Scott and Berger (2010) for a similar visualization of prior odds ratios in a single-response context.

p	50	250	500	1,000	2,500
Mean # of FP					
Uncorrected	1.06 (0.11)	6.70 (0.33)	16.52 (0.64)	35.22 (0.79)	73.55 (1.09)
Corrected	0.39 (0.08)	0.30 (0.07)	0.36 (0.07)	0.34 (0.08)	0.31 (0.08)
Mean # of TP					
Uncorrected	9.67 (0.07)	9.69 (0.07)	9.72 (0.06)	9.77 (0.05)	9.80 (0.06)
Corrected	9.69 (0.06)	9.42 (0.09)	9.20 (0.09)	9.25 (0.10)	8.98 (0.10)

Table 3.1 – Multiplicity adjustment for the predictor dimensionality. The average numbers of false positives (FP) and true positives (TP) obtained with the uncorrected and corrected regimes are compared for $p_\gamma = 10$ active predictors and an increasing number of noise predictors, $p - p_\gamma$. Selection is performed using the *median probability model* rule, pr($\gamma_{st} = 1 \mid y$) > 0.5 (Barbieri and Berger, 2004). The total number of responses is $q = 25$. 64 replicates were performed; standard errors are in parentheses.

gap by considering the prior probability that $\boldsymbol{X}_s$ is associated with at least one response,

$$\text{pr}\left(\cup_{t=1}^{q}\{\gamma_{st}=1\}\right) = 1 - \prod_{t=1}^{q} \frac{b_s + q - t}{a_s + b_s + q - t}, \tag{3.8}$$

and setting this probability to be equal to p^*/p, where $p^* \ll p$ is the average number of predictors expected to be included in the model. This can be achieved by choosing

$$a_s \equiv 1\,, \qquad b_s \equiv q(p - p^*)/p^*\,, \qquad 0 < p^* < p, \tag{3.9}$$

assuming exchangeability.

Figure 3.2 displays (3.7) for $q_s = 1,\ldots,5$ as a function of p and indicates that, when a_s and b_s are specified as in (3.9), the penalty does increase with the total number of predictors, p; in other words, the prior now also controls for the predictor dimensionality. Moreover, the penalties are not uniform when moving from one to two responses associated with $\boldsymbol{X}_s$, or from four to five, for instance; we will discuss the implications of this in Chapter 5.

The experiment reported in Table 3.1 confirms that adjustment takes place in practice. It considers problems with $p_\gamma = 10$ "active" predictors, i.e., associated with at least one response, and an increasing number of "noise" predictors and it compares the regime with a_s and b_s set according to (3.9) to an "uncorrected" regime with $a_s \equiv 1$, $b_s \equiv 2q-1$, so $\mathrm{E}(\omega_s) \equiv (2q)^{-1}$, meaning that the prior mean number of responses associated with $\boldsymbol{X}_s$ is 0.5. The number of false positives grows linearly with p when the uncorrected model is used but remains roughly constant and close to zero with correction (3.9), giving a clear multiplicity adjustment. The number of true positives is only weakly affected by this correction.

An important caveat here relates to the additional degree of freedom entailed when choosing the pair of hyperparameters a_s and b_s so that (3.8) equals p^*/p for a given p^*. It turns out that inference can be sensitive to different specifications, especially when q is large. This will be the subject of Chapter 5, where we will characterize the sensitivity and propose a solution to it.

3.2 Structured variational inference

Section 3.1.2 described some differences between our model and those of Jia and Xu (2007), Richardson et al. (2010) and Scott and Berger (2010), but a more fundamental distinction concerns the inference procedure.

Joint inference on molecular QTL models is particularly difficult, a serious complication being the high dimensionality of the predictor and the response spaces. In our proposal, as well as in all above-cited proposals, the binary latent matrix $\boldsymbol{\Gamma} = \{\gamma_{st}\}$ creates a discrete search space of dimension $2^{p\times q}$, with $p, q \gg n$, and the quality of inference hinges on the successful exploration of this space. Several sampling schemes have been proposed for spike-and-slab models. Most of them involve drawing each latent component from its marginal posterior distribution, and therefore require costly evaluation of marginal likelihoods at each iteration. Mixing problems also arise, mainly caused by the difficulty that the sampler has in jumping between the states defined by the spike and the slab components. The resulting sample autocorrelations are high, so many iterations are usually needed to collect enough independent samples.

As discussed in Chapter 2, scaling up Bayesian inference algorithms may be attempted by designing more efficient Markov Chain Monte Carlo (MCMC) algorithms. However, research is rather scarce for the high-dimensional case, apart from work on approximating transition kernels (O'Brien and Dunson, 2004; Bhattacharya and Dunson, 2010; Guhaniyogi et al., 2018), so effectively scaling MCMC methods for dimensions such as those involved in the molecular QTL analyses is still largely out of reach. The methods of Jia and Xu (2007), Richardson et al. (2010) and Scott-Boyer et al. (2012) all rely on MCMC inference and hence do not escape this difficulty. Even the enhanced posterior exploration by the adaptive parallel tempering/evolutionary Monte Carlo of Richardson et al. (2010) is not practicable for large molecular QTL data, and we are unaware of any fully multivariate approach that can deal with such data within a reasonable time. Scalable inference is crucial to the uptake of our model in practice. This initiated our interest in variational inference as a deterministic alternative to sampling-based approaches on our model. Carbonetto and Stephens (2012) already proposed a variational algorithm for single-response genome-wide association problems. We next recall the bases of variational inference which we will be relying upon, and outline our approximation.

Let $\boldsymbol{\nu}$ be the parameter vector of interest. Variational posterior approximations are obtained by considering a tractable analytical approximation, $q(\boldsymbol{\nu})$, to the true posterior distribution, $p(\boldsymbol{\nu} \mid \boldsymbol{y})$. The *mean-field* approximation (Opper and Saad, 2001) assumes that $q(\boldsymbol{\nu})$ factorizes over some partition of

$\boldsymbol{v}$, $\{v_j\}_{j=1,\dots,J}$, i.e.,

$$q(\boldsymbol{v}) = \prod_{j=1}^{J} q(v_j)\,, \tag{3.10}$$

with no assumption on the functional forms of the $q(v_j)$. One then performs inference by maximizing the *variational lower bound* on the marginal log-likelihood,

$$\mathcal{L}(q) = \int q(\boldsymbol{v}) \log\left\{\frac{p(\boldsymbol{y},\boldsymbol{v})}{q(\boldsymbol{v})}\right\} \mathrm{d}\boldsymbol{v}\,, \tag{3.11}$$

which is a tractable alternative to minimizing the Kullback–Leibler divergence,

$$\mathrm{KL}\big(q\|p\big) = -\int q(\boldsymbol{v}) \log\left\{\frac{p(\boldsymbol{v}\mid\boldsymbol{y})}{q(\boldsymbol{v})}\right\} \mathrm{d}\boldsymbol{v}\,, \tag{3.12}$$

see Section 2.4.

With each v_j modelled as independent *a posteriori* of the other parameters given the observations and the hyperparameters, mean-field variational inferences (3.10) trade off posterior dependence assumptions and computational complexity. For our model, independence assumptions between β_{st} and γ_{st} would be particularly problematic: they would make $q(\boldsymbol{\beta}_t)$ a unimodal representation of the marginal distribution $p(\boldsymbol{\beta}_t \mid \boldsymbol{y})$, and thus a poor proxy for the highly multimodal posterior distribution implied by the spike-and-slab prior on $\boldsymbol{\beta}_t$. We instead employ a structured factorization, whereby we model β_{st} and γ_{st} jointly, i.e., for each fixed $t \in \{1,\dots,q\}$, we seek a variational distribution of the form

$$\prod_{s=1}^{p} q(\beta_{st},\gamma_{st}). \tag{3.13}$$

This structured factorization induces point mass mixture factors and hence retains the multimodal behaviour of the spike-and-slab distribution. It is also a faithful representation of the true posterior distribution when predictors are only weakly dependent, since the latter factorizes as (3.13) when using an orthogonal design matrix, as pointed out by Carbonetto and Stephens (2012). Indeed, if $\boldsymbol{X}$ is such that $\boldsymbol{X}^T\boldsymbol{X} = n\boldsymbol{I}_n$,

$$\begin{aligned} p\left(\boldsymbol{\beta}_t,\boldsymbol{\gamma}_t \mid \sigma^2,\tau_t,\boldsymbol{\omega},\boldsymbol{y}_t\right) &\propto \exp\Big(-\frac{\tau_t}{2}\|\boldsymbol{y}_t - \boldsymbol{X}\boldsymbol{\beta}_t\|^2\Big)\, p(\boldsymbol{\beta}_t,\boldsymbol{\gamma}_t \mid \sigma^2,\boldsymbol{\omega}) \\ &\propto \exp\left\{-\frac{n\tau_t}{2}\sum_{s=1}^{p}\left(\beta_{st}-\hat{\beta}_{st}\right)^2\right\} \prod_{s=1}^{p} p\left(\beta_{st}\mid\gamma_{st},\sigma^2,\tau_t\right) p(\gamma_{st}\mid\omega_s), \end{aligned}$$

where $\hat{\beta}_{st}$ is the ordinary least squares estimator of β_{st}, $\hat{\beta}_{st} = \boldsymbol{X}_s^T\boldsymbol{y}_t/n$.

In *large n* regimes, the scalability of variational algorithms can often greatly benefit from data-subsampling, which may be implemented generically in stochastic gradient ascent schemes; this is not the case in high dimensions. In this latter regime, we believe that tailored, model-specific, derivations aiming for closed-form updates are important. Taking advantage of the conditional conjugacy properties of our model and of the form of our structured variational approximation, we obtain all the variational updates analytically. The prior distributions of all parameters are preserved by the variational distributions; for instance, we recover a spike-and-slab distribution with modified parameters at posterior level, $q(\beta_{st},\gamma_{st}) = q(\beta_{st}\mid\gamma_{st})q(\gamma_{st})$, with

$$\beta_{st}\mid\gamma_{st}=1,\boldsymbol{y} \sim \mathcal{N}\left(\mu_{\beta,st},\sigma^2_{\beta,st}\right), \qquad \beta_{st}\mid\gamma_{st}=0,\boldsymbol{y} \sim \delta_0\,, \qquad \gamma_{st}\mid\boldsymbol{y} \sim \text{Bernoulli}\left(\gamma^{(1)}_{st}\right),$$

where $\mu_{\beta,st}$, $\sigma^2_{\beta,st}$, $\gamma^{(1)}_{st}$ are *variational parameters* to be updated.

We optimize the variational parameters using a block coordinate ascent scheme, that is, the parameters are updated in turn and by batches for all the responses, taking advantage of the concavity of $\mathcal{L}(q)$ in each of these batches. This scheme combined with the rapidly computable updates produces a highly effective algorithm, which is detailed in Appendix A.2.

3.3 Empirical quality assessment of the variational approximation

3.3.1 Tightness of the variational lower bound

Our variational algorithm is meant to be faster than MCMC inference on our model. However, scalability should not come at the expense of accurate inference; the remainder of the chapter clarifies this through extensive empirical studies. In this section, we evaluate the "closeness" of the variational density q to the target posterior distribution by approximating the Kullback–Leibler divergence $\mathrm{KL}\big(q\big\|p\big)$. This amounts to assessing the tightness of the variational lower bound $\mathcal{L}(q)$ for the marginal log-likelihood, because (recall Section 2.4.2)

$$\mathrm{KL}\big(q\big\|p\big) = \log p(\boldsymbol{y}) - \mathcal{L}(q).$$

For small problems, the marginal likelihood $p(\boldsymbol{y})$ may be accurately approximated using simple Monte Carlo sums. We have

$$p(\boldsymbol{y}) = \int\cdots\int \mathrm{d}\boldsymbol{\omega}\,\mathrm{d}\sigma^{-2} \left\{\prod_{s=1}^{p} p(\omega_s)\right\} p\left(\sigma^{-2}\right) \prod_{t=1}^{q} \left\{\sum_{\gamma_t\in\{0,1\}^p} p\left(\boldsymbol{y}_t \,|\, \boldsymbol{\gamma}_t, \sigma^{-2}\right) \prod_{s=1}^{p} p\left(\gamma_{st} \,|\, \omega_s\right)\right\},$$

with

$$p\left(\boldsymbol{y}_t \,|\, \boldsymbol{\gamma}_t, \sigma^{-2}\right) = \begin{cases} (2\pi)^{-n/2}\,\Gamma\left(\dfrac{n}{2}+\eta_t\right)\dfrac{\kappa_t^{\eta_t}}{\Gamma(\eta_t)}\left(\kappa_t + \dfrac{\|\boldsymbol{y}_t\|^2}{2}\right)^{-n/2-\eta_t}, & q_{\gamma_t} = 0, \\ (2\pi)^{-n/2}\left|\boldsymbol{V}_{\gamma_t,\sigma^{-2}}\right|^{-1/2}\Gamma\left(\dfrac{n}{2}+\eta_t\right)\dfrac{\kappa_t^{\eta_t}}{\Gamma(\eta_t)}\left(\kappa_t + \dfrac{\boldsymbol{S}^2_{\gamma_t}}{2}\right)^{-n/2-\eta_t}\left(\sigma^{-2}\right)^{q_{\gamma_t}/2}, & \\ & \text{otherwise,} \end{cases}$$

where

$$q_{\gamma_t} = \sum_{s=1}^{p}\gamma_{st}, \qquad \boldsymbol{V}_{\gamma_t,\sigma^{-2}} = \boldsymbol{X}_{\gamma_t}^T\boldsymbol{X}_{\gamma_t} + \sigma^{-2}\boldsymbol{I}_{q_{\gamma_t}}, \qquad \boldsymbol{S}^2_{\gamma_t,\sigma^{-2}} = \|\boldsymbol{y}_t\|^2 - \boldsymbol{y}_t^T\boldsymbol{X}_{\gamma_t}\boldsymbol{V}^{-1}_{\gamma_t,\sigma^{-2}}\boldsymbol{X}_{\gamma_t}^T\boldsymbol{y}_t;$$

see Appendix A.3 for details. As no closed form is available for the remaining integrals, we use

$$p(\boldsymbol{y}) \approx \frac{1}{I}\sum_{i=1}^{I}\prod_{t=1}^{q}\left\{\sum_{\gamma_t\in\{0,1\}^p} p\left(\boldsymbol{y}_t \,|\, \boldsymbol{\gamma}_t, \left(\sigma^{-2}\right)^{(i)}\right)\prod_{s=1}^{p} p\left(\gamma_{st} \,|\, \omega_s^{(i)}\right)\right\},$$

where we independently generate

$$\left(\sigma^{-2}\right)^{(i)} \sim \mathrm{Gamma}(\lambda,\nu), \qquad \omega_s^{(i)} \sim \mathrm{Beta}(a_s,b_s), \qquad s=1,\ldots,p, \qquad i=1,\ldots,I. \tag{3.14}$$

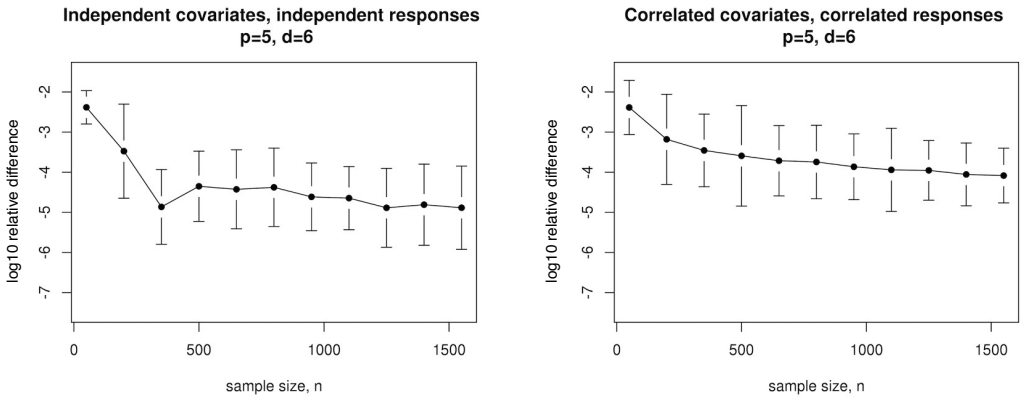

Figure 3.3 – Log_{10} relative difference between the marginal log-likelihood and the variational lower bound. Left: independent predictors and responses. Right: correlated predictors and responses, $\rho = 0.75$. Problems with $p = 5$ predictors, of which $p_\gamma = 3$ randomly selected as "active" (associated with at least one response), and $q = 6$ responses, of which $q_\gamma = 3$ "active" (associated with at least one predictor). Each active predictor is associated with an additional active response with probability 0.25 and explains on average 3.5% of the variance of its corresponding response(s). The number of draws for the simple Monte Carlo approximations is $I = 50,000$; the number of replicates for each sample size is 150.

Figure 3.3 displays the relative difference $\{\log p(\boldsymbol{y}) - \mathcal{L}(q)\} / \log p(\boldsymbol{y})$ for problems with $p = 5$ predictors, $q = 6$ responses and increasing sample sizes, n. In the left panel, the predictors are independent of each other, and so are the responses. In the right panel, the predictors are equicorrelated with correlation coefficient $\rho = 0.75$, and so are the responses. In both cases, the mean relative difference is below 1% with $n = 50$ and tends to decrease as n grows. Although we are not aware of any such study with which to benchmark our results, these values seem very small, suggesting that our variational distribution q adequately reflects the target distribution p for small problems, which is an encouraging sign. These results also suggest using the variational lower bound $\mathcal{L}(q)$ as a proxy for the marginal log-likelihood when performing model selection; we will illustrate this use in Section 3.6. Finally, the fact that the variational lower bound remains tight in the correlated data case is also reassuring, as it suggests that the independence assumptions underlying the mean-field factorization of q may only weakly impact the quality of the approximation.

3.3.2 Comparison with Markov Chain Monte Carlo

We complement our quality assessment by comparing several variational posterior quantities with those for MCMC inference on problems of small sizes. A fair comparison is not straightforward, as these two types of inference rely on stopping rules and convergence diagnostics of very different natures. While the convergence criterion for variational inference comes down to a tolerance to be prescribed, the ability of MCMC sampling to adequately explore the model space for a given chain length can be difficult to evaluate, and usually varies greatly with the problem size. To alleviate the risk of inaccurate MCMC inference, we run 10^5 iterations and discard the first half. We also support our comparison with selected quantities approximated by simple Monte Carlo sums, namely, the marginal posterior

			Active			Inactive
10×	$\beta_{1,2}$	$\beta_{2,1}$	$\beta_{3,2}$	$\beta_{4,1}$	$\beta_{4,2}$	β_{rest} (avg)
Truth	−1.75	2.87	2.37	3.73	−4.76	0.00
VB	−1.74 (0.01)	1.86 (0.01)	1.70 (0.02)	2.26 (0.01)	−3.48 (0.01)	0.02 (0.04)
MCMC	−1.74 (0.33)	1.86 (0.32)	1.69 (0.34)	2.26 (0.32)	−3.48 (0.34)	0.02 (0.14)
			Active			Inactive
	ω_1	ω_2	ω_3	ω_4		ω_{rest} (avg)
True prop. of active resp.	0.2	0.2	0.2	0.4		0
VB	0.21 (0.14)	0.25 (0.15)	0.19 (0.14)	0.33 (0.17)		0.03 (0.06)
MCMC	0.23 (0.18)	0.26 (0.18)	0.21 (0.16)	0.35 (0.18)		0.04 (0.08)
Simple Monte Carlo	0.25	0.26	0.21	0.35		0.05

Table 3.2 – Variational Bayes (VB), MCMC and simple Monte Carlo estimates for $\boldsymbol{\beta}$ (×10) and $\boldsymbol{\omega}$ (components corresponding to noise averaged). Standard deviations are in parentheses.

probability of inclusion of a predictor $\boldsymbol{X}_s$ for a response $\boldsymbol{y}_t$,

$$p(\gamma_{st}=1 \mid \boldsymbol{y}) = \frac{1}{p(\boldsymbol{y})}\frac{1}{I}\sum_{i=1}^{I}\left[\prod_{k\neq t}\left\{\sum_{\boldsymbol{\gamma}_k\in\{0,1\}^p} p\left(\boldsymbol{y}_k \mid \boldsymbol{\gamma}_k, \left(\sigma^{-2}\right)^{(i)}\right)\prod_{j=1}^{p} p\left(\gamma_{jk} \mid \omega_j^{(i)}\right)\right\} \times \left\{\sum_{\boldsymbol{\gamma}_t\in\{0,1\}^p:\,\gamma_{st}=1} p\left(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t, \left(\sigma^{-2}\right)^{(i)}\right)\prod_{j=1}^{p} p\left(\gamma_{jt} \mid \omega_j^{(i)}\right)\right\}\right],$$

and the posterior mean of ω_s, controlling the proportion of responses associated with predictor $\boldsymbol{X}_s$,

$$\mathrm{E}(\omega_s \mid \boldsymbol{y}) = \frac{1}{p(\boldsymbol{y})}\frac{1}{I}\sum_{i=1}^{I}\omega_s^{(i)}\prod_{t=1}^{q}\left\{\sum_{\boldsymbol{\gamma}_t\in\{0,1\}^p} p\left(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t, \left(\sigma^{-2}\right)^{(i)}\right)\prod_{j=1}^{p} p\left(\gamma_{jt} \mid \omega_j^{(i)}\right)\right\},$$

with the samples $\left(\sigma^{-2}\right)^{(i)}$ and $\{\omega_s^{(i)}\}$ generated as in (3.14), with $I = 2\times 10^5$ draws.

We simulated a problem with $p = 8$ predictors, $q = 5$ responses for $n = 250$ samples, and with the nonzero associations explaining on average 13.5% of response variance (data not simulated from the model). Figures 3.4 displays the posterior distributions of $\boldsymbol{\omega}$, $\boldsymbol{\beta}$, σ^{-2} and $\boldsymbol{\tau}$ approximated by MCMC and variational inferences (the latter are obtained in closed form), along with the posterior means obtained by simple Monte Carlo approximations and the true, simulated, values. All distributions agree closely. Those corresponding to inactive β_{st} coefficients all have zero posterior mode. Moreover, in this case, the variational distributions are usually solely made up of a clear spike at zero, whereas the MCMC histograms correspond roughly to a centered Gaussian distribution with average standard deviation 0.014. Table 3.2 summarizes the posterior mean estimates for $\boldsymbol{\omega}$ and $\boldsymbol{\beta}$. Those of the five active regression coefficients, $\beta_{1,2}$, $\beta_{2,1}$, $\beta_{3,2}$, $\beta_{4,1}$ and $\beta_{4,2}$, are significantly different from zero, despite being shrunk under both the MCMC and variational inferences. This shrinkage is a consequence of the spike-and-slab prior but we checked that it does not hamper the detection of the association signals: the marginal posterior probabilities of inclusion of the true nonzero associations are concentrated around 1, while those corresponding to noise are usually much lower, whether obtained by MCMC, variational or simple Monte Carlo procedures; see Figure A.1 of Appendix A.3.2. Finally, the estimates of ω_s provide a fair approximation to the actual proportion of responses associated with a given predictor.

The problems considered thus far were small enough to allow accurate and tractable MCMC inference; the good performance of variational inference is satisfactory but unsurprising for such sizes. The remainder of the chapter considers larger problems, with setups tailored to molecular QTL studies.

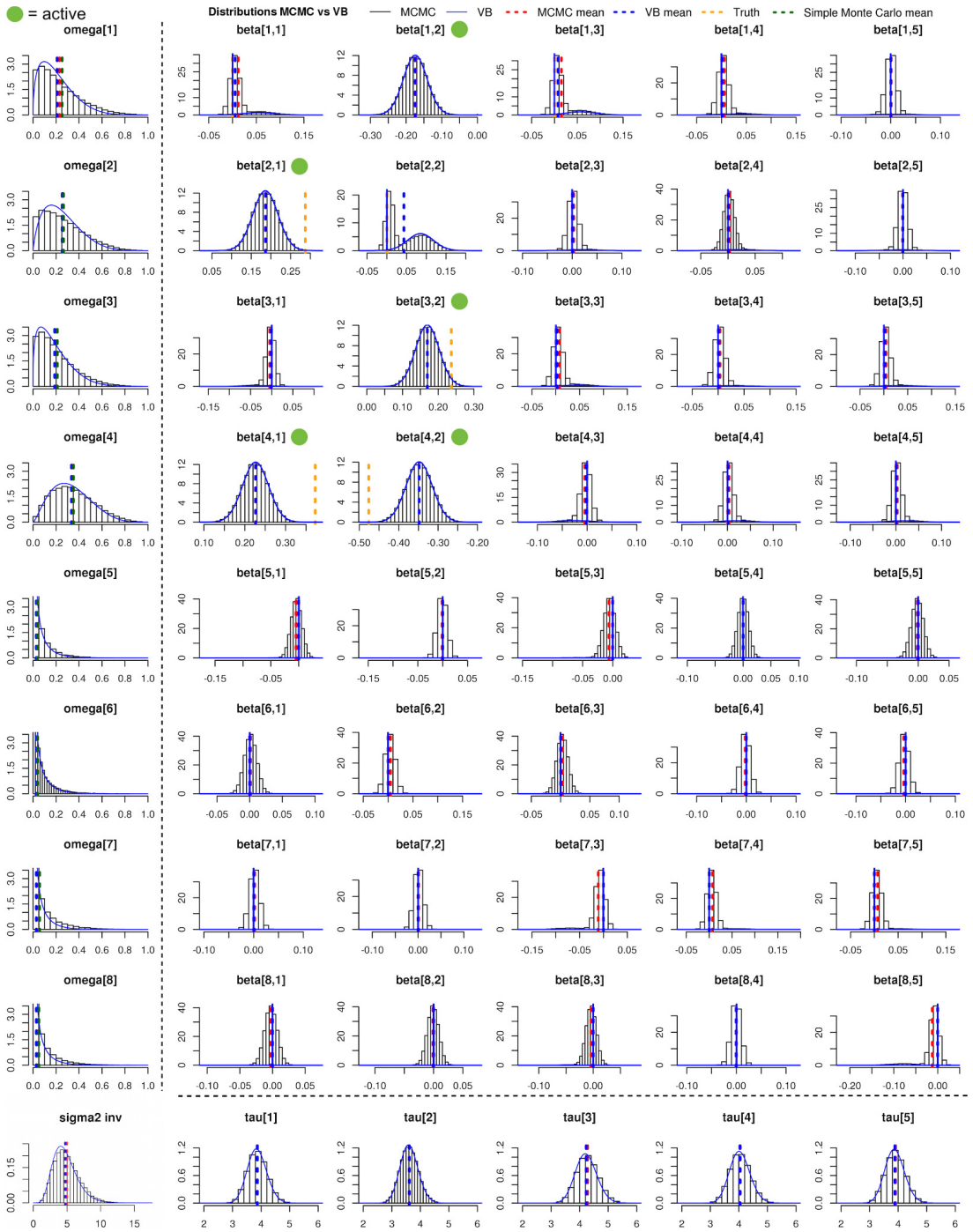

Figure 3.4 – MCMC histograms and variational Bayes (VB, blue) posterior densities for parameters $\boldsymbol{\omega}$ (far left panels), $\boldsymbol{\beta}$ (central panels), σ^{-2} (bottom far left panel) and $\boldsymbol{\tau}$ (bottom panels). The MCMC means (dashed red) and variational means (dashed blue) are displayed, along with the simulated values for the $\boldsymbol{\beta}$ plots (dashed orange) and the simple Monte Carlo approximation of the posterior mean of $\boldsymbol{\omega}$ (dashed green). Most of the dashed vertical lines overlap. The problem has of $p = 8$ independent predictors and $q = 5$ responses for $n = 250$ samples. The five green dots indicate the simulated nonzero associations. We used the software OpenBUGS (Spiegelhalter et al., 2007) and the R package coda (Plummer et al., 2006) for the MCMC inference and convergence diagnostics.

3.4 Variable selection performance

3.4.1 Data-generation design

Our data-generation schemes are meant to embody accepted principles of population genetics. We simulate SNPs as autocorrelated by blocks and under Hardy–Weinberg equilibrium. To this end, we form the blocks using realisations from multivariate Gaussian latent variables and with autocorrelation coefficient drawn uniformly at random in a preselected interval. We then use a quantile thresholding rule to code the number of minor alleles as 0, 1 or 2 according to a SNP-specific minor allele frequency drawn from a uniform distribution, Unif(0.05, 0.5). We also generate responses using multivariate normal variables whose (residual) dependence is either enforced block-wise with preselected auto- or equicorrelation coefficients or chosen to be that of real data. We pick the labels of the active SNPs and outcomes randomly, and associate each active SNP to one (randomly selected) active outcome and to each of the remaining active outcomes with a prescribed probability; some outcomes are therefore under pleiotropic control, i.e., associated with a same hotspot SNP.

We effectively generate the associations under an additive dose-effect scheme, whereby each copy of the minor allele results in a uniform and linear increase in risk, and we draw the proportion of response variance explained by individual SNPs from a left-skewed Beta distribution to favour the generation of smaller effects. We then rescale these proportions so that the response variance attributable to genetic variants matches a given average proportion; the magnitude of SNP effects derives from this value, and the sign of the effects is altered with probability 0.5. These choices imply an inverse relationship between minor allele frequencies and effect sizes, as expected under natural selection (selection against SNPs with large penetrance is stronger, see, e.g., Park et al., 2011). The corresponding data-generating functions are gathered in the R package `echoseq` available at https://github.com/hruffieux/echoseq.

3.4.2 Predictor selection

In this section, we compare our approach, hereafter called LOCUS, with six variable selection methods. These methods are Bayesian or frequentist, and implement either a joint modelling of outcomes and candidate predictors (elastic net for multivariate Gaussian responses), or a joint modelling of candidate predictors only (Bayesian multiple regression based on MCMC inference, "BAS", or variational inference, "varbvs"), or, finally, a fully marginal modelling (univariate ordinary least squares and "lmBF" Bayesian regressions). Complete descriptions and references are in Appendix A.3.3.

We evaluate the ability of each method to identify the "active" predictors, i.e., to determine which candidate predictors are associated with at least one response. For our variational approach, this task is achieved by ranking the posterior means of the hotspot propensity parameters ω_s, which control the proportion of responses associated with a given predictor.

The ROC curves in Figure 3.5 report performance for three simulation configurations based 48 replicates following the design of Section 3.4.1. The first configuration has moderate numbers of predictors ($p = 5,000$) and outcomes ($q = 50$), and allows time-consuming methods to run within hours. The second has many outcomes ($q = 20,000$) and the third has many predictors ($p = 150,000$); these numbers approach those encountered in molecular QTL studies. The remaining settings (numbers of active outcomes and predictors, of observations, effect sizes, etc) are detailed in the caption to Figure 3.5.

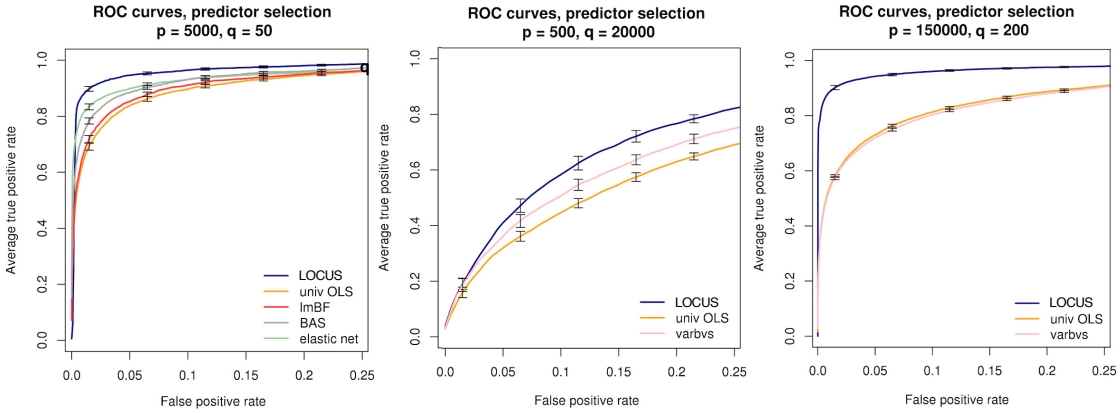

Figure 3.5 – Truncated average receiver operating characteristic (ROC) curves with 95% confidence intervals for predictor selection obtained from 48 replicates. The competing methods are used in three studies of different sizes, based on their computational tractability. Left: $p = 5,000$ predictors spatially autocorrelated with correlation coefficient $\rho_X = 0.75$, $q = 50$ outcomes with residual equicorrelation by blocks with four blocks of equal sizes and correlation coefficients $\rho_Y = 0.8, 0.3, 0.2$ and 0.5, $p_\gamma = 100$ active predictors, $q_\gamma = 40$ active outcomes, $n = 250$ observations, probability of association between an active predictor and an active outcome $p_{\text{add}} = 0.15$, average outcome variance percentage explained by the active predictors $p_{\text{ve}} = 30.0\%$. Middle: $p = 500$ independent predictors, $q = 20,000$ outcomes with residual equicorrelation by blocks of size 10 with $\rho_Y \in \{0.5, \ldots, 0.8\}$, $p_\gamma = 300$, $q_\gamma = 12,500$, $n = 300$, $p_{\text{add}} = 0.01$, $p_{\text{ve}} = 56\%$. Right: $p = 150,000$ predictors autocorrelated by blocks of size 100 with $\rho_X \in \{0.5, \ldots, 0.9\}$, $q = 200$ outcomes with same residual correlation structure as real protein expression levels (DiOGenes study, Larsen et al., 2010), $p_\gamma = 500$, $q_\gamma = 150$, $n = 200$, $p_{\text{add}} = 0.05$, $p_{\text{ve}} = 63\%$. The univariate ordinary least squares and varbvs curves overlap.

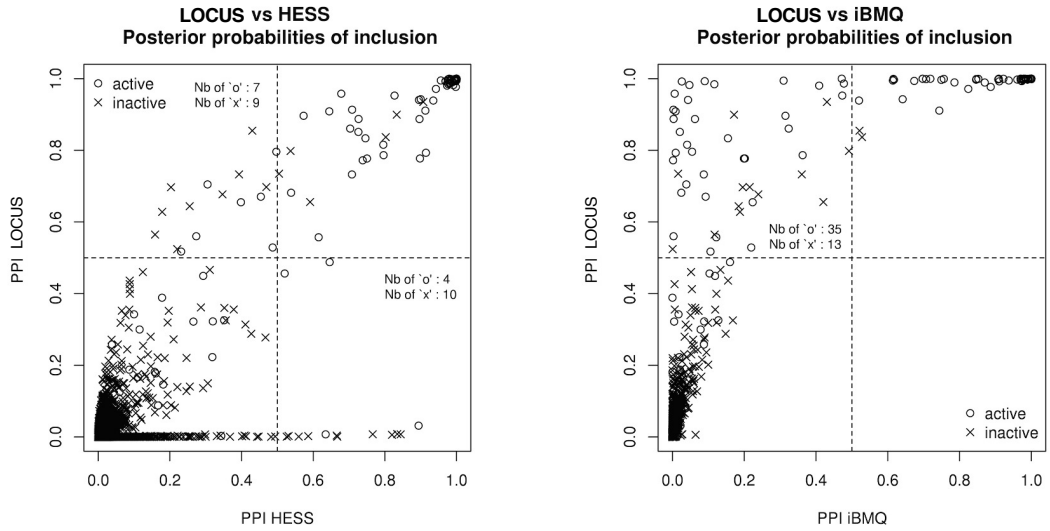

Figure 3.6 – Marginal posterior probabilities of inclusion (PPI) obtained by LOCUS, and those of HESS (left) and those of iBMQ (right), for a problem with $p = 250$ predictors of which $p_\gamma = 50$ are active, with $q = 100$ outcomes, of which $q_\gamma = 50$ are active, and $n = 250$ samples. The probability of association between an active predictor and an active response is $p_{\text{add}} = 0.05$. On average, the active predictors account for $p_{\text{ve}} = 22\%$ of the variance of an outcome with which they are associated.

Our approach outperforms the other methods; its ability to exploit the similarity across outcomes improves substantially the detection of hotspot predictors, associated with multiple outcomes. As already suggested by the results of Section 3.3.1, the correlation structure among the predictors and the outcomes doesn't not seem to have a substantial impact on inferences, despite the independence assumptions implied by the mean-field approximation. The marginal ordinary least squares and marginal lmBF regressions appear to miss many associations because of their univariate modelling of predictors, but jointly accounting for the predictors may not suffice, as suggested by the rather poor performances of the Bayesian multiple regression approaches BAS and varbvs, which apply separate multiple linear regressions for each outcome. Finally, even though the multivariate elastic net models jointly the predictor and outcome variables, its inference suffers from the assumption that to each predictor corresponds a single regression coefficient, common to all responses. As a consequence, regression estimates of predictors with weak or few associations with the responses may be shrunk to zero.

3.4.3 Combined selection of predictors and responses

Unlike the classical variable selection methods used as comparators in Section 3.4.2, our approach, LOCUS, and those of Richardson et al. (2010), HESS, and Scott-Boyer et al. (2012), iBMQ, are tailored to molecular QTL problems: they quantify the associations between each predictor-response pair in a single model, and thus provide flexible and unified frameworks for detecting pairs of associated SNP-outcomes, as well as hotspot SNPs. In this section, we compare the posterior quantities used to perform such selection for the three approaches. As both HESS and iBMQ rely on MCMC sampling, we consider smaller problems than in Section 3.4.2 in order to ensure convergence within a reasonable time; the simulated datasets have $p = 250$ predictors and $q = 100$ outcomes, for $n = 250$ samples (see caption to Figure 3.6 for details). We ran HESS with three MCMC chains, the number selected by the authors for their simulations and with 50,000 iterations of which 25,000 were discarded as burn-in. For iBMQ, we saved 50,000 iterations, after removal of 50,000 burn-in samples, as suggested in the package documentation for a problem of comparable dimensions. Inference for one replicate took on average 10 seconds with our method, around 21 minutes with iBMQ and 4 hours with HESS on an Intel Xeon CPU at 2.60 GHz with 64 GB RAM.

Figure 3.6 compares the marginal posterior probabilities of inclusion obtained by LOCUS with those of HESS and iBMQ. We observe a strong correlation between our approach and HESS, with a quite good ability to discriminate between active and inactive predictor-response pairs. There is a discrepancy at the zero ordinate, where HESS signals a series of false positives and few true positives. The comparison with iBMQ is more uneven, as its posterior probabilities of inclusion for many true associations are below 0.1 and indistinguishable from noise. We reached the same conclusions when running the three methods on 47 additional datasets; see Table 3.3, which gathers sensitivity and specificity measures based on *median probability models* (Barbieri and Berger, 2004, recall Section 2.2).

Figure 3.7 compares the patterns uncovered by the best two methods, namely HESS and LOCUS, again based on the marginal posterior probabilities of inclusion from the first replicate. Visual comparison of the true positive rates suggests that the abilities of the two approaches to detect the true associations are very similar. Our approach indicates the presence of associations in the region of active predictors only, whereas the HESS pattern is blurrier in regions of inactive predictors. The posterior means of ω_s from our approach discriminate quite well between active and inactive predictors, and so do, for HESS,

100×	TPR	TNR
LOCUS	58.9 (0.8)	99.9 (0.0)
HESS	57.9 (0.8)	99.9 (0.0)
iBMQ	35.5 (0.8)	100.0 (0.0)

Table 3.3 – Mean true positive rate (TPR) and true negative rate (TNR) for LOCUS, HESS and iBMQ based on median probability models. Settings: $p = 250$, $p_\gamma = 50$, $q = 100$, $q_\gamma = 50$, $n = 250$, $p_{\text{add}} = 0.05$, $p_{\text{ve}} = 22\%$, 48 replicates. Standard errors are in parentheses.

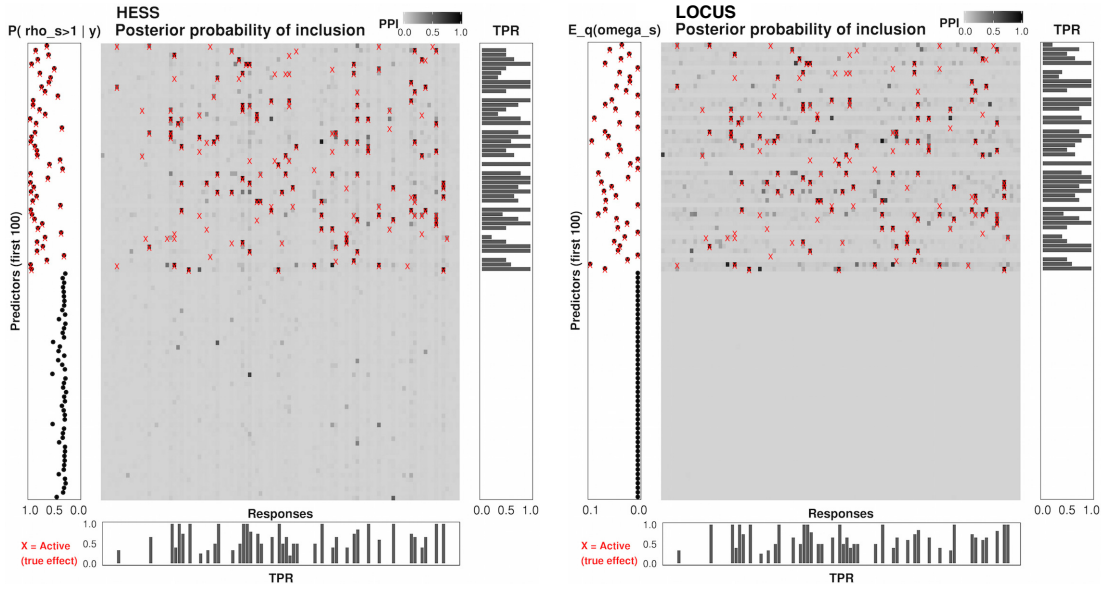

Figure 3.7 – Posterior quantities for detection of associations with HESS (left) and with LOCUS (right), for the simulated datasets described in the caption of Figure 3.6 (only the first 100 predictors are shown). Marginal posterior probabilities of inclusion (central panel), true positive rates for predictor and response selection based on posterior probability of inclusion being > 0.5 (bottom and right panels), posterior probability $\text{pr}(\rho_s > 1 \mid \boldsymbol{y})$ for HESS and posterior mean $\text{E}_q(\omega_s)$ for LOCUS (left panel). The simulated associations are shown by red crosses.

the tail posterior probabilities $\text{pr}(\rho_s > 1 \mid \boldsymbol{y})$, used by Richardson et al. (2010) as measures of predictor hotspot propensities.

3.5 Computational efficiency

Our primary motivation for implementing variational inference schemes for QTL models was to enhance scalability; it is therefore of interest to quantify the actual gain in this respect. While a naive implementation of the algorithm would not scale linearly with the number of predictors (see the variational updates for β_{st} in the algorithm of Appendix A.2.3), we obtained linear scaling for the predictor dimension p, response dimension q and sample size n, with some algebra, by locally updating quantities stored once for all. We also improved the runtime using the optimized linear algebra C++ library *Eigen* (Guennebaud and Jacob, 2010), and we managed memory consumption by recomputing certain objects on demand and by using zero-copy techniques across the R/C++ interface.

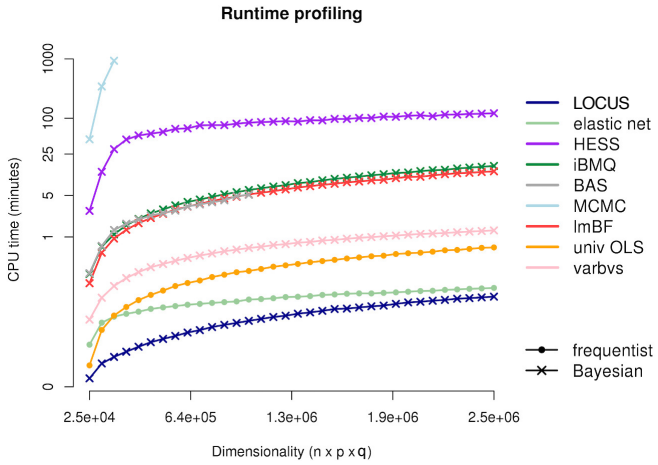

Figure 3.8 – Serial runtime profiling for all methods discussed, on an Intel Xeon CPU at 2.60 GHz with 64 GB RAM.

We report a runtime profiling of the different methods assessed in Section 3.4 and for a range of problem sizes ($n \times p \times q$). Figure 3.8 displays the run times in minutes, averaged over 24 replicates. We aim for overall evaluations rather than precise and exhaustive comparisons, since the methods all depend on parallelism and convergence characteristics that are not directly comparable. We ran all methods serially, in an attempt to treat them on an equal footing. This choice could be challenged, as some approaches (e.g., any marginal screening) are more parallelizable than others (e.g., LOCUS or HESS), but the number of cores used for the latter represents an additional setting that may have a large impact on the measures. The number of chains for MCMC inference also matters: we ran HESS (Richardson et al., 2010) with three chains, following its authors' choice made in their simulations; the other MCMC inferences are based on a single chain. Finally, the runtime may also greatly vary depending on the chosen chain length: the Bayesian multiple regression method BAS (Clyde, 2016) selects it adaptively, and we ran the remaining MCMC methods for 50,000 iterations, based on preliminary convergence diagnostics. In practice, the number of samples needed until convergence typically increases with the problem size, a fact that we didn't take into account in this profiling. Hence, if one were to properly adapt the chain lengths to the dimensionality, the curves of Figure 3.8 corresponding to the MCMC approaches would tend to deviate more from that of LOCUS, whose coordinate ascent scheme stops only once convergence is reached.

With these serial settings, LOCUS is the fastest method; it usually converged in tens of iterations. At the other extreme, MCMC inference for our model is the slowest, which underlines the intractability of MCMC sampling for large problems. The evolutionary stochastic search of HESS does better, but still more than 650 times slower than our variational approach. Our method is also about 10 times faster than q applications (one for each outcome) of the varbvs method (Carbonetto and Stephens, 2012) but these parallel regressions can run independently on multiple cores, which is an important advantage. Likewise, the runtime of the univariate methods is to be (roughly) divided by the number of available cores.

In terms of absolute figures, LOCUS scales to any proteomic, metabolomic and lipidomic QTL problems that involve several hundreds of thousands of SNPs ($p = 500,000$ or more), thousands of levels ($q = 5,000$ or more) and thousands of samples ($n = 5,000$ or more). The memory requirements for a typical pQTL analysis with $p = 500,000$, $q = 1,000$ and $n = 1,000$ is about 16 GB of RAM. For expression QTL

analyses with around twenty-thousand transcripts, the current implementation of the method requires running it chromosome by chromosome; this can be done in parallel and should not have important consequences on inference as SNPs tend to be weakly correlated across chromosomes. Improvements should consider both time and memory aspects.

3.6 Application to metabolite quantitative trait locus data

We end these numerical experiments by illustrating our approach on data from a large multicenter dietary intervention study (DiOGenes; Larsen et al., 2010). The study collected genetic, genomic, proteomic and metabolomic data at different stages of a dietary treatment provided to the cohort. Its goal is to uncover molecular mechanisms underlying the metabolic status of overweight individuals and improve understanding of the factors predisposing weight regain after a diet. Here, we perform a (lipid-)metabolite quantitative trait locus (mQTL) analysis; in this context, one may hypothesize that some metabolites and their variation during the intervention could serve as proxies for the clinical outcome of interest, weight maintenance. We also use this illustration on real data to further highlight the benefits of modelling the outcomes jointly via an extensive permutation-based comparison with the single-response variational method varbvs (Carbonetto and Stephens, 2012).

After quality control, the data consist of $p = 215,907$ tag SNPs and $q = 125$ metabolite levels, adjusted for age, center and gender, for $n = 317$ individuals. The SNPs were genotyped by Illumina HumanCore technology and the metabolites were quantified in plasma using liquid chromatography-mass spectrometry (LC-MS). They span cholesterol esters (CholE), phosphatidylcholines (PC), phosphatidylethanolamines (PE), sphingomyelins (SM), di- (DG) and triglycerides (TG).

We control for predictor multiplicity by specifying the hyperparameters for $\boldsymbol{\omega}$ according to the discussion of Section 3.1 and choose the prior average number of active SNPs, p^*, by grid search within a 3-fold cross-validation procedure that maximizes the variational lower bound. The entire procedure took about 10 hours on an Intel Xeon CPU at 2.60 GHz, and the final run converged in 83 iterations. The posterior means of ω_s suggest the presence of several hotspot SNPs, spread across the chromosomes (Figure 3.9).

We compare varbvs and LOCUS on these data based on the number of associations declared by each method at specific false discovery rates estimated by permutations. We apply Efron's Bayesian interpretation of the false discovery rate (Efron, 2008) to marginal posterior probabilities of inclusion, $\text{PPI}_{st} = \text{pr}(\gamma_{st} = 1 \mid \boldsymbol{y})$, and use an empirical null distribution based on $B = 400$ permutations to compute the estimate

$$\widehat{\text{FDR}}(\tau) = \frac{\text{median}_{b=1,\ldots,B} \#\{\text{PPI}_{st}^{(b)} > \tau\}}{\#\{\text{PPI}_{st} > \tau\}}, \qquad 0 < \tau < 1, \tag{3.15}$$

for a grid of thresholds τ; we then fit a cubic spline to the resulting false discovery rates to find thresholds for specific rates (Appendix A.4.1). At estimated FDR of 5%, LOCUS declares 21 associations and varbvs 19, and at FDR 25%, these numbers are 89 and 47 respectively; see Figure 3.9, and Table 3.4, which provides the numbers for further FDR choices. These associations also partly agree with those obtained with marginal screening at Benjamini–Hochberg FDR of 25%.

Database searches on the functional relevance of the detected associations provide hints of promising biological functions related to metabolic activities for 12 of the 25 SNPs selected by our procedure.

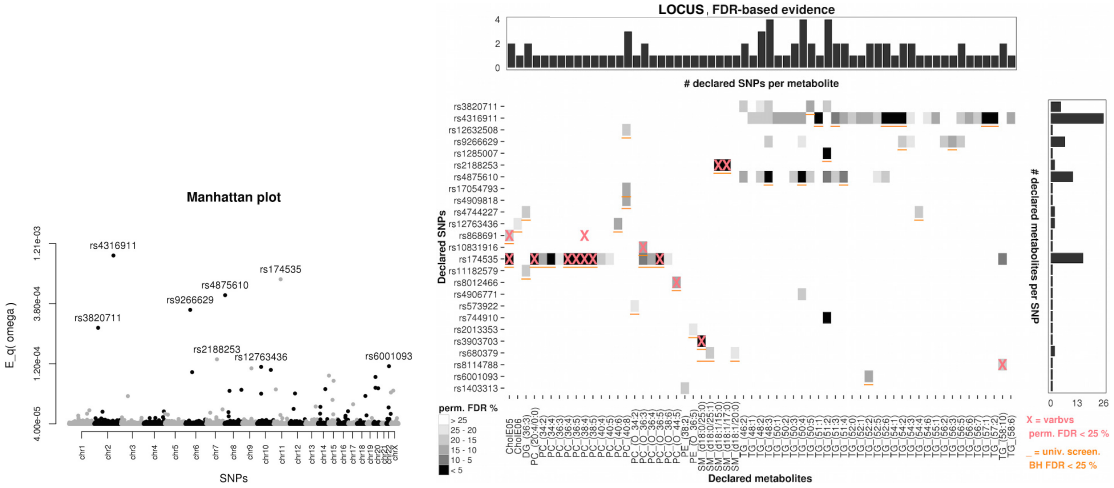

Figure 3.9 – SNPs and pairwise associations identified by our approach for the DiOGenes study. Left: Manhattan plot for SNP association and evidence of pleiotropy. Right: marginal posterior probabilities of inclusion for SNP-metabolite associations found at estimated FDR of 25% and overlap with the associations found by the varbvs method at same FDR level (pink crosses) and found by univariate screening at Benjamini–Hochberg FDR of 25% (orange underscores).

	# detected:		
Permutation-based FDR (%)	LOCUS	varbvs	LOCUS ∩ varbvs
5	21	19	8
10	26	19	8
15	47	21	10
20	76	31	12
25	89 (48 univ.)	47 (19 univ.)	14 (13 univ.)

Table 3.4 – Numbers of associations detected by LOCUS (our method) and by varbvs, and numbers of signals in common at selected permutation-based false discovery rates. In each case, the number of associations also detected by univariate screening at Benjamini–Hochberg FDR of 25% is in parentheses.

For instance, the top hotspot SNP, rs4316911, has many associations with triglyceride levels. It is also located less than 150 kilobases upstream of the *ITGA6* gene, which has probable implications in diabetic kidney disease (Iyengar et al., 2015); the region therefore appears as a candidate risk locus, whose possible mechanisms of action need to be clarified. The second most prominent hotspot identified by LOCUS, rs174535, associates with phospholipids, more precisely with 14 different phosphatidylcholine levels, of which four are ether-linked/plasmalogen (PC-O). Interestingly, this SNP has known links with metabolite levels, in particular with trans fatty acid levels and plasma phospholipid levels (Mozaffarian et al., 2015), in line with our findings. This SNP is also an eQTL for the fatty acid desaturase genes *FADS1* and *FADS2*. Finally, SNP rs3903703 is known as being associated with very long-chain fatty acid levels (Lemaitre et al., 2015). This seems to agree with our findings, in which rs3903703 exhibits associations with sphingomyelin, a type of lipid containing fatty acids of different chain lengths. The complete list of SNPs with metabolism-related associations found by LOCUS is given in Appendix A.4.

A replication analysis using simulated data can be found in the supplementary material of Ruffieux et al. (2017); we used the R package `echoseq` for generating the SNPs based on the minor allele frequencies

and linkage disequilibrium structure of the real SNPs, as well as for generating the outcomes based the empirical dependence structure of the metabolites.

3.7 Some direct extensions

Several extensions naturally come to mind. First, molecular QTL datasets are often complemented with other variables that may correspond to potential confounding factors, like gender, age, lifestyle, or demographic characteristics. We extended the model to include such variables as covariates that are not subject to selection. This is straightforward using a centered normal prior for their regression coefficient, $\boldsymbol{\alpha}_t$, i.e., for d such covariates gathered in the $n \times d$ matrix $\boldsymbol{Z}$,

$$\boldsymbol{y}_t = \boldsymbol{Z}\boldsymbol{\alpha}_t + \boldsymbol{X}\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \quad \alpha_{rt} \sim \mathcal{N}\left(0, \zeta_r^2\right), \quad \zeta_r^{-2} \sim \text{Gamma}\left(\phi_r, \xi_r\right), \quad r = 1, \ldots, d,$$

keeping the rest of the hierarchy untouched.

Second, although it is primarily intended for molecular QTL problems, our model may be considered in genome-wide associations problems with a few clinical phenotypes (clinical QTL problems). Since these may involve case-control designs, we also implemented variants of the algorithm for binary responses, based on logit and probit link functions, as well as for mixed responses, based on combined linear and probit link functions. While probit regressions pose no difficulties for deriving closed-form variational updates, extending our algorithm for logistic regression requires a further level of approximation, because the likelihood does not have a conjugate prior in the exponential family. Although it is possible to resort to Monte Carlo approximations for estimating the variational lower bound, this may substantially affect scalability. We instead choose to stick to analytical expressions and follow Jaakkola and Jordan (2000) to further bound the intractable expectation using a local Gaussian approximation to the sigmoid function, $\text{Sig}(z) = \left\{1 + \exp(-z)\right\}^{-1}$, i.e.,

$$\text{Sig}(z) \geq \text{Sig}(\eta) \exp\left\{\frac{z-\eta}{2} - \rho(\eta)\left(z^2 - \eta^2\right)\right\}, \qquad \rho(\eta) = \frac{1}{2\eta}\left(\text{Sig}(\eta) - \frac{1}{2}\right).$$

The price to pay is the introduction of the auxiliary parameter η (actually, n of them as there is one copy η_i per observation), which we optimize using an expectation-maximization step, see Appendix A.5.2; credits go to Loris Michel, who derived all logistic variational updates in his masters thesis. The algorithms for the above-cited models are in Appendix A.5 and are available in the R package `locus`.

Finally, the Bayesian framework allows natural extensions to missing responses; obtaining these is future work. As a provisional ersatz, the current implementation of our algorithms accept missing data by simply discarding the samples with missing responses in the incriminated parallel regression. This is better than discarding all samples with at least one missing response but is clearly a misuse of the flexibility entailed by Bayesian modelling. As for missing values in the SNP data, imputing them from the model is a lower priority given the wealth of imputation algorithms available to practitioners (Halperin and Stephan, 2009; Howie et al., 2009; Marchini and Howie, 2010); these algorithms are often fine-tuned to the specificities of these data.

3.8 Summary

This chapter presented our hierarchical regression approach to the joint modelling of large predictor and response vectors, and applied it to data emulating molecular QTL data. It confirmed earlier demonstrations (Jia and Xu, 2007; Richardson et al., 2010; Scott-Boyer et al., 2012) that borrowing information across responses greatly improves variable selection. In particular, it showed that capturing dependence through the model hierarchy is a flexible alternative to directly modelling the correlation structure of the responses.

To the central question "is variational inference a good surrogate for MCMC inference for variable selection from our model?", our numerical experiments suggest an affirmative answer. This agrees with the conclusions of Carbonetto and Stephens (2012) on the appropriateness of variational inference on single-outcome genome-wide association data. Carbonetto and Stephens (2012) point out, however, that selection performance can suffer from highly correlated SNPs; we will discuss and attempt addressing this in Chapter 4. Finally, we highlighted that model-specific derivations for our variational algorithm, although somewhat tedious, are crucial to its efficacy in high dimensions, and we illustrated the computational gain over competing variable selection methods, including MCMC-based approaches.

4 Dependence structures

This chapter considers two separate but related themes: modelling for spatially dependent predictors and inference for multimodal parameter spaces. These themes are linked because dependence structures tend to exacerbate posterior multimodality.

Correlation among genetic variants is a central characteristic of genome-wide and molecular quantitative trait locus data, and is generically termed *linkage disequilibrium*; we provide more context on this in Section 4.1. In regression settings, the effects of highly correlated predictors on response variables can be difficult to infer and interpret. When these effects can be regarded as related by some underlying structure, it is natural to try to incorporate this information in the model, which may be easily achieved in the Bayesian framework. Section 4.2 provides two illustrations of this, based on the model presented in Chapter 3.

Instead of injecting structural assumptions in the model, one may attempt to make inference more robust to strong multimodality in general. This is the aim of Section 4.3, where we propose coupling our variational algorithm with a simulated annealing procedure that permits better exploration of multimodal parameter spaces. This simulated annealing procedure has been developed in collaboration with Leonardo Bottolo and Sylvia Richardson and is part of Ruffieux et al. (2018a, submitted).

4.1 Problem statement

Although every individual has a unique DNA sequence, certain combinations of genetic variants are inherited together. The extent of the resulting nonrandom assortment of alleles at two or more polymorphisms on a chromosome is termed *linkage disequilibrium* and tends to give rise to block dependence structures among variants along the genome (Balding et al., 2008, Chap. 27). Such blocks can span large portions of a chromosome, and their size and location depend on factors such as the times of mutation events and population history (Goode, 2011; International HapMap Consortium, 2005). A number of algorithms have been proposed to estimate pairwise or block linkage disequilibrium; see, e.g., Barrett (2009), Berisa and Pickrell (2016) and Zheng et al. (2017).

Linkage disequilibrium structures have important effects on inference, regardless of whether marginal or joint approaches are used. The motivating example of Chapter 1 indicates that marginal screening generates spurious associations in regions of high linkage disequilibrium. If two nearby SNPs are almost perfectly correlated and only one of them is causal, as often assumed at a given locus (Li and

Zhang, 2010), then both are likely to be assigned a high measure of support for association. Because of the high multiplicity burden entailed by molecular QTL problems, we saw that this hamper the detection of distal *trans* effects, which are typically weaker than proximal *cis* effects. A conventional yet simplistic strategy to account for linkage disequilibrium is to report, for each locus, the SNP whose effect is most significant.

Sparse regression methods also suffer from linkage disequilibrium. When there is insufficient information to pinpoint the relevant SNP from its correlated neighbours, some methods tend to choose one SNP interchangeably among all the candidates. This is a well-known characteristic of the LASSO method (Tibshirani, 1996), that can also concern Bayesian variable selection (Chipman, 1996). Our variational inference approach can tolerate some dependence in the data (recall the experiments of Sections 3.3 and 3.4), but is also affected by strong correlation structures, as we will explain in Section 4.3. We will also see that model averaging can often mitigate this by providing a better grasp of the uncertainty entailed by the selection of correlated SNPs. However, this requires rerunning the analysis multiple times, which can represent a substantial computational overhead.

A complication is when the causative variant is not among those analysed. This can occur when the variant was sequenced or genotyped but discarded after the quality control or other preprocessing steps, or when the variant was not even measured on the genotyping chip. Such chips are made of probes for few hundred thousand so-called *tag SNPs* out of the millions of SNPs entailed by human genomes, and the criteria for selecting tag SNPs usually depend on the chip type and manufacturer; typically, they are chosen to cover the major linkage disequilibrium blocks of genes in a given population (Rogers and Weiss, 2017). A consequence is that the hits observed may relate to tag SNPs acting as proxies for unmeasured causal SNPs.

In all the above cases, fine mapping and external annotation data on targeted regions may help to clarify the functional roles. This is a difficult task, however; it has been suggested that most of the SNPs collected in biomarker databases are at best surrogates for a nearby SNP in linkage disequilibrium, rather than genuine functional entities (Donnelly, 2008).

4.2 Structured modelling

4.2.1 Group sparsity model

Consider a regression setting where the candidate predictors can be naturally arranged into G disjoint groups and let

$$\begin{aligned}
\boldsymbol{y}_t &\mid \boldsymbol{\beta}_t, \tau_t \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}_t, \tau_t^{-1}\boldsymbol{I}_n\right), && t = 1,\dots,q,\\
\boldsymbol{\beta}_{gt} &\mid \gamma_{gt}, \sigma^2, \tau_t \sim \gamma_{gt}\mathcal{N}_{|g|}\left(\boldsymbol{0}, \sigma^2\tau_t^{-1}\boldsymbol{I}_{|g|}\right) + (1-\gamma_{gt})\,\delta_0, && g = 1,\dots,G, \qquad (4.1)\\
\gamma_{gt} &\mid \omega_g \sim \text{Bernoulli}(\omega_g), \qquad \omega_g \sim \text{Beta}(a_g, b_g),
\end{aligned}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1,\dots,\boldsymbol{y}_q)$ is an $n \times q$ matrix of q centered responses, $\boldsymbol{X}$ is an $n \times p$ matrix of p centered candidate predictors for each of n samples, $|g|$ is the cardinality of group $g \in \{1,\dots,G\}$, and δ_0 is a point mass at $\boldsymbol{0} \in \mathbb{R}^{|g|}$. We assign Gamma priors to the precision parameters τ_t and σ^{-2}.

Model (4.1) is a variant of model (3.1)–(3.2)–(3.3) whereby the binary latent indicator γ_{gt} selects groups of variables. It therefore applies to settings where it makes sense to let related predictors enter

the model simultaneously rather than selectively. When the group structure is the result of strong block dependences among predictors, the model also has the practical advantage of bypassing the instability of sparse regression methods when attempting to select individual predictors from correlated candidates. Chipman (1996) was probably the first to use structural grouping information under a (continuous) spike-and-slab prior for single-response regression. He argues

> "Not only does the grouping principle reduce the size of the total model space, but it makes headway in dealing with the pitfalls of multiple comparisons";

indeed, the dimension of each parallel regression latent indicator $\boldsymbol{\gamma}_t$ has been reduced from $p \times 1$ to $G \times 1$, where G may be one or two orders of magnitude smaller than p, depending on the context. Chipman suggests that interpretation from this model should be in two steps: one starts by identifying components of $\boldsymbol{\gamma}_t$ for which $\text{pr}(\gamma_{gt} = 1 \mid \boldsymbol{y})$ is large, then, for these, one identifies the components of the $|g| \times 1$ regression vector $\boldsymbol{\beta}_{gt}$ with large posterior means.

Model (4.1) does not impose sparsity on the regression coefficients, conditional on their group being selected. This is like the classical group selection extension of the LASSO by Yuan and Lin (2006), where an ℓ_2 penalty is applied uniformly within the selected groups. Friedman et al. (2010) proposed a *sparse-within-group* alternative to this original group LASSO by replacing the ℓ_2 penalty with a ℓ_1 penalty.

In the present molecular QTL context, it is natural to base the groups on linkage disequilibrium blocks, whose estimation is thus prerequisite to the use of model (4.1). The dense-within-group assumption of (4.1) is convenient here, as selecting SNPs from blocks in a sparse-within-group fashion may result in the collinearity issues met initially. Hence, when doing inference with model (4.1), we take the view that the model should provide evidence at the level of loci and that finer within-loci selection should be deferred to follow-up studies. This makes further sense if the functional SNP from an identified locus is absent from the SNP panel.

Returning to model (3.1)–(3.2)–(3.3) where $\boldsymbol{\gamma}_t$ is $p \times 1$, we saw in Section 3.2 that the posterior $p(\boldsymbol{\beta}_t, \boldsymbol{\gamma}_t \mid \sigma^2, \tau_t, \boldsymbol{\omega}, \boldsymbol{y}_t)$ factorizes across the predictors in the orthogonal design case, and is then faithfully reproduced by the structured variational mean-field distribution

$$q(\boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) = \prod_{s=1}^{p} q(\beta_{st}, \gamma_{st}).$$

Model (4.1) lends itself better to an approximation that accounts for the block linkage disequilibrium structure of the SNPs, since

$$\begin{aligned} q(\boldsymbol{\beta}_t, \boldsymbol{\gamma}_t) &= \prod_{g=1}^{G} q(\boldsymbol{\beta}_{gt} \mid \gamma_{gt}) q(\gamma_{gt}) \\ &= \prod_{g=1}^{G} \mathcal{N}_{|g|}\left(\boldsymbol{\beta}_{gt}; \boldsymbol{\mu}_{gt}, \boldsymbol{\Sigma}_{gt}\right)^{\gamma_{gt}} \delta_0^{1-\gamma_{gt}} \text{Bernoulli}\left(\gamma_{gt}; \gamma_{gt}^{(1)}\right), \end{aligned}$$

where $\boldsymbol{\mu}_{\beta,gt}$, $\boldsymbol{\Sigma}_{\beta,gt}$, $\gamma_{gt}^{(1)}$ are the variational parameters for group g. In particular, the update for $\boldsymbol{\Sigma}_{\beta,gt}$ explicitly involves the empirical covariance structure of the candidate predictors in group g,

$$\boldsymbol{\Sigma}_{\beta,gt}^{-1} = \tau_t^{(1)} \left\{ \boldsymbol{X}_g^T \boldsymbol{X}_g + \left(\sigma^{-2}\right)^{(1)} \boldsymbol{I}_g \right\}. \tag{4.2}$$

In order words, conditionally on $\gamma_{gt} = 1$, τ_t and σ^2, the regression coefficients of the SNPs in group g are independent *a priori* but dependent *a posteriori*, with a structure that reflects that in the data. The full variational updates and objective function are given in Appendix B.1.

4.2.2 Similarity sparsity model

The assumption that genetic variants along the genome form disjoint groups is stringent, as the boundaries of linkage disequilibrium blocks may be blurred. Under model (4.1), two slightly different group partitions may generate very different hypotheses for prospective functional analyses, since the inclusion of a given variant as candidate regulatory marker hinges on the selection of the entire group to which it has been assigned. To alleviate this, we describe a model that replaces the partition-based structural information with a continuous measure of similarity between variants that acts as a "smoothness" penalty. For q centered responses and p centered candidate predictors, we consider

$$\begin{aligned} \boldsymbol{y}_t &\mid \boldsymbol{\beta}_t, \tau_t \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}_t, \tau_t^{-1}\boldsymbol{I}_n\right), && t = 1, \dots, q, \\ \beta_{st} &\mid \gamma_{st}, \sigma^2, \tau_t \sim \gamma_{st}\mathcal{N}\left(0, \sigma^2\tau_t^{-1}\right) + (1-\gamma_{st})\,\delta_0, && s = 1, \dots, p, \\ \gamma_{st} &\mid \theta_s \sim \text{Bernoulli}\{\Phi(\theta_s)\}, && \boldsymbol{\theta} \sim \mathcal{N}_p(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0), \end{aligned} \tag{4.3}$$

where τ_t and σ^{-2} are assigned Gamma priors, $\Phi(\cdot)$ is the standard normal cumulative distribution function and δ_0 is the Dirac distribution. Pairwise similarity between predictors is encoded by the $p \times p$ (positive definite) matrix $\boldsymbol{\Sigma}_0$, and we choose the hyperparameter $\boldsymbol{m}_0$ to induce sparsity; see the simulation settings of Section 4.2.3.

Engelhardt and Adams (2014) propose a similar structured probit-link formulation in a single-response context,

$$\boldsymbol{\beta} \mid \boldsymbol{\Gamma}, \sigma^2, \tau \sim \mathcal{N}_p\left(\boldsymbol{0}, \tau^{-1}\sigma^2\boldsymbol{\Gamma}\right), \qquad \boldsymbol{\theta} \sim \mathcal{N}_p\left(\boldsymbol{0}, \boldsymbol{\Sigma}_0\right), \qquad \Gamma_{s,s} = \mathbb{1}\left(\theta_s > \theta_0\right), \qquad s = 1, \dots p, \tag{4.4}$$

where $\boldsymbol{\Gamma}$ is a $p \times p$ degenerate diagonal covariance matrix and where θ_0 has a Gaussian prior. They provide a list of kernel functions that may be used for $\boldsymbol{\Sigma}_0$, but mention that this choice had little impact in their experiments. They don't discuss the inversion of $\boldsymbol{\Sigma}_0$ and the storage of its $O(p^2)$ entries, which both become computationally prohibitive for large p. We consider the simple specification $\boldsymbol{\Sigma}_0 = \alpha\boldsymbol{X}^T\boldsymbol{X}$, with a scaling factor $\alpha > 0$ and with $\boldsymbol{X}$ standardized to have zero-mean and unit-norm columns. We then build a sparse block estimate of $\boldsymbol{\Sigma}_0$ by observing that $\boldsymbol{X}^T\boldsymbol{X}$ is approximately banded and block diagonal, owing to the local nature of linkage disequilibrium structures. Unlike for the group sparsity model where groups are typically chosen to be small linkage disequilibrium blocks, the blocks used for estimating $\boldsymbol{\Sigma}_0$ should be as large as computationally feasible; they may therefore cover several dense linkage disequilibrium blocks. A second practical concern is on ensuring that our estimate of $\boldsymbol{\Sigma}_0$ (or equivalently each of its blocks) is positive definite. If this criterion is not met, we replace our estimate by the positive definite covariance matrix closest in Frobenius norm, using the algorithm of Higham (2002).

The data-dependent prior specification through $\boldsymbol{\Sigma}_0$ may raise concerns regarding using the data twice: the first time in describing the prior belief and the second time when updating the prior using the likelihood. Engelhardt and Adams (2014) acknowledge this and propose building an estimate of $\boldsymbol{\Sigma}_0$ using reference genome data from the same population as that of the study. A non-data-based alternative would be to use a conditional autoregressive matrix for $\boldsymbol{\Sigma}_0$ to enforce spatial similarity across regression coefficients for nearby locations. However, since the entire model is conditional on

$\boldsymbol{X}$, using data-dependent priors should not be a problem; another such prior is the g-prior (Zellner, 1986), to which we will return briefly in Section 4.2.3.

Our variational algorithm for model (4.3) relies on the data augmentation strategy first employed by Albert and Chib (1993) in the context of probit regression (and implicit used in (4.4)). We reparametrize the top level of our model by introducing the auxiliary variable z_{st},

$$\gamma_{st} = \mathbb{1}(z_{st} > 0), \qquad z_{st} \mid \theta_s \sim \mathcal{N}(\theta_s, 1), \qquad \boldsymbol{\theta} \sim \mathcal{N}_p(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0),$$

and use the factorization

$$\left\{ \prod_{t=1}^{q} \prod_{s=1}^{p} q(\beta_{st}, \gamma_{st}, z_{st}) \right\} q(\boldsymbol{\theta}),$$

from which we obtain

$$q(\beta_{st}, \gamma_{st}, z_{st}) = q(\beta_{st} \mid z_{st}) q(z_{st} \mid \gamma_{st}) q(\gamma_{st}),$$

given by

$$\begin{aligned} \beta_{st} \mid z_{st} > 0, \boldsymbol{y} &\sim \mathcal{N}\left(\mu_{\beta,st}, \sigma^2_{\beta,st}\right), \quad \beta_{st} \mid z_{st} \leq 0, \boldsymbol{y} \sim \delta_0, \\ z_{st} \mid \gamma_{st} = \delta, \boldsymbol{y} &\sim \mathcal{TN}\left(\mu_{\theta,s}, 1; \left\{0 < (-1)^{1-\delta} z_{st}\right\}\right), \\ \gamma_{st} \mid \boldsymbol{y} &\sim \text{Bernoulli}\left(\gamma^{(1)}_{st}\right), \end{aligned}$$

where $\mathcal{TN}\left(\mu, \sigma^2; \{a < x < b\}\right)$ denotes the truncated normal distribution, and $\mu_{\beta,st}$, $\sigma^2_{\beta,st}$, $\gamma^{(1)}_{st}$ and $\mu_{\theta,s}$ are variational parameters. Moreover, $\boldsymbol{\theta}$ (or each of its blocks) is approximated using a multivariate normal distribution, $\boldsymbol{\theta} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta\right)$, with variational parameters

$$\boldsymbol{\mu}_\theta = \boldsymbol{\Sigma}_\theta \left(\boldsymbol{Z}^{(1)} \mathbb{1}_q + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{m}_0\right), \qquad \boldsymbol{\Sigma}_\theta^{-1} = q \boldsymbol{I}_p + \boldsymbol{\Sigma}_0^{-1}, \tag{4.5}$$

where $\boldsymbol{Z}^{(1)}$ is the $p \times q$ matrix with entries

$$z^{(1)}_{st} = \gamma^{(1)}_{st} \left\{ \mathrm{M}(\mu_{\theta,s}, 1) - \mathrm{M}(\mu_{\theta,s}, 0) \right\} + \mu_{\theta,s} + \mathrm{M}(\mu_{\theta,s}, 0),$$

and

$$\mathrm{M}(u, \delta) = (-1)^{1-\delta} \frac{\varphi(u)}{\Phi(u)^\delta \left[1 - \Phi(u)\right]^{1-\delta}}, \qquad u \in \mathbb{R},\ \delta = 0, 1,$$

is the inverse Mills ratio (Mills, 1926).

4.2.3 Simulations

In this section, we describe a small numerical experiment to illustrate the type of posterior inferences obtained from the group and similarity sparsity models (4.1) and (4.3), and to compare them with that of our reference model (3.1)–(3.2)–(3.3) presented in Chapter 3. We simulated $p = 2{,}000$ SNPs, from which we randomly designated $p_\gamma = 10$ SNPs as "active", i.e., associated with at least one of $q = 100$ responses. The SNPs are autocorrelated in blocks of size 50, with correlation coefficients for the blocks drawn from a right-skewed Beta distribution. Figure 4.1 shows the empirical distribution of these coefficients, and ROC curves for assessing the selection of SNPs involved in associations under each of the models.

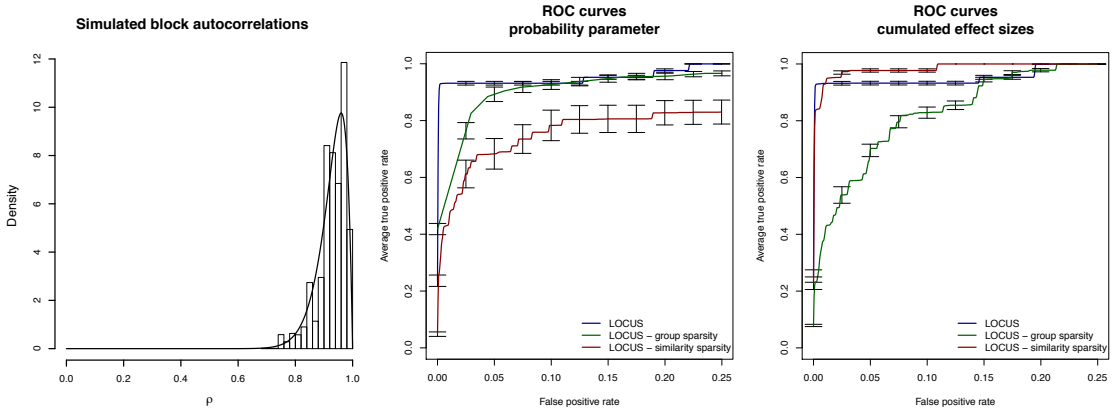

Figure 4.1 – Distributions of linkage disequilibrium block autocorrelations and selection performance using group and similarity sparsity models, for problems with $p = 2{,}000$ SNPs, autocorrelated by blocks of size 50, and $q = 100$ responses, for $n = 200$ samples. There are $p_\gamma = 10$ SNPs associated with at least one response, and each of them is associated with 3.25 responses, on average. The cumulated proportion of variance explained by the SNPs for a given response does not exceed 50%, and $q - q_\gamma = 75$ responses have no association with the SNPs. We simulated 200 datasets following the data generation design of Section 3.4.1. Left: histogram of the block autocorrelation coefficients simulated from a Beta(25,2) distribution (density also shown). Middle: truncated average ROC curves for predictor selection based on the probability parameters. Right: truncated average ROC curves for predictor selection based on the cumulated effect sizes.

In an attempt to treat the hyperparameter specification of all three models on an equal footing, we based it on the multiplicity control procedure described in Section 3.1.3 for both the reference model (3.1)–(3.2)–(3.3) and group sparsity model (4.1), and matched the first moment of the marginal distribution of γ_{st} under the similarity sparsity model (4.3) to that under the reference model; details are in Appendix B.3. For simplicity, we defined the groups for the group sparsity model to be the simulated linkage disequilibrium blocks, and also used this pattern to define a block empirical covariance matrix $\mathbf{\Sigma}_0$ for the similarity sparsity model; further investigation would be needed to assess the sensitivity of inferences to these choices.

Figure 4.2 displays two types of posterior summaries to quantify the support for each SNP to be involved in associations: the estimated probability parameters, at the top level of the model hierarchy, and the estimated effect sizes cumulated for all responses, closer to the data in the hierarchy. While these two summaries provide essentially the same information for the reference model, they are qualitatively very different for each of the structured sparsity models. In particular, for the similarity sparsity model, the dependence structure of SNPs is apparent in the probability estimates, but does not propagate to the cumulated effect sizes, which clearly discriminate the signal from the noise. For the group sparsity model, the probability parameter estimates quantify associations for groups of SNPs, and the within-group structure appears in the cumulated effect sizes, reflecting the dense-within-group assumption discussed in Section 4.2.1. These observations could have been deduced from the variational updates: the matrix $\mathbf{\Sigma}_0$ of the similarity sparsity model appears in the top-level updates (4.5), and the group empirical covariance used in the group sparsity model appears closer to the data, in the updates for β_{st}, see (4.2). As their effect size specification also involves the predictor covariance matrix, g-prior regressions should yield estimated effect sizes like those of the group sparsity model but somewhat sparser, as the $p \times 1$ indicator parameter $\boldsymbol{\gamma}_t$ selects predictors and not groups.

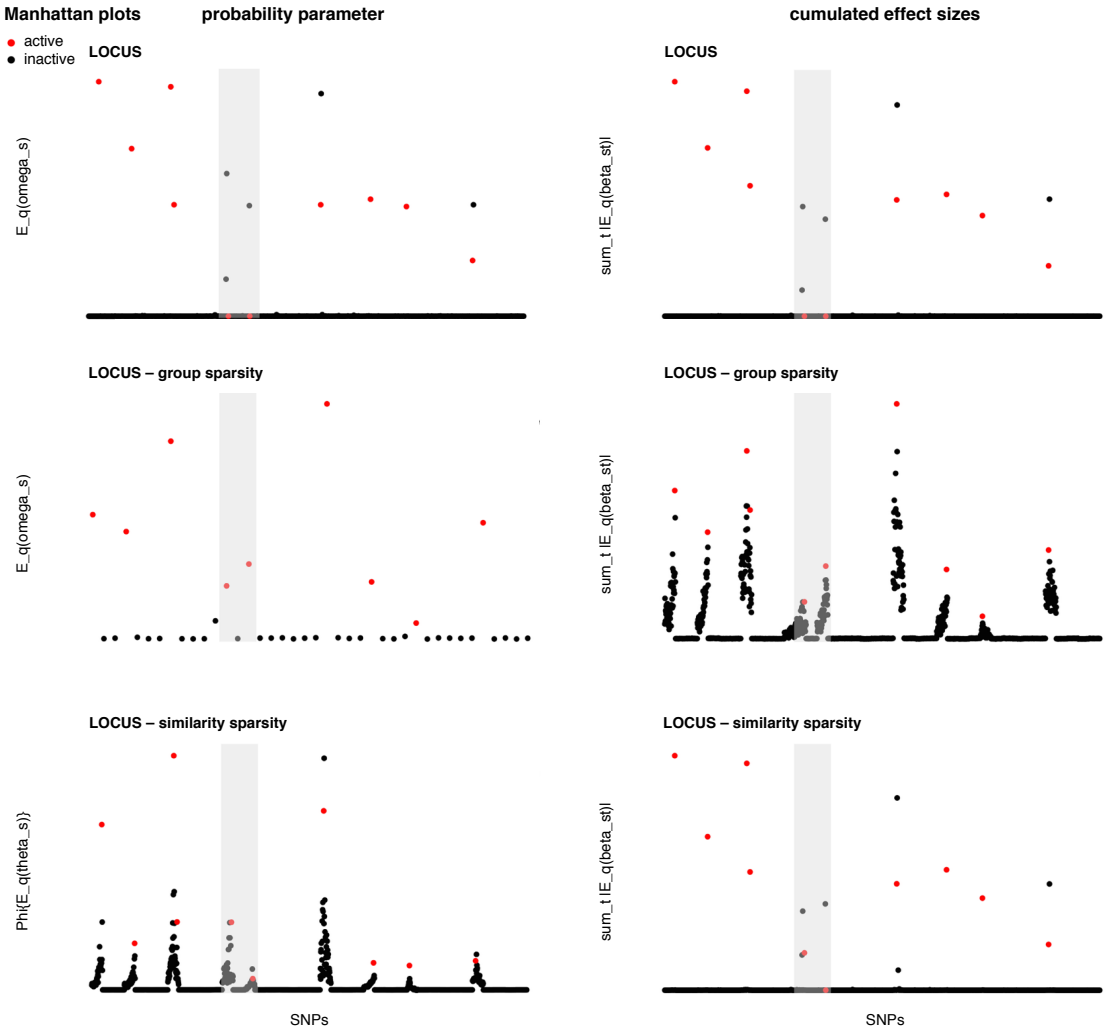

Figure 4.2 – Manhattan plots using group and similarity sparsity models. The data correspond to the first replicate used for Figure 4.1; see its caption. The SNPs with simulated associations are marked in red. Left column: evidence based on the inferred probability parameters $\omega_s^{(1)}$ for the reference model (3.1)–(3.2)–(3.3), $\omega_g^{(1)}$ for group sparsity model (4.1), and $\Phi(\mu_{\theta,s})$ for similarity sparsity model (4.3). Right column: evidence based on the cumulated effect sizes $\sum_{t=1}^{q} |\mu_{\beta,st}| \times \gamma_{st}^{(1)}$ for models (3.1)–(3.2)–(3.3) and (4.3), and $\sum_{t=1}^{q} |\mu_{\beta,st}| \times \gamma_{gt}^{(1)}$, $g \ni s$, for model (4.1).

The Manhattan plots exhibiting dependence structure indicate that even when the active SNP in a block is not attributed the highest evidence, it appears as a potential candidate among its correlated neighbours, since its posterior summary is inflated. This output is similar to that of marginal screening, and is qualitatively appealing in weakly informative and highly structured data, where the reference model may be more likely to wrongly choose one of several correlated SNPs. The shaded areas show how this encoded structure may also help to trigger the activation of signals: two active SNPs are not picked up by the reference model, but receive some support under the similarity and group sparsity models. Moreover, even though most groups involve a single active SNP out of fifty, the group-model posterior estimates of ω_g identify the correct groups.

While Figure 4.2 shows a satisfactory recovery of signals for all three models, the ROC curves of Figure 4.1, based on 200 replicates, indicate that predictor selection performance depends on the chosen posterior summary. It is worse when the summaries are inflated as a result of correlation, as this produces false positive signals; recall the motivating example of Chapter 1. When using the cumulated effect sizes for selection, the similarity sparsity model slightly outperforms the reference model, as a result of activations being triggered, as explained in the previous paragraph.

More experiments would be needed to fully assess the potential of our structured models and evaluate the sensitivity of selection to the choice of grouping or posterior summary. In particular, the ROC curves of Figure 4.1 are based on 10 active predictors only, on given effect strengths, predictor autocorrelation levels and distribution of effects, and we suspect that the comparison of models and posterior summaries will depend on these data-generation settings. Finally, while encoding spatial dependence structures may help to combat artefacts caused by multicollinearity, the dependence of effect sizes may not mimic that of the SNPs. A seminal example pertains to the *FTO* gene, for which associations with body mass have been reported, and to genetic variants lying within introns of *FTO*, which have been linked to an increased risk for obesity; despite the proximity between the gene and the risk variants, recent evidence suggests that the variants influence human adipocyte functions via the distal *IRX3* gene (a half-megabase downstream of *FTO*), rather than via *FTO* (Smemo et al., 2014). To explicitly account for such long-range mechanisms, it may be sensible to build the groups or the covariance graph $\boldsymbol{\Sigma}_0$ by also incorporating other types of SNP similarity information. Such information could be based on the involvement of SNPs in common biological pathways, on chromatin interaction using Hi-C data, or on any other prior knowledge suggesting similar regulatory properties.

4.3 Variational inference for multimodal problems

4.3.1 Simulated annealing variational inference

Another approach to better handling data dependence structures is to enhance the inference strategy, leaving the model untouched. To this end, it is important to understand the weaknesses of the variational algorithm presented in Section 3.2 when applied to highly-correlated data.

Consider a bivariate posterior distribution $p(\boldsymbol{\nu} \mid \boldsymbol{y})$, $\boldsymbol{\nu} = (\nu_1, \nu_2) \in \mathbb{R}^2$, and a mean-field variational approximation to it,

$$q(\boldsymbol{\nu}) = q(\nu_1)\, q(\nu_2). \tag{4.6}$$

Under the posterior independence assumptions entailed by (4.6), the covariance structure of $q(\boldsymbol{\nu})$ is decoupled by construction, and the marginal variances are smaller than those of $p(\boldsymbol{\nu} \mid \boldsymbol{y})$ as a result of optimizing the *reverse* Kullback–Leibler divergence,

$$\mathrm{KL}\big(q \,\|\, p\big) = -\int q(\boldsymbol{\nu}) \log\left\{\frac{p(\boldsymbol{\nu} \mid \boldsymbol{y})}{q(\boldsymbol{\nu})}\right\} \mathrm{d}\boldsymbol{\nu},$$

which penalizes putting mass in regions of $q(\cdot)$ where $p(\cdot)$ has little mass; recall Section 2.4.2. This underestimation of posterior variances clearly generalizes to mean-field approximations for parameter vectors of any dimension, and happens to affect the selection performance of our algorithm in presence of strong dependence. The variational objective function,

$$\mathcal{L}(q) = \mathrm{E}_q \log p(\boldsymbol{y}, \boldsymbol{\nu}) - \mathrm{E}_q \log q(\boldsymbol{\nu}), \tag{4.7}$$

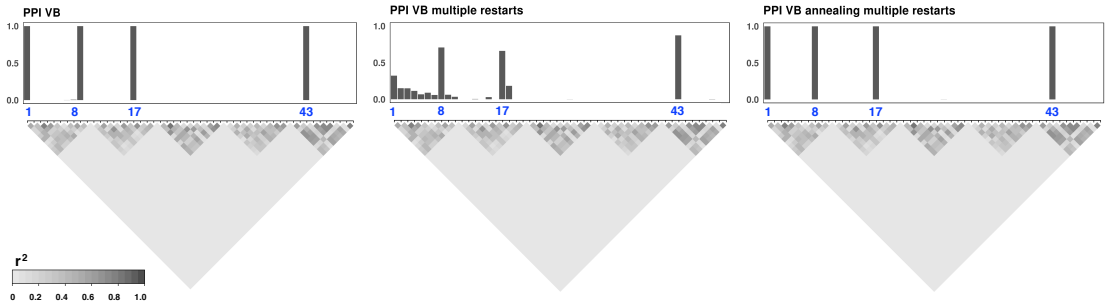

Figure 4.3 – Variable selection under high multicollinearity. Problem with a single response and 1,000 genetic variants (SNPs) autocorrelated by blocks as candidate predictors (first 50 shown). The SNPs simulated as associated with the response explain 30% of its variance; their positions are marked by the blue labels. The bars show the marginal posterior probabilities of inclusion produced by the variational algorithm of Chapter 3, one run (left) and average of 1,000 runs with different starting values (middle), and by the annealed variational algorithm with initial temperature $T = 5$ and grid of 100 temperatures, average of 1,000 runs with different starting values (right).

for our model $p(\boldsymbol{y}, \boldsymbol{v})$ and approximation $q(\boldsymbol{v})$ tends to be highly multimodal, which exposes our coordinate ascent algorithm to entrapment in local modes, corresponding to suboptimal configurations of variables. This risk is increased by the posterior variance underestimation, which encourages the approximation to concentrate most of its mass on a single hypothesis. Figure 4.3 considers a problem with blocks of highly correlated SNPs; the algorithm completely misses one active SNP (SNP 8) and instead picks one of its correlated neighbours with high confidence (SNP 9). It also shows that averaging posterior probabilities across multiple runs with different starting values can produce "diluted" posterior summaries that better reflect the uncertainty of SNP selection in regions of high linkage disequilibrium. Such averaging strategies aiming for more stable and accurate estimates have been proposed in different contexts; two examples are the bootstrap aggregating (bagging, Breiman, 1996) and M-posterior (Minsker et al., 2017) algorithms. But although they mitigate the problem, they typically result in increased computational costs, which can become quite substantial for typical molecular QTL problems. We instead augment our algorithm with a simulated annealing procedure that directly targets improving the exploration of multimodal posteriors.

The central idea of simulated annealing is essentially the same as that of simulated tempering for MCMC algorithms (recall Section 2.3). It consists in introducing a so-called *temperature* parameter T which indexes a series of *heated* distributions,

$$p_T(\boldsymbol{y}, \boldsymbol{v}) \propto p(\boldsymbol{y}, \boldsymbol{v})^{1/T},$$

and controls the degree of separation of their modes. The procedure starts with large temperatures that flatten the density of interest, thereby sweeping most of its local modes away and facilitating the search for the global optimum. Temperatures are then progressively decreased until the *cold* distribution, corresponding to the original multimodal distribution, is reached.

Like variational inference, simulated annealing was first used in statistical physics. Its name and heuristic come from metallurgy, where annealing consists in heating a material and cooling it down in a controlled fashion to limit its defects and achieve certain crystallization properties (Brunger and Rice, 1997). Optimization via simulated annealing was first described in Metropolis et al. (1953) and Kirkpatrick et al. (1983) for Metropolis algorithms, and was then adapted for expectation-maximization

by Ueda and Nakano (1998) and for variational inference by Katahira et al. (2008). Variational inference lends itself to simulated annealing principles. Indeed, recall that the objective function (4.7) entails a tradeoff between the first term of its sum, the expected log joint distribution, which encourages the approximation to put mass on configurations of the variables that best explain the data, and the second term, the entropy, which prefers the approximation to be more dispersed. Annealing inflates the entropy term by multiplying it by the temperature parameter,

$$\mathcal{L}_T(q) = \int q_T(\boldsymbol{v}) \log p(\boldsymbol{y}, \boldsymbol{v}) \mathrm{d}\boldsymbol{v} - T \int q_T(\boldsymbol{v}) \log q_T(\boldsymbol{v}) \mathrm{d}\boldsymbol{v}, \qquad T \geq 1, \tag{4.8}$$

where $q_T(\cdot)$ is the heated variational distribution; it penalizes the first term (when $T > 1$) and gradually relaxes this penalty until the original variational algorithm is obtained (when $T = 1$).

The variational updates for each heated variational distribution factor $q_T(v_j)$ are obtained as for the vanilla mean-field algorithm. We rewrite (4.8) with respect to v_j as

$$\begin{aligned}
\mathcal{L}_T(q) &= \mathrm{E}_j \left[\mathrm{E}_{-j} \left\{ \log p(\boldsymbol{v}, \boldsymbol{y}) \right\} - T \log q_T(v_j) \right] + \text{cst} \\
&= \mathrm{E}_j \left[\log \left\{ \frac{\exp \left\{ \mathrm{E}_{-j} \log p(\boldsymbol{v}, \boldsymbol{y}) \right\}}{q_T(v_j)^T} \right\} \right] + \text{cst} \\
&= T \, \mathrm{E}_j \left[\log \left\{ \frac{p_{T,-j}(v_j, \boldsymbol{y})}{q_T(v_j)} \right\} \right] + \text{cst},
\end{aligned}$$

where we introduced the distribution $p_{T,-j}(v_j, \boldsymbol{y}) \propto \exp \left\{ T^{-1} \mathrm{E}_{-j} \log p(\boldsymbol{v}, \boldsymbol{y}) \right\}$, and where $\mathrm{E}_j(\cdot)$ denotes expectation with respect to the distribution $q_T(v_j)$, $\mathrm{E}_{-j}(\cdot)$, the expectation with respect to the distributions $q_T(v_k)$, for all the variables v_k, $(k \neq j)$, and cst does not depend on v_j. The expectation in (4.9) corresponds to the negative Kullback–Leibler divergence between $q_T(v_j)$ and $p_{T,-j}(v_j, \boldsymbol{y})$; $\mathcal{L}_T(q)$ is therefore maximal when $q_T(v_j) = p_{T,-j}(v_j, \boldsymbol{y})$, i.e., when

$$\log q_T(v_j) = T^{-1} \mathrm{E}_{-j} \{ \log p(\boldsymbol{y}, \boldsymbol{v}) \} + \text{cst}, \qquad j = 1, \ldots, J.$$

Our annealed variational algorithm for the reference model (3.1)–(3.2)–(3.3) is given in Appendix B.4.

There is no consensus on the type of temperature schedule to use. In the numerical experiments of Section 4.3.2, we will compare three types of schedules proposed by Gramacy et al. (2010) on the inverse temperature scale, namely, a linear schedule,

$$T_l^{-1} = T_L^{-1} + \Delta(L - l), \qquad \Delta = \frac{1 - T_L^{-1}}{L - 1},$$

a harmonic schedule,

$$T_l = 1 + \Delta(l - 1), \qquad \Delta = \frac{T_L - 1}{L - 1},$$

and a geometric schedule

$$T_l = (1 + \Delta)^{l-1}, \qquad \Delta = T_L^{1/(L-1)} - 1,$$

where $l = L, \ldots, 1$, and T_L is the hottest temperature. We decrease the temperature at each iteration, so L corresponds to the number of iterations required by the annealing scheme, after which the classical

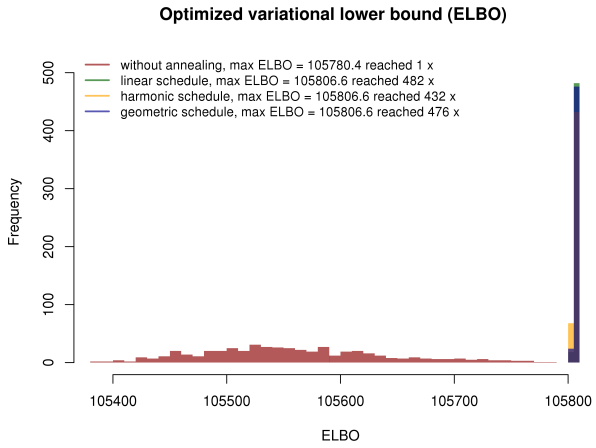

Figure 4.4 – Histograms of optimized variational lower bound values obtained by classical and annealed variational inference using linear, harmonic and geometric schedules. The initial temperature is set to 5 and the temperature grid size is 100. Each algorithm was run 500 times using different starting points, on the first simulated dataset. The three histograms corresponding to annealed variational inference overlap.

variational algorithm is run until convergence. As $L = 10 - 100$ typically, the computational overhead is limited.

A final purpose of Figure 4.3 is to illustrate the benefits of annealed variational inference over classical variational inference: while selection based on the former may suffer from poorly chosen starting values, selection based on the latter consistently identifies the relevant SNPs across 1,000 restarts. The next section provides more empirical support for our annealed variational algorithm.

4.3.2 Simulations

The following numerical experiments borrow simulated data from Bottolo et al. (2011, dataset "SIM3"). The data consist of $p = 498$ SNPs for $n = 120$ individuals from the Yoruba population, obtained from the HapMap project (Consortium et al., 2005). The SNPs span 500 kilobases on chromosome 7 and are in high linkage disequilibrium; their dependence structure and the positions of the "active" SNPs, i.e., with simulated associations, are displayed in Figure 4.5. The six active SNPs are hotspots of different sizes; they are associated with 4, 10 or 26 responses. The total number of responses is $q = 1,000$, of which 950 were generated from Gaussian noise. The dataset comprises 25 replicates of simulated responses and effects sizes.

We first examine sensitivity to the choice of schedule (linear, harmonic or geometric) by comparing the values of the optimized variational lower bound on the marginal log-likelihood using the data from the first replicate and performing 500 runs, using different starting values. We also benchmark these values against those obtained by classical variational inference. For the annealed algorithms, we use an initial temperature of 5 and a grid of 100 temperatures; all algorithms are based on the reference model (3.1)–(3.2)–(3.3) discussed in Chapter 3. Figure 4.4 indicates that the optimized variational lower bounds obtained by annealed variational inference are consistently higher and less variable than under vanilla variational inference. It also shows little differences across schedules, with the maximum value

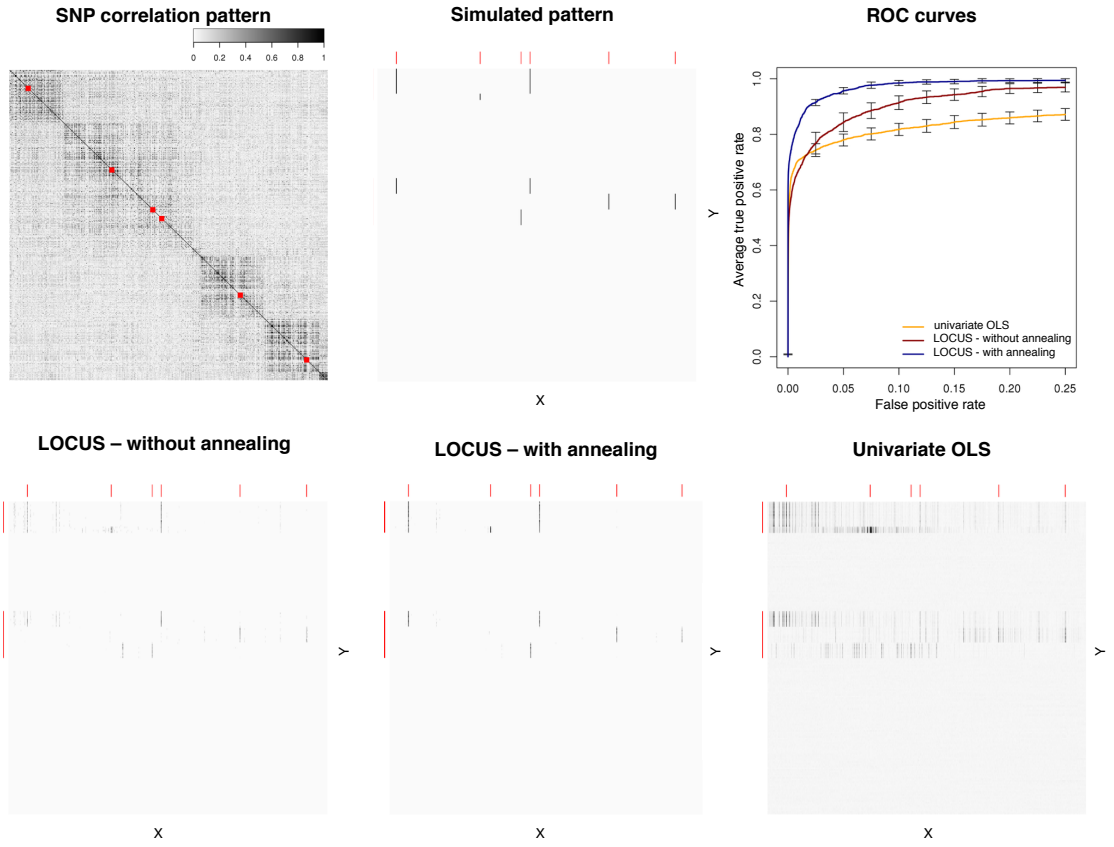

Figure 4.5 – Selection performance by classical and annealed variational inference (LOCUS method of Chapter 3) and univariate screening based on 25 simulated datasets. The annealing uses a geometric schedule with initial temperature 5 and grid size 100. Top, left: correlation pattern of the SNPs (Yoruba population HapMap, ENm014 region, chromosome 7, Consortium et al., 2005), with location of SNPs with simulated associations; reproduced from Bottolo et al. (2011). Top, middle: simulated association pattern (Bottolo et al., 2011, dataset "SIM3"). Top right: truncated average ROC curves based on marginal posterior probabilities of inclusion for the variational algorithms and p-values for the univariate method. Bottom: association patterns uncovered for the first 200 responses (the remaining 800 are not involved in associations). Left: average marginal posterior inclusion probabilities by classical variational inference. Middle: average marginal posterior inclusion probabilities by annealed variational inference. Right: average $-\log_{10}$ p-values by univariate screening. The red marks indicate the rows and columns containing simulated signals.

of $\mathcal{L}(q) = 105,806.6$ reached in most of the runs under the three schedules: this value is obtained $\approx 86\%$ of times for the harmonic schedule, which appears to be slightly inferior to the geometric and linear schedules, which reach it ≈ 95 and 96% of times. The three schedules also result in equivalent variable selection performance (their ROC curves overlap, not shown). We follow Kirkpatrick et al. (1983) and use the geometric schedule for the rest of our experiments.

We next compare our classical and annealed variational inference algorithms on all 25 replicates, also using a plain univariate screening for reference. Figure 4.5 shows that annealing substantially improves signal recovery. The association pattern uncovered is close to that simulated, whereas that obtained with no annealing presents spurious marks resulting from entrapments in the multiple local modes caused by the linkage disequilibrium. As expected, the pattern produced by univariate screening is

riddled with false positives. The ROC curves translate these observations into selection performance. Additional simulations showing the benefits of annealing for our variational schemes will be presented in Section 5.4.4 for a model tailored to the detection of pleiotropic effects in very large response settings.

4.3.3 Towards an adaptive temperature schedule?

The annealing schedule entails several presets (schedule type, initial and number of temperatures) whose choices are left to the practitioner. Our simulations of Section 4.3.2 suggested that inferences may not be sensitive to the schedule type, and further experimentation indicated that initial temperatures between 2 and 20, and grids of 10 to 100 temperatures are sufficient for good exploration. However, these choices have no theoretical basis and optimal set-ups may be model-dependent. Intuitively, if the number of temperatures is small or the initial temperature is very low, the procedure may become inefficient in avoiding local optima, whereas if the number of temperatures is large, the number of iterations before convergence may be unnecessarily large. In Section 5.3, we will discuss a ill-behaved scenario, where poor schedule choices can be particularly damaging.

A natural solution would be to develop a procedure that embeds the temperature as an auxiliary parameter to be inferred. This would permit adaptive and dynamic control of the temperature schedule and may help to balance the number of temperatures used, and hence the resource usage, with the level of entropy needed for good exploration. Mandt et al. (2016) have described a procedure based on this idea: they assign a discrete uniform prior over the temperature assignments on the grid and estimate them jointly with the model parameters using a mean-field variational formulation. An important drawback of their proposal is that it requires the precomputation of an approximation to the joint distribution normalizing constant

$$\left\{ \int \int p(\boldsymbol{y}, \boldsymbol{\nu})^c \mathrm{d}\boldsymbol{\nu} \mathrm{d}\boldsymbol{y} \right\}^{-1}, \tag{4.9}$$

where $c = 1/T$ is the inverse temperature. Cheap estimates may be envisioned for *large n* cases, but are unrealistic for high-dimensional models.

An alternative to computing (4.9) may be to couple the variational algorithm with a Metropolis–Hastings rule for sampling the temperatures. Inspiration may come from simulated tempering, and in particular from the adaptive procedure of Geyer and Thompson (1995). Writing the unnormalized densities

$$h_l(\boldsymbol{\nu}) = p(\boldsymbol{y}, \boldsymbol{\nu})^{c_l}, \qquad l = 1, \ldots, L,$$

they assume having, for each inverse temperature $c_l = 1/T_l$, a procedure for updating $\boldsymbol{\nu}$, and couple it with a Metropolis–Hastings step that implements transitions between consecutive temperatures of the grid. Their procedure for sampling $(\boldsymbol{\nu}, c_l)$ has stationary distribution proportional to

$$h_l(\boldsymbol{\nu})\pi(c_l), \tag{4.10}$$

for some auxiliary numbers $\pi(c_l)$, $l = 1, \ldots, L$. They call $\pi(\cdot)$ an (unnormalized) *pseudo-prior* as (4.10) resembles the product of a likelihood and a prior, and choose it to obtain a uniform marginal distribution for the assignments, that is,

$$p(c_l) \propto \pi(c_l) \int h_l(\boldsymbol{\nu}) \mathrm{d}\boldsymbol{\nu} = 1,$$

giving

$$\pi(c_l) = \left\{ \int h_l(\boldsymbol{\nu}) \mathrm{d}\boldsymbol{\nu} \right\}^{-1}, \qquad l = 1, \ldots, L; \tag{4.11}$$

by doing so, the sampler will spend roughly equal time sampling each heated distribution and no temperature will be visited too infrequently. The algorithm then chooses $k = l \pm 1$ with equal probability and accepts the transition from c_l to c_k, using the Metropolis rule $\min(1, r)$ where r is the Hastings ratio

$$r = \frac{h_k(\boldsymbol{\nu})}{h_l(\boldsymbol{\nu})} \frac{\pi(c_k)}{\pi(c_l)}. \tag{4.12}$$

The pseudo-prior (4.11) may be more amenable to rough estimation than (4.9); Geyer and Thompson (1995) propose several estimation strategies for (4.9) based on preliminary runs, such as an iterative adjustment method or a stochastic approximation that updates the computation as the chain progresses.

Instead of directly estimating the $\pi(c_l)$, we may exploit the fact that they only appear in ratios for consecutive temperatures, $\pi(c_k)/\pi(c_l)$, $k = l \pm 1$, in Geyer and Thompson (1995)'s Metropolis–Hastings rule, and implement an *annealed importance sampling* approximation (Neal, 2001). For instance, with $k = l - 1$, we have

$$\begin{aligned} \frac{\pi(c_l)}{\pi(c_{l-1})} &= \pi(c_l) \int \frac{h_{l-1}(\boldsymbol{\nu})}{h_l(\boldsymbol{\nu})} h_l(\boldsymbol{\nu}) \mathrm{d}\boldsymbol{\nu} \\ &= \int p(\boldsymbol{y}, \boldsymbol{\nu})^{(c_{l-1} - c_l)} p_{c_l}(\boldsymbol{\nu} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\nu}, \end{aligned} \tag{4.13}$$

since $p_{c_l}(\boldsymbol{\nu} \mid \boldsymbol{y}) = \pi(c_l) h_l(\boldsymbol{\nu})$ is the posterior of the parameter vector under temperature $1/c_l$.

For a fine grid of temperatures, the difference $c_{l-1} - c_l > 0$ is small, so the first term in the integrand of (4.13) is approximately constant. In our variational inference settings, we may thus obtain good importance sampling estimates by approximating $p_{c_l}(\boldsymbol{\nu} \mid \boldsymbol{y})$ with the variational distribution $q_{c_l}(\boldsymbol{\nu})$ under the current inverse temperature c_l, and drawing from the latter. This readily provides importance distributions whose importance weights have small variance; these distributions are otherwise difficult to specify in high dimensions. Estimating ratios (4.13) might be done during the course of the algorithm.

The above ingredients may serve to develop an adaptive annealed variational algorithm for our model, but a number of important concerns still need to be addressed. For instance, we need to be able to control the error introduced by the importance sampling approximation in the Metropolis–Hastings step. Some error may be tolerated in the temperature chain, as it would merely lead to a suboptimal schedule choice. A greater concern is on how to embed the Metropolis–Hastings procedure in the variational algorithm, without violating the chain's convergence properties. The algorithm of Geyer and Thompson (1995) involves constructing a Markov chain for $\boldsymbol{\nu}$ having stationary distribution proportional to $p(\boldsymbol{y}, \boldsymbol{\nu})^{c_l}$, and using samples from this chain to update the Hastings ratio (4.12) for c_l. Using samples from the variational distribution $q_{c_l}(\boldsymbol{\nu})$ instead may be interpreted as using a modified Metropolis–Hastings rule for $\boldsymbol{\nu}$ with very good proposal (the variational distribution) and acceptance ratio forced to unity. While there is hope that the entire procedure may be valid, deriving the theory to prove this may be hard, as reflected by the absence of strong research results on hybrid variational and MCMC algorithms (recall Section 2.3). Moreover, sampling from the variational density may generate important additional computational costs.

4.4 Summary

In this chapter, we tackled tailoring our approach to data with marked dependence. We first presented two attempts to explicitly leverage known spatial structure among the predictors, in order to enhance selection performance and interpretability of posterior quantities in highly-correlated data scenarios. We saw that the hierarchical probit model on the probability of associations provides a flexible avenue to incorporate such structure, yet it is not limited to this use case. For instance, Quintana and Conti (2013) employed this formulation to introduce a second-stage regression model to exploit external information on the predictors. We will see other extensions in Chapters 5 and 6.

Several other approaches exist: for instance, Li and Zhang (2010) propose modelling structured predictors using an Ising model in which they inject prior information on pairwise predictor similarity. Stingo et al. (2011) describe a spike-and-slab model to encode pathway membership or other network information by specifying a second-stage logistic model on the probability of inclusion. Kwon et al. (2011) embed structural information at the algorithmic level; they use information about the dependence of the predictors to accept or reject moves in a Metropolis–Hastings algorithm.

In the second part of the chapter, we proposed improving exploration multimodal spaces without resorting to any prior structural belief. We described a simulated annealing algorithm which readily extends our original variational algorithm, at little added computational cost.

Genome-wide association studies always require follow-up work to pinpoint actual causal variants; this diminishes the practical relevance of identifying the most promising candidate SNPs inside linkage disequilibrium blocks. From this perspective, the group and similarity structure models, which provide association evidence at the level of linkage disequilibrium blocks, could suffice. However, it is always desirable to narrow down the range of candidate functional variants, as this may save substantial investment in prospective research. Our experiments suggest that annealed variational inference can achieve this, as it produces improved selection over classical variational inference in highly-correlated settings. Moreover, it can robustly and agnostically handle any structure in the data, and is applicable to any model and mean-field approximation. For all these reasons, we will use annealed variational inference implementations for the rest of the thesis.

5 A global-local approach to modelling hotspots

In this chapter we address a parameter sensitivity issue that can be especially damaging in large response settings. The sensitivity concerns a top-level variance parameter controlling the propensity of predictors to be hotspots, and was alluded in Section 3.1.3, when discussing hyperparameter choices for ω_s. We develop a solution based on a second-stage continuous shrinkage model that allows automatic discrimination of hotspots. Our proposal entails both global and local scale parameters to shrink noise globally and hence flexibly accommodate the highly sparse nature of genetic analyses, while being robust to individual signals, thus leaving the effects of hotspots unshrunk. Even though all previous chapters already tackle efficient modelling hotspot genetic variants, whose *trans*-regulatory effects are difficult to uncover, the work presented in this chapter tailors inference to very large response settings, in which the detection of hotspots is a prominent task. The discussion focuses on modelling aspects but numerical experiments show that simulated annealing (adapted from Chapter 4 for the present model) is critical for efficient selection and size estimation of hotspots.

The chapter is organized as follows. Section 5.1 states the problem in light of Jia and Xu (2007), Richardson et al. (2010) and our proposal of Chapter 3, and formalizes its consequences for sensitivity and multiplicity control. Section 5.2 presents our modelling framework and discusses its properties, and Section 5.3 comments on a detail of inference. Section 5.4 assesses the performance of our approach in simulations, and Section 5.5 applies it to real eQTL data.

The work presented below has been developed with Leonardo Bottolo and Sylvia Richardson, and has been submitted for publication, with minor changes, in Ruffieux et al. (2018a); Section 5.3 is an addition.

5.1 Problem statement

Recall our general modelling framework (3.1)–(3.2): consider a series of hierarchically related regressions, with q centered responses, $\boldsymbol{y} = \left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q\right)$, and p centered candidate predictors, $\boldsymbol{X} = \left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p\right)$, for n samples ($n \ll p$),

$$\begin{aligned} \boldsymbol{y}_t \quad &| \quad \boldsymbol{\beta}_t, \tau_t \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}_t, \tau_t^{-1}\boldsymbol{I}_n\right), && t = 1, \ldots, q, \\ \beta_{st} \quad &| \quad \gamma_{st}, \sigma^2, \tau_t \sim \gamma_{st}\mathcal{N}\left(0, \sigma^2\tau_t^{-1}\right) + (1-\gamma_{st})\,\delta_0, && s = 1, \ldots, p, \\ \gamma_{st} \quad &| \quad \omega_{st} \sim \text{Bernoulli}\left(\omega_{st}\right), \end{aligned} \tag{5.1}$$

where δ_0 is the Dirac distribution, and where τ_t and σ^{-2} are assigned Gamma priors.

In Section 3.1.2 we saw that the proposals of Jia and Xu (2007) and Richardson et al. (2010) involve variants of (5.1) and that the primary difference between these models is in the prior specification for the probability of association parameter, ω_{st}. Richardson et al. (2010) decouple the predictor and response effects by setting $\omega_{st} = \omega_s \times \omega_t$, and place prior distributions on each of ω_s and ω_t, whereas Jia and Xu (2007) and our model of Chapter 3 are based on the simpler formulation $\omega_{st} \equiv \omega_s$. A suitable specification of the predictor-specific parameter ω_s is crucial, as ω_s controls the propensity of each predictor $\boldsymbol{X}_s$ to be a hotspot, i.e., to be simultaneously associated with several responses. As we now explain, the discrimination of hotspots can be very sensitive to the choice of prior distribution for ω_s and this sensitivity becomes particularly severe in very large response settings, where the detection of hotspots is a key task.

For the sake of discussion, we illustrate our point with formulation (3.3), whereby

$$\omega_{st} \equiv \omega_s \overset{\text{iid}}{\sim} \text{Beta}(a,b), \qquad a,b>0, \qquad s=1,\ldots,p, \tag{5.2}$$

assuming exchangeability; the same considerations apply to the models of Jia and Xu (2007) and Richardson et al. (2010). We discuss the choice of the hyperparameters a and b through the prior expectation and variance for ω_s. The expectation corresponds to the prior base rate of associated pairs, $\mu_\omega = \text{E}(\omega_s) = \text{pr}(\gamma_{st}=1)$. Its value should be small to induce sparsity, typically $\mu_\omega \ll 1$ for $p,q \gg n$, and may be fixed using an estimate of the overall signal sparsity. In contrast, there is no prior state of knowledge about $\sigma^2_\omega = \text{Var}(\omega_s)$, and its choice turns out to impact the prior size of hotspots when q is large. To formalize this, it is helpful to study prior odds ratios, as for the discussion on predictor multiplicity adjustment in Section 3.1.3. Recall that, for a given predictor $\boldsymbol{X}_s$, and a model $\mathcal{M}_{q_s}$ in which $\boldsymbol{X}_s$ is associated with $1 \leq q_s \leq q$ responses, the prior odds ratio

$$\text{POR}(q_s-1:q_s) = \frac{\text{pr}\left(\mathcal{M}_{q_s-1}\right)}{\text{pr}\left(\mathcal{M}_{q_s}\right)} = \frac{b+q-q_s}{a+q_s-1}, \tag{5.3}$$

quantifies the penalty induced by the prior when moving from q_s-1 to q_s responses associated with $\boldsymbol{X}_s$. The penalty increases with the total number of responses in the model (for fixed a, b and q_s), but it also decreases monotonically as q_s increases, so that it is *a priori* easier to add a response when $\boldsymbol{X}_s$ is already associated with many responses. More insight into this phenomenon can be obtained by looking at the quantity

$$\frac{\text{POR}(0:1)}{\text{POR}(q_s-1:q_s)}, \tag{5.4}$$

which compares the cost of adding a further response association with $\boldsymbol{X}_s$ when moving from the null model or from a model with q_s-1 associations already.

In molecular QTL problems, q_s is typically much smaller than q, as each SNP is believed to control just a few molecular entities. For $q_s \ll q$, (5.4) behaves roughly linearly in q_s with slope $\approx a^{-1} = \sigma^2_\omega \left\{\mu^2_\omega(1-\mu_\omega) - \mu_\omega \sigma^2_\omega\right\}^{-1}$. Hence, large σ^2_ω favours large hotspots while small σ^2_ω tends to give an association pattern that is more scattered across predictors. In the latter case, strong shrinkage towards $\mu_\omega \ll 1$ may be induced and the resulting hotspot sizes may be underestimated, whereas, in the former case artifactual hotspots may appear when data are insufficiently informative to dominate the prior specification. Table 5.1 shows that the penalties (5.4) can differ drastically for different choices of σ^2_ω.

σ^2_ω \ q_s	5	10	50	100
10^{-4}	1.0	1.1	1.5	2.1
10^{-3}	1.4	2.0	6.5	12.2
10^{-2}	6.0	12.3	62.4	125.4

Table 5.1 – Ratios (5.4) for a grid of variances σ^2_ω and numbers of associated responses q_s. The total number of responses is $q = 20,000$ and the base rate is $\mu_\omega = 0.1$. The penalty varies greatly depending on the chosen value for σ^2_ω and increases roughly linearly with q_s.

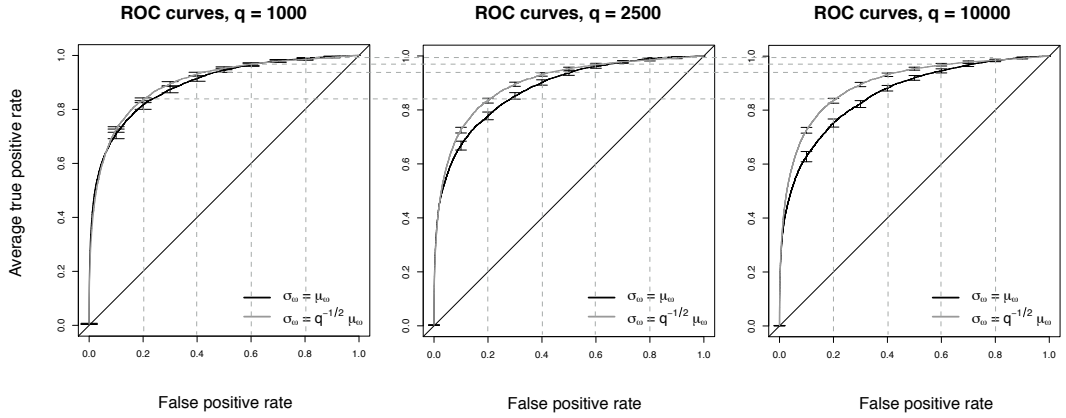

Figure 5.1 – Variable selection performance with and without multiplicity adjustment, measured by average ROC curves with 95% confidence intervals obtained from 100 replicates. Three problems are simulated, with an increasing number of response variables, $q = 1,000$ (left), $q = 2,500$ (middle), $q = 10,000$ (right), and $p = 100$ candidate predictors for $n = 100$ samples. The pattern of associations is the same for all three scenarios: 50 responses are chosen randomly among the first $1,000$ responses to be associated with at least one of 10 predictors; the rest of the responses are drawn from Gaussian noise. For a given response, the proportion of its variance explained by the predictors doesn't exceed 15%. Two implementations of model (5.1)–(5.2) are compared: one uses a fixed choice of variance $\sigma^2_\omega = \mu^2_\omega$ (black curves); its performance deteriorates as q increases, from left to right. The other uses the proposed adjustment for the total number of responses q, i.e., $\sigma^2_\omega = q^{-1}\mu^2_\omega$ (grey curves); its performance remains unchanged as q increases (see grid). The base rate is fixed to the simulated proportion of associated predictors, i.e., $\mu_\omega = 0.1$.

To evaluate the extent to which this could impact inference in flat likelihood scenarios, it is helpful to also study the case where q_s is of order q, even though this is unlikely to be encountered in our applications. When $q_s \sim q$ (i.e., when q_s/q tends to a strictly positive constant as $q \to \infty$), (5.4) is of order $O(q)$, so that, in weakly informative data settings, the sensitivity may lead to the manifestation of massive spurious hotspots associated with nearly all responses. Such undesired "pile-up" effects highlight the need to adjust for the dimensionality of the response.

The sensitivity of inferences to the hotspot propensity variance relates to the well-known issue of specifying prior distributions for variance components, as ω_s can be viewed as a random effect. While this sensitivity and its related response multiplicity burden are important problems that affect any hierarchically related regression model such as (5.1), they have been neither formalized nor investigated in the literature. In fact, the number of responses presented in numerical experiments is usually rather small (10–1,000), mainly limited by the heavy computational load of MCMC sampling, so that this sensitivity issue may go unnoticed. Another aspect is that "pile-up" effects can be avoided by choosing a small hotspot propensity variance, at the risk of giving up substantial hotspot selection performance.

The very sparse nature of molecular QTL analyses also rules out the use of simple empirical Bayes estimates, which typically collapse to the degenerate case $\hat{\sigma}^2_\omega = 0$, see, e.g., van de Wiel et al. (2018). Thus, a tailored solution is needed.

Our proposal resolves the above issues, based on two considerations. First, we argue that "pile-up" effects can be prevented by suitably linking the hotspot propensity variance to the number of responses, in effect performing multiplicity adjustment. Indeed, choosing $\sigma^2_\omega = O\left(q^{-1}\right)$, ratio (5.4) is $O(1)$ when $q_s \sim q$. For small values of μ_ω, typically chosen in sparse association problems, this adjustment amounts to enforcing small and similar numbers of response associations for all predictors, with the degree of shrinkage depending on the number of responses q (in the limiting case $q \to \infty$, we obtain $\omega_s \equiv \mu_\omega$). Figure 5.1 illustrates the degradation of the variable selection performance in moderately informative problems with increasing q, and shows how the proposed penality addresses the issue.

Second, we embed and *relax* this multiplicity adjustment in a fully Bayesian framework involving a second-stage model on the probability of association, ω_{st}, and hence infer the hotspot propensity variances from the data in a fully automatic way, with no ad-hoc choice or compromise that would bias the hotspot sizes.

5.2 Global-local modelling framework

5.2.1 Second-stage probit model on the probability of association

As a first step in detailing our proposal, we complement (5.1) with a hierarchical probit model on the probability of association, i.e.,

$$\omega_{st} = \Phi(\theta_s + \zeta_t), \qquad \zeta_t \overset{\text{iid}}{\sim} \mathcal{N}(n_0, t_0^2), \qquad s = 1, \ldots, p,\ t = 1, \ldots q, \tag{5.5}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function, and where we assume for now that $\theta_s \overset{\text{iid}}{\sim} \mathcal{N}(0, s_0^2)$. This second-stage model offers an interpretable representation of the association probability in multi-response settings: it involves a response-specific parameter, ζ_t, which adapts to the sparsity pattern corresponding to each response, and a propensity parameter, θ_s, which encodes predictor-specific modulations of the probability of association, as in Richardson et al. (2010); it is therefore more flexible than specification (5.2). The hyperparameters n_0 and t_0^2 are set to match a selected expectation and variance for the prior number of associated predictors per response (see Appendix C.1). The variance parameter s_0^2 essentially plays the role of σ^2_ω, presented in Section 5.1, in influencing the prior odds ratios; in particular, an application of the delta method shows that if $s_0^2 \sim O\left(q^{-1}\right)$ as $q \to \infty$, then $\text{Var}\{\Phi(\theta_s)\} \sim O\left(q^{-1}\right)$. While no closed form can be obtained for prior odds ratios (5.3) based on model (5.5), numerical experiments suggest that (5.4) indeed behaves independently of q when $s_0^2 = q^{-1}$, for $q_s \approx q$ large. Formulation (5.5) sets the stage for introducing our new multiplicity-adjusted hotspot model, which combines the benefits of both global and local control and adaptation.

5.2.2 Horseshoe prior on hotspot propensities

Our proposed specification for the hotspot propensity adds flexibility in modelling the scale of θ_s in (5.5) by letting

$$\theta_s \mid \lambda_s, \sigma_0 \sim \mathcal{N}\left(0, \sigma_0^2 \lambda_s^2\right), \qquad \lambda_s \overset{\text{iid}}{\sim} \text{C}^+(0,1), \qquad s = 1, \ldots, p, \tag{5.6}$$

where $\mathrm{C}^+(\cdot,\cdot)$ is a half-Cauchy distribution. This corresponds to placing a horseshoe prior (Carvalho et al., 2010) on the hotspot propensities, $\theta_s \overset{\text{iid}}{\sim} \mathrm{HS}(0,\sigma_0)$. The global scale σ_0 adapts to the overall sparsity pattern, while the Cauchy tails of the predictor-specific scale parameters λ_s flexibly capture the hotspot effects.

Gelman (2006) discusses the relevance of several noninformative and weakly informative priors on random effect variances. When the variance is close to zero, which is the case in sparse scenarios such as molecular QTL studies, Gelman cautions that badly chosen priors may severely distort posterior inferences. This observation is at the heart of work on scale-mixture priors such as the Strawderman–Berger prior (Strawderman, 1971; Berger, 1980), the Student-t prior (Gelman et al., 2008) or the horseshoe prior (5.6). These shrinkage priors differ in the modelling of the scale parameter, and all have substantial mass near zero in order to achieve good recovery of the sparsity pattern, while being sufficiently heavy-tailed to be robust to strong individual signals (recall Section 2.1.3). In particular, the horseshoe prior belongs to the class of global-local shrinkage priors that have an infinite spike at the origin and regularly-varying tails (Polson and Scott, 2010; Bhadra et al., 2016), and has newly established theoretical guarantees, such as near-minimaxity in estimation (van der Pas et al., 2017). Fully noninformative priors (e.g., whereby the scale parameter would be assigned a Jeffreys' prior) are ruled out, as they would fail to regularize.

5.2.3 Multiplicity-adjusted shrinkage profile

While the local scale parameters λ_s are essential to suitably detect the few large signals, the choice of the global scale σ_0 is no less important, as σ_0 controls the ability of the model to discriminate signal from noise. Piironen and Vehtari (2017) propose to choose σ_0 based on specific sparsity assumptions; we extend their considerations to our multi-response setting and further highlight how the dimension of the response needs to be accounted for in order to recover the beneficial shrinkage properties conferred by the horseshoe prior when used in the classical normal means model. For a given predictor $\boldsymbol{X}_s$, we reparametrize the probit link formulation,

$$\gamma_{st} \mid \theta_s, \zeta_t \sim \text{Bernoulli}\{\Phi(\theta_s + \zeta_t)\}\,, \qquad t = 1,\dots,q\,,$$

by introducing a q-variate auxiliary variable $\boldsymbol{z}_s = (z_{s1},\dots,z_{sq})$, as

$$\gamma_{st} = \mathbb{1}\{z_{st} > 0\}\,, \qquad z_{st} \mid \theta_s, \zeta_t \sim \mathcal{N}(\theta_s + \zeta_t, 1)\,, \qquad t = 1,\dots,q\,. \tag{5.7}$$

In this second-stage probit model, z_{st} can be understood as data, and $\boldsymbol{\theta}$ as a sparse parameter. Given the hyperparameters n_0 and t_0^2 for ζ_t, we have

$$z_{st} \mid \theta_s \sim \mathcal{N}(n_0 + \theta_s, 1 + t_0^2),$$

so that

$$\mathrm{E}(\theta_s \mid \boldsymbol{z}_s, \sigma_0, \lambda_s) = (1-\kappa_s)\frac{1}{q}\sum_{t=1}^{q}(z_{st} - n_0) + \kappa_s \times 0 = (1-\kappa_s)\bar{z}_s',$$

where $\bar{z}_s' = \bar{z}_s - n_0$ and

$$\kappa_s = \frac{1}{1 + \alpha(\sigma_0)\lambda_s^2}$$

is the *shrinkage factor* for hotspot propensities, with $\alpha(\sigma_0) = q(1+t_0^2)^{-1}\sigma_0^2$.

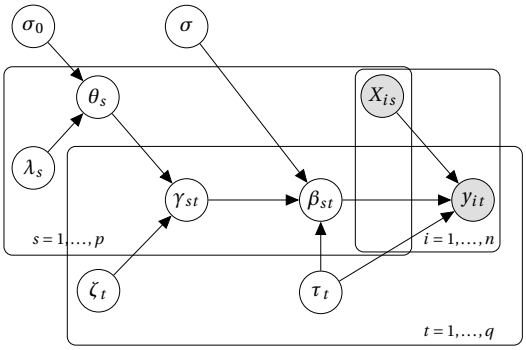

Figure 5.2 – Graphical representation of model (5.9). The shaded nodes are observed, the others are inferred. The probability of association is decoupled in response-specific, ζ_t, and predictor-specific, θ_s, contributions. The latter entails the global-local second-stage model for hotspots.

In the horseshoe prior literature, with half-Cauchy priors on the local scales as well as unit global scale and error variance, this factor has a Beta$(1/2, 1/2)$ prior whose shape resembles a horseshoe, hence the name. As this prior density is unbounded at 0 and 1, *a priori* one expects, either large effects, with κ_s close to zero, or no effects, with κ_s close to one (recall Section 2.1.3). In our case, it can be shown that

$$p(\kappa_s \mid \sigma_0) = \pi^{-1}\alpha(\sigma_0)^{1/2}\kappa_s^{-1/2}(1-\kappa_s)^{-1/2}\left[1+\kappa_s\{\alpha(\sigma_0)-1\}\right]^{-1}, \qquad 0<\kappa_s<1,$$

using $\lambda_s \overset{\text{iid}}{\sim} \text{C}^+(0,1)$; this prior density reduces to Beta$(1/2,1/2)$ when $\alpha(\sigma_0)=1$, that is, when $\sigma_0^2 \approx q^{-1}$, as $t_0^2 \ll 1$ under sparse assumptions. This formulation therefore enjoys the shrinkage properties of the horseshoe prior. Critically, using a default choice of $\sigma_0^2 = O(1)$ as $q \to \infty$ would yield $\text{E}(\kappa_s \mid \sigma_0) \approx 0$ for q large, so that on average, θ_s would be unregularized. These two choices can be read in light of the discussion in Section 5.1: the latter mirrors the absence of any correction for the dimensionality of the response, possibly creating spurious "pile-up" effects, whereas the former satisfies the multiplicity adjustment condition with the proposed scaling factor q^{-1} for σ_0^2.

Fixing $\sigma_0^2 = q^{-1}$ would stop the global scale from adapting to the degree of signal sparsity. We instead place a hyperprior on σ_0 which embeds the penalty. Following Carvalho et al. (2010), we choose a half-Cauchy prior,

$$\sigma_0 \sim \text{C}^+(0, q^{-1/2}). \tag{5.8}$$

An equivalent parametrization of (5.6) and (5.8) is

$$\theta_s \mid \lambda_s, \sigma_0 \sim \mathcal{N}\left(0, q^{-1}\lambda_s^2\sigma_0^2\right), \quad \lambda_s \overset{\text{iid}}{\sim} \text{C}^+(0,1), \quad \sigma_0 \sim \text{C}^+(0,1),$$

from which one clearly sees how the multiplicity factor rescales the hotspot propensity variance. For clarity, we gather the complete specification of our global-local hierarchical model; it combines (5.1) and the decomposition of the probability parameter (5.5) with (5.6) and (5.8):

$$\begin{aligned}
\boldsymbol{y}_t &\mid \boldsymbol{\beta}_t, \tau_t \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}_t, \tau_t^{-1}\boldsymbol{I}_n\right), && t = 1,\ldots,q,\\
\beta_{st} &\mid \gamma_{st}, \sigma, \tau_t \sim \gamma_{st}\,\mathcal{N}\left(0, \sigma^2\tau_t^{-1}\right) + (1-\gamma_{st})\,\delta_0, && s = 1,\ldots,p,\\
\gamma_{st} &\mid \theta_s, \zeta_t \sim \text{Bernoulli}\{\Phi(\theta_s+\zeta_t)\}, \qquad \zeta_t \overset{\text{iid}}{\sim} \mathcal{N}(n_0, t_0^2),\\
\theta_s &\mid \lambda_s, \sigma_0 \sim \mathcal{N}\left(0, \lambda_s^2\sigma_0^2\right), \qquad \lambda_s \overset{\text{iid}}{\sim} \text{C}^+(0,1), \qquad \sigma_0 \sim \text{C}^+(0, q^{-1/2}),
\end{aligned} \tag{5.9}$$

with Gamma prior distributions for τ_t and σ^{-2}; a graphical representation is provided in Figure 5.2.

5.3 A remark on inference

We implemented an annealed variational inference algorithm for model (5.9), resorting to suitable reparametrisations and data augmentations to obtain the updates analytically (albeit using the incomplete gamma and exponential integral functions). In particular, obtaining closed-form updates for the horseshoe's half-Cauchy scale parameters hinged on introducing auxiliary variables to arrive at variational distributions in the Gamma family or involving cheap-to-compute special functions, and this was somewhat complicated by the annealing. The full derivation of the annealed variational updates is in Appendix C.2. We will see in the numerical experiments of Section 5.4.4 that annealing is critical to accurate hotspot size inferences when SNPs are in high linkage disequilibrium.

We next discuss a pathology that can arise in noninformative data settings when strong annealing is employed (initial temperature $T = 20$ or larger, according to our experiments), and we evaluate several options to address it. The issue relates to the prior dependency between the horseshoe parameters θ_s and λ_s, as parametrized in model (5.9). The joint prior distribution $p(\theta_s \mid \lambda_s, \sigma_0)p(\lambda_s)$ creates a marked "funnel" shape, which reflects the concentration of θ_s around zero as the scale λ_s decreases (Figure 5.3). When the likelihood is nearly flat, this shape propagates to the posterior distribution, so the inference algorithm must manage important variations in curvature to fully explore the posterior space. Somewhat counterintuitively, practical difficulties arise only when the distributions are strongly annealed: by traversing larger regions of the parameter space, our algorithm can find itself exploring in the bulk of the funnel, with little chance of reaching back the neck of it. As a result, some λ_s may have unusually large variational means, which translates in spurious hotspot effects.

Such pathologies are very common in hierarchical models, owing to their complex geometries and dependencies; they are not limited to horseshoe distributions, although these are particularly at risk because of their heavy tails. Moreover, they may concern any type of algorithm, sampling-based or not: Gibbs sampling may mix poorly (Yu and Meng, 2011), Hamiltonian Monte-Carlo (HMC) chains may encounter divergent transitions resulting from inadequate step sizes (smaller steps are needed in the neck of the funnel, but larger in its bulk; Betancourt and Girolami, 2015), and variational algorithms may suffer from poor mean-field representations of the posterior (Ingraham and Marks, 2016).

Papaspiliopoulos et al. (2007) proposed using *noncentered* reparametrizations, such as

$$\theta_s = \lambda_s \tilde{\theta}_s, \qquad \tilde{\theta}_s \mid \sigma_0 \sim \mathcal{N}\left(0, \sigma_0^2\right), \qquad \lambda_s \sim \mathrm{C}^+(0,1), \tag{5.10}$$

as a replacement for the *centered* parametrization in (5.9). Formulation (5.10) trades prior independence between $\tilde{\theta}_s$ and λ_s, which diminishes the risk of strongly-curved posterior shapes and facilitates parameter exploration (Figure 5.3). Noncentered parametrizations have since attracted substantial interest (see, e.g., Betancourt and Girolami, 2015; Ingraham and Marks, 2016), yet a drawback is the loss of conjugacy, which, for variational inference, prevents closed-form updates. Rudimentary tests with a stochastic gradient implementation of the noncentered reparametrization in our model indicated that inferences are sensitive to the choice of the step size, which also impacts computational times. We did not pursue this approach further, as it has been suggested that centered parametrizations are more efficient when data informativeness increases (Yu and Meng, 2011; Betancourt and Girolami, 2015).

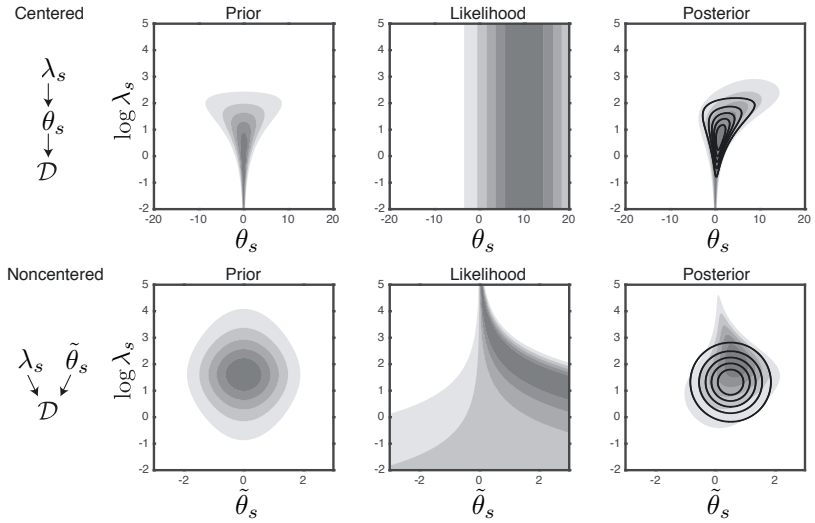

Figure 5.3 – Centered and noncentered parametrizations for horseshoe prior distribution. Top: "funnel"-shaped prior (left) under the centered parametrization; the dependency propagates to the posterior (right), as the likelihood (middle) is close to flat. Bottom: decoupled prior (left) under the noncentered parametrization with $\tilde{\theta}_s = \theta_s/\lambda_s$ which allows the mean-field distribution for $\tilde{\theta}_s$ and λ_s (black contour, bottom right) to implicitly reproduce the dependence structure of θ_s and λ_s (black contours, top right). Figure adapted from Ingraham and Marks (2016).

Concurrently to reparametrization ideas, many have advocated using lighter tails for local scale parameters in case of weakly informative data or separation in logistic regression (Gelman et al., 2008; Piironen and Vehtari, 2015; Ghosh et al., 2017). In particular, Piironen and Vehtari (2015) proposed replacing the half-Cauchy distribution in the horseshoe prior with a half-t distribution, $\mathrm{t}_\nu^+(\cdot,\cdot)$, with degrees of freedom $\nu > 1$,

$$\theta_s \mid \lambda_s, \sigma_0 \sim \mathcal{N}\left(0, \lambda_s^2\sigma_0^2\right), \qquad \lambda_s \sim \mathrm{t}_\nu^+(0,1), \tag{5.11}$$

and called this new prior specification *hierarchical shrinkage*. They argued that the runs of their HMC algorithm encountered a reduced number of divergent transitions compared to the horseshoe case $\nu = 1$. We implemented a modification of our algorithm to accommodate (5.11); no generic derivation was possible for any ν, but we could use a systematic approach for $\nu = 1, 3, 5, 7$, after which computational stability could not be guaranteed (see Appendix C.3.1). We considered real eQTL data for which our original algorithm with initial temperature $T = 20$ produced suspiciously large values for a few local scale parameters. We ran the algorithm using (5.11) and $\nu = 7$ and observed that the issue was transferred to the global scale parameter, which grew unexpectedly large owing to its heavy-tailed distribution, $\sigma_0 \sim \mathrm{C}^+(0, q^{-1/2})$; a quick-fix was to force $\sigma_0 = q^{-1/2}$. Lowering the initial temperature to $T = 5$ solved the problem for all algorithms, including our original algorithm for model (5.9), and yielded inferences comparable to that from the half-t_7-distribution with fixed σ_0; see Appendix C.3.2. Finally, simulations indicated that the half-Cauchy version (5.9) and the half-t version (5.11) achieve the same selection performance when the hotspots are relatively small, but the former outperforms the latter in presence of large hotspots.

The above observations therefore suggest that leaving model (5.9) untouched is best. We did not encounter issues when controlling the initial temperature T. In case of any doubt as to whether a chosen T is appropriate, a convenient diagnostic can be used that doesn't require running the

algorithm to convergence: the traces of the local scale variational means can be inspected once the cold temperature $T = 1$ has been reached, to make sure that no divergences appear.

5.4 Simulations

5.4.1 Data generation for pleiotropic QTL problems

The numerical experiments presented below are meant to closely reproduce real genetic data scenarios demonstrating pleiotropy, i.e., the control of several outcomes by a single SNP, for which our method is primarily designed. They also broadly illustrate the characteristic features of the method when applied to association studies with a large number of correlated responses. We either extract SNPs from real datasets or simulate them as autocorrelated by blocks of size 50 and in Hardy–Weinberg equilibrium, as described in Section 3.4.1. We also generate block-dependent responses using multivariate normal variables; the blocks consist of 10 equicorrelated responses.

The pleiotropic association pattern is constructed as follows. To model large and functionally inert genomic regions, we partition the SNPs into N chunks of size 200 and leave $\lfloor N/2 \rfloor$ chunks with no associations. From the remaining chunks, we randomly select labels for the "active" predictors, namely, SNPs associated with at least one response. Similarly, we select labels for the "active" responses, namely, responses associated with at least one SNP. We then randomly associate each active predictor with one active response, and each active response with one active predictor. For each active SNP s, we draw a "propensity" parameter π_s from a Beta$(1,5)$ distribution, and further associate the SNP with other active responses whose labels are sampled with probability π_s; these SNP-specific propensities $\{\pi_s\}$ therefore create hotspots of different "sizes". We effectively generate the associations under an additive dose-effect scheme with prescribed proportion of response variance attributable to genetic variants, following the procedure outlined in Section 3.4.1. For a given experiment, we keep the same association pattern across all replicates, but we regenerate the SNPs (if not real), responses and effect sizes for each replicate. The remaining settings (e.g., numbers of variables and of samples, effect sizes) vary, so will be detailed in the text corresponding to each experiment. Data-generation functions are implemented in the R package `echoseq` available at https://github.com/hruffieux/echoseq.

5.4.2 Variable selection performance with global-local modelling

In this section we evaluate the performance of our proposal for discriminating hotspots and selecting pairs of associated predictor and response variables. We simulated a "reference" data scenario with hotspots associated with approximately 35 responses on average and whose cumulated effect sizes are responsible for at most 25% of the variability of each response. We also generated four variants of this scenario: with smaller or larger hotspots (average sizes ≈ 17 and 85, respectively), and with weaker or stronger effects (response variance explained by hotspots below 20% and 30%, respectively). Each problem involves $p = 1,000$ SNPs and $q = 20,000$ responses (which corresponds the estimated number of protein-coding genes in humans), for $n = 300$ samples. We simulated 20 hotspots, and, depending on the hotspot size scenario, 100, 200 or 500 responses had at least one association.

We benchmarked our global-local model (5.9) against four alternatives. The first three are based on the proposal of Chapter 3, re-stated in (5.1)–(5.2), with three choices of hotspot propensity variance, σ_ω^2. These choices were made without assuming any prior state of knowledge, as would be faced in real data situations: we set the base rate of associated pairs to $\mu_\omega = 0.002$, so that two predictors are *a priori*

	$\sigma_\omega = \mu_\omega \times 0.1$	$\sigma_\omega = \mu_\omega \times 0.5$	$\sigma_\omega = \mu_\omega \times 1$	global	global-local
Pairwise selection					
Reference	55.5 (1.2)	74.1 (1.3)	**85.9 (0.7)**	69.5 (1.2)	**93.6 (0.4)**
Smaller hotspots	56.7 (1.0)	67.7 (1.3)	**82.6 (0.8)**	74.4 (0.9)	**90.4 (0.6)**
Larger hotspots	57.7 (1.0)	84.4 (0.5)	**89.6 (0.3)**	65.3 (0.8)	**96.7 (0.1)**
Weaker hotspots	44.7 (1.0)	57.5 (1.4)	**77.3 (0.8)**	53.0 (1.2)	**81.5 (1.0)**
Stronger hotspots	64.0 (1.1)	82.6 (0.7)	**90.2 (0.4)**	78.6 (0.7)	**96.2 (0.1)**
Predictor selection					
Reference	65.8 (2.7)	68.2 (2.6)	68.2 (2.2)	**71.7 (2.1)**	**74.6 (1.6)**
Smaller hotspots	53.1 (3.4)	54.7 (3.3)	54.9 (3.2)	**61.0 (3.1)**	**64.0 (3.0)**
Larger hotspots	80.2 (2.9)	**84.3 (2.6)**	84.0 (2.7)	83.7 (2.4)	**87.1 (2.1)**
Weaker hotspots	52.9 (3.1)	54.6 (3.2)	56.0 (2.8)	**59.2 (2.9)**	**63.2 (2.5)**
Stronger hotspots	75.6 (3.1)	78.3 (2.8)	77.4 (2.7)	**81.3 (2.4)**	**84.7 (2.1)**

Table 5.2 – Average standardized partial areas under the curve ×100 with false positive threshold 0.01 for predictor-response selection performance and predictor (hotspot) selection performance. Different hotspot size and effect size scenarios are reported, each based on 64 replicates; the "reference" case is displayed in Figure 5.4. Standard errors are in parentheses and, for each scenario, the best two performances are in bold.

associated with each response, on average. Then, for each model, we set the hotspot propensity scale to a different fraction of this base rate. The fourth model places a *global* Gamma prior on the hotspot propensity precision and embeds the multiplicity penalty used in our proposal, i.e.,

$$\theta_s \,|\, \sigma_0 \sim \mathcal{N}\left(0, q^{-1}\sigma_0^2\right), \quad \sigma_0^{-2} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right),$$

which can be reparametrized as

$$\theta_s \,|\, \sigma_0 \sim \mathcal{N}\left(0, \sigma_0^2\right), \quad \sigma_0^{-2} \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2q}\right).$$

With this choice, the propensity parameter has a Cauchy marginal prior distribution, $\theta_s \sim \text{C}(0, q^{-1/2})$. Both the Cauchy and the horseshoe models rely on the base rate level used for the three fixed-variance models to define the prior expected number of predictors associated with each response as $\text{E}_p = \mu_\omega \times p = 2$; the prior variance for this number is set to $\text{V}_p = 100$, which is large enough to cover a wide range of models. We use annealed variational inference on all five models; the geometric schedule consists of a grid of 100 temperatures, with initial temperature $T = 5$ (recall Section 4.3.1).

Figure 5.4 and Table 5.2 compare the five models in terms of selection of associated pairs of predictors and responses, selection of predictors (in our case, hotspots) and hotspot size estimation. They suggest several comments.

First, they illustrate our motivating statement: selection is sensitive to the choice of hotspot propensity variance; the pairwise selection performance of the three models with fixed variances varies greatly. The model with small variance strongly shrinks the hotspot sizes, which prevents the detection of many associations. The model with large variance identifies more pairs but fails to uncover the smallest hotspots; their estimated signals are overwhelmed by noise as a result of insufficient sparsity being induced (also see Appendix C.4.1). Moreover, arbitrarily fixing hotspot propensity variances to large values may trigger artifactual "pile-up" effects when the data are less informative, as discussed in Section 5.1.

Second, the Cauchy model (global shrinkage only) is often able to discriminate the small hotspot signals from the noise, thanks to its global scale inferred from the data, but is not as good for pairwise

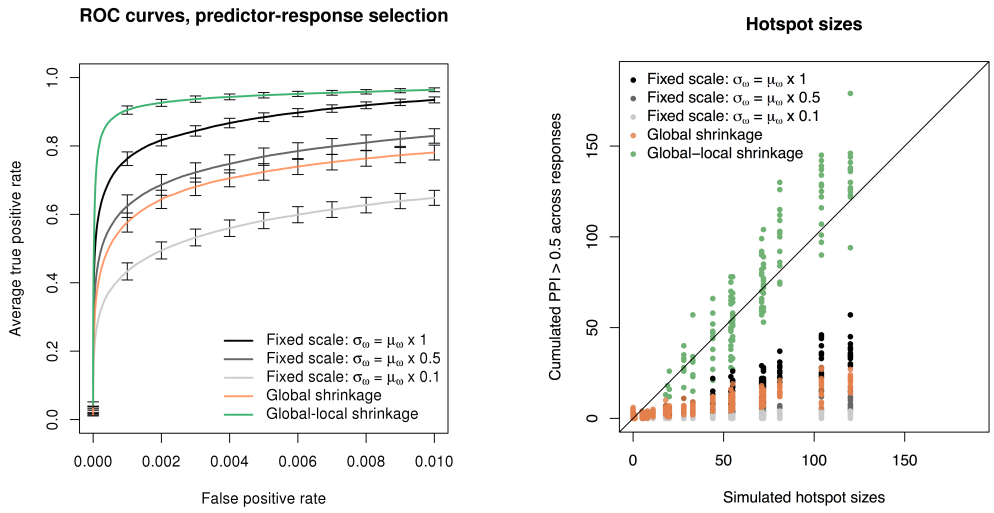

Figure 5.4 – Performance of five hotspot modelling approaches, for the "reference" data generation case. Left: truncated average ROC curves for predictor-response selection with 95% confidence intervals obtained from 64 replicates. Right: sizes of recovered hotspots based on the median probability model rule (Barbieri and Berger, 2004); 16 replicates are superimposed. The data comprise $p = 1,000$ simulated SNPs with 20 hotspots, $q = 20,000$ responses, of which 200 are associated with at least one hotspot, leaving the rest of the responses unassociated. The block-autocorrelation coefficients for SNPs were drawn from the interval $(0.75, 0.95)$, and the residual block-equicorrelation coefficients for responses were drawn from the interval $(0, 0.25)$. At most 25% of each response variance is explained by the hotspots. For the fixed-variance models, we used a base rate $\mu_\omega = 0.002$, and scales $\sigma_\omega = \mu_\omega \times \{1, 0.5, 0.1\}$, as explained in the text.

selection and estimation of hotspot sizes; because it is mostly informed by SNPs with no simulated associations, the global scale concentrates towards zero, which over-penalizes large hotspots, hampering the detection of pairwise associations with these hotspots. This phenomenon is of particular concern when signals are extremely sparse, as is thought to be the case in molecular QTL problems. One may attempt to improve the Cauchy specification by acknowledging the presence of genomic regions with diverse degrees of functional plausibility and introducing region-specific variance parameters to adapt to these degrees. Although inference may then be marginally impacted by the overall signal sparsity, such a formulation raises questions on the sensitivity to the chosen genome partition.

Our horseshoe-based proposal performs well for selection of both response-predictor pairs and hotspots. Unlike the fixed-scale models, it can clearly separate the small hotspots from the noise. Moreover, the hotspot sizes are well inferred overall: there is some variability depending on the simulated effects (re-drawn for each replicate), with the very small hotspots often underestimated, but the estimated sizes are much closer to the truth than those of the other models, which all strongly overshrink. We obtained the hotspot sizes by thresholding the marginal posterior probabilities of association at 0.5, a threshold which corresponds to the *median probability model* rule described by Barbieri and Berger (2004) as having optimal prediction performance. Hence, the flexibility offered by the horseshoe's heavy-tailed local scale parameters improves on global scale parameter formulations, whether the parameter values are fixed or inferred.

5.4.3 Null model scenario

We examine the behaviour of our approach on data with neither hotspots nor individual associations. We took the data simulated for the first replicate of the "reference" scenario discussed in Section 5.4.2, but randomly shuffled the response sample labels, thus leaving the response correlation structure untouched. We ran the method on eight such permuted datasets and observed no hotspot using the 0.5-thresholding rule on the marginal posterior probabilities of inclusion: there were at most four associated responses per predictor. The average proportion of false positive pairwise associations was 2×10^{-5}.

5.4.4 The benefits of annealing the local scales

The present numerical experiment focuses on data exhibiting strong predictor and response multicollinearity. To best reproduce conditions encountered in molecular QTL studies, we used real SNPs from eQTL data (see application Section 5.5). We considered a 1.7 megabase (Mb) region located ≈ 1 Mb upstream of the MHC region and comprising 200 variants for which $n = 413$ observations were available. We distributed five active SNPs across the blocks and simulated 500 "active" responses. Effects were small, with each response having at most 10% of its variability explained by genetic variation. We added another 19,500 inactive responses, drawn from Gaussian noise. The residual correlation of the responses spanned larger values than in Section 5.4.2, with block-correlation coefficients $\rho \in (0, 0.5)$.

Figure 5.5 indicates that the annealed variational algorithm clearly discriminates hotspots. Moreover, when declaring associations using a threshold of 0.5 on the marginal posterior probabilities, the hotspot sizes were well estimated, except for SNP id 105. In contrast, the non-annealed version of the algorithm struggled to single out the relevant SNPs from their correlated neighbours, especially around SNP id 110. We also applied the algorithm with and without annealing on the data from the first replicate, performing 500 runs each using different starting values, and reached conclusions identical to those of Section 4.3.2: the optimal value reached by the variational objective function is larger and less variable in the annealed case (Figure 5.5).

5.4.5 Comparison with other approaches

We conclude this series of simulation experiments by comparing the method with existing approaches. We choose two competing methods, MatrixEQTL (Shabalin, 2012) and HESS (Richardson et al., 2010) as representative of two types of approaches: a univariate screening algorithm that tests the SNP-response pairs one by one, and joint hierarchical modelling coupled with parallel chain MCMC inference.

We restrict the number of simulated responses to 10,000 in order to ensure a reasonable convergence time for the HESS MCMC run, and involve 15 SNPs in associations. We rely on the default settings proposed in the MatrixEQTL and HESS implementations: these correspond, for the former, to using an additive linear model for the genotype effects and t-statistics for significance tests, and for the latter, to running three parallel chains for 22,000 iterations, discarding the first 2,000 as burn-in. Our annealed variational inference procedure was about 30 times faster than the MCMC inference implemented in HESS, with an average runtime for one replicate of 4 hours and 17 minutes for the former and 5 days and 10 hours for the latter on an Intel Xeon CPU, 2.60 GHz.

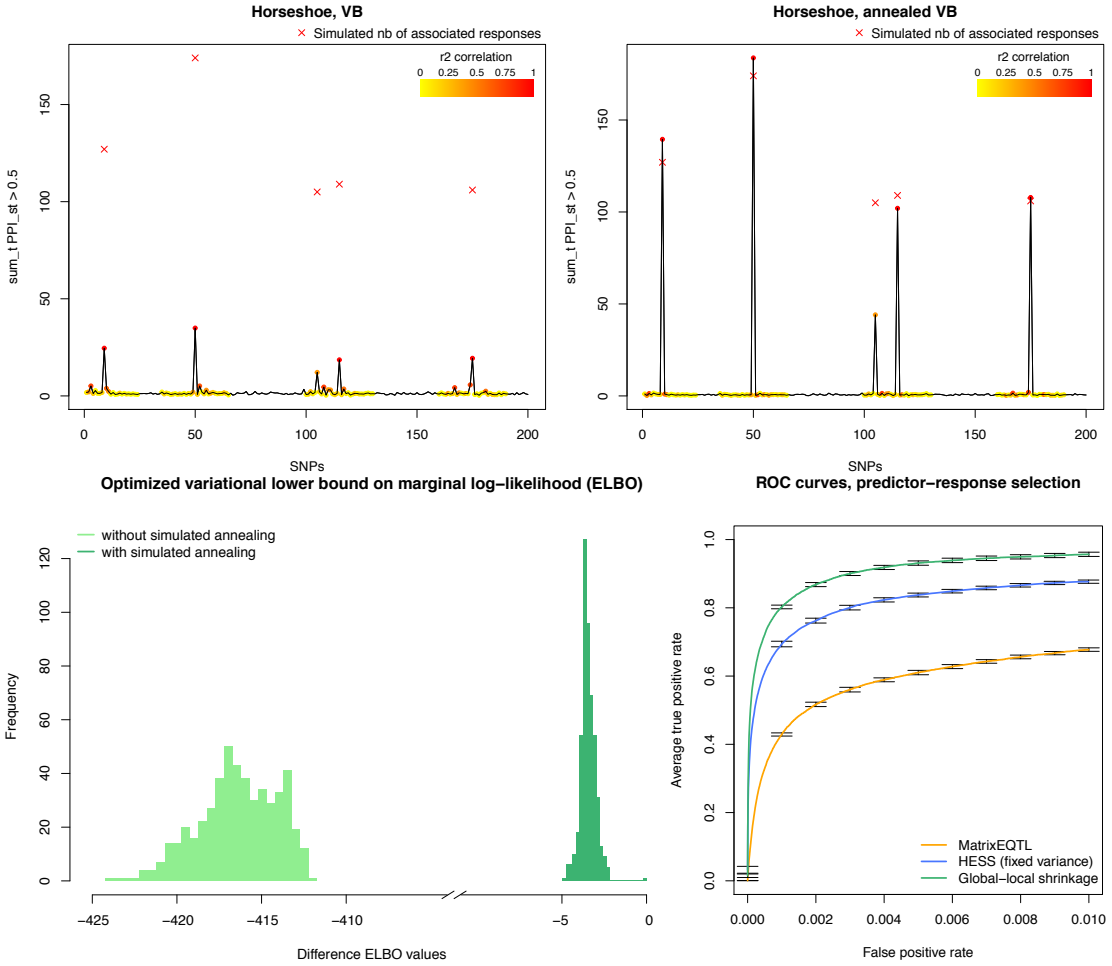

Figure 5.5 – Performance comparisons between classical and annealed variational inferences, and with competing methods. Problem with responses equicorrelated by blocks with residual correlation $\rho \in (0, 0.5)$; 500 of them are under genetic control. Candidate predictors are $p = 200$ SNPs from a cohort of European ancestry, for $n = 413$ individuals. Top: Hotspot discrimination achieved by classical (left) and annealed (right) variational inferences, for a problem with $q = 20,000$ responses. The plots show the cumulated number of responses associated per SNP, after thresholding marginal posterior probabilities at 0.5, and averaging over 16 replicates. The red crosses show the simulated sizes of the five hotspots (whose cumulated effects account for at most 10% of the variability of a response). The coloured regions quantify the linkage disequilibrium structure in r^2 computed with respect to hotspots 9, 50, 115 and 175, respectively. Bottom, left: Histograms of optimized lower bound on the marginal log-likelihood (ELBO) with classical and annealed variational inferences, 500 replicates; the x-axis shows the maximum ELBO value subtracted from all other values. Bottom, right: Truncated average ROC curves with 95% confidence intervals for the MatrixEQTL and HESS methods, and our proposal. The same settings as above are used, except for the number of responses, limited to $q = 10,000$, and the number of hotspots, 15, whose cumulated effects account for at most 20% of the variability of a response. Both HESS and our proposal have prior expectation $\mathrm{E}_p = 1$ and variance $\mathrm{V}_p = 10$ for the number of SNPs associated with a response; the value of E_p is smaller than in Sections 5.4.2 and 5.4.3 because there are fewer candidate predictors, and so is the value of V_p, to limit the computational costs of HESS.

As expected, the ROC curves of Figure 5.5 indicate that MatrixEQTL performs worse than the two joint approaches. It correctly identifies the strong associations but also declares many spurious associations involving SNPs in high linkage disequilibrium. This agrees with the motivating example in Section 1.1; marginal screening often provides satisfactory answers when the aim is to highlight *cis* associations at the level of loci, but, because of the multiplicity burden, it often fails to declare weaker effects such as those involved in *trans* associations. By borrowing information across all SNPs and responses, HESS achieves much better association recovery. The HESS run is based on a specific choice of hotspot propensity variance, which is hard-coded and not accessible to the user; we expect the performance to vary with other choices of variances, similarly to what was shown in Figure 5.4 for the fixed-scale approach of Chapter 3. With its global and local variances inferred from the data, our proposal performs best. Confronting this performance with MCMC inference further suggests that the independence assumptions underlying the variational mean-field formulation do not degrade the quality of variable selection, as seen in Chapter 3. The coupling with simulated annealing results in an excellent selection in our experiments, and in a fraction of the time required by MCMC techniques; this is particularly remarkable in highly multimodal settings.

5.5 A targeted study of hotspot activity with stimulated monocyte expression

In this section, we apply our approach to the eQTL data introduced in the motivating example of Section 1.1. These data differ from most molecular QTL data, as they entail expression from $CD14^+$ monocytes before and after immune stimulation, performed by exposing the monocytes to the inflammation proxies interferon-γ (IFN-γ) or differing durations of lipopolysaccharide (LPS 2h or LPS 24h). The genetic variants are single nucleotide polymorphisms (SNPs) determined using Illumina arrays and the samples were obtained from 432 healthy European individuals.

Related work (Fairfax et al., 2014; Kim et al., 2014; Lee et al., 2014) has suggested that gene stimulation may trigger substantial *trans*-regulatory activity, creating favourable conditions for the manifestation of hotspot genetic variants; the analysis of stimulated eQTL data should therefore benefit from a method tailored to the detection of hotspots. In addition to monocyte expression, we consider B-cell expression data for the same samples, to contrast the hotspot activity for the two cell types. Here, we analyse three genomic regions comprising genes thought to play a central role in the pathogenesis of immune disorders (Fairfax et al., 2012, 2014): *NFE2L3* on chromosome 7, *IFNB1* on chromosome 9, and *LYZ* on chromosome 12. Each region involves 1,500 SNPs and spans from 7.5 to 12 Mb.

The following quality control steps were performed prior to the analyses. For the genotyping, we applied standard filters that exclude SNPs with call rate < 95%, violate the Hardy–Weinberg equilibrium assumption (at nominal p-value level 10^{-4}), or have minor allele frequency < 5%. For the transcripts, we considered the top 30% quantile of the interquartile range distributions in each (un)stimulated condition. In order to work with a common set of transcripts across conditions, we then retained the intersection of the transcripts selected in each condition, and checked that no highly varying transcript was dropped in this process. Finally, we discarded samples with unusual transcript values, separately for each condition; the numbers of individuals thus retained were 413 for unstimulated monocytes, 366 for IFN-γ, 260 for LPS 2h, 321 for LPS 24h and 275 for B-cells, and the number of transcripts was 24,461.

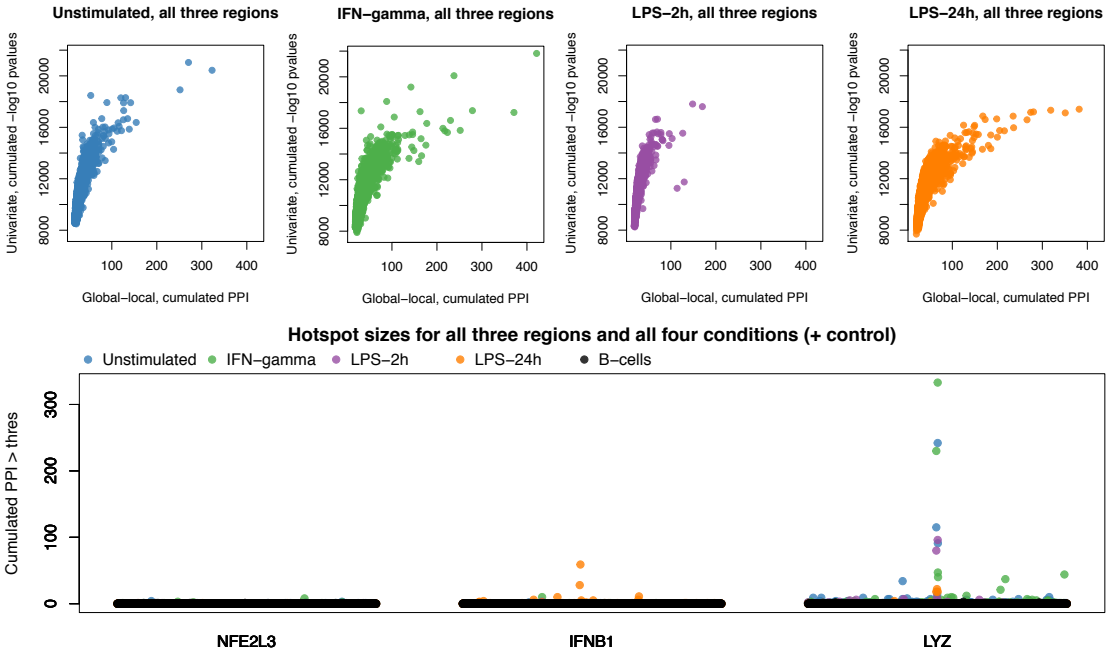

Figure 5.6 – Evidence for hotspots from stimulated eQTL analyses, for the *NFEL2L3*, *IFNB1* and *LYZ* genomic regions with the four monocyte conditions and the B-cell negative controls. Top: for each condition, raw hotspot evidence for all three regions comprising *NFEL2L3*, *IFNB1* and *LYZ*. Scatterplots with $-\log_{10}$ p-values of univariate screening, summed across responses, versus marginal posterior probabilities of inclusion obtained by our proposal, summed across responses. Bottom: Hotspot sizes, as declared using a permutation-based FDR of 20%.

We ran our method on each of the three regions, and for all four monocyte conditions, as well as for the B cells, resulting in 15 separate analyses. We employed the same prior base rate of associated pairs as in the simulation of Sections 5.4.2 and 5.4.3, giving a prior average of $\mathrm{E}_p = 0.002 \times p = 3$ SNPs associated with each transcript, and used a variance of $\mathrm{V}_p = 25$. Figure 5.6 compares the evidence for hotspots produced by our proposal and plain univariate screening. It shows the nominal $-\log_{10}$ p-values of a univariate screening against the raw posterior probabilities, both summed across responses, and suggests that the two approaches agree on the small or moderate evidence but also that our proposal appears to boost and better distinguish hotspot effects.

In order to derive empirical false discovery rates, we ran a permutation analysis with 30 replicates for each region and condition, by shuffling the sample labels of the expression matrix; this was computationally feasible thanks to the efficiency of our variational procedure. We then obtained Bayesian false discovery rates for a fine grid of thresholds on the posterior probabilities, and fitted a spline in order to derive thresholds corresponding to a false discovery rate of 20%.

Figure 5.6 indicates increased *trans*-regulatory activity under stimulation with IFN-γ and LPS 24h. This activity was endorsed by the absence of hotspots in the B-cell analysis; indeed, previous studies comparing B cells and monocytes on the three regions suggested that QTL activity was specific to the latter (Fairfax et al., 2012), so the former may be used as negative controls in our analyses. The degree of activity also varies greatly across the three regions: the *NFE2L3* region is essentially inactive; its largest hotspot is of size 8 and appears under IFN-γ stimulation, in line with previous observations (Fairfax et al., 2014). The *IFNB1* region shows more activity under LPS 24h stimulation; this confirms

existing work (Fairfax et al., 2014), but also reveals more associations with transcripts. The top LPS 24h hotspot in the *IFNB1* region, rs3898946, is an eQTL reported in the GTEx database, for genes *FOCAD* and *MLLT3* in the tibial artery, and for gene *PTPLAD2* in skin tissues; this provides further support for a mechanistic role of this hotspot (to be confirmed in further work). The *LYZ* region is known for its high degree of pleiotropy (Rotival et al., 2011) and is indeed very active in our analyses.

Although Fairfax et al. (2014) mostly report stimuli-specific *trans*-regulatory activities, our top hotspot hit, rs6581889, located only 9 Kb downstream of the *LYZ* gene, is persistent across all four conditions: it is the largest hotspot in the unstimulated condition with size 242, in the IFN-γ condition with size 333, and in the LPS 2h condition with size 96, and it is the second largest hotspot in the LPS 24h condition with size 18; a Venn diagram showing the transcript overlap across conditions is given in Appendix C.5. Hence, SNP activity was triggered by the IFN-γ stimulation, but was also substantial after 2 hours and 24 hours of LPS stimulation. The B-cell data provide a good negative control as they show no activity in the *LYZ* region; the largest number of responses associated with a given SNP is three, and the signal does not colocalize with any hotspot uncovered in monocytes. Finally, rs6581889 is a known *cis* eQTL for *LYZ* and *YEATS4* in multiple tissues, two associations which our analyses confirmed.

5.6 Summary

We have introduced a new approach for the efficient detection of hotspots in regression problems with tens of thousands of response variables. Our proposal accommodates three essential characteristics of molecular QTL problems: extreme sparseness of association patterns, strong multimodality induced by locally correlated genetic variants, and very high dimensions of both the predictor and the response vectors.

Our simulations indicate that severe sparsity renders both fixed and inferred global hotspot propensity variance formulations ineffective. Our global-local model provides sufficient refinement to properly identify the locations and sizes of individual hotspots; it is free of ad-hoc variance choices and automatically adapts to different signal sparsity degrees. Informativeness in the hotspot propensity prior is limited to the embedded penalty adjusting for the response dimension. This penalty prevents the manifestation of artifactual hotspots when the likelihood is relatively flat; we provided two justifications for its choice.

Our experiments also demonstrate that the simulated annealing scheme for this new model yields satisfactory estimates of hotspot sizes in situations where classical variational inference would strongly overshrink.

Finally, our application to multiple-condition monocyte data highlighted a strong candidate eQTL hotspot that was persistent across conditions, and whose activity was supported by B-cell negative control analyses and public eQTL annotations.

6 Leveraging predictor-level information

The joint inference approaches presented thus far aimed at improving selections from large predictor and response spaces of genetic variants and molecular outcomes. A natural enhancement would be to refine these selections by leveraging additional data that may provide insights into the propensity of predictors to be involved in associations. In genetics, *epigenomic annotations* on genetic variants are increasingly collected and used as a source of information on the functional potential of these variants. In this chapter we present a second-stage hierarchical regression extension that can encode such variables and let them modulate the probabilities of associations.

Section 6.1 provides some background on epigenomic annotations and sets our goals. Section 6.2 presents our model in light of previous proposals and comments on an important modelling assumption. Section 6.3 describes our variational-expectation-maximization approach to scale up inference to large numbers of annotations, genetic variants, molecular outcomes and samples. Section 6.4 evaluates our proposal in simulations.

The material of this chapter is ongoing work in collaboration with Leonardo Bottolo and Sylvia Richardson.

6.1 Motivation

Molecular datasets and annotation databases are growing in diversity. On one hand, the biological entities involved in processes of interest are more likely to be covered by the collected measurements and their identification may be facilitated by the use of rich complementary data sources to the primary data. On the other hand, a number of obstacles remain or worsen: the multiplicity burden increases, sample sizes grow much more slowly than the number of variables analyzed, and the local dependence of genetic variants (linkage disequilibrium) confounds the identification of biological signals.

In Chapter 4, we considered embedding spatial information in the model to improve the interpretability of posterior quantities for regions of high linkage disequilibrium. Here, we exploit another type of structural information, which is based on the functional potential of genetic variants, capitalizing on the wealth of available *epigenomic annotation* sources. Suitable use of this information may boost the detection of weak associations and help in discriminating genuine signals from spurious ones caused by linkage disequilibrium.

As suggested by its Greek prefix, the "epigenome" is a generic term referring to chemical compounds that exert control on the regulatory functions of the genome "on top of" the basic genetic principles of inheritance. A more precise definition is given in Morgensztern et al. (2018):

> "[The] epigenome is the complete description of all the chemical modifications to DNA and histone proteins that regulate the expression of genes within the genome. [...] The most common mechanisms of epigenetic modification include DNA methylation, histone modifications, and transcription of small noncoding RNA."

DNA methylation is a process that attaches a methyl group to the bases (typically cytosine) of a DNA molecule, thereby possibly repressing or activating gene expression (Kulis and Esteller, 2010). Histone modifications encompass multiple types of chemical modifications of amino acids of histone proteins, which are primary components of eukaryotic chromatin. These modifications can alter the chromatin structure and function of chromatin-associated proteins (Shogren-Knaak and Peterson, 2003). Finally, small noncoding RNAs are RNA molecules that do not code for proteins, yet can be important regulators of gene expression and impact the organization of chromatin (van Wolfswinkel and Ketting, 2010). Several studies have confirmed that genome-wide association and molecular QTL hits are enriched in epigenomic marks, as well as in other types of functional annotation, such as DNase-I hypersensitive sites, transcription factor binding sites or location of genetic variants (intronic, intergenic), see, e.g., Gaffney et al. (2012), Maurano et al. (2012), Karczewski et al. (2013) and Trynka et al. (2013). Although, in a strict sense, epigenomic marks represent a subset of functional marks, the former terminology is often used generically in place of the latter, so we hereafter follow this practice.

Practitioners resort to epigenomic annotations mostly for prioritization of hits obtained from marginal screening. They typically loop through all the loci with significant associations, and, for each such locus, they manually inspect a few marks to decide on "a most promising" functional candidate SNP among all those in linkage disequilibrium. This approach is unsatisfactory for several reasons: first, publicly available databases nowadays contain several hundreds of epigenomic annotations for each variant, and selecting a few may involve omitting many relevant others and thus may bias the conclusions. Second, even if a comprehensive inspection were feasible, the degrees of relevance of the marks may be very uneven and may depend on the conditions, tissues, even genomic regions considered, so it is not clear how to weight each contribution.

To avoid arbitrariness and to best leverage this information, we propose a molecular QTL model where the propensity of SNPs to be involved in associations is moderated by annotation variables. The relevance and effects of these variables are inferred in a top-level regression. The resulting two-stage hierarchical model enables information to be borrowed from the three data sources (annotations, SNPs and expression outcomes) in a unified manner. It is designed to flexibly accommodate important features of the data and to provide interpretable posterior quantities for selection of annotation, SNP and outcome variables. In particular, it can handle several hundred annotation variables by enforcing sparsity for their effects. Another particularity of our model is that the annotation effects are modelled as specific to QTL blocks, each formed by a set of SNPs and of expression outcomes (*module*) whose possible associations may be governed by common epigenomic mechanisms.

We intend to use our new framework with genome-wide annotations from the 1,000 Genome Project or ENCODE consortium, and hope it will both enhance the discovery of risk variants and help to disentangle the corresponding functional processes, thanks to the selected epigenomic marks.

6.2 Two-stage hierarchical regression model

6.2.1 Model and earlier proposals

Consider the usual base regression model for q centered responses, $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q)$, p centered candidate predictors for them, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, and n samples,

$$\begin{aligned} \boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t &\sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}_t, \tau_t^{-1}\boldsymbol{I}_n\right), && t = 1, \ldots, q, \\ \beta_{st} \mid \gamma_{st}, \sigma^2, \tau_t &\sim \gamma_{st}\mathcal{N}\left(0, \sigma^2\tau_t^{-1}\right) + (1-\gamma_{st})\,\delta_0, && s = 1, \ldots, p, \end{aligned} \tag{6.1}$$

where δ_0 is the Dirac distribution, and τ_t and σ^{-2} are precision parameters having Gamma priors. To let the effects modelled in (6.1) be informed by r predictor-level covariates, we introduce a second-stage probit regression on the probability of effects,

$$\begin{aligned} \gamma_{st} \mid \theta_{b,s}, \zeta_t, \boldsymbol{\xi}_b &\sim \text{Bernoulli}\left\{\Phi(\theta_{b,s} + \zeta_t + \boldsymbol{V}_s^T\boldsymbol{\xi}_b)\right\}, \\ \xi_{b,l} \mid \rho_{b,l} &\sim \rho_{b,l}\mathcal{N}(0, s_b^2) + (1-\rho_{b,l})\,\delta_0, \qquad \theta_{b,s} \sim \mathcal{N}(0, s_{0b}^2), \qquad \zeta_t \sim \mathcal{N}(n_0, t_0^2), \\ \rho_{b,l} &\sim \text{Bernoulli}\left(\omega_{b,l}\right), \qquad l = 1, \ldots, r, \end{aligned} \tag{6.2}$$

where $\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_r)$ is the $p \times r$ matrix of (centered) predictor-level covariates, $b \in \mathscr{P}$ is a block of predictor and response variables, with $\mathscr{P}$ a partition of $\{1, \ldots, p\} \times \{1, \ldots, q\}$ and $b \ni (s, t)$.

In our molecular QTL setting, the primary regression model (6.1) describes the regulation of q molecular expression outcomes by p candidate SNP predictors, while the secondary regression model (6.2) leverages information from r epigenomic annotation marks on the involvement of the SNPs in the primary associations. The partition $\mathscr{P}$ into pairs of response and predictor subsets enables block-specific predictor-level effects, $\boldsymbol{\xi}_b$. A given block b should ideally correspond to a *module* of expression outcomes under genetic control and a set of SNPs from the genomic region exerting this control. Model (6.2) then represents the possible epigenomic effects underlying the functional mechanisms in the block. Most pairs of SNPs and outcomes are likely to present no association and form a *null block*.

The effects of the predictor-level covariates on the association probabilities are modelled using a spike-and-slab prior to induce sparsity. This allows the incorporation of a large number of epigenomic annotations, even though a fraction may be related to functional mechanisms between the SNPs and the molecular outcomes. Moreover, selecting the relevant annotations is easily achieved using the posterior means of $\rho_{b,l}$, $\text{pr}(\rho_{b,l} = 1 \mid \boldsymbol{y})$.

The values of the hyperparameters n_0 and t_0^2 are chosen to induce sparsity, using the procedure described in Appendix C.1 (the same as for model (5.9)). The specification of hyperparameters s_{0b}^2, s_b^2 and $\boldsymbol{\omega}_b$ is the object of Section 6.3; in particular, we aim for a unified treatment the SNP and annotation effect variances, s_{0b}^2 and s_b^2, and hence don't place a global-local hyperprior on the former, as in Chapter 5. Moreover, for the same reason, the hotspot propensity parameter θ_s of Chapter 5 is now module-specific, $\theta_{b,s}$, i.e., only informed by the expression outcomes from a same block b, which may be sufficient under the assumption that different modules have different genetic bases. A graphical representation of model (6.1)–(6.2) is given in Figure 6.1.

The rationale for block-specific annotation effects is threefold. First, it is unlikely that all molecular QTL associations are uniformly governed by the same functional mechanisms, involving a common set of epigenomic marks. Second, the QTL association pattern is very sparse, so, if the effects of annotations were modelled globally, they would be diluted by the large, functionally inert, genomic regions, as

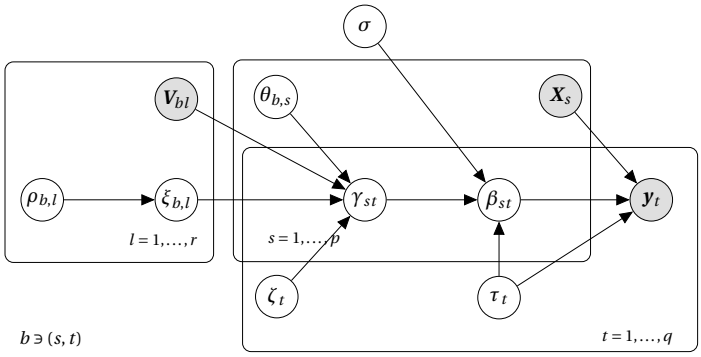

Figure 6.1 – Graphical representation of model (6.1)–(6.2). The shaded nodes are observed, and the others are inferred. The block-specific second-stage regression corresponds to the left plate. $\boldsymbol{V}_{bl}$ is the vector gathering the observations of predictor-level covariate l for the predictors belonging to block b.

will be illustrated in the simulations of Section 6.4. Finally, the possibility of using different types of candidate annotation variables for different blocks is a side-benefit. Moreover, as most such variables entail binary information, the number of candidate annotations effectively considered can be reduced by discarding, in each block b, the constant annotation variables, that are all zero or all unity for the SNPs in b, and those that are perfectly collinear. An important question however concerns the specification of a suitable partition; we will discuss this in Section 6.2.2.

Several proposals exist to leverage epigenomic information in genome-wide association studies. Most rely on genetic association summary statistics instead of raw SNP and expression data, and focus on fine-mapping tasks, thereby restricting to genomic loci of interest (e.g., Kichaev et al., 2014; Pickrell, 2014; Chen et al., 2015). The approach of Yang et al. (2017) instead infers association from raw QTL data; it involves preprocessing the data to assign a single annotation feature to each SNP (e.g., coding, noncoding, intergenic), and models SNP effects using a spike-and-slab prior with annotation-specific probability parameter and effect-size variance. The proposal of Quintana and Conti (2013) is closer to ours: it also resorts to a second-stage regression with a probit link on the probability parameter. However, the effect of annotations is modelled using a Gaussian prior, while we separate the active and inactive annotations with a spike-and-slab prior that better lends itself to selection. Finally, van de Wiel et al. (2018) provide a general discussion on using external information in empirical Bayes frameworks. They claim that empirical Bayes estimation is particularly suited to settings with lots of variables and auxiliary information on them; this is indeed supported by Efron (2010)'s empirical Bayes approach to multiplicity adjustment which quickly became a reference (recall Section 2.2). They formalize the task of leveraging external information using the concept of "co-data", originally introduced in te Beest et al. (2017):

> "Co-data are defined as any type of information that is available on the variables of the primary data, but does not use its response labels",

and mention two earlier contributions: van de Wiel et al. (2016) use co-data to define groups of predictors and perform inference using a Bayesian ridge regression with group penalties estimated via empirical Bayes, and te Beest et al. (2017) use co-data to inform the sampling probabilities of a random forest algorithm, which they estimate from the data. Finally, van de Wiel et al. (2018) discuss moderating predictor-specific spike-and-slab probabilities with co-data using a link function g (in our

case, the probit link), i.e.,

$$\begin{aligned} \beta_s \mid \gamma_s &\sim \gamma_s f_\beta + (1-\gamma_s) f_0, \qquad \gamma_s \mid \pi_s \sim \text{Bernoulli}(\pi_s), \qquad s = 1, \ldots, p, \\ \pi_s &= g^{-1}\left(\boldsymbol{V}_s^T \boldsymbol{\xi}\right), \end{aligned}$$

where f_β and f_0 are "signal" and "noise" distributions. They suggest coupling MCMC or variational inference with an expectation-maximization (EM) algorithm for inferring $\boldsymbol{\xi}$ but provide no implementation.

We independently developed a variational-EM strategy to obtain empirical Bayes estimates for the block-annotation hyperparameters, s_{0b}^2, s_b^2 and $\omega_{b,l}$ (see Section 6.3). This allows the scaling of inference to several hundred annotations, which the above earlier proposals, based on MCMC or MCMC-EM procedures, fail to achieve. Moreover, we can handle thousands of response variables simultaneously, whereas all existing methods apply to a single response; the exception is Li and Kellis (2016), who jointly model summary statistics for a handful of related responses.

6.2.2 Partition choice

In Section 6.2.1, we motivated the use of block-specific predictor-levels effects; this assumption raises the question of the choice of partition $\mathscr{P}$. We have already argued that molecular QTL blocks should ideally comprise molecular entities with similar functional properties, e.g., SNPs from a genomic region and related expression outcomes controlled by this region.

A natural idea would be to revise our approach to adaptively infer the partition along with the QTL associations and annotation effects. This seems very challenging, however, given the complexity of the modelling task, so it may be more reasonable to define the partition before applying the model in its current state. However this exposes us to the problem of "using the data twice". Indeed, seeking a partition into blocks of associated SNPs and outcomes conflicts with our primary goal of uncovering the QTL association pattern, by already attempting to provide partial answers. Hence, existing proposals whose main objective is to partition molecular QTL data based on association information (e.g., Monni and Tadesse, 2009; Zhang et al., 2010; Jiang and Liu, 2015) are of no help to us.

A simpler solution may be to define the partition as a grid, by splitting the SNP and outcome spaces independently, i.e., without using the conditional information of $\boldsymbol{y}$ given $\boldsymbol{X}$. For instance, we may group the SNPs based on spatial structures, e.g., by chromosome, haplotypes, or regions flanked by recombination hotspots. We may also separately seek modules of co-expressed molecular outcomes, assuming that they may be co-regulated. This can be done in an unsupervised manner, for instance using hierarchical clustering or graphical models. Functional information gathered in external databases may also provide guidance (e.g., using pathways such as GO, KEGG or Reactome).

Finding an approach tailored to our needs is future work. It should be sufficiently flexible to accommodate complex functional patterns, yet simple enough to be seamlessly incorporated into our already complicated framework. The numerical experiments of Section 6.4.3 attempt to provide some insights into the sensitivity of our approach to the choice of $\mathscr{P}$.

6.3 Annealed variational-EM inference

Let $\boldsymbol{\nu}$ be the parameter vector for model (6.1)–(6.2), and let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_B)$ be the second-stage model hyperparameter vector, with $\boldsymbol{\eta}_b = (\boldsymbol{\omega}_b, s^2_{0b}, s^2_b)$ for block b. We propose estimating $\boldsymbol{\eta}$ via empirical Bayes, by finding

$$\hat{\boldsymbol{\eta}} = \arg\max_{\boldsymbol{\eta}} \ell\left(\boldsymbol{\eta}; \boldsymbol{y}\right), \tag{6.3}$$

where $\ell\left(\boldsymbol{\eta}; \boldsymbol{y}\right) = \log p\left(\boldsymbol{y} \mid \boldsymbol{\eta}\right)$ is the marginal log-likelihood. Computing (6.3) analytically for our model ~~is not feasible as this~~ would require high-dimensional integration. Because this is a general issue for variable selection models, George and Foster (2000) proposed two strategies referred to as the *marginal maximum likelihood criterion* (MML) and *conditional maximum likelihood criterion* (CML) to approximate the solution in the context of a spike-and-slab model with g-prior (Zellner, 1986). The MML criterion corresponds to direct optimization, expressing the marginal log-likelihood as a sum over all possible models $\boldsymbol{\gamma}$ and maximizing numerically. As this is infeasible even for moderate dimensions, the proposed workaround is simply the observation that the sum becomes a product under orthogonal designs. For the majority of cases where the orthogonality condition is not fulfilled, George and Foster (2000) propose the CML criterion, which approximates (6.3) by maximizing, rather than marginalizing, over the model space, thereby avoiding summing over $\boldsymbol{\gamma}$. However they observe in numerical experiments that this can affect the quality of inferences, sometimes substantially. Moreover, the CML criterion still requires a search through the model space, which can have a high computational price.

To circumvent computing marginal likelihoods, Casella (2001) proposed coupling the empirical Bayes estimation of the hyperparameter $\boldsymbol{\eta}$ with a Gibbs sampling scheme that simultaneously infers the model parameter vector $\boldsymbol{\nu}$. His procedure corresponds to a *Monte Carlo EM algorithm* which alternates between constructing an estimate $\hat{\boldsymbol{\eta}}$ using the samples from a (fully converged) Gibbs sampling for $p(\boldsymbol{\nu} \mid \boldsymbol{y}, \boldsymbol{\eta})$, and obtaining Gibbs samples from $p(\boldsymbol{\nu} \mid \boldsymbol{y}, \hat{\boldsymbol{\eta}})$.

A similar strategy can be implemented within variational inference frameworks; the variational-EM (or "VBEM") algorithm was proposed by Blei et al. (2003) in their seminal work on variational inference for latent Dirichlet allocation. The procedure results in alternating optimizations of the variational lower bound

$$\mathcal{L}(q; \boldsymbol{\eta}) = \mathrm{E}_q \log p(\boldsymbol{y}, \boldsymbol{\nu} \mid \boldsymbol{\eta}) - \mathrm{E}_q \log q(\boldsymbol{\nu}), \tag{6.4}$$

where $q(\boldsymbol{\nu})$ is the variational approximation for $p\left(\boldsymbol{\nu} \mid \boldsymbol{y}, \hat{\boldsymbol{\eta}}\right)$ for a current estimate $\hat{\boldsymbol{\eta}}$. More precisely, it initializes the parameter and hyperparameter vectors $\boldsymbol{\nu}^{(0)}$ and $\boldsymbol{\eta}^{(0)}$, and alternates between the E-step,

$$q^{(t)} = \arg\max_{q} \mathcal{L}\left(q; \boldsymbol{\eta}^{(t-1)}\right),$$

using the variational algorithm for obtaining $q^{(t)}$ at iteration t, and the M-step,

$$\boldsymbol{\eta}^{(t)} = \arg\max_{\boldsymbol{\eta}} \mathcal{L}\left(q^{(t)}; \boldsymbol{\eta}\right),$$

until convergence of $\boldsymbol{\eta}^{(t)}$. In our case, the updates for the M-step can be obtained analytically by setting to zero the first derivative of $\mathcal{L}\left(q^{(t)}; \boldsymbol{\eta}\right)$ with respect to each component of $\boldsymbol{\eta}$. This only requires computing and differentiating the joint likelihood term $\mathrm{E}_q \log p(\boldsymbol{y}, \boldsymbol{\nu} \mid \boldsymbol{\eta})$ in (6.4), as the entropy term $-\mathrm{E}_q \log q(\boldsymbol{\nu})$ is a function of $\boldsymbol{\eta}^{(t-1)}$ and is constant with respect to $\boldsymbol{\eta}$.

Algorithm 2: Variational-EM algorithm

Define: Parameters $\boldsymbol{v}$, hyperparameters $\boldsymbol{\eta}_b = (\boldsymbol{\omega}_b, s_{0b}^2, s_b^2)$, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_B)$

1. Block variational-EM runs for hyperparameter estimation

for $b = 1, \ldots, B$ *(parallel loop)* **do**

Input: Predictors, responses and annotations for block b: $\boldsymbol{y}_b$, $\boldsymbol{X}_b$, $\boldsymbol{V}_b$.
Output: Empirical Bayes hyperparameter estimate: $\hat{\boldsymbol{\eta}}_b$

initialize: $\boldsymbol{\eta}_b^{(0)}$, $t \leftarrow 0$

repeat

$t \leftarrow t + 1$

E-step (details in Appendix D.1):

Input: Current hyperparameter value: $\boldsymbol{\eta}_b^{(t-1)}$
Output: Second-stage model variational parameters: $\boldsymbol{\mu}_\theta$, $\boldsymbol{\sigma}_\theta^2$, $\boldsymbol{\mu}_\xi$, $\boldsymbol{\sigma}_\xi^2$, $\boldsymbol{\rho}^{(1)}$ (dropping label b)

repeat

for $j = \text{shuffle}(1, \ldots, J_b)$ **do**

$$q_b\left(v_j; \boldsymbol{\eta}_b^{(t-1)}\right) \propto \exp\left\{\mathrm{E}_{-j} \log p\left(\boldsymbol{v}, \boldsymbol{y}_b \mid \boldsymbol{\eta}_b^{(t-1)}\right)\right\},$$

end

until convergence of all variational parameters;

M-step:

Input: Current variational parameter values: $\boldsymbol{\mu}_\theta$, $\boldsymbol{\sigma}_\theta^2$, $\boldsymbol{\mu}_\xi$, $\boldsymbol{\sigma}_\xi^2$, $\boldsymbol{\rho}^{(1)}$
Output: Updated hyperparameter value: $\boldsymbol{\eta}_b^{(t)}$

$$s_{0b}^2 \leftarrow \frac{1}{p_b} \sum_{s \in b} \left(\mu_{\theta,s}^2 + \sigma_{\theta,s}^2\right), \qquad p_b = |\{s : (s,t) \in b\}|$$

$$s_b^2 \leftarrow \frac{\sum_{l=1}^r \rho_l^{(1)} \left(\mu_{\xi,l}^2 + \sigma_{\xi,l}^2\right)}{\sum_{l=1}^r \rho_l^{(1)}}$$

$$\omega_{b,l} \leftarrow \rho_l^{(1)}, \quad l = 1, \ldots, r,$$

$$\boldsymbol{\eta}_b^{(t)} \leftarrow (\boldsymbol{\omega}_b, s_{0b}^2, s_b^2)$$

until convergence of $\boldsymbol{\eta}_b^{(t)}$;

$\hat{\boldsymbol{\eta}}_b \leftarrow \boldsymbol{\eta}_b^{(t)}$

end

2. Final variational run

Input: All predictors, responses and annotations: $\boldsymbol{y}$, $\boldsymbol{X}$, $\boldsymbol{V}$, empirical Bayes hyperparameter: $\hat{\boldsymbol{\eta}}$
Output: Variational parameters

repeat

for $j = \text{shuffle}(1, \ldots, J)$ **do**

$$q\left(v_j; \hat{\boldsymbol{\eta}}\right) \propto \exp\left\{\mathrm{E}_{-j} \log p\left(\boldsymbol{v}, \boldsymbol{y} \mid \hat{\boldsymbol{\eta}}\right)\right\}$$

end

until convergence of all variational parameters;

Each E-step of a variational-EM algorithm requires running the variational algorithm until convergence, and even if a large tolerance may be sufficient, these multiple complete variational runs represent a substantial overhead compared to a vanilla variational algorithm. Moreover, the two regression

levels of our model (6.1)–(6.2) necessitate the exploration of a very large parameter space, which is complex and time-consuming for any type of inference. Fortunately, the block-specification in (6.2) suggests that its hyperparameters may be estimated reasonably well by restricting the variational-EM scheme to subproblems corresponding to each block, i.e., applying model (6.1)–(6.2) to the subsets of responses $\boldsymbol{y}_b$ and predictors $\boldsymbol{X}_b$ of a given block b for obtaining the corresponding empirical Bayes estimates $\hat{\boldsymbol{\omega}}_b$, $\hat{s}^2_{0b}$ and $\hat{s}^2_b$. In addition to accelerating hyperparameter estimation for each block (as the model is much smaller), this has the advantage of allowing parallelization across blocks. Once all block hyperparameters are estimated, they are inserted into model (6.1)- (6.2) and variational inference is run on the entire dataset. A sketch of the procedure is given in Algorithm 2, and the full derivation is in Appendix D.1. We augmented all variational schemes (in the E-step and the final run) with simulated annealing, although this is not described in Algorithm 2 for brevity.

6.4 Simulations

6.4.1 Data-generation design

Generating realistic QTL data with epigenome-induced associations involves several steps, which we describe here. We first define the block association pattern and then successively build the epigenomic annotation matrix, SNPs, molecular outcomes and effect sizes. An example pattern is given in Figure 6.2; it corresponds the first replicate of a dataset used in the experiments of Sections 6.4.2 and 6.4.3.

We obtain the block association pattern as follows. We first partition the $p \times q$ SNP-outcome space into blocks of identical sizes. We then choose a proportion of "active" blocks, that is, blocks in which epigenome-induced associations will be simulated for at least one SNP-outcome pair. Finally, we choose the proportions of "active" predictors and outcomes in active blocks, and randomly select their labels in each block.

We next generate a binary epigenomic annotation matrix $\boldsymbol{V}$ as follows. We start with a $p \times r$ matrix of zeros. Then, for each active block b, we draw the number of "active" annotations (responsible for at least one SNP-response association) from a zero-truncated Poisson distribution with parameter 0.1, to obtain mostly one or two active annotations per block, and we randomly select their labels. We then assign the value unity to each entry of $\boldsymbol{V}_{b\cdot}$ corresponding to an active SNP and an active annotation. Once this is done for all blocks, we add some noise by randomly transforming zero entries into unity in the $\boldsymbol{V}$ matrix, making sure that no annotation variable is constant across all SNPs.

Given this matrix $\boldsymbol{V}$, we generate the full dataset: we first obtain the $n \times p$ matrix of SNPs $\boldsymbol{X}$, with minor allele frequencies $> 5\%$, as described in Section 3.4.1. For each block b, we then simulate the $r \times 1$ annotation regression vector $\boldsymbol{\xi}_b$, such that its nonzero entries correspond to the labels of the active annotations for b and have a log-normal distribution. Hence, the resulting non-negative annotation effects can only increase the functional potential of SNPs. We next draw auxiliary variables $z_{st} \sim \mathcal{N}(\zeta + \boldsymbol{V}_s^T \boldsymbol{\xi}_b, 1)$, for all $(s,t) \in b$, with a large negative baseline $\zeta = -2.5$ to induce overall sparsity. We build the $p \times q$ matrix $\boldsymbol{\Gamma}$ with entries $\gamma_{st} = \mathbb{1}\left(z_{st} > q_{1-\alpha}\right)$, where $q_{1-\alpha}$ is the $1-\alpha$ empirical quantile of z_{st} ($s = 1,\dots,p$, $t = 1,\dots,q$), with α, a chosen proportion of pairwise associations; this creates QTL associations that either result from epigenomic marks or are independent of these marks (possibly outside the active blocks). Finally, we generate a $p \times q$ QTL effect matrix $\boldsymbol{B}$, whose nonzero entries match those of $\boldsymbol{\Gamma}$ and are drawn from a centered normal distribution with variance set to reach desired

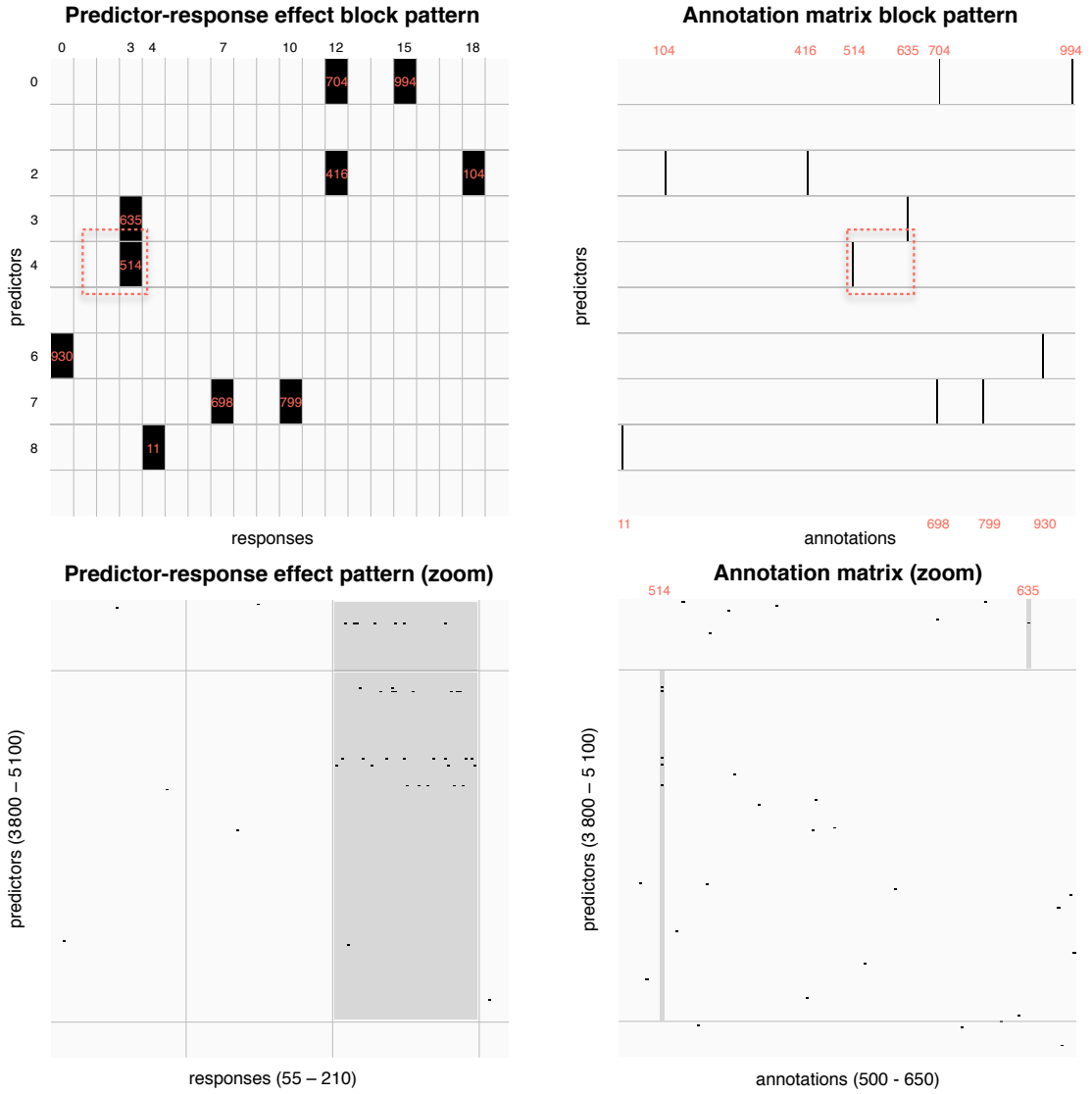

Figure 6.2 – Simulated block pattern for QTL associations and matrix of epigenomic annotations, for the first replicate of the first data-generation scenario. There are $p = 10,000$ SNPs (predictors) divided in blocks of size $1,000$, $q = 1,000$ outcomes (responses) divided in blocks of size 50, and $r = 1,000$ annotation variables (predictor-level covariates). Top left: $p \times q$ block QTL association pattern. The black rectangles indicate the active blocks, i.e., containing at least one epigenome-induced QTL association. The red labels on the blocks indicate which annotation variable is responsible for the QTL effects in the block. Top, right: $p \times r$ block pattern for the annotation matrix $\boldsymbol{V}$. The vertical lines indicate which annotation variables (columns) act on which SNP blocks (rows). Bottom left: zoom on the dashed red square (top left) of the QTL association pattern. The black marks locate the associated SNPs and outcomes. Many simulated associations are located in the active blocks (with SNPs associated with several related responses), but there are also associations located outside them, that were not induced by epigenomic annotations. Bottom, right: zoom on the dashed red square (top right) of the annotation matrix. The matrix is binary, with the black marks locating the one entries. The active variables (514 and 635) have ones at the rows corresponding to the active SNPs (see bottom left). The simulated pattern for the second data-generation scenario is similar but comprises more blocks, i.e., $1,000$ blocks of size 100×100.

effect strengths, and we build the $n \times q$ matrix $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$, where the entries of $\boldsymbol{E}$ are normal with variance unity.

In the numerical experiments of Sections 6.4.2 and 6.4.3, we consider two data-generation scenarios. The first has $q = 1{,}000$ outcomes and $p = 10{,}000$ SNPs, and the second has $q = 20{,}000$ outcomes and $p = 500$ SNPs. Both are based on $r = 1{,}000$ candidate annotation variables and $n = 100$ samples. Representative examples of simulated association pattern and annotation matrix are shown in Figure 6.2. Each experiment involves 16 replicates based on the same active blocks and annotations, but with re-simulated QTL and annotation effects.

We run all algorithms using a geometric annealing schedule with grid size 100 and initial temperature $T = 5$ (recall Section 4.3.1), and use an expected number of predictors associated with each response of $\mathrm{E}_p = 5$ (arbitrarily set to the number of active predictors per active block), and a variance of $\mathrm{V}_p = 25$ (see Appendix C.1).

6.4.2 Variable selection performance

In this section, we assess the ability of our approach to leverage the annotation data for informing the selection of SNPs and associated outcomes. We first make the simplifying assumption that the partition used for generating the data is known and provided it as input to our algorithm. We compare our proposal with two approaches; the first is the global-local method of Chapter 5, which doesn't incorporate annotation marks, and the second is based on a variant of our annotation model, that doesn't involve any partitioning or empirical Bayes estimation, namely, where the second-stage model (6.2) is replaced by

$$\begin{aligned} \gamma_{st} \mid \theta_s, \zeta_t, \boldsymbol{\xi} &\sim \text{Bernoulli}\left\{\Phi(\theta_s + \zeta_t + \boldsymbol{V}_s^T \boldsymbol{\xi})\right\}, \\ \xi_l \mid \rho_l &\sim \rho_l \mathcal{N}(0, s^2) + (1 - \rho_l)\,\delta_0, \qquad \theta_s \sim \mathcal{N}(0, s_0^2), \qquad \zeta_t \sim \mathcal{N}(n_0, t_0^2), \\ \rho_l &\sim \text{Bernoulli}(\omega_l), \qquad \omega_l \sim \text{Beta}(a_l, b_l), \qquad l = 1, \ldots, r, \end{aligned} \tag{6.5}$$

fixing $s_0^2 = s^2 = 0.1$ and using $a_l = b_l = 0.5$ (Jeffrey hyperprior), so the inclusion or exclusion of annotation variables is *a priori* equally likely.

Figure 6.3 indicates that our proposal perform best, which is unsurprising given that it relies on the splitting used for data generation. This nevertheless confirms that the model could effectively exploit the annotation data to improve the estimation of molecular QTL associations. This performance is the result of a good recovery of the annotation variables relevant to each active block. Figure 6.4 shows the marginal posterior inclusion probabilities for annotations, $\mathrm{pr}(\rho_{b,l} = 1 \mid \boldsymbol{y})$, averaged over the 16 replicates of the first data scenario. The method produced a few false positives for some replicates, but the correct annotations had the highest detection rates. No false positives were produced for the second data scenario, and all active annotations had average probabilities very close to unity (not shown). Hence, inferences from our model not only serve to improve the detection of QTL associations, but they also offer appealing posterior summaries to answer the additional biological question of which annotation variables are important for which blocks of SNPs and expression outcomes. A drawback of the empirical Bayes estimation of $\boldsymbol{\omega}_b$ is that uncertainty is not properly propagated to the annotation posterior inclusion probabilities, which typically collapsed to either zero or unity.

Variant (6.5) improves upon the global-local approach with no annotation information in the large-response scenario, but loses power in the large-predictor scenario; in this latter case, the method

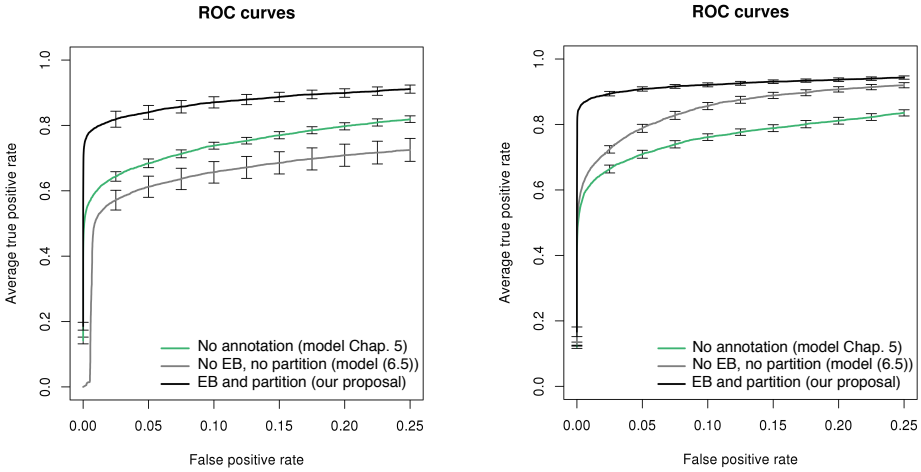

Figure 6.3 – Truncated average ROC curves for predictor-response selection with 95% confidence intervals obtained from 16 replications. Our proposal is compared with the global-local approach based on model (5.9) (no annotation information) and the model variant (6.5) (annotation effects across all predictor and response variables). Left: first dataset with $p = 10,000$ predictors, $q = 1,000$ responses, $r = 1,000$ annotations and $n = 100$ samples. Right: second dataset with $p = 500$ predictors, $q = 20,000$ responses, $r = 1,000$ annotations and $n = 100$ samples.

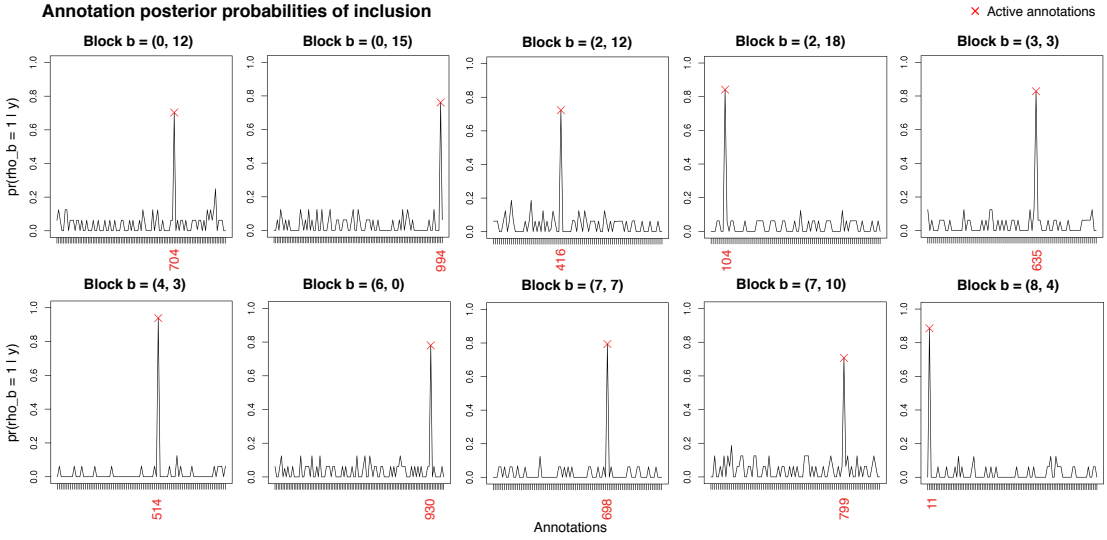

Figure 6.4 – Marginal posterior probabilities of inclusion for annotation variables, $\text{pr}(\rho_{b,l} = 1 \mid \boldsymbol{y})$, averaged over the 16 replicates of the first data scenario. Only the values for the ten active blocks are shown; those for the inactive blocks are all very close to zero. The locations of the active blocks on the QTL association pattern are shown in Figure 6.2. The problem involves $r = 1,000$ annotations, but, as mentioned in Section 6.2.1, a practical side advantage of using a partition-based model is that constant and collinear annotations for a given block can be removed prior to estimating the effects for each block; this reduced the dimensionality of the latent indicator $\boldsymbol{\rho}_b$ by one order of magnitude in our experiments (number of ticks of x-axis).

fails to distinguish the relevant annotations from the noise and includes many irrelevant annotations in the model. The simulated QTL association effects are weak and the patterns are very sparse; in particular, the simulated proportion of active SNPs is smaller in the large-predictor scenario than

in the large-response scenario. As a result, very little information effectively triggers the annotation parameters when these are modelled across all SNP and outcome variables. Moreover, the annotation effects were simulated as block-specific, while the model assumes that the effects are shared across all blocks.

Finally, we have also considered a model where the spike-and-slab prior in (6.5) is replaced by a pure slab prior, similarly to Quintana and Conti (2013). Besides being less appealing as direct selection of annotations is impractical, this approach struggles to separate relevant from irrelevant annotations when the number of annotations exceeds ≈ 100. Moreover, the algorithm takes very long to converge, owing to the absence of sparsity constraints in the second-stage regression.

6.4.3 Sensitivity to model misspecifications

We next assess the sensitivity of our proposal to the choice of partition. We ran the method on the second data scenario of Section 6.4.2, providing to the algorithm a partition which is either finer or coarser than the "true" partition used to generate the data. More precisely, the data-generation design partitioned the SNPs into $b_x = 5$ batches and the responses into $b_y = 200$ batches, resulting in 1,000 blocks, and the misspecified partitions use $b_x \pm b_x/5$ and $b_y \pm b_y/5$, yielding 1,440 blocks for the finer partition and 640, for the coarser partition. We also evaluate making inference using an annotation matrix whose entries are drawn from noise, more precisely, from a Bernoulli distribution with probability 0.5.

Figure 6.5 indicates that the misspecified partitions substantially impact inferences, and more so with the finer partition. Further experiments are needed to fully assess this, on data with various sparsity levels, patterns and strengths of QTL and annotation effects, yet these preliminary results highlight the importance of obtaining good partition estimates.

It is reassuring that inferences based on the annotation matrix with $r = 1{,}000$ noise variables did not result in poorer selections than the global-local model of Chapter 5, with no annotation information. Hence, were the annotation data irrelevant or the partition suboptimal, there seems to be no risk of worsening inferences using the former model, compared to the latter model; this alleviates slightly our concerns on the choice of partition.

6.4.4 Empirical Bayes estimation without partitioning

Finally, we describe a small experiment to illustrate a degeneracy phenomenon that can affect empirical Bayes estimation in highly sparse association problems. Scott and Berger (2010) highlighted such degeneracy in variable selection contexts and formalized it as a consequence of marginal likelihood maximization, causing the model posterior distribution to collapse to either the null model or the full model. In our case, the issue arises when considering an empirical Bayes variant of our proposal (6.2), with no partition, i.e.,

$$\begin{aligned}
\gamma_{st} \mid \theta_s, \zeta_t, \boldsymbol{\xi} &\sim \text{Bernoulli}\left\{\Phi(\theta_s + \zeta_t + \boldsymbol{V}_s^T \boldsymbol{\xi})\right\}, \\
\xi_l \mid \rho_l &\sim \rho_l \mathcal{N}(0, s^2) + (1-\rho_l)\,\delta_0, \qquad \theta_s \sim \mathcal{N}(0, s_0^2), \qquad \zeta_t \sim \mathcal{N}(n_0, t_0^2), \\
\rho_l &\sim \text{Bernoulli}\,(\omega_l), \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad l = 1, \ldots, r,
\end{aligned} \tag{6.6}$$

with ω_l, s_0^2, and s^2 obtained by variational-EM estimation.

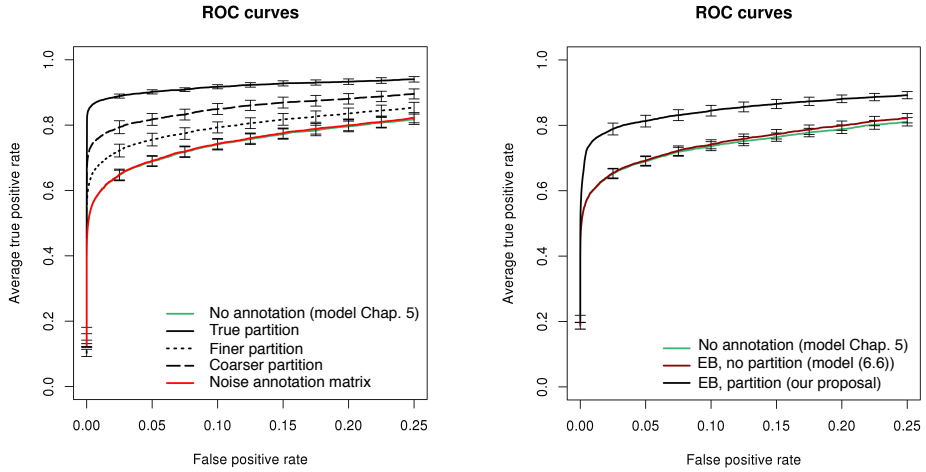

Figure 6.5 – Truncated average ROC curves for predictor-response selection with 95% confidence intervals obtained from 16 replications. Left: comparisons on the second data scenario of our proposal using the true partition and two misspecified partitions (finer and coarser), and using a noise annotation matrix. The global-local approach based on model (5.9) is also shown (the green and red curves overlap). Right: comparisons of our proposal with the variational-EM approach on model variant (6.6) and the global-local approach (the green and dark red curves overlap), on a problem with $q = 1{,}000$ responses, $p = 500$ predictors, $r = 100$ annotations and $n = 100$ samples.

The empirical Bayes variance of the annotation spike-and-slab model collapses to zero, reaching $\hat{s}^2 \approx 7 \times 10^{-7}$. The resulting estimates for the annotation effects are close to null, so selection does not benefit from modelling the annotations; see Figure 6.5. Hence, both non-partition-based models (6.5) and (6.6) are ineffective, owing the high sparsity entailed by the simulated QTL patterns. Model (6.6) has the added drawback of requiring prohibitive computational times: indeed, at each iteration of the EM algorithm, the variational algorithm has to run on the full model until convergence, while our proposal based on model (6.2) restricts the variational-EM scheme to the variables in the block under consideration. Simulations could only complete on small problems, here, with $q = 1{,}000$ responses, $p = 500$ predictors, $r = 100$ annotations and $n = 100$ samples.

6.5 Summary

We proposed a framework that accommodates annotation information on the candidate SNPs to enhance the estimation of QTL associations. Our proposal entails a second-stage regression model on the probability of association. It models the effects of annotation variables using block-specific spike-and-slab priors. This specification is biologically, statistically and computationally appealing: it accounts for the localized nature of functional mechanisms, prevents degeneracy issues caused by extreme sparsity, and permits efficient computations thanks to our parallel variational-EM scheme.

However, the model is complex, and so are the molecular mechanisms that it aims at representing. Substantial efforts are still needed to test our approach on a variety of plausible data scenarios and on real data, as well as to compare it with existing approaches. In particular, we saw in Section 6.4.3 that selection performance hinges on the adequacy of the block partition. We also saw, based on two model variants, that omitting the partition doesn't work, hence even if finding a good partition may be difficult, this seems to be a necessary prerequisite.

7 A pQTL study sheds light on the genetic architecture of obesity

This chapter presents a genome-wide application of our approach on proteomic QTL data from two clinical obesity cohorts. It reproduces the content of a paper in preparation (Ruffieux et al., 2018b), which is based on joint work with Armand Valsesia.

The model used for the proteomic analyses is that of Chapter 1, and inference is done with the annealed variational algorithm presented in Chapter 4; we hereafter call it LOCUS, as in Chapter 1. The aim of the chapter is twofold. First, it illustrates the applicability of the method for large-scale QTL analyses, and second, it evaluates the biological relevance of the uncovered pQTL associations.

Section 7.1 introduces the study and sets our goals. Section 7.2 presents our two-stage pQTL study design and provides general results. Section 7.3 discusses the findings based on associations between proteins under genetic control and clinical variables on metabolic traits, and Section 7.4 focuses on *trans*-acting pQTL associations. Section 7.5 summarizes the results and opens some perspectives. A "Methods" section, detailing the data collection, processing and analyses, is in Appendix E. The Supplementary Tables (S1–9) can be browsed online at https://heleneruffieux.com/locus-pqtl.

7.1 Introduction

Hundreds of genome-wide association studies (GWAS) have assessed the role of thousands of loci on obesity susceptibility, yet the action of genetic variation on metabolism remains poorly understood. In particular, most identified risk loci lie in intergenic regions (Ward and Kellis, 2012; Tak and Farnham, 2015), which complicates their functional interpretation. The analysis of intermediate expression traits or *endophenotypes*, via molecular quantitative trait locus (QTL) studies, may provide greater insight into the biology underlying clinical traits. While gene expression QTL (eQTL) studies are nowadays routinely performed, protein expression QTL (pQTL) studies have emerged more recently, prompted by new large-scale production capabilities for proteomic samples. These studies may be particularly appropriate for exploring the functional bases of obesity. Indeed, previous work reported substantial variations in protein expression between obese and normal-weight individuals, as well as among obese individuals, and suggested that proteins may act as proxies for diverse metabolic endpoints (e.g., López-Villar et al., 2015; Garrison et al., 2016; Hess et al., 2018; Thrush et al., 2018).

However two major hurdles hamper pQTL analyses. First, owing to the number of tests that they entail, conventional univariate approaches fail to uncover weak effects, such as those between a variant

trans-acting on remote gene products. This is troublesome because *trans* and related pleiotropic effects are believed to have important roles in regulating complex traits (Solovieff et al., 2013; Yao et al., 2017). Practitioners are well aware of this weakness, yet existing multivariate methods, suited to the detection of *trans* effects, fail to scale to the sizes required by current molecular QTL studies, so simplistic univariate approaches still prevail. Second, the clinical data complementing QTL data are often very limited, so the biomedical relevance of QTL signals is often examined by cross-matching external information from unrelated populations, health status or study designs, which renders some degree of speculation unavoidable.

Here we attempt to address both concerns in an integrative study of two obesity clinical cohorts, the Optifast Ottawa ($n = 1,644$, Dent et al., 2002) and DiOGenes ($n = 789$, Larsen et al., 2010) cohorts; two pQTL datasets are available for each cohort, with protein expression levels quantified in plasma by mass-spectrometry and aptamer-based SomaLogic assays (Kraemer et al., 2011), respectively.

We present the first multivariate genome-wide pQTL analysis, tailored to the detection of weak *trans* regulatory effects. We use our variational Bayesian method LOCUS (Ruffieux et al., 2017), which simultaneously accounts for all the genetic variants and proteomic outcomes, thereby leveraging the similarity across proteins controlled by shared regulatory mechanisms. We analyze the pQTL data in a two-stage design: we apply LOCUS on both pQTL datasets of the Ottawa cohort, and replicate > 80% of the hits in the independent DiOGenes cohort; our rich mass-spectrometry and SomaLogic proteomic data permit both cross- and intra-platform validations.

Pertinent clinical interpretation of pQTL effects for obese individuals hinges on a careful examination of metabolic endpoints in the population in question. Here, we highlight the biomedical potential of several validated pQTL hits, using comprehensive clinical data from the two pQTL obesity cohorts. Our methods and results shed light on obesity and obesity co-morbidities, and suggest novel candidate biomarkers for the metabolic syndrome.

A workflow of the pQTL study is given in Figure 7.1d, and all pQTL and clinical association results are freely available from our online browser (https://heleneruffieux.com/locus-pqtl). The performance of our multivariate method LOCUS was extensively assessed in Ruffieux et al. (2017) via comparisons with state-of-the-art QTL methods. Figure 7.1a–c and Appendix E.1.6 also give an overview of LOCUS, and Appendix E.2 provides an illustration of its increased statistical power over univariate analyses in simulations involving genetic variants from the Ottawa cohort and synthetic outcomes emulating the proteomic data.

7.2 Two-stage multivariate pQTL analyses

7.2.1 Discovery with the Ottawa cohort

We applied LOCUS on two pQTL datasets, involving plasma samples from the Ottawa obesity cohort. The first dataset comprised untargeted mass spectrometry (MS) measurements for 133 proteins, and the second comprised measurements for 1,096 proteins obtained with the aptamer-based multiplex technology SomaLogic (Kraemer et al., 2011). We analyzed about 275,000 common variants (minor allele frequency > 5%), for nearly 400 subjects, adjusting for age, gender and body mass index (BMI); see Figure 7.1d.

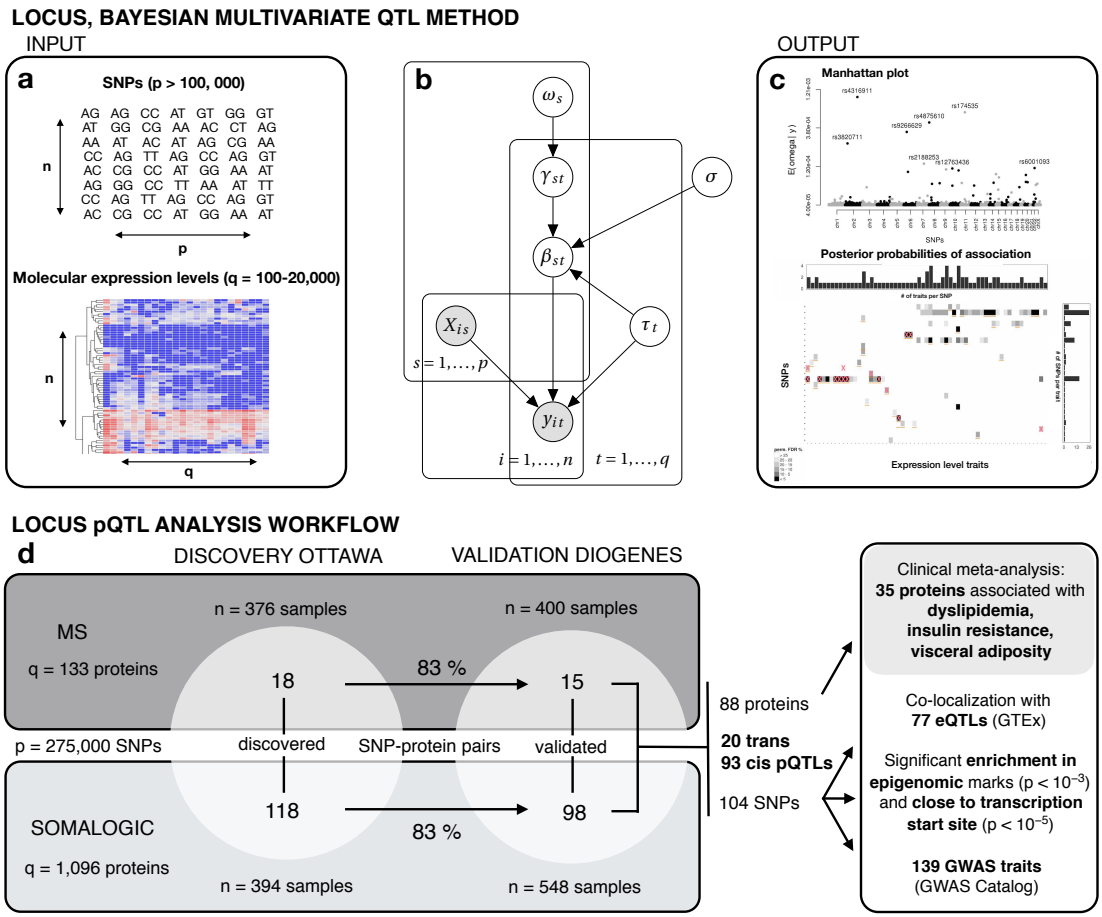

Figure 7.1 – LOCUS model overview and study workflow. (a) Inputs to LOCUS are an $n \times p$ design matrix $\boldsymbol{X}$ of p SNPs, and an $n \times q$ outcome matrix $\boldsymbol{y}$ of q molecular traits, e.g., gene, protein, lipid or methylation levels, for n individuals. The model is multivariate; it accounts for all the SNPs and molecular traits jointly. (b) Graphical model representation of LOCUS. The effect size between a SNP s and a trait t is modelled by β_{st}, and γ_{st} is a latent variable taking value unity if they are associated, and zero otherwise. The parameter ω_s controls the pleiotropic level of each SNP, i.e., the number of traits with which it is associated. The parameter σ represents the typical size of effects, and the parameter τ_t is a precision parameter that relates to the residual variability of each trait t. LOCUS enforces sparsity on the QTL effects, so it identifies just one or few markers per relevant locus, even in regions of high linkage disequilibrium (LD). By design, univariate screening approaches do not exploit association patterns common to multiple outcomes or markers; they analyze the outcomes one by one, and do not account for LD structures, thereby highlighting redundant signals at loci with strong LD structures (see, e.g., Figure 7.2). (c) Outputs of LOCUS are posterior probabilities of associations, $\text{pr}(\gamma_{st} = 1 \mid \boldsymbol{y})$, for each SNP and each trait ($p \times q$ panel), and posterior means for the pleiotropy propensity of each SNP, $\text{E}\left(\omega_s \mid \boldsymbol{y}\right)$ (Manhattan plot). (d) Workflow of the pQTL study. The mass-spectrometry and SomaLogic pQTL data are analysed in parallel. LOCUS is applied on the Ottawa data for discovery, and 83% of the 18 and 118 pQTL associations discovered with the MS and SomaLogic data replicate in the independent study DiOGenes. The relevance of the validated pQTLs in the obese population is assessed via analyses of clinical parameters from the Ottawa and DiOGenes cohorts. Further support is obtained by evaluating colocalization with eQTLs, epigenomic marks and GWAS traits.

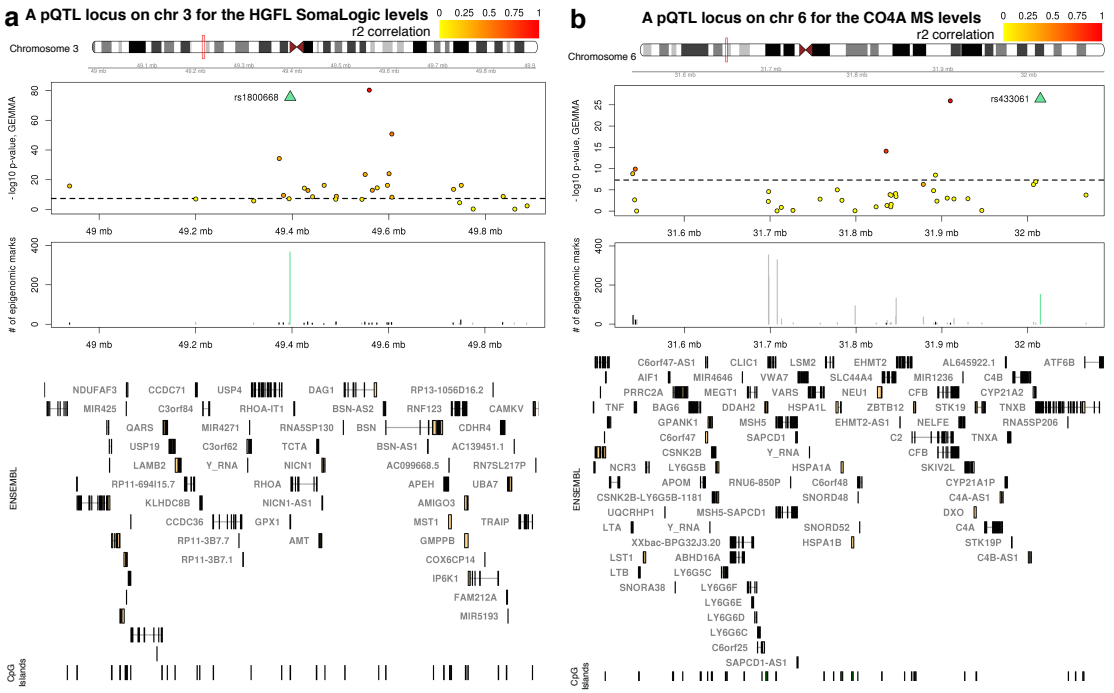

Figure 7.2 – Regional association plots for two loci, identified by the (a) SomaLogic and (b) MS pQTL analyses. In each case, the top panel displays the nominal $-\log_{10} p$-values obtained when re-analyzing the region with GEMMA; the dashed horizontal line corresponds to $p = 5 \times 10^{-8}$. The top SNP identified by LOCUS is marked with a green triangle, and its correlation in r^2 with the surrounding SNPs is indicated by the yellow to red colors. The second-row panel shows the cumulated numbers of annotation marks for each SNP. The green bars correspond to LOCUS top SNPs, the black bars, to the SNPs with GEMMA p-values $< 5 \times 10^{-8}$. The bottom panel shows the transcript and CpG island positions.

At false discovery rate (FDR) 5%, the MS analysis identified 18 pQTL associations, corresponding to 14 unique proteins and 18 SNPs, while the SomaLogic analysis identified 118 pQTLs, corresponding to 99 proteins and 111 SNPs; the full list is in Suppl. Table S1.

To illustrate the relevance of LOCUS multivariate selections from regions with strong linkage disequilibrium (LD) structures, we briefly confronted them on two complex *cis*-pQTL loci with the univariate selections of GEMMA, a single-SNP/single-outcome linear mixed model approach by Zhou and Stephens (2014).

The first locus, on chromosome 3, displayed associations with the SomaLogic levels of the HGFL (hepatocyte growth factor-like) protein, encoded by the macrophage-stimulating *MST1* gene (Figure 7.2a). At FDR 5%, LOCUS reported two variants associated with the HGFL levels, namely rs1800668 and rs56116382. Re-analysing the region with GEMMA highlighted a large block of correlated SNPs with very low p-values; twenty had nominal p-value $< 5 \times 10^{-8}$. The pQTLs selected by LOCUS corresponded to the second and third most significant hits of GEMMA. One of these two, rs1800668, is located 326 Kb upstream of the *MST1* gene, within a gene-dense region (> 40 genes). It had the highest overlap in epigenomic annotation marks, 336 out of 450 marks, which corresponded to significant enrichment ($p = 2.06 \times 10^{-3}$). SNP rs1800668 is also a known eQTL for many genes (25 including *MST1*) in several tissues, including adipose, brain, muscle, skin and blood (Suppl. Table S2). Interestingly, public pQTL studies reported associations of this SNP with 23 distinct proteins (Suppl. Table S3), but not with HGFL.

The second example concerns a locus in the MHC region, with evidence of *cis* regulation on the CO4A MS protein levels (Figure 7.2b). The effects reported by GEMMA were less extreme than in the previous example; they were also sparser, as the SNPs were in lower LD (six SNPs had $p < 5 \times 10^{-8}$). At FDR 5%, LOCUS selected two variants associated with the CO4A protein levels, one of which, rs433061, corresponded to the top GEMMA hit ($p = 3.94 \times 10^{-27}$). This variant colocalized with 156 epigenomic marks (enrichment, $p = 0.0184$) and was 442 base pairs away from a transcription start site (enrichment, $p = 0.0253$). This SNP is also a known *cis* eQTL for the *C4A* gene in many tissues, including liver, arteries and adipose tissue, as well as for > 70 other transcripts (Suppl. Table S2). Moreover, it has already been described as associated with the CO4A plasma levels, as well as with 27 other proteins (Suppl. Table S3), suggesting a pleiotropic role.

Our comparisons of LOCUS and GEMMA at these two loci indicated that the parsimonious selections of LOCUS retained SNPs that colocalize with many epigenomic marks and eQTLs, which supports possible regulatory roles. These SNPs were also top hits of GEMMA; comparisons are more difficult at loci with *trans* associations, which are typically weaker and more likely to be missed by univariate approaches.

7.2.2 Replication with the DiOGenes cohort

We next undertook to replicate the uncovered pQTLs in the independent DiOGenes cohort, using two datasets with proteins quantified by MS and SomaLogic for $n = 400$ and $n = 548$ subjects, respectively (Figure 7.1d). The DiOGenes cohort differed substantially from the discovery Ottawa cohort; in particular, the Ottawa subjects were significantly more obese, and had higher lipid levels and insulin resistance (Suppl. Table S4). These differences were expected, as the Ottawa trial enrolled patients from a specialized obesity clinic whereas the DiOGenes trial included overweight and obese non-diabetic subjects.

We validated 15 of the 18 discovered MS pQTLs, and 98 of the 118 discovered SomaLogic pQTLs at FDR 5% (Suppl. Table S5), yielding a replication rate of 83% in both cases. While the two platforms had inherent technological differences, 72 proteins were quantified by both and were used in a cross-platform replication. Eight pQTLs identified with our MS analyses could be assessed with SomaLogic (i.e., had available protein levels), and 7 of these pQTLs replicated at FDR 5%. Likewise, of the 20 SomaLogic associations having MS measurements, 14 were confirmed, demonstrating appreciable replication with distinct technologies.

We also evaluated replication rates separately for *cis* and *trans* effects. With the MS data, all 15 *cis* pQTLs could be replicated, while the 3 *trans* pQTLs could not. With the SomaLogic data, 78 of 81 *cis* and 20 of 37 *trans* pQTLs could be validated; the latter are shown in Figure 7.3a. We reached overall replication rates of 97% for *cis*-pQTL and 50% for *trans* pQTLs; the *trans*-pQTL rate is in line with other pQTL studies (Suhre et al., 2017; Sun et al., 2018; Yao et al., 2018a).

Finally, we found that 73 of our validated pQTLs overlap with pQTLs previously identified in the general population (using proxy search $r^2 > 0.8$, and reporting associations at $p < 1 \times 10^{-5}$; Suppl. Table S6), yet several published hits had no independent replication. The remaining 40 pQTLs are, to our knowledge, new.

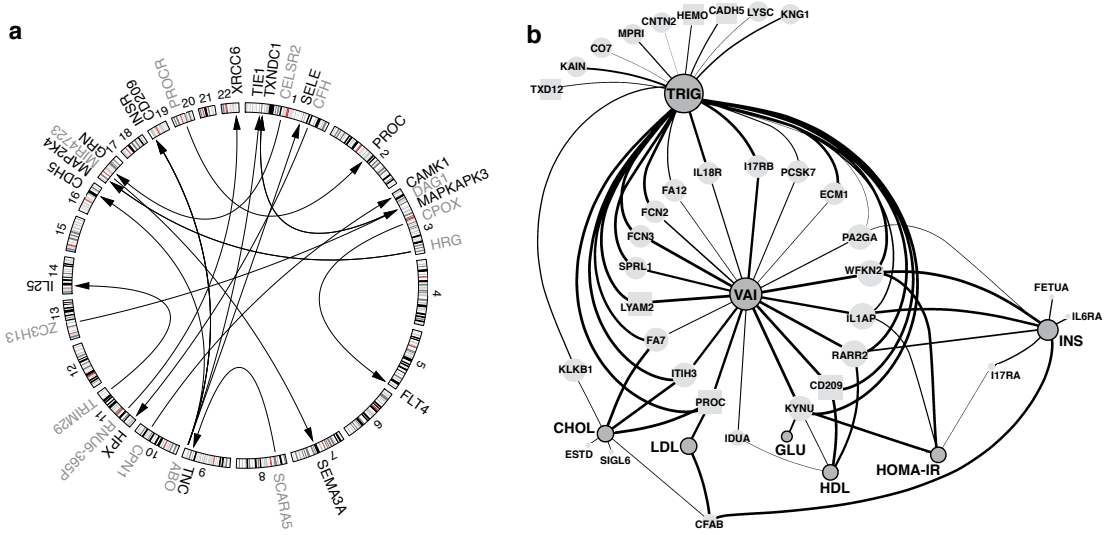

Figure 7.3 – *Trans*-pQTLs and associations of proteins under genetic control with clinical parameters. (a) Circular plot showing the *trans*-pQTL associations uncovered by LOCUS (FDR < 5%). Each arrow starts from the pQTL SNP with label indicating its closest gene (grey) and points to the gene (black) coding for the controlled protein. (b) Network displaying the associations (FDR < 5%) between protein levels and clinical variables obtained by meta-analysis, adjusting for age, gender and BMI. Nodes for clinical parameters are in dark grey with black borders; proteins are in light grey, and type of genetic control, *cis* or *trans*, is depicted with round or square nodes, respectively. The edge thickness is proportional to the significance of association, and the node size is proportional to its connectivity.

7.2.3 Colocalization with eQTLs and evidence for regulatory impact

We next assessed the overlap of the 113 validated pQTLs with known eQTLs using the GTEx portal (GTEx Consortium, 2015). Seventy-seven of the 104 SNPs involved in our pQTL associations had one or more eQTL associations in at least one tissue. These SNPs have been implicated in 83 eQTL associations, and this represented a significant enrichment ($p < 2.2 \times 10^{-16}$). Forty-nine of the 77 SNPs were eQTL variants for the gene coding for the protein with which they were associated in our datasets; they were involved in 68 pairwise associations out of the 83. These colocalization patterns lend support to potential regulatory consequences.

Further evidence emerges from inspecting the epigenome: our pQTLs were enriched in epigenome annotation marks ($p = 9.20 \times 10^{-4}$) and significantly closer to transcription start sites compared to randomly chosen SNP sets ($p = 9.99 \times 10^{-6}$).

7.2.4 Colocalization with disease GWAS loci

A total of 217 previously reported genome-wide associations overlapped the pQTL regions (GWAS Catalog, Welter et al., 2014); this represented 139 unique traits or diseases mapping to 68 distinct regions (based on LD $r^2 > 0.8$). Nineteen *sentinel* SNPs, i.e., SNPs specifically identified by LOCUS pQTL analyses, were directly involved in these associations (Suppl. Table S8). These results generate useful hypotheses to be explored in future research. We now provide two examples of these.

The first concerns our aforementioned HGFL pQTL, rs1800668; this SNP is in strong LD ($r^2 > 0.95$) with rs9858542 and rs3197999, which are known to associate with Crohn's disease (Wellcome Trust Case Control Consortium, 2007; Franke et al., 2010; Liu et al., 2015). While gene causality remains to be demonstrated, our pQTL finding may be of clinical relevance given the prevalence of Crohn's disease in overweight and obese subjects (Nic Suibhne et al., 2013; Singh et al., 2017); the region would merit additional follow-up in inflammatory bowel disease cohorts.

Our second example concerns an association between rs3865444 and the Siglec-3 protein, whose coding gene, *CD33*, has been reported as a risk factor for Alzheimer's disease (Naj et al., 2011; Lambert et al., 2013; Sims et al., 2017). As subjects obese in midlife are more at risk to develop late-life Alzheimer's (Xu et al., 2011; Lambert et al., 2013), this pQTL may help to understand the genetic bases of Alzheimer's disease and dementia; its potential as a prognosis biomarker should be studied in Alzheimer's cohorts, ideally using weight records.

7.3 Proteins as endophenotypes for the genetics of obesity

We next performed a more systematic evaluation of the clinical relevance of the replicated pQTLs for obesity by studying the links of all proteins under genetic control with different clinical parameters.

Queries in existing databases and literature searches suggested that most of these proteins have implications in inflammation, insulin resistance, lipid metabolism or cardiovascular disorders. While such public GWAS results often help to generate hypotheses on the clinical relevance of QTL hits, we next performed meta-analyses using clinical data from the DiOGenes and Ottawa cohorts to directly assess the associations with metabolic endpoints in the obese population considered, based on the same study design and proteomic measurements (tissue, technology, samples) as those used for discovering the pQTLs. The data include glycemic, total lipid measurements and values of the *visceral adiposity index* (VAI), a clinically established measure of visceral fat (Amato et al., 2010). Our analyses indicated that, of the 88 proteins under genetic control, 35 were associated with at least one such endpoint at FDR 5%, adjusting for age, gender and BMI, with consistent directions of effects in the two cohorts (Suppl. Table S9). Moreover, these associations should be attributable metabolic factors independent of overall adiposity, since we controlled for BMI as a potential confounder. Figure 7.3b represents the associations as a network; most of the proteins were associated with several clinical endpoints. The triglyceride and VAI variables had the highest degree of connectivity and were connected with measures of insulin resistance and other lipid traits via proteins such as FA7, IL1AP, KYNU, PROC, RARR2 and WFKN2. The proteins CFAB, FETUA, PA2GA had lower connectivity or significance, but were nevertheless highlighted in the context of obesity and its complications (Goustin and Abou-Samra, 2011; Monroy-Muñoz et al., 2017; Matsunaga et al., 2018).

Finally, several *trans*-regulated proteins were implicated in clinical associations: CADH5, CD209 and LYAM2, all controlled by the pleiotropic *ABO* locus; HEMO (Hemopexin), a liver glycoprotein controlled by the *CFH* locus, itself coding for another liver glycoprotein; PROC controlled by its own receptor *PROCR*; and TXD12 (thioredoxin domain containing 12), controlled by the *DAG1/BSN* locus.

The pQTL associations involving proteins with clinical associations at FDR 5% are listed in Table 7.1. The subsequent sections discuss the possible functional and biomedical relevance of a selection of pQTL associations based on their connection with clinical variables, as summarized by Figure 7.3b. Forest plots for this selection are given in Figure 7.4 to help visualize the directions of effects. Unless

Protein	Protein name	Clin.	SNP	Chr	Position	LOCUS PPI	pQTL validation p-value
CADH5	**Cadherin-5**	**L**	**rs8176741**	**9**	**136131461**	**1.00**	$\mathbf{4.68 \times 10^{-30}}$
CD209	**DC-SIGN**	**L/V**	**rs8176741**	**9**	**136131461**	**1.00**	$\mathbf{7.74 \times 10^{-10}}$
			rs2519093	**9**	**136141870**	**1.00**	$\mathbf{6.24 \times 10^{-26}}$
CFAB	Factor B	G/L	rs150132450	6	31906334	0.85	9.61×10^{-4}
			rs641153	6	31914180	1.00	3.92×10^{-12}
CNTN2	CNTN2	L	rs11240396	1	205205081	1.00	6.82×10^{-14}
CO7	C7	L	rs71623870	5	40966676	0.83	4.03×10^{-4}
ECM1	ECM1	L/V	rs34964511	1	150298015	1.00	3.77×10^{-6}
			rs71578487	1	150340059	1.00	1.07×10^{-11}
			rs72696900	1	150425256	0.82	1.5×10^{-6}
			rs11802612	1	150427279	1.00	3.7×10^{-6}
			rs35094010	1	150449557	1.00	3.76×10^{-6}
ESTD	Esterase D	L	rs73193065	13	47383681	0.90	2.31×10^{-15}
FA12	Coagulation factor XII	L/V	rs55785724	5	176817583	1.00	1.34×10^{-5}
FA7	Coagulation Factor VII	L/V	rs3093233	13	113758130	1.00	3.11×10^{-88}
FCN2	FCN2	L/V	rs3811140	9	137772111	1.00	9.66×10^{-14}
FCN3	Ficolin-3	L/V	rs10902652	1	27558522	1.00	1.62×10^{-3}
FETUA	a2-HS-Glycoprotein	G	rs2593813	3	186332571	1.00	2.47×10^{-10}
			rs2593813	3	186332571	1.00	4.51×10^{-8}
HEMO	**Hemopexin**	**L**	**rs10801560**	**1**	**196714600**	**1.00**	$\mathbf{2.36 \times 10^{-26}}$
I17RA	IL-17 sR	G	rs738035	22	17594886	1.00	1.48×10^{-20}
I17RB	IL-17B R	L/V	rs35518479	3	53873814	0.76	9.98×10^{-6}
IDUA	IDUA	L/V	rs10017289	4	943534	1.00	1.22×10^{-11}
IL18R	IL-18 Ra	L/V	rs3836108	2	103037742	1.00	5.22×10^{-26}
IL1AP	IL-1 R AcP	G/L/V	rs724608	3	190348810	1.00	8.7×10^{-114}
IL6RA	IL-6 sRa	G	rs4845372	1	154415396	1.00	1.72×10^{-81}
ITIH3	Inter-alpha-trypsin inhibitor heavy chain H3	L/V	rs736408	3	52835354	0.97	1.46×10^{-6}
KAIN	Kallistatin	L	rs5511	14	95033595	1.00	9.9×10^{-24}
KLKB1	Prekallikrein	L	rs80177406	4	187166024	0.99	3.54×10^{-6}
KNG1	Kininogen HMW	L	rs1621816	3	186439173	1.00	1.44×10^{-13}
KYNU	KYNU	G/L/V	rs6741488	2	143793701	1.00	3.22×10^{-20}
LYAM2	**sE-Selectin**	**L/V**	**rs2519093**	**9**	**136141870**	**1.00**	$\mathbf{6.81 \times 10^{-62}}$
LYSC	Lysozyme	L	rs71094714	12	69790495	1.00	8.41×10^{-19}
MPRI	IGF-II receptor	L	rs3777411	6	160476945	1.00	4.95×10^{-11}
PA2GA	NPS-PLA2	G/L/V	rs6672057	1	20293791	1.00	3.86×10^{-15}
PCSK7	PCSK7	L/V	rs11216284	11	117003060	1.00	8.17×10^{-31}
PROC	**Protein C**	**L/V**	**rs141091409**	**20**	**33739915**	**0.43**	$\mathbf{1.66 \times 10^{-18}}$
RARR2	TIG2	G/L/V	rs1047586	7	150035459	0.96	2.39×10^{-11}
SIGL6	Siglec-6	L	rs77561179	19	52029477	1.00	3.39×10^{-14}
SPRL1	SPARCL1	L/V	rs7681694	4	88462729	0.99	5.70×10^{-14}
TXD12	**TXD12**	**L**	**rs13062429**	**3**	**49559485**	**1.00**	$\mathbf{2.26 \times 10^{-5}}$
			rs34519883	**3**	**49575831**	**1.00**	$\mathbf{5.39 \times 10^{-33}}$
WFKN2	WFKN2	G/L/V	rs9303566	17	48922281	1.00	3.38×10^{-11}

Table 7.1 – Proteins associated with clinical parameters (Figure 7.3b) and controlled by pQTL variants. All associations were detected at FDR < 5%. Associations with glycemic traits (fasting glucose, insulin, HOMA-IR) are indicated by *G*, with total lipid traits (HDL, LDL, triglycerides, total cholesterol), by *L*, and with visceral fat (visceral adiposity index), by *V*. *Trans* pQTL associations are in bold.

otherwise specified, all associations described have meta-analysis FDR corrected p-value below 5%, and we provide their nominal p-values in parentheses.

7.3.1 CFAB and RARR2, mediators of adipogenesis are under genetic control

Our analyses suggested that the CFAB (complement factor B) and RARR2 (Retinoic acid receptor responder protein 2) levels associate with distinct clinical parameters (Figures 7.3b and 7.4), yet both

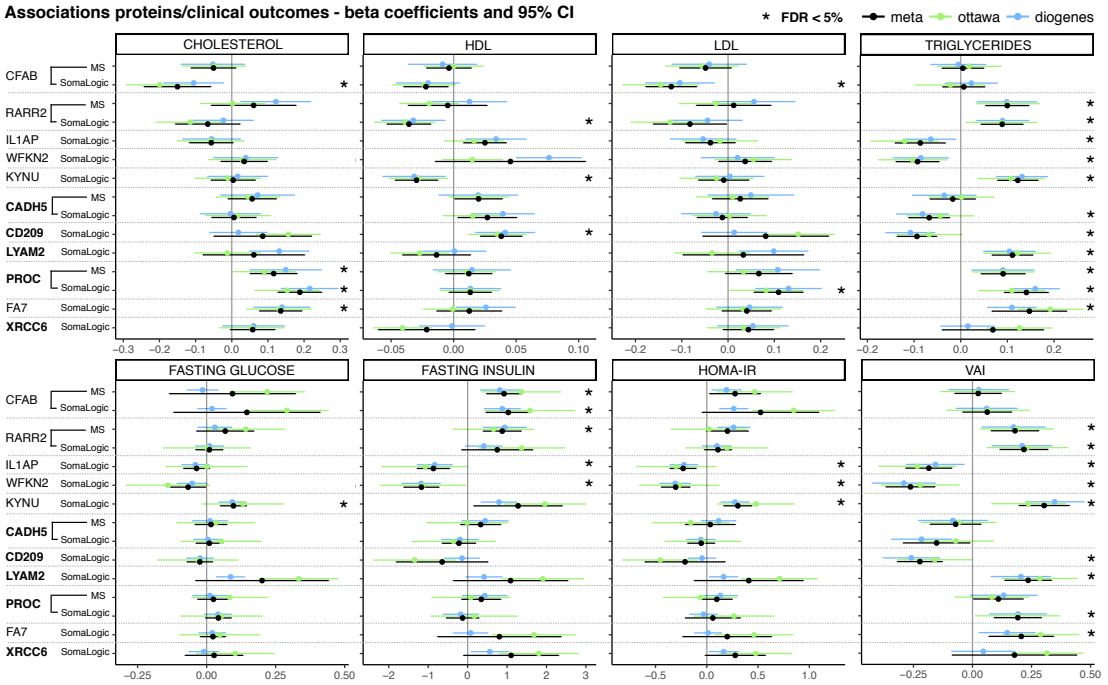

Figure 7.4 – Forest plots for associations between a selection of proteins under genetic control and clinical parameters, adjusting for age, gender and BMI (Appendix E.1.11). All endpoints are measured in both the Ottawa and DiOGenes cohorts; they correspond to total lipid levels (first row: total cholesterol, HDL, LDL, triglycerides), glucose/insulin resistance (second row: fasting glucose, fasting insulin, HOMA-IR) and VAI. In each case, regression coefficients with 95% confidence intervals are shown for the Ottawa and DiOGenes analyses, and for the meta-analysis. The stars indicate associations with meta-analysis FDR < 5% (correction applied across all proteins under genetic control, not only those displayed; see also Figure 7.3b). The order of appearance of the proteins follows that in the text. For proteins with measurements in the MS and SomaLogic platforms, association results are displayed for both; *trans*-regulated proteins are in bold.

play a role in adipogenesis and hence are particularly interesting in the context of obesity and related co-morbidities.

The CFAB protein controls the maturation of adipocytes (Matsunaga et al., 2018). Existing literature describes the *CFB* gene, coding for CFAB, as positively associated with BMI and insulin resistance, and negatively associated with HDL (Moreno-Navarrete et al., 2010; Coan et al., 2017). We confirmed and complemented these observations using the MS and SomaLogic measurements. Both were positively associated with BMI (MS: $p = 2.08 \times 10^{-8}$, SomaLogic: $p = 2.23 \times 10^{-13}$) and with fasting insulin (adjusting for BMI; MS: $p = 4.45 \times 10^{-5}$, SomaLogic: $p = 3.44 \times 10^{-4}$). The CFAB SomaLogic levels were also negatively associated with cholesterol ($p = 1.43 \times 10^{-3}$), LDL ($p = 1.30 \times 10^{-5}$), and with HDL at higher FDR (nominal $p = 1.47 \times 10^{-2}$, corrected $p = 0.11$).

Our pQTL study allows us to delve into the genetic roots of these associations; both the MS and SomaLogic analyses independently highlighted the same *cis*-acting locus as putative regulator of the CFAB protein. In particular, the sentinel pQTL SNP reported in the SomaLogic analysis, rs641153, is a missense variant located in the MHC region, 180 base pairs away from a transcription binding site (significantly closer than other SNPs, $p = 1.16 \times 10^{-2}$). Further investigation using JASPAR (Khan et al.,

2018) and SNP2TFBS (Kumar et al., 2017) indicated that this SNP may affect the binding sites of four transcription factors (EBF1, TFAP2A, TFAP2C and HNFA), which lends support to a regulatory role. In GTEx, rs641153 is described as an eQTL for the *NELFE* and *SKIV2L* genes in multiple tissues (including adipose tissue, aorta, colon, coronary artery and muscle), but not for the *CFB* gene.

RARR2 is encoded by an essential adipogenesis gene, *RARRES2*, on chromosome 7. *RARRES2* regulates glucose and lipid metabolism by altering the expression of adipocyte genes (Müssig et al., 2009). A recent study for the Taiwanese population reported pleiotropic associations of *RARRES2* variants with circulating chemerin, triglyceride levels and several measurements pertaining to inflammation (Er et al., 2018). In line with these results, we found significant associations with triglycerides, fasting insulin and HDL (Figure 7.4, Suppl. Table S9). We attempted to clarify the as-yet unclear relation between *RARRES2* and visceral fat mass in obese subjects (Müssig et al., 2009): the MS and SomaLogic RARR2 levels were strongly associated with BMI, but also positively associated with visceral fat, even when controlling for BMI (Figure 7.4; Suppl. Table S9). We further found that the association was not gender-specific (data not shown).

Our pQTL analyses also indicated a *cis* association between a missense variant, rs1047586, and RARR2. This variant was described as an eQTL for multiple genes, in several tissues (including adipose tissue, coronary artery, aorta, colon and muscle), and as highly associated with DNA methylation in whole blood, T cells, neutrophils and monocytes, and with different histone modification marks (including H3K27ac and H3K4me1 enhancers; Bonder et al., 2017).

Our analyses therefore support the relevance of CFAB and RARR2 for better understanding metabolic complications in obese subjects, and provide evidence in favour of their genetic control; both pQTL loci colocalize with several epigenomic marks.

7.3.2 The importance of IL1AP for Metabolic Syndrome

The IL-1 pathway plays a critical role in the immune-response associated with obesity and type 2 diabetes (Tack et al., 2012; Banerjee and Saxena, 2012); other IL-1 related cytokines, such as IL-1ra, are also well documented in the context of type 1 and type 2 diabetes (Pfleger et al., 2008; Böni-Schnetzler et al., 2018). The IL1AP (IL-1 receptor accessory) protein is a co-receptor of the IL-1 receptor, whose relevance has been discussed for other diseases, such as cancer (Barreyro et al., 2012) and Alzheimer's (Ramanan et al., 2015). The plasma levels of the soluble isoform of IL1AP have also been described as reduced in obese subjects, and were found as associated with SNPs rs2885373 and rs6444444 (Bozaoglu et al., 2014). Sun et al. (2018) identified an association with rs724608, a SNP in strong LD with these two SNPs ($r^2 = 0.93$), but did not have independent replication.

Our two-stage analysis confirmed the association of rs724608 with the SomaLogic plasma levels of IL1AP. We also found associations between IL1AP expression and measures of fasting insulin levels ($p = 3.88 \times 10^{-5}$), HOMA-IR ($p = 3.89 \times 10^{-4}$), triglycerides ($p = 1.61 \times 10^{-3}$) and visceral fat ($p = 2.1 \times 10^{-4}$); see Figures 7.3b and 7.4. To try relating IL1AP with cardio-metabolic risk, we derived the *metabolic syndrome severity score* (Alberti et al., 2009), and found that higher scores, i.e., increased risk factor severity, were associated with lower protein levels ($p = 1.20 \times 10^{-3}$ in Ottawa and $p = 2.50 \times 10^{-4}$ in DiOGenes). Our uncovered pQTL could therefore have implications in the obese population, as well as in pre-type 2 diabetic and type 2 diabetic populations.

7.3.3 WFKN2, a TGFβ-activity protein with protective effect against metabolic disorders

The role of the WFKN2 protein and of its coding gene, *WFIKKN2*, in regulating TGFβ activity has been extensively studied in muscle and skeletal muscle (Monestier and Blanquet, 2016), but, to our knowledge, not in other tissues. Here, we describe it in the context of obesity and metabolic disorders. We found that higher protein levels were associated with lower levels of fasting insulin, triglycerides, HOMA-IR and visceral fat (Figure 7.4), suggesting a protective role against metabolic dysregulation.

Our analyses also suggested that the WFKN2 levels are controlled by rs9303566, which is consistent with reports from other p- and eQTL studies (Suppl. Tables S6-7). This specific SNP was also found to be associated with DNA methylation and histone marks (Chen et al., 2016; Bonder et al., 2017), and is located within 100 base pairs of a transcription factor binding site, with numerous factors such as MYBL2, NFIC, EP300 and MXI1. It is in strong LD with other SNPs with potential regulatory impact; for instance, it is located 9Kb upstream to rs8072476 ($r^2 = 0.97$), which overlaps another cluster of transcription factor binding sites (FOXA1, ESR1, USF1 & 2, TFAP2A & 2C).

7.3.4 Inflammation mediated proteins and their role in insulin resistance

Our analyses suggested a *cis* effect of rs6741488 on KYNU (Kynureninase) plasmatic levels. KYNU is an enzyme involved in the biosynthesis of nicotinamide adenine dinucleotide (NAD) cofactors from tryptophan. This protein and its pathway have been found to be particularly relevant for obesity and associated metabolic disorders. For instance, Favennec et al. (2015) reported that KYNU was up-regulated by pro-inflammatory cytokines in human primary adipocytes, and more so in the omental adipose tissue of obese compared to lean control subjects. Other studies indicated that the kynurenine pathway (KP) may act as an inflammatory sensor, and that increased levels of its catabolites may be linked with several cardiometabolic defects, including cardiovascular diseases, diabetes, obesity and hypertension (Song et al., 2017; Rebnord et al., 2017). In our cohorts, higher KYNU levels were associated with decreased HDL levels ($p = 6.66 \times 10^{-4}$), and increased triglycerides levels ($p = 3.43 \times 10^{-8}$), visceral fat ($p = 2.51 \times 10^{-8}$) and insulin resistance (marginally, nominal $p = 2.53 \times 10^{-2}$, corrected $p = 0.17$), see Figure 7.4; as expected, higher protein levels also increased the metabolic syndrome score (Ottawa $p = 8.23 \times 10^{-5}$; DiOGenes $p = 3.62 \times 10^{-6}$).

Additional elements support the role of KYNU in inflammation; in particular, a recent work suggested a possible causal connection between obesity and cancer, mediated by KP activation through inflammatory mechanisms (Stone et al., 2018). Interestingly, our analyses highlighted two soluble interleukin receptor antagonist proteins, namely IL6RA and I17RA, that were both under genetic control and associated with insulin resistance (Figure 7.3b). We did not find significant correlation between the I17RA and KYNU protein levels, but we did observe a significant anti-correlation between IL6RA and KYNU in both cohorts (Ottawa $p = 0.01$ and DiOGenes $p = 4 \times 10^{-3}$). We also found a link between the plasma levels of KYNU and pro-inflammatory molecules, namely, IL6, IFNG and TNF alpha. In the severely obese Ottawa cohort, where subjects displayed high low-grade inflammation status, KYNU was positively associated with IL6 and IFNG at FDR 5%, while in the overweight/obese cohort DiOGenes, where subjects had a milder inflammation status, we found a positive association with IFNG only, at FDR 5%. Finally, metabolic dysfunctions mediated via KP may relate to another inflammatory pathology, namely, psoriasis (Harden et al., 2016), a skin disease with stronger prevalence in obese subjects (Armstrong et al., 2012).

Our results thus highlighted proteins under genetic control with probable roles in inflammation and subsequent metabolic dysfunctions; this reinforces previous discussions (Song et al., 2017; Jacobs et al., 2017) about the potential of KP therapeutic inhibitors against cardiovascular and metabolic diseases.

7.4 *Trans*-pQTLs in a stratified obese population

In this section, we focus on *trans*-regulatory mechanisms that may be of particular relevance for studying metabolic disorders in the obese population. Indeed, owing to its multivariate modelling tailored to the detection of weak effects, LOCUS could identify several *trans* and pleiotropic effects that suggest novel metabolic pathways (Figure 7.3a). We next discuss three important examples among the validated *trans* pQTLs, in light of clinical associations; as before, all associations mentioned have corrected meta-analysis p-values < 5% and nominal p-values are provided, unless noted otherwise.

7.4.1 Pleiotropic effects from the ABO locus onto CADH5, CD209, INSR, LYAM2 and TIE1

ABO is a well-known pleiotropic locus associated with molecular and clinical phenotypes, including coronary artery diseases, type 2 diabetes, liver enzyme levels (alkaline phosphatase), LDL and total cholesterol (Sivakumaran et al., 2011; Pickrell et al., 2016; Carayol et al., 2017; Suhre et al., 2017; Sun et al., 2018). Here, we complement these results by highlighting two independent candidate hotspots in the ABO region: rs2519093 and rs8176741 ($r^2 = 0.03$). Our results suggested that rs2519093 is *trans*-acting on E-selectin (protein LYAM2, coding gene SELE, on chromosome 1), the Insulin Receptor (protein and coding gene name INSR, on chromosome 19) and the CD209 antigen (protein and coding gene name CD209, on chromosome 19), and that rs817674 is *trans*-acting on the Tyrosine-protein kinase receptor Tie-1 (protein and coding gene name TIE1, on chromosome 1), Cadherin-5 (protein CADH5, coding gene CDH5, on chromosome 16) and the CD209 antigen (protein and coding gene name CD209, on chromosome 19). Both rs2519093 and rs8176741 were also reported as *cis*-acting eQTL variants for ABO, OBP2B and SURF1, in multiple tissues. Further queries in public databases suggested that rs8176741 may affect the binding sites for three transcription factors (Myc, MYC-MAX and Arnt), which suggests effects on different genes.

We then studied possible associations between the above proteins and the clinical parameters of our cohorts. Three proteins stood out from these analyses, CADH5, CD209 and LYAM2 (Figure 7.4). All were associated with triglyceride levels, and LYAM2 and CD209 also were also associated with visceral fat (Figures 7.3 and 7.4). Moreover, the macrophage protein CD209 may have an important role in controlling lipid levels as it was also associated with HDL ($p = 7.58 \times 10^{-6}$). More precisely, we observed higher CD209 protein levels with higher HDL, lower triglyceride levels, and, consistently with these effects, with lower visceral fat index. Dyslipidemia is a risk factor for Non-Alcoholic Fatty Liver Disease (NAFLD; Bass et al., 2010), and the CD209 gene levels have been reported as differentially expressed in patients with Non-Alcoholic Steatohepatitis (NASH) compared to healthy subjects (Sheldon et al., 2016). The role of circulating protein levels of CD209 could be further studied in NASH/NAFLD cohorts.

Finally, the LYAM2 levels were also associated with all the glycemic variables in the Ottawa cohort (fasting glucose: $p = 6.43 \times 10^{-6}$, fasting insulin: $p = 3.54 \times 10^{-4}$, HOMA-IR: $p = 1.8 \times 10^{-4}$), but only with fasting glucose in the DiOGenes cohort ($p = 8.91 \times 10^{-4}$), although we did observe a suggestive association with HOMA-IR (nominal $p = 0.02$, corrected $p = 0.15$). Since the Ottawa subjects are more

insulin-resistant than the DiOGenes subjects (average HOMA-IR with standard deviation: 4.97(3.88) versus 3.00(1.71), $p = 2.52 \times 10^{-18}$; Suppl. Table S4), LYAM2 might represent a marker of insulin-resistance severity. Consistent with this hypothesis, the plasma levels of LYAM2 are employed as a biomarkers of endothelial dysfunction and risk of type 2 diabetes (Meigs et al., 2004; Song et al., 2007).

7.4.2 Complement/coagulation: a *trans*-acting insertion linking PROC and its receptor

PROC (Protein C, coding gene PROC on chromosome 2) and its paralog protein FA7 (Coagulation Factor 7, coding gene F7 on chromosome 13) regulate the complement and the coagulation systems. Both systems promote inflammation (Ricklin et al., 2010) and contribute to metabolic dysfunction in the adipose tissue and liver (Phieler et al., 2013). Our analyses suggested novel pQTLs for these proteins (Suppl. Table S5): FA7 was associated with rs3093233, which is also a known eQTL of *F7* and *F10* in several tissues (Suppl. Table S7). PROC may be controlled by *trans*-regulatory mechanisms, initiated in its receptor gene, *PROCR*, on chromosome 20; it was indeed associated with an insertion, rs141091409, located 20 Kb upstream to *PROCR*. We validated this association using both the MS and SomaLogic measurements of the protein ($p = 2.16 \times 10^{-5}$ and $p = 1.66 \times 10^{-18}$, respectively). Previous studies found associations between cardiovascular diseases and variants located in the *PROC* or *PROCR* genes (Reiner et al., 2008; Howson et al., 2017; van der Harst and Verweij, 2018). Interestingly, our hit, rs141091409, was in strong LD ($r^2 > 0.95$) with the missense variant rs867186, previously identified as associated with coronary heart disease (van der Harst and Verweij, 2018).

Our clinical analyses support the relation of PROC and FA7 levels with lipid traits: both proteins were positively associated with cholesterol, triglycerides and visceral fat (Figures 7.3b and 7.4). PROC levels were quantified by both the MS and SomaLogic platforms, and displayed consistent results. The SomaLogic measurements of PROC were also positively associated with LDL ($p = 5.39 \times 10^{-5}$). The role of these proteins for cardiovascular and NAFLD diseases in the overweight/obese population would merit further investigation.

7.4.3 XRCC6, a DNA repair protein as putative biomarker for metabolic disorders

We identified a novel *trans* pQTL on chromosome 11 for the protein XRCC6 (X-Ray Repair Complementing Defective Repair In Chinese Hamster Cells; also known as Ku70), whose coding gene lies on chromosome 22. *XRCC6* gene has repair functions for double-stranded DNA breaks, to which it binds, thereby activating DNA-dependent protein kinases (DNA-PK) to repair DNA by nonhomologous end joining. DNA-PKs have been linked to lipogenesis in response to feeding and insulin signaling (Wong et al., 2009). DNA-PK inhibitors may also reduce the risk of obesity and type 2 diabetes by activating multiple AMPK targets (Park et al., 2017). A recent review discussed further the role of DNA-PK in energy metabolism, and in particular, the conversion of carbohydrates into fatty acids in the liver, in response to insulin (Chung, 2018). It described increased DNA-PK activity with age, and links with mitochondrial loss in skeletal muscle and weight gain. Finally, *XRCC6* functions have been reported as associated with regulation of beta-cell proliferation, islet expansion, increased insulin levels and decreased glucose levels (Park et al., 2017; Tavana et al., 2013).

We observed significant associations between the XRCC6 protein levels and several clinical variables in the Ottawa cohort (FDR < 5%). Higher expression was associated with decreased HDL ($p = 5.83 \times 10^{-4}$), as well as with higher triglycerides ($p = 4.39 \times 10^{-4}$), insulin levels ($p = 4.50 \times 10^{-4}$) and visceral adiposity

($p = 5.94 \times 10^{-5}$; Figure 7.4). We only found marginal associations using the DiOGenes data for insulin levels (nominal $p = 0.02$, corrected $p = 0.14$) and HOMA-IR (nominal $p = 0.02$, corrected $p = 0.16$). The directionality of these effects was consistent in both cohorts. As the Ottawa subjects were more severely obese, the effects might be larger for subjects with pronounced metabolic syndrome, but this would require confirmation.

Our pQTL sentinel SNP, rs4756623, is intronic and located within the *LRRC4C* gene, a binding partner for Netrin G1 and member of the axon guidance (Lin et al., 2003). To our knowledge, *LRRC4C* has not been previously described in the context of obesity, insulin resistance or type 2 diabetes. However, its partner Netrin G1 is known to promote adipose tissue macrophage retention, inflammation and insulin resistance in obese mice (Ramkhelawon et al., 2014). The underlying regulatory mechanisms between rs4756623 and the *XRCC6* locus need to be clarified, and functional studies will be required to understand its physiological impact. Given the possible clinical relevance of this novel pQTL, this result could be of interest to both clinicians and scientists.

7.5 Conclusion

Despite important technological advances, genome-wide association studies of intermediate proteomic expression phenotypes remain infrequent, owing to their high costs (Folkersen et al., 2017; Suhre et al., 2017; Carayol et al., 2017; Sun et al., 2018; Yao et al., 2018a). To date, all but our recent study (Carayol et al., 2017) have focused on the general population, and had therefore no possibility to assess links between the uncovered pQTLs and diseases in the cohort used for pQTL discovery. Hence, to evaluate the relevance, and sometimes causality, of the pQTLs for a given clinical condition, these studies have mostly relied on overlaps with published GWAS risk loci and eQTLs, although such results may concern very different tissues or populations.

Here, we described the first integrative pQTL study that relates the discovered associations with metabolic disorders, such as insulin resistance and dyslipidemia, in a stratified obese population. Our study is also the first to present fully multivariate pQTL analyses at genome-wide scale. Our Bayesian hierarchical method LOCUS identified pQTL associations with sound evidence for functional relevance, and achieved very high independent-replication rates, despite sample sizes 2.5 to 18 times smaller than in other studies and the different degrees of metabolic disorder severity in the discovery and validation cohorts.

Access to clinical phenotypes from these cohorts allowed the identification of novel pQTLs with biomedical potential. Our integrated approach also provided molecular insights into obesity and associated co-morbidities, including the development of the metabolic syndrome. Our complete pQTL and clinical association results offer opportunities to generate further hypotheses about therapeutic options; they are accessible from the searchable online database https://heleneruffieux.com/locus-pqtl.

7.6 Summary

We presented a detailed pQTL study using our approach LOCUS. This study demonstrated that LOCUS is scalable, and finds pertinent QTL associations, both biologically and clinically. Several improvements may be considered, ~~though~~ however.

First, many of the detected effects were strong, and had been reported in previous studies. Moreover, the number of validated *trans* associations seems rather small (twenty). It would be tempting to try using a looser false discovery rate threshold (> 5%), or different threshold choices for detecting *cis* or *trans* associations similarly as in Peterson et al. (2016), and see if additional discoveries can be validated in the independent cohort.

Second, we did not use the global-local approach of Chapter 5 (which was developed after these data analyses); its use may bring some performance improvement for the detection of *trans* associations. However, the gain may not be substantial if there are no large hotspots, which may be the case given the small number of measured outcomes compared to eQTL data ($q \approx 130$ and $q \approx 1,000$) .

Finally, we evaluated the enrichment of the pQTL loci in epigenomic annotation marks. It would interesting to compare the results with the method of Chapter 6 that directly models them. As this approach selects annotation marks, this may also permit a finer interpretation of the mechanisms underpinning the pQTL associations.

8 Discussion and future work

This thesis has two main themes. The first concerns devising expressive hierarchical models for molecular QTL data, and the second tackles efficient variational inference for them. We demonstrated the importance of addressing these themes together; models should lend themselves to efficient inference, and inference algorithms should be model-specific. The purpose of inference is variable selection. In particular, the approaches presented in this thesis all attempt to leverage complex dependencies across molecular entities to enhance the detection of *trans* associations and hotspot genetic variants.

Effective solutions require modelling tailored to important features of molecular QTL data. In Chapter 3, we presented our hierarchical regression framework for joint modelling of q expression outcomes (responses) and p genetic variants (predictors) for n samples. It accommodates the $p, q \gg n$ regime using a series of parallel regressions that are linked through the model hierarchy. In particular, our proposal borrows information across expression outcomes via a parameter that controls hotspot propensity. This parameter is central to our work: it permits a direct modelling of hotspots by influencing the probabilities of association of each genetic variant with the expression outcomes, and can accommodate diverse modelling extensions, which we explored in the subsequent chapters.

In Chapter 3, we also considered extending the bottom of the model hierarchy using logit and probit link functions to model binary or mixed outcome data. Another natural enhancement would be to propose a negative-binomial adaptation for RNA-seq count data, possibly relying on the Pólya–Gamma data augmentations of Polson et al. (2013) to obtain closed-form variational updates. It would also be useful to jointly account for multiple conditions, tissues or cell-types using multivariate link functions (see, e.g., Petretto et al., 2010; Lewin et al., 2015).

Our simulations suggested that variable selection doesn't suffer much from not accounting for the dependence of the outcomes beyond that induced by the model hierarchy, though it would be interesting to quantify this better using dedicated experiments. One may envision modelling residual correlation for co-expressed molecular outcomes, although this would require defining modules, which may be unsatisfactory. Finally, real QTL data may exhibit substantial population structure or relatedness, and we would need to assess its impact on inferences. We may also consider extensions to involve a random-effect component capturing sample structure, as in the hybrid sparse linear mixed model of Zhou et al. (2013).

In Chapter 4, we explored encoding genetic structural information via the hotspot propensity parameter to better represent the uncertainty entailed by the selection of correlated genetic variants. We proposed a group and a similarity sparsity model that use the empirical correlation of genetic variants, thereby improving the interpretability of association estimates in regions with marked dependence. However, selection may be hampered by the numerous false positive signals, as with marginal screening. It may be worth trying to incorporate other types of similarity information that may relate more directly to functional processes, such as those provided by shared pathways or chromatin interactions.

In Chapter 5, we formulated a fully Bayesian second-stage model on the hotspot propensity, based on a flexible global-local representation that can retain large hotspot effects in highly sparse settings. We placed particular emphasis on hotspot detection in very large response settings, for which we proposed a suitable response multiplicity adjustment. Further work is needed to provide recommendations for selecting hotspots and estimating their sizes from the obtained posterior summaries. Current practice often uses the 0.5 optimal-prediction threshold of Barbieri and Berger (2004) on the posterior probabilities of inclusion, or runs permutation analyses when feasible (which was the case for us). However, a neater approach may be to devise a decision rule that would exploit the full variational distributions of the hotspot propensity parameters.

In Chapter 6, we tackled guiding the selection of SNPs by encoding functional information under the form of epigenomic marks. We used a second-stage spike-and-slab model on the probabilities of association to infer the relevance and effects of the marks in SNP-outcome blocks. These block-specific effects reflect common biological assumptions on the localized nature of regulation mechanisms on the genome and the existence of co-regulated expression outcomes. Our approach is not yet ready for practical uses; coming efforts will concern adequately defining the QTL block partition and testing the sensitivity of inferences to its choice. This should be kept simple to seamlessly adapt to our modelling framework, which is already quite complex. We will also need to compare our annotation-based model with our original QTL model on real data.

The effectiveness of the above approaches hinges on the accuracy of variational inference for variable selection in our QTL modelling framework. Evaluating and enhancing variational inference was a connecting thread of this thesis, and we found that the algorithms should adapt to the specificities of each model variant. In particular, our structured approximation of Chapter 3 relaxed the strong posterior independence assumptions of vanilla mean-field factorizations by retaining the spike-and-slab multimodal structure. The group and similarity sparsity models of Chapter 4 allowed the restoration of additional structure into the approximation while maintaining analytical updates. Closed-form algorithms were critical for scalability, and could be obtained by resorting to appropriate reparametrizations (e.g., Chapter 5) or local approximations (e.g., the logistic regression extension of Chapter 3).

We also saw that variational inference lends itself to simulated annealing and expectation-maximization (EM) augmentations, which proved useful in several respects. Simulated annealing enabled improved exploration of multimodal posteriors (Chapter 4). In particular, this allowed increasing robustness to different algorithm initializations and permitted accurate estimation of hotspot sizes by effectively handling linkage disequilibrium structures (Chapter 5). We discussed adaptive ideas for learning the temperature schedule to optimize the annealing procedure; developing an effective procedure for this is future work. The coupling of variational updates with EM steps permitted empirical Bayes estimation of the epigenomic annotation hyperparameters (Chapter 6). Indeed, the variational-EM algorithm bypasses intractable maximization of the marginal log-likelihood by alternating optimizations of the variational lower bound. The partition-based approach allowed running the empirical Bayes estimation in parallel, which resulted in a highly effective algorithm.

Finally, the computational efficiency of our approaches relates to the nature of deterministic algorithms, but also hinges on their effective implementation. The current version of the code is written in R and C++, but it may be interesting to evaluate the benefit of GPU computing or technologies for large-scale applications, such as Hadoop MapReduce (Dean and Ghemawat, 2008), Apache Spark (Zaharia et al., 2010) or TensorFlow (Abadi et al., 2016).

A central ambition of this thesis was to bridge the gap between Bayesian joint inference and its practical use for analyzing current molecular QTL data. Most practitioners still resort to marginal screening, despite its known defects for genome-wide association analyses and its inability to answer questions related to pleiotropy. In Chapter 7, we provided a practical illustration of the advantages of our hierarchical approach via a detailed study of two pQTL datasets. We confronted our approach with marginal screening on these data, highlighting the benefits of joint modelling based on external evidence. The resulting analyses identified novel candidate biomarkers, whose roles merit further investigation. This, and our various numerical experiments on synthetic and real data, demonstrated that our approach is scalable, yields interpretable posterior summaries for selection of QTL pairs and hotspots, and requires very little input (an expected number of associated genetic variants per outcomes, and a variance for this number); they also indicated that it greatly outperforms marginal screening approaches for identifying *trans* and pleiotropic effects, thereby effectively managing the inherent tension between expressive modelling and scalability of inference. We hope that these advantages will be sufficiently appealing for practitioners to try our approach on their data, especially as there are increasing opportunities for joint inference on transcriptomic, proteomic, lipidomic, metabolomic or even clinical datasets.

Appendices

A Appendix for Chapter 3

A.1 Predictor multiplicity control

Sparsity control at predictor level can be induced through the prior distribution of $\boldsymbol{\omega}$, by carefully selecting its hyperparameters. The prior probability that $\boldsymbol{X}_s$ is "active" (i.e., associated with at least one response) is

$$\text{pr}\left(\cup_{t=1}^{q}\{\gamma_{st}=1\}\right)=\int\left\{1-\prod_{t=1}^{q}\text{pr}\left(\gamma_{st}=0\mid\omega_s\right)\right\}\text{pr}(\omega_s)\text{d}\omega_s=1-\frac{\text{B}(a_s,b_s+q)}{\text{B}(a_s,b_s)},$$

where $\text{B}(\cdot,\cdot)$ is the beta function, and after a little algebra this equals

$$1-\prod_{t=1}^{q}\frac{b_s+q-t}{a_s+b_s+q-t},$$

so assuming exchangeability and setting

$$a_s\equiv 1,\qquad b_s\equiv q(p-p^*)/p^*,\qquad 0<p^*<p, \tag{A.1}$$

implies that

$$\text{pr}\left(\cup_{t=1}^{q}\{\gamma_{st}=1\}\right)=\frac{q}{b_s+q}=\frac{p^*}{p},$$

where p^* is interpreted as a prior average number of active predictors. The choice (A.1) yields a multiplicity adjustment as suggested by a plot of the prior odds ratios (Figure 3.2), indicating the penalty induced by the prior when moving from q_s-1 to q_s responses associated with $\boldsymbol{X}_s$,

$$\text{POR}(q_s-1:q_s)=\frac{\text{pr}\left(\mathcal{M}_{q_s-1}\right)}{\text{pr}\left(\mathcal{M}_{q_s}\right)}=\frac{\text{B}(a_s+q_s-1,b_s+q-q_s+1)}{\text{B}(a_s+q_s,b_s+q-q_s)}=\frac{b_s+q-q_s}{a_s+q_s-1},$$

where $\mathcal{M}_{q_s}$ is a model in which predictor $\boldsymbol{X}_s$ is associated with $1\leq q_s\leq q$ responses, so

$$\text{pr}\left(\mathcal{M}_{q_s}\right)=\int\omega_s^{q_s}(1-\omega_s)^{q-q_s}\text{pr}(\omega_s)\text{d}\omega_s.$$

A.2 Derivation of the variational algorithm

A.2.1 Variational distributions

We provide the detailed derivation of our variational algorithm for the model and approximation presented in Chapter 3. For q centered responses, $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q)$, and p centered predictors, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, for n samples, we have

$$\begin{aligned} p\left(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\tau}, \sigma^{-2}\right) &= \left\{\prod_{t=1}^{q} p\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t\right)\right\}\left\{\prod_{s=1}^{p}\prod_{t=1}^{q} p\left(\beta_{st} \mid \gamma_{st}, \tau_t, \sigma^{-2}\right)\right\}\left\{\prod_{s=1}^{p}\prod_{t=1}^{q} p\left(\gamma_{st} \mid \omega_s\right)\right\} \\ &\quad \times \left\{\prod_{s=1}^{p} p(\omega_s)\right\}\left\{\prod_{t=1}^{q} p(\tau_t)\right\} p\left(\sigma^{-2}\right), \end{aligned} \tag{A.2}$$

where

$$\begin{aligned} \boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t &\sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}_t, \tau_t^{-1}\boldsymbol{I}_n), \\ \beta_{st} \mid \gamma_{st}, \tau_t, \sigma^{-2} &\sim \gamma_{st}\mathcal{N}\left(0, \sigma^2\tau_t^{-1}\right) + (1-\gamma_{st})\,\delta_0, \\ \gamma_{st} \mid \omega_s &\sim \text{Bernoulli}\left(\omega_s\right), \\ \omega_s &\sim \text{Beta}\left(a_s, b_s\right), \\ \tau_t &\sim \text{Gamma}\left(\eta_t, \kappa_t\right), \\ \sigma^{-2} &\sim \text{Gamma}\left(\lambda, \nu\right), \end{aligned}$$

with δ_0, the Dirac distribution.

Let $\boldsymbol{v} = \left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\tau}, \sigma^{-2}\right)$ and consider the following mean-field variational approximation,

$$q\left(\boldsymbol{v}\right) = \left\{\prod_{s=1}^{p}\prod_{t=1}^{q} q(\beta_{st}, \gamma_{st})\right\}\left\{\prod_{s=1}^{p} q(\omega_s)\right\}\left\{\prod_{t=1}^{q} q(\tau_t)\right\} q\left(\sigma^{-2}\right).$$

We obtain each component of this factorization using the formula

$$\log q_j(v_j) = \mathrm{E}_{-j}\{\log p(\boldsymbol{y}, \boldsymbol{v})\} + \text{cst}, \qquad j = 1, \ldots, J,$$

with $p(\boldsymbol{y}, \boldsymbol{v})$ given in (*A*.2), where $\mathrm{E}_{-j}(\cdot)$ is the expectation with respect to the distributions q_k over all variables v_k ($k \neq j$), and where cst does not depend on v_j (recall Lemma 2.4.1). Hereafter, we also write $\mathrm{E}_q(\cdot)$ for the expectation with respect to the complete variational distribution, and $v_j^{(r)}$ for the r^{th} moment with respect to the distribution q_j of v_j. We have

$$\begin{aligned} \log q(\beta_{st}, \gamma_{st}) &= \sum_{k=1}^{q} \mathrm{E}_{-(\beta_{st}, \gamma_{st})}\left\{\log p(\boldsymbol{y}_k \mid \boldsymbol{\beta}_k, \tau_k)\right\} + \sum_{j=1}^{p}\sum_{k=1}^{q} \mathrm{E}_{-(\beta_{st}, \gamma_{st})}\left\{\log p(\beta_{jk} \mid \gamma_{jk}, \tau_k, \sigma^{-2})\right\} \\ &\quad + \sum_{j=1}^{p}\sum_{k=1}^{q} \mathrm{E}_{-(\beta_{st}, \gamma_{st})}\left\{\log p(\gamma_{jk} \mid \omega_j)\right\} + \text{cst}, \end{aligned}$$

where cst does not depend on β_{st} and γ_{st}. Completing the square yields

$$\begin{aligned} q(\beta_{st},\gamma_{st}) &= \text{cst}\left[\left(2\pi\sigma^2_{\beta,st}\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2_{\beta,st}}\left(\beta_{st}-\mu_{\beta,st}\right)^2\right\}\right]^{\gamma_{st}} \\ &\quad\times\sigma_{\beta,st}^{\gamma_{st}}\exp\left\{\frac{1}{2}\mathrm{E}_q\left(\log\sigma^{-2}\right)+\frac{1}{2}\mathrm{E}_q\left(\log\tau_t\right)+\frac{1}{2}\mu^2_{\beta,st}\sigma^{-2}_{\beta,st}+\mathrm{E}_q\left(\log\omega_s\right)\right\}^{\gamma_{st}} \\ &\quad\times\left\{\delta_0(\beta_{st})\right\}^{1-\gamma_{st}}\exp\left\{\mathrm{E}_q\log(1-\omega_s)\right\}^{1-\gamma_{st}}, \end{aligned}$$

with

$$\mu_{\beta,st}=\sigma^2_{\beta,st}\tau_t^{(1)}\boldsymbol{X}_s^T\left\{\boldsymbol{y}_t-\sum_{j=1,j\neq s}^{p}\gamma_{jt}^{(1)}\mu_{\beta,jt}\boldsymbol{X}_j\right\},\qquad \sigma^{-2}_{\beta,st}=\tau_t^{(1)}\left\{\|\boldsymbol{X}_s\|^2+\left(\sigma^{-2}\right)^{(1)}\right\}.$$

We therefore observe that

$$q(\beta_{st},\gamma_{st})=q(\beta_{st}\mid\gamma_{st})q(\gamma_{st}),$$

with

$$\beta_{st}\mid\gamma_{st}=1,\boldsymbol{y}\sim\mathcal{N}\left(\mu_{\beta,st},\sigma^2_{\beta,st}\right),\quad \beta_{st}\mid\gamma_{st}=0,\boldsymbol{y}\sim\delta_0,\quad \gamma_{st}\mid\boldsymbol{y}\sim\text{Bernoulli}\left(\gamma_{st}^{(1)}\right),$$

and with

$$\frac{\gamma_{st}^{(1)}}{1-\gamma_{st}^{(1)}}=\sigma_{\beta,st}\exp\left[\mathrm{E}_q\left(\log\omega_s\right)-\mathrm{E}_q\log(1-\omega_s)+\frac{1}{2}\mathrm{E}_q\left(\log\tau_t\right)+\frac{1}{2}\mathrm{E}_q\left(\log\sigma^{-2}\right)+\frac{1}{2}\mu^2_{\beta,st}\sigma^{-2}_{\beta,st}\right],$$

i.e.,

$$\gamma_{st}^{(1)}=\left[1+\sigma^{-1}_{\beta,st}\exp\left\{\mathrm{E}_q\log(1-\omega_s)-\mathrm{E}_q\left(\log\omega_s\right)-\frac{1}{2}\mathrm{E}_q\left(\log\tau_t\right)-\frac{1}{2}\mathrm{E}_q\left(\log\sigma^{-2}\right)-\frac{1}{2}\mu^2_{\beta,st}\sigma^{-2}_{\beta,st}\right\}\right]^{-1}. \quad \text{(A.3)}$$

We then find that

$$\begin{aligned} \log q(\tau_t) &= \mathrm{E}_{-\tau_t}\left\{\log p\left(\boldsymbol{y}_t\mid\boldsymbol{\beta}_t,\tau_t\right)\right\}+\sum_{s=1}^{p}\mathrm{E}_{-\tau_t}\left\{\log p\left(\beta_{st}\mid\gamma_{st},\tau_t,\sigma^{-2}\right)\right\}+\log p(\tau_t)+\text{cst} \\ &= \frac{n}{2}\log\tau_t-\frac{\tau_t}{2}\mathrm{E}_q\left(\|\boldsymbol{y}_t-\boldsymbol{X}\boldsymbol{\beta}_t\|^2\right)+\frac{1}{2}\log\tau_t\sum_{s=1}^{p}\gamma_{st}^{(1)}-\frac{\tau_t}{2}\left(\sigma^{-2}\right)^{(1)}\sum_{s=1}^{p}\beta_{st}^{(2)}+(\eta_t-1)\log\tau_t \\ &\quad -\kappa_t\tau_t+\text{cst} \\ &= \left(\eta_t+\frac{n}{2}+\frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)}-1\right)\log\tau_t-\tau_t\left[\kappa_t+\frac{1}{2}\|\boldsymbol{y}_t\|^2-\boldsymbol{y}_t^T\sum_{s=1}^{p}\mu_{\beta,st}\gamma_{st}^{(1)}\boldsymbol{X}_s\right. \\ &\quad \left.+\sum_{s=1}^{p-1}\mu_{\beta,st}\gamma_{st}^{(1)}\boldsymbol{X}_s^T\sum_{j=s+1}^{p}\mu_{\beta,jt}\gamma_{jt}^{(1)}\boldsymbol{X}_j+\frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)}\left(\sigma^2_{\beta,st}+\mu^2_{\beta,st}\right)\left\{\|\boldsymbol{X}_s\|^2+\left(\sigma^{-2}\right)^{(1)}\right\}\right]+\text{cst}. \end{aligned}$$

Therefore

$$\tau_t\mid\boldsymbol{y}\sim\text{Gamma}\left(\eta_t^*,\kappa_t^*\right),\qquad \tau_t^{(1)}=\eta_t^*/\kappa_t^*,$$

where

$$\begin{aligned}
\eta_t^* &= \eta_t + \frac{n}{2} + \frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)}, \\
\kappa_t^* &= \kappa_t + \frac{1}{2}\|\boldsymbol{y}_t\|^2 - \boldsymbol{y}_t^T\sum_{s=1}^{p}\mu_{\beta,st}\gamma_{st}^{(1)}\boldsymbol{X}_s + \sum_{s=1}^{p-1}\mu_{\beta,st}\gamma_{st}^{(1)}\boldsymbol{X}_s^T\sum_{j=s+1}^{p}\mu_{\beta,jt}\gamma_{jt}^{(1)}\boldsymbol{X}_j \\
&\quad + \frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right)\left\{\|\boldsymbol{X}_s\|^2 + \left(\sigma^{-2}\right)^{(1)}\right\}.
\end{aligned}$$

Since τ_t has a Gamma distribution, the expectation $\mathrm{E}\left(\log\tau_t\right)$ appearing in (A.3) can be rewritten in terms of η_t^* and κ_t^* using the digamma function,

$$\Psi(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log\Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)},$$

as

$$\mathrm{E}_q\left(\log\tau_t\right) = \Psi(\eta_t^*) - \log(\kappa_t^*). \tag{A.4}$$

We also find that

$$\begin{aligned}
\log q\left(\sigma^{-2}\right) &= \sum_{s=1}^{p}\sum_{t=1}^{q}\mathrm{E}_{-\sigma^{-2}}\left\{\log p\left(\beta_{st}\mid\gamma_{st},\tau_t,\sigma^{-2}\right)\right\} + \log p(\sigma^{-2}) + \mathrm{cst} \\
&= \left(\frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{q}\gamma_{st}^{(1)}\right)\log\sigma^{-2} - \sigma^{-2}\sum_{t=1}^{q}\frac{\tau_t^{(1)}}{2}\sum_{s=1}^{p}\beta_{st}^{(2)} + (\lambda-1)\log\sigma^{-2} - \nu\sigma^{-2} + \mathrm{cst} \\
&= \left(\lambda + \frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{q}\gamma_{st}^{(1)} - 1\right)\log\sigma^{-2} - \sigma^{-2}\left\{\nu + \sum_{t=1}^{q}\frac{\tau_t^{(1)}}{2}\sum_{s=1}^{p}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right)\gamma_{st}^{(1)}\right\} + \mathrm{cst}.
\end{aligned}$$

Thus

$$\sigma^{-2}\mid\boldsymbol{y} \sim \mathrm{Gamma}\left(\lambda^*,\nu^*\right), \qquad \left(\sigma^{-2}\right)^{(1)} = \lambda^*/\nu^*,$$

where

$$\lambda^* = \lambda + \frac{1}{2}\sum_{s=1}^{p}\sum_{t=1}^{q}\gamma_{st}^{(1)}, \qquad \nu^* = \nu + \frac{1}{2}\sum_{t=1}^{q}\tau_t^{(1)}\sum_{s=1}^{p}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right)\gamma_{st}^{(1)},$$

and, as before, we have

$$\mathrm{E}_q\left(\log\sigma^{-2}\right) = \Psi(\lambda^*) - \log\nu^*. \tag{A.5}$$

Finally, we have

$$\begin{aligned}
\log q(\omega_s) &= \sum_{t=1}^{q}\mathrm{E}_{\gamma_{st}}\left\{\log p(\gamma_{st}\mid\omega_s)\right\} + \log p(\omega_s) + \mathrm{cst} \\
&= \left(a_s + \sum_{t=1}^{q}\gamma_{st}^{(1)} - 1\right)\log\omega_s + \left(b_s - \sum_{t=1}^{q}\gamma_{st}^{(1)} + q - 1\right)\log(1-\omega_s) + \mathrm{cst},
\end{aligned}$$

that is,

$$\omega_s\mid\boldsymbol{y} \sim \mathrm{Beta}\left(a_s^*, b_s^*\right), \qquad \omega_s^{(1)} = \frac{a_s^*}{a_s^* + b_s^*},$$

where

$$a_s^* = a_s + \sum_{t=1}^{q}\gamma_{st}^{(1)} \qquad\qquad b_s^* = b_s - \sum_{t=1}^{q}\gamma_{st}^{(1)} + q.$$

As ω_s has a Beta distribution, we also get

$$\mathrm{E}_q\left(\log\omega_s\right) = \Psi(a_s^*) - \Psi(a_s^* + b_s^*), \qquad \mathrm{E}_q \log(1-\omega_s) = \Psi(b_s^*) - \Psi(a_s^* + b_s^*). \tag{A.6}$$

A.2.2 Variational lower bound

We provide the expression of the variational lower bound, $\mathcal{L}(q)$, on the marginal log-likelihood, $\log p(\boldsymbol{y})$. It is evaluated at each iteration of the algorithm, in order to monitor its convergence:

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{\nu}) \log\left\{\frac{p(\boldsymbol{y},\boldsymbol{\nu})}{q(\boldsymbol{\nu})}\right\} \mathrm{d}\boldsymbol{\nu} \\
&= \sum_{t=1}^{q} \mathcal{L}_y\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t\right) + \sum_{s=1}^{p}\sum_{t=1}^{q} \mathcal{L}_\beta\left(\beta_{st}, \gamma_{st} \mid \tau_t, \sigma^{-2}\right) + \sum_{t=1}^{q} \mathcal{L}_\tau(\tau_t) + \mathcal{L}_\sigma\left(\sigma^{-2}\right) + \sum_{s=1}^{p} \mathcal{L}_\omega(\omega_s),
\end{aligned}$$

with

$$\begin{aligned}
\mathcal{L}_y\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t\right) &= \mathrm{E}_q\left\{\log p(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t)\right\} \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\mathrm{E}\left(\log\tau_t\right) - \frac{1}{2}\tau_t^{(1)}\left\{\|\boldsymbol{y}_t\|^2 - 2\boldsymbol{y}_t^T \sum_{s=1}^{p} \mu_{\beta,st}\gamma_{st}^{(1)}\boldsymbol{X}_s \right. \\
&\quad \left. + 2\sum_{s=1}^{p-1} \mu_{\beta,st}\gamma_{st}^{(1)}\boldsymbol{X}_s^T \sum_{j=s+1}^{p} \mu_{\beta,jt}\gamma_{jt}^{(1)}\boldsymbol{X}_j + \sum_{s=1}^{p} \|\boldsymbol{X}_s\|^2\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right)\gamma_{st}^{(1)}\right\} \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\mathrm{E}_q\left(\log\tau_t\right) - \tau_t^{(1)}\left\{\kappa_t^* - \frac{1}{2}\left(\sigma^{-2}\right)^{(1)} \sum_{s=1}^{p} \gamma_{st}^{(1)}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right) - \kappa_t\right\},
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}_\beta\left(\beta_{st}, \gamma_{st} \mid \tau_t, \sigma^{-2}\right) &= \mathrm{E}_q\left\{\log p(\beta_{st} \mid \gamma_{st}, \tau_t, \sigma^{-2})\right\} + \mathrm{E}_q\left\{\log p(\gamma_{st} \mid \omega_s)\right\} - \mathrm{E}_q\left\{\log q(\beta_{st}, \gamma_{st})\right\} \\
&= \frac{1}{2}\gamma_{st}^{(1)}\left\{\mathrm{E}_q\left(\log\sigma^{-2}\right) + \mathrm{E}_q\left(\log\tau_t\right)\right\} - \frac{1}{2}\left(\sigma^{-2}\right)^{(1)}\tau_t^{(1)}\gamma_{st}^{(1)}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right) \\
&\quad + \gamma_{st}^{(1)}\mathrm{E}_q\left(\log\omega_s\right) + \left(1-\gamma_{st}^{(1)}\right)\mathrm{E}_q\left\{\log(1-\omega_s)\right\} + \frac{1}{2}\gamma_{st}^{(1)}\left(\log\sigma_{\beta,st}^2 + 1\right) \\
&\quad - \gamma_{st}^{(1)}\log\gamma_{st}^{(1)} - \left(1-\gamma_{st}^{(1)}\right)\log\left(1-\gamma_{st}^{(1)}\right),
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}_\tau(\tau_t) &= \mathrm{E}_q\left\{\log p(\tau_t)\right\} - \mathrm{E}_q\left\{\log q(\tau_t)\right\} \\
&= \left(\eta_t - \eta_t^*\right)\mathrm{E}_q\left(\log\tau_t\right) - \left(\kappa_t - \kappa_t^*\right)\tau_t^{(1)} + \eta_t\log\kappa_t - \eta_t^*\log\kappa_t^* - \log\Gamma(\eta_t) + \log\Gamma(\eta_t^*),
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}_\sigma\left(\sigma^{-2}\right) &= \mathrm{E}_q\left\{\log p\left(\sigma^{-2}\right)\right\} - \mathrm{E}_q\left\{\log q\left(\sigma^{-2}\right)\right\} \\
&= \left(\lambda - \lambda^*\right)\mathrm{E}_q\left(\log\sigma^{-2}\right) - \left(\nu - \nu^*\right)\left(\sigma^{-2}\right)^{(1)} + \lambda\log\nu - \lambda^*\log\nu^* - \log\Gamma(\lambda) + \log\Gamma\left(\lambda^*\right),
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}_\omega(\omega_s) &= \mathrm{E}_q\left\{\log p(\omega_s)\right\} - \mathrm{E}_q\left\{\log q(\omega_s)\right\} \\
&= \left(a_s - a_s^*\right)\mathrm{E}_q\left(\log\omega_s\right) + \left(b_s - b_s^*\right)\mathrm{E}_q\left\{\log(1-\omega_s)\right\} - \log\mathrm{B}(a_s, b_s) + \log\mathrm{B}\left(a_s^*, b_s^*\right),
\end{aligned}$$

where $\mathrm{E}_q\left(\log\tau_t\right)$, $\mathrm{E}_q\left(\log\sigma^{-2}\right)$, $\mathrm{E}_q\left(\log\omega_s\right)$ and $\mathrm{E}_q\log(1-\omega_s)$ are given by (A.4), (A.5) and (A.6).

Hence, $\mathcal{L}(q)$ and all variational updates are obtained in closed form, albeit using special functions, such as the digamma function. To optimize computational efficiency, updates are made by blocks, in a

vectorized fashion, for all responses; convergence under such a scheme is guaranteed by the concavity of the objective function, $\mathcal{L}(q)$, in each of the subvectors composing the blocks. Moreover, $\mathcal{L}(q)$ is guaranteed to increase monotonically at every iteration, which provides a useful check against mistakes in the computations or the implementation.

A.2.3 Variational algorithm

Algorithm 3: Structured mean-field variational algorithm

Input: y (centered), X (standardized using the usual unbiased estimator of the variance), $a, b, \eta, \kappa, \lambda, \nu$, tol, maxit, seed

initialize: $M = \{\mu_{\beta,st}\}$, $\Sigma = \left\{\sigma^2_{\beta,st}\right\}$, $\Gamma^{(1)} = \left\{\gamma^{(1)}_{st}\right\}$, $\tau^{(1)} = \left\{\tau^{(1)}_t\right\}$
$\mathcal{L}(q) \leftarrow -\infty$, it $\leftarrow 0$

repeat

$\left(\sigma^{-2}\right)^{(1)} \leftarrow \lambda^*/\nu^*$,
where $\lambda^* = \lambda + \frac{1}{2}\mathbb{1}_p^T \Gamma^{(1)} \mathbb{1}_q$, $\nu^* = \nu + \frac{1}{2}\mathbb{1}_p^T \left\{(\Sigma + M \odot M) \odot \Gamma^{(1)}\right\} \tau^{(1)}$; // $\mathbf{E_q}\left(\boldsymbol{\sigma^{-2}}\right)$

$\tau^{(1)} \leftarrow \eta^* \oslash \kappa^*$,
where $\eta^* = \eta + \frac{n}{2}\mathbb{1}_q + \frac{1}{2}\left(\Gamma^{(1)}\right)^T \mathbb{1}_p$, $A_s = X_s\left(M \odot \Gamma^{(1)}\right)_s$.
$\kappa^* = \kappa + \frac{1}{2}\left(y \odot y\right)^T \mathbb{1}_n - \left\{X\left(M \odot \Gamma^{(1)}\right) \odot y\right\}^T \mathbb{1}_n + \left(\sum_{s=1}^{p-1} A_s \odot \sum_{j=s+1}^{p} A_j\right)^T \mathbb{1}_n$
$+ \frac{1}{2}\left\{n - 1 + \left(\sigma^{-2}\right)^{(1)}\right\}\left\{\Gamma^{(1)} \odot (\Sigma + M \odot M)\right\}^T \mathbb{1}_p$; // $\mathbf{E_q}\left(\boldsymbol{\tau_t}\right)$

$\Sigma \leftarrow \mathbb{1}_p \mathbb{1}_q^T \oslash B$,
where $B = \left\{n - 1 + \left(\sigma^{-2}\right)^{(1)}\right\} \mathbb{1}_p \left(\tau^{(1)}\right)^T$; // $\mathbf{Var_q}\left(\boldsymbol{\beta_{st}} \,|\, \boldsymbol{\gamma_{st}} = \mathbf{1}\right)$

$\log(\tau)^{(1)} \leftarrow \Psi(\eta^*) - \log(\kappa^*)$,
$\log\left(\sigma^{-2}\right)^{(1)} \leftarrow \Psi(\lambda^*) - \log(\nu^*)$,
$\log(\omega)^{(1)} \leftarrow \Psi\left(a + \Gamma^{(1)}\mathbb{1}_q\right) - \Psi\left(a + b + q\mathbb{1}_p\right)$, $\log(1-\omega)^{(1)} \leftarrow \Psi\left(b - \Gamma^{(1)}\mathbb{1}_q + q\mathbb{1}_p\right) - \Psi\left(a + b + q\mathbb{1}_p\right)$;

for $s = \text{shuffle}(1, \ldots, p)$ **do**
for $t \in \{1, \ldots, q\}$ *(parallel)* **do**
$M_{st} \leftarrow \Sigma_{st}\tau^{(1)}_t X_s^T \left(y_t - \sum_{j=1, j\neq s}^{p} \Gamma^{(1)}_{jt} M_{jt} X_j\right)$; // $\mathbf{E_q}\left(\boldsymbol{\beta_{st}} \,|\, \boldsymbol{\gamma_{st}} = \mathbf{1}\right)$
$\Gamma^{(1)}_{st} \leftarrow \left[\exp\left\{\log(1-\omega_s)^{(1)} - \log(\omega_s)^{(1)} - \frac{1}{2}\log(\tau_t)^{(1)} - \frac{1}{2}\log(\sigma^{-2})^{(1)} - \frac{1}{2}\left(M_{st}\right)^2 \left(\Sigma_{st}\right)^{-1}\right\}\right.$
$\left. \times \left(\Sigma_{st}\right)^{-1/2} + 1\right]^{-1}$; // $\mathbf{E_q}\left(\boldsymbol{\gamma_{st}}\right)$
end
end

$\omega^{(1)} \leftarrow a^* \oslash (a^* + b^*)$,
where $a^* = a + \Gamma^{(1)}\mathbb{1}_q$, $b^* = b - \Gamma^{(1)}\mathbb{1}_q + q\mathbb{1}_p$; // $\mathbf{E_q}\left(\boldsymbol{\omega_s}\right)$

$\mathcal{L}^{\text{old}}(q) \leftarrow \mathcal{L}(q)$
it $\leftarrow$ it $+ 1$
Compute $\mathcal{L}(q)$ based on the current parameter updates (see Appendix A.2.2)

until $\left|\mathcal{L}(q) - \mathcal{L}^{\text{old}}(q)\right| <$ tol *or* it = maxit;

The symbols $\odot$ and $\oslash$ are the Hadamard operators standing for element-wise multiplication and division of two matrices of the same dimension.

A.3 Details on the empirical quality assessment of the variational approximation

A.3.1 Marginal likelihood computation

We have

$$\begin{aligned} p(\boldsymbol{y}) &= \int\cdots\int \mathrm{d}\boldsymbol{\omega}\,\mathrm{d}\sigma^{-2}\, p(\boldsymbol{\omega})\, p\left(\sigma^{-2}\right) \\ &\qquad \times \prod_{t=1}^{q}\left\{\sum_{\boldsymbol{\gamma}_t\in\{0,1\}^p} p\left(\boldsymbol{\gamma}_t \mid \boldsymbol{\omega}\right)\int\cdots\int \mathrm{d}\boldsymbol{\beta}_t\,\mathrm{d}\tau_t\, p\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t,\tau_t\right)\, p\left(\boldsymbol{\beta}_t \mid \boldsymbol{\gamma}_t,\tau_t,\sigma^{-2}\right)\, p(\tau_t)\right\} \\ &= \int\cdots\int \mathrm{d}\boldsymbol{\omega}\,\mathrm{d}\sigma^{-2}\left\{\prod_{s=1}^{p} p(\omega_s)\right\} p\left(\sigma^{-2}\right)\prod_{t=1}^{q}\left\{\sum_{\boldsymbol{\gamma}_t\in\{0,1\}^p} p\left(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t,\sigma^{-2}\right)\prod_{s=1}^{p} p\left(\gamma_{st} \mid \omega_s\right)\right\}, \end{aligned}$$

and one can obtain a closed-form expression for $p\left(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t,\sigma^{-2}\right)$, after integrating out $\boldsymbol{\beta}_t$ and then τ_t. Indeed, proceeding similarly as in George and McCulloch (1997),

$$\begin{aligned} p\left(\boldsymbol{y}_t \mid \tau_t,\boldsymbol{\gamma}_t,\sigma^{-2}\right) &= \int (2\pi)^{-n/2}\tau_t^{n/2}\exp\left\{-\frac{\tau_t}{2}\|\boldsymbol{y}_t-\boldsymbol{X}\boldsymbol{\beta}_t\|^2\right\}(2\pi)^{-q_{\gamma_t}/2}\sigma^{-q_{\gamma_t}}\tau_t^{q_{\gamma_t}/2} \\ &\qquad \times\exp\left\{-\frac{\tau_t}{2}\sigma^{-2}\|\boldsymbol{\beta}_t\|^2\right\}\mathrm{d}\boldsymbol{\beta}_t \\ &= \int (2\pi)^{-n/2-q_{\gamma_t}/2}\tau_t^{n/2+q_{\gamma_t}/2}\exp\left\{-\frac{\tau_t}{2}\left(\boldsymbol{\beta}_{\gamma_t}-\boldsymbol{\mu}_{\beta_t}\right)^T\boldsymbol{V}_{\gamma_t,\sigma^{-2}}\left(\boldsymbol{\beta}_{\gamma_t}-\boldsymbol{\mu}_{\beta_t}\right)\right\} \\ &\qquad \times\exp\left(-\frac{\tau_t}{2}\boldsymbol{S}^2_{\gamma_t}\right)\sigma^{-q_{\gamma_t}}\mathrm{d}\boldsymbol{\beta}_t, \end{aligned}$$

where

$$\begin{aligned} q_{\gamma_t} &= \sum_{s=1}^{p}\gamma_{st}, \qquad \tilde{\boldsymbol{X}}_{\gamma_t} = \begin{pmatrix} \boldsymbol{X}_{\gamma_t} \\ \sigma^{-1}\boldsymbol{I}_{q_{\gamma_t}} \end{pmatrix}, \qquad \tilde{\boldsymbol{y}}_t = \begin{pmatrix} \boldsymbol{y}_t \\ \boldsymbol{0} \end{pmatrix}, \\ \boldsymbol{S}^2_{\gamma_t,\sigma^{-2}} &= \|\tilde{\boldsymbol{y}}_t\|^2 - \tilde{\boldsymbol{y}}_t^T\tilde{\boldsymbol{X}}_{\gamma_t}\left(\tilde{\boldsymbol{X}}_{\gamma_t}^T\tilde{\boldsymbol{X}}_{\gamma_t}\right)^{-1}\tilde{\boldsymbol{X}}_{\gamma_t}^T\tilde{\boldsymbol{y}}_t = \|\boldsymbol{y}_t\|^2 - \boldsymbol{y}_t^T\boldsymbol{X}_{\gamma_t}\boldsymbol{V}^{-1}_{\gamma_t,\sigma^{-2}}\boldsymbol{X}_{\gamma_t}^T\boldsymbol{y}_t, \\ \boldsymbol{V}_{\gamma_t,\sigma^{-2}} &= \tilde{\boldsymbol{X}}_{\gamma_t}^T\tilde{\boldsymbol{X}}_{\gamma_t} = \boldsymbol{X}_{\gamma_t}^T\boldsymbol{X}_{\gamma_t} + \sigma^{-2}\boldsymbol{I}_{q_{\gamma_t}}, \qquad \boldsymbol{\mu}_{\beta_t} = \boldsymbol{V}^{-1}_{\gamma_t,\sigma^{-2}}\tilde{\boldsymbol{X}}_{\gamma_t}^T\tilde{\boldsymbol{y}}_t. \end{aligned}$$

Hence, if $q_{\gamma_t}\neq 0$,

$$p\left(\boldsymbol{y}_t \mid \tau_t,\boldsymbol{\gamma}_t,\sigma^{-2}\right) = (2\pi)^{-n/2}\tau_t^{n/2}\det\left(\boldsymbol{V}_{\gamma_t,\sigma^{-2}}\right)^{-1/2}\exp\left(-\frac{\tau_t}{2}\boldsymbol{S}^2_{\gamma_t}\right)\sigma^{-q_{\gamma_t}}.$$

Now,

$$\begin{aligned} p\left(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t,\sigma^{-2}\right) &= \int p\left(\boldsymbol{y}_t \mid \tau_t,\boldsymbol{\gamma}_t,\sigma^{-2}\right)p(\tau_t)\mathrm{d}\tau_t \\ &= \int (2\pi)^{-n/2}\tau_t^{n/2}\det\left(\boldsymbol{V}_{\gamma_t,\sigma^{-2}}\right)^{-1/2}\exp\left(-\frac{\tau_t}{2}\boldsymbol{S}^2_{\gamma_t}\right)\sigma^{-q_{\gamma_t}}\frac{\kappa_t^{\eta_t}}{\Gamma(\eta_t)}\tau_t^{\eta_t-1}\exp\{-\kappa_t\tau_t\}\,\mathrm{d}\tau_t \\ &= (2\pi)^{-n/2}\det\left(\boldsymbol{V}_{\gamma_t,\sigma^{-2}}\right)^{-1/2}\Gamma\left(\frac{n}{2}+\eta_t\right)\frac{\kappa_t^{\eta_t}}{\Gamma(\eta_t)}\left(\kappa_t+\frac{\boldsymbol{S}^2_{\gamma_t}}{2}\right)^{-n/2-\eta_t}\left(\sigma^{-2}\right)^{q_{\gamma_t}/2}. \end{aligned}$$

If $q_{\gamma_t}=0$, then

$$p\left(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t, \sigma^{-2}\right) = (2\pi)^{-n/2}\,\Gamma\left(\frac{n}{2}+\eta_t\right)\frac{\kappa_t^{\eta_t}}{\Gamma(\eta_t)}\left(\kappa_t + \frac{\|\boldsymbol{y}_t\|^2}{2}\right)^{-n/2-\eta_t}.$$

A.3.2 Simple Monte Carlo posterior quantities

The marginal posterior probability of inclusion for predictor $\boldsymbol{X}_s$ and response $\boldsymbol{y}_t$ can be approximated using simple Monte Carlo sums, as follows,

$$\begin{aligned}
p(\gamma_{st}=1 \mid \boldsymbol{y}) &= \frac{p(\gamma_{st}=1,\boldsymbol{y})}{p(\boldsymbol{y})} \\
&= \frac{1}{p(\boldsymbol{y})}\int\cdots\int \mathrm{d}\boldsymbol{\omega}\,\mathrm{d}\sigma^{-2}\left\{\prod_{s=1}^{p} p(\omega_s)\right\} p\left(\sigma^{-2}\right) \\
&\quad\times\left[\prod_{k\neq t}\left\{\sum_{\boldsymbol{\gamma}_k\in\{0,1\}^p} p\left(\boldsymbol{y}_k \mid \boldsymbol{\gamma}_k,\sigma^{-2}\right)\prod_{j=1}^{p} p\left(\gamma_{jk}\mid\omega_j\right)\right\}\right. \\
&\quad\left.\times\left\{\sum_{\boldsymbol{\gamma}_t\in\{0,1\}^p:\,\gamma_{st}=1} p\left(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t,\sigma^{-2}\right)\prod_{j=1}^{p} p\left(\gamma_{jt}\mid\omega_j\right)\right\}\right] \\
&= \frac{1}{p(\boldsymbol{y})}\,\frac{1}{I}\sum_{i=1}^{I}\left[\prod_{k\neq t}\left\{\sum_{\boldsymbol{\gamma}_k\in\{0,1\}^p} p\left(\boldsymbol{y}_k \mid \boldsymbol{\gamma}_k,\left(\sigma^{-2}\right)^{(i)}\right)\prod_{j=1}^{p} p\left(\gamma_{jk}\mid\omega_j^{(i)}\right)\right\}\right. \\
&\quad\left.\times\left\{\sum_{\boldsymbol{\gamma}_t\in\{0,1\}^p:\,\gamma_{st}=1} p\left(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t,\left(\sigma^{-2}\right)^{(i)}\right)\prod_{j=1}^{p} p\left(\gamma_{jt}\mid\omega_j^{(i)}\right)\right\}\right],
\end{aligned}$$

where the samples are generated independently from

$$\left(\sigma^{-2}\right)^{(i)} \sim \text{Gamma}(\lambda,\nu), \qquad \omega_s^{(i)} \sim \text{Beta}(a_s,b_s), \quad s=1,\ldots,p, \qquad i=1,\ldots,I. \tag{A.7}$$

Similarly, we approximate the posterior mean for ω_s as

$$\begin{aligned}
\mathrm{E}(\omega_s \mid \boldsymbol{y}) &= \int \omega_s\, p(\omega_s\mid\boldsymbol{y})\mathrm{d}\omega_s = \frac{1}{p(\boldsymbol{y})}\int \omega_s\, p(\omega_s,\boldsymbol{y})\mathrm{d}\omega_s \\
&= \frac{1}{p(\boldsymbol{y})}\int\cdots\int \mathrm{d}\boldsymbol{\omega}\,\mathrm{d}\sigma^{-2}\,\omega_s\left\{\prod_{j=1}^{p} p(\omega_j)\right\} p\left(\sigma^{-2}\right) \\
&\quad\times\prod_{t=1}^{q}\left\{\sum_{\boldsymbol{\gamma}_t\in\{0,1\}^p} p\left(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t,\sigma^{-2}\right)\prod_{j=1}^{p} p\left(\gamma_{jt}\mid\omega_j\right)\right\} \\
&= \frac{1}{p(\boldsymbol{y})}\,\frac{1}{I}\sum_{i=1}^{I}\omega_s^{(i)}\prod_{t=1}^{q}\left\{\sum_{\boldsymbol{\gamma}_t\in\{0,1\}^p} p\left(\boldsymbol{y}_t \mid \boldsymbol{\gamma}_t,\left(\sigma^{-2}\right)^{(i)}\right)\prod_{j=1}^{p} p\left(\gamma_{jt}\mid\omega_j^{(i)}\right)\right\}.
\end{aligned}$$

Figure A.1 displays and compares the posterior inclusion probabilities obtained by variational, MCMC or simple Monte Carlo approximations, for the experiment of Section 3.3.2.

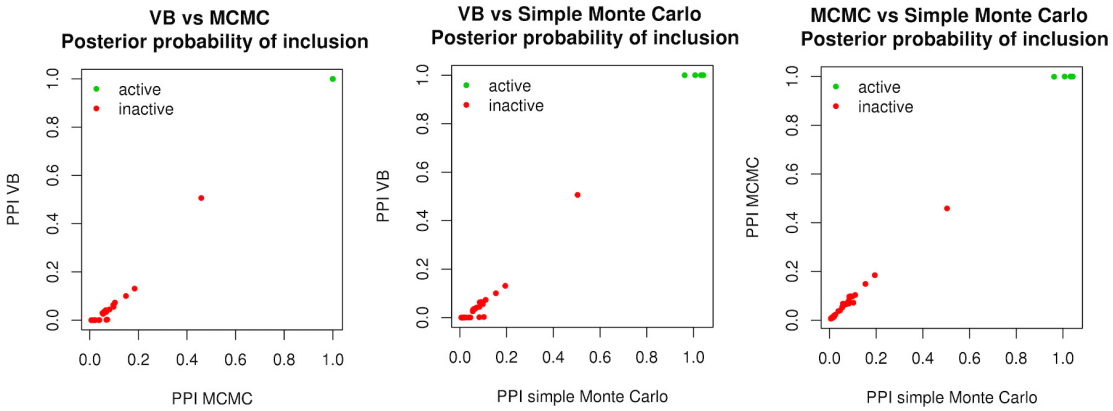

Figure A.1 – Comparison of marginal posterior probabilities of inclusion (PPI) for variational Bayes (VB), MCMC, and simple Monte Carlo approximations. The posterior probabilities of inclusion corresponding to the true signals all overlap.

A.3.3 Competing predictor selection methods

The performance of our approach in terms of predictor selection is compared with the following regression procedures:

1. univariate ordinary least squares: each response $\boldsymbol{y}_t$ is regressed on each predictor $\boldsymbol{X}_s$ and the statistics $\max_{t=1,\ldots,q} t_{st}$, where t_{st} is the t-statistic for the significance of β_{st}, are gathered and ranked;

2. elastic net regression for multivariate Gaussian responses ($\alpha = 0.5$) with 10-fold cross-validation for choosing the tuning parameter λ (glmnet, Friedman et al., 2009). The estimates $|\beta_s|$ ($s = 1,\ldots,p$), where β_s is the regression coefficient for $\boldsymbol{X}_s$ and common to all responses, are gathered and ranked;

3. univariate Bayesian regressions, lmBF (Morey and Rouder, 2015): each response $\boldsymbol{y}_t$ is regressed on each predictor $\boldsymbol{X}_s$ with all computations made analytically. The (average) Bayes factors, $\sum_{t=1}^{q} \mathrm{BF}_{st}/q$ ($s = 1,\ldots,p$), are gathered and ranked;

4. q Bayesian multiple regressions, BAS (Clyde, 2016), one for each response, using MCMC inference. A g-prior is used for the regression coefficients. The (average) Bayes factors, $\sum_{t=1}^{q} \mathrm{BF}_{st}/q$ ($s = 1,\ldots,p$), are gathered and ranked; and

5. q Bayesian multiple regressions, varbvs (Carbonetto and Stephens, 2012), one for each response, using variational inference. The posterior probabilities of inclusion are summed across responses and ranked.

A.4 Details on the real data problem

A.4.1 Permutation-based Bayesian false discovery rate estimation

We detail the false discovery rate estimation procedure applied in Section 3.6 to compare our method with the varbvs method of Carbonetto and Stephens (2012) on the real data. We use the two-group

mixture approach proposed by Efron (2008) in the context of microarray data analysis (recall Section 2.2): we simultaneously consider N null hypotheses and their corresponding test statistics, which we assume to follow a mixture distribution

$$F = (1-\pi_0)F_1 + \pi_0 F_0,$$

where π_0 is the prior probability for a null case, and F_0 and F_1 are the null and non-null cumulative distribution functions. We derive Bayesian false discovery rate values for thresholds τ,

$$\mathrm{FDR}(\tau) = \frac{\pi_0 \bar{F}_0(\tau)}{\bar{F}(\tau)}, \tag{A.8}$$

where $\bar{F} = 1-F$ and $\bar{F}_0 = 1-F_0$, using permutation-based estimates. Specifically, we obtain an empirical null distribution by running our algorithm (with the same hyperparameters as those used for the actual inference) on B datasets with permuted outcome sample labels and compute, for a grid of thresholds $0 < \tau_1 < \cdots < \tau_K < 1$,

$$\widehat{\mathrm{FDR}}(\tau_k) = \frac{\mathrm{median}_{b=1,\ldots,B} \#\{\mathrm{PPI}_{st}^{(b)} > \tau_k;\ s = 1,\ldots,p;\ t = 1,\ldots,q\}}{\#\{\mathrm{PPI}_{st} > \tau_k;\ s = 1,\ldots,p;\ t = 1,\ldots,q\}}, \qquad k = 1,\ldots,K, \tag{A.9}$$

where PPI_{st} is the posterior probability of inclusion $\mathrm{pr}(\gamma_{st} = 1 \mid \boldsymbol{y})$, $\mathrm{PPI}_{st}^{(b)}$ is the corresponding posterior probability of inclusion when running the method on the permuted dataset b, and where we conservatively set π_0 to 1 in (A.8). To find thresholds $\hat{\tau}$ corresponding to preselected false discovery rates, we fit a cubic spline to the estimates (A.9) obtained for $\tau_1,\ldots,\tau_K$.

A.4.2 Biological evidence for the mQTL analysis findings

We used public association results from the following databases to support the mQTL findings of Section 3.6: GWAS Catalog (Welter et al., 2014), UCSC genome browser (Karolchik et al., 2003), GTEx (GTEx Consortium, 2015) and GeneCards (Rebhan et al., 1998). Twelve of the 25 SNPs identified by LOCUS (rs3820711, rs4316911, rs4909818, rs4744227, rs174535, rs680379, rs8012466, rs4906771, rs573922, rs3903703, rs8114788 and rs6001093) have been reported as associated with BMI, diverse diabetic or obesity diseases, fatty acid, sphingolipid or phospholipid levels.

A.5 Variational algorithms for some model extensions

We provide general results about the variational algorithms for the model extensions described in Section 3.7.

A.5.1 Confounding variables not subject to selection

We consider including d covariates $\boldsymbol{Z}$ in the model as

$$\boldsymbol{y}_t = \boldsymbol{Z}\boldsymbol{\alpha}_t + \boldsymbol{X}\boldsymbol{\beta}_t + \boldsymbol{\varepsilon}_t, \qquad \alpha_{rt} \sim \mathcal{N}\left(0, \zeta_r^2\right), \qquad \zeta_r^{-2} \sim \mathrm{Gamma}\left(\phi_r, \xi_r\right), \qquad r = 1,\ldots,d,$$

keeping the rest of the hierarchy untouched.

The corresponding mean-field factors are

$$q(\boldsymbol{\alpha}, \boldsymbol{\zeta}) = \left\{ \prod_{t=1}^{q} \prod_{r=1}^{d} q(\alpha_{rt}) \right\} \left\{ \prod_{r=1}^{d} q\left(\zeta_r^{-2}\right) \right\},$$

and since

$$\log q(\alpha_{rt}) = -\frac{1}{2} \mathrm{E}_{-\alpha_{rt}} \| \boldsymbol{y}_t - \boldsymbol{X}\boldsymbol{\beta}_t - \boldsymbol{Z}\boldsymbol{\alpha}_t \|^2 - \frac{\mathrm{E}_q\left(\zeta_r^{-2}\right)}{2} \alpha_{rt}^2 + \mathrm{cst},$$

we identify a normal distribution

$$\alpha_{rt} \mid \boldsymbol{y} \sim \mathcal{N}\left(\mu_{\alpha,rt}, \sigma_{\alpha,rt}^2\right),$$

where

$$\sigma_{\alpha,rt}^{-2} = \| \boldsymbol{Z}_r \|^2 + \left(\zeta_r^{-2}\right)^{(1)}, \qquad \mu_{\alpha,rt} = \sigma_{\alpha,rt}^2 \boldsymbol{Z}_r^T \left(\boldsymbol{y}_t - \sum_{l=1, l\neq r}^{d} \mu_{\alpha,lt} \boldsymbol{Z}_l - \sum_{s=1}^{p} \gamma_{st}^{(1)} \mu_{\beta,st} \boldsymbol{X}_s \right),$$

with

$$\mu_{\beta,st} = \sigma_{\beta,st}^2 \tau_t^{(1)} \boldsymbol{X}_s^T \left(\boldsymbol{y}_t - \sum_{r=1}^{q} \mu_{\alpha,rt} \boldsymbol{Z}_r - \sum_{j=1, j\neq s}^{p} \gamma_{jt}^{(1)} \mu_{\beta,jt} \boldsymbol{X}_j \right),$$

and the updates for $\sigma_{\beta,st}^2$ and $\gamma_{st}^{(1)}$ as for the reference model.

Then we have

$$\log q\left(\zeta_r^{-2}\right) = \left(\frac{q}{2} + \phi_r - 1\right) \log \zeta_r^{-2} - \left\{ \frac{1}{2} \sum_{t=1}^{q} \left(\mu_{\alpha,rt}^2 + \sigma_{\alpha,rt}^2\right) + \xi_r \right\} \zeta_r^{-2} + \mathrm{cst},$$

we identify a Gamma distribution

$$\zeta_r^{-2} \mid y \sim \mathrm{Gamma}\left(\phi_r^*, \xi_r^*\right),$$

with

$$\phi_r^* = \phi_r + \frac{q}{2}, \qquad \xi_r^* = \xi_r + \frac{1}{2} \sum_{t=1}^{q} \left(\mu_{\alpha,rt}^2 + \sigma_{\alpha,rt}^2\right).$$

The rest of the updates are unchanged and computations for the variational lower bound is straightforwardly adapted from those for the reference model.

A.5.2 Logistic regression model

We replace the linear link of model (3.1)–(3.2)–(3.3) with a logit link specification

$$y_{it} \mid \boldsymbol{\beta}_t \sim \mathrm{Bernoulli}\left\{ \mathrm{Sig}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t\right) \right\}, \qquad i = 1, \ldots, n,\ t = 1, \ldots, q, \tag{A.10}$$

y_{it} is the response t for sample i, $\boldsymbol{X}_i$ is the $p \times 1$ candidate predictor vector for sample i, and $\mathrm{Sig}(z) = (1 + e^{-z})^{-1}$ is the sigmoid function; the rest of the model hierarchy is unchanged. The resulting non-conjugacy of the model prevents the derivation of coordinate ascent updates in closed form. We resort to a local approximation that seeks a lower bound on the conditional distribution $p(\boldsymbol{y} \mid \boldsymbol{\beta})$ (Jaakkola

and Jordan, 2000); the bound is expressed as the exponential of a quadratic form, and thus gives rise to a Gaussian approximation.

We have

$$\begin{aligned} p(y_{it} \mid \boldsymbol{\beta}_t) &= \mathrm{Sig}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t\right)^{y_{it}} \left\{1 - \mathrm{Sig}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t\right)\right\}^{1-y_{it}} \\ &= \exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t y_{it}\right) \mathrm{Sig}\left(-\boldsymbol{X}_i^T \boldsymbol{\beta}_t\right) \\ &\geq \exp\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t y_{it}\right) \mathrm{Sig}\left(\eta_{it}\right) \exp\left[-\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t + \eta_{it}\right)/2 - \rho(\eta_{it}) \left\{\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t\right)^2 - \eta_{it}^2\right\}\right] \\ &=: h(\boldsymbol{\beta}_t, \eta_{it}), \end{aligned}$$

where η_{it} is an auxiliary parameter, and where

$$\rho(\eta) = \frac{1}{2\eta}\left\{\mathrm{Sig}(\eta) - \frac{1}{2}\right\};$$

see Bishop (2006, , Chap. 10) for the derivation of this bound. Writing $h(\boldsymbol{\beta}, \boldsymbol{\eta}) = \prod_{t=1}^{q} \prod_{i=1}^{n} h(\boldsymbol{\beta}_t, \eta_{it})$, where $\boldsymbol{\eta} = \{\eta_{it}\}_{i=1,\ldots,n, t=1,\ldots,q}$, we thus obtain a lower bound on the variational objective function, $\mathcal{L}(q)$,

$$\begin{aligned} \log p(\boldsymbol{y}) &\geq \mathcal{L}(q) \\ &= \mathrm{E}_q \log\left\{p(\boldsymbol{y} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^{-2}) p(\boldsymbol{\gamma} \mid \boldsymbol{\omega}) p(\boldsymbol{\omega}) p(\sigma^{-2})\right\} - \mathrm{E}_q \log q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \sigma^{-2}) \\ &\geq \mathrm{E}_q \log\left\{h(\boldsymbol{\beta}, \boldsymbol{\eta}) p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^{-2}) p(\boldsymbol{\gamma} \mid \boldsymbol{\omega}) p(\boldsymbol{\omega}) p(\sigma^{-2})\right\} - \mathrm{E}_q \log q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \sigma^{-2}) \\ &=: \mathcal{L}(q, \boldsymbol{\eta}), \end{aligned}$$

where $q(\cdot)$ is the mean field variational approximation, whose factors can now be obtained using $\mathcal{L}(q, \boldsymbol{\eta})$ as objective function. Inference requires optimizing the auxiliary parameter $\boldsymbol{\eta}$ using an expectation-maximization algorithm. The algorithm cycles between updating the mean-field factors of $q(\cdot)$, keeping $\boldsymbol{\eta}$ fixed, and optimizing the bound $\mathcal{L}(q, \boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$. This latter step is obtained by developing $\mathcal{L}(q, \boldsymbol{\eta})$ and omitting all the terms that do not depend on $\boldsymbol{\eta}$, i.e.,

$$\begin{aligned} \mathcal{L}(q, \boldsymbol{\eta}) &= \mathrm{E}_q\left\{\log h(\boldsymbol{\beta}, \boldsymbol{\eta})\right\} + \mathrm{cst} \\ &= \sum_{t=1}^{q} \sum_{i=1}^{n} \left[\log \mathrm{Sig}(\eta_{it}) - \frac{\eta_{it}}{2} - \rho(\eta_{it}) \left\{\boldsymbol{X}_i^T \mathrm{E}_q\left(\boldsymbol{\beta}_t \boldsymbol{\beta}_t^T\right) \boldsymbol{X}_i - \eta_{it}^2\right\}\right] + \mathrm{cst}, \end{aligned}$$

where cst is constant with respect to $\boldsymbol{\eta}$. Maximizing this yields, after a little algebra,

$$(\eta_{it})^2 = \boldsymbol{X}_i^T \mathrm{E}_q\left(\boldsymbol{\beta}_t \boldsymbol{\beta}_t^T\right) \boldsymbol{X}_i = \sum_{s=1}^{p} X_{is}^2 \gamma_{st}^{(1)} \left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right) + \sum_{s=1}^{p} X_{is} \gamma_{st}^{(1)} \mu_{\beta,st} \sum_{j=1, j \neq s}^{p} X_{ij} \gamma_{jt}^{(1)} \mu_{\beta,jt},$$

where $\gamma_{st}^{(1)}$, $\mu_{\beta,st}$ and $\sigma_{\beta,st}^2$ are the variational parameters for β_{st} and γ_{st} given by

$$\mu_{\beta,st} = \sigma_{\beta,st}^2 \sum_{i=1}^{n} \left\{X_{is}\left(y_{it} - \frac{1}{2}\right) - 2\rho(\eta_{it}) X_{is} \sum_{j=1, j \neq s}^{p} \gamma_{jt}^{(1)} \mu_{\beta,jt} X_{ij}\right\}, \quad \sigma_{\beta,st}^{-2} = 2 \sum_{i=1}^{n} \rho(\eta_{it}) X_{is}^2 + \left(\sigma^{-2}\right)^{(1)},$$

and $\gamma_{st}^{(1)}$ is given by (A.3), dropping the term involving τ_t. The rest of the variational updates are unchanged and derivation of the variational lower bound poses no difficulty.

A.5.3 Probit regression model

We replace the linear link of model (3.1)–(3.2)–(3.3) with the probit link specification (Albert and Chib, 1993)

$$y_{it} \mid \boldsymbol{\beta}_t \sim \text{Bernoulli}\left\{\Phi\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t\right)\right\}, \qquad i = 1,\ldots,n,\ t = 1,\ldots,q, \tag{A.11}$$

where $\Phi(\cdot)$ is the standard normal cumulative function, y_{it} is the response t for sample i and $\boldsymbol{X}_i$ is the $p \times 1$ candidate predictor vector for sample i; the rest of the model hierarchy is unchanged. We employ the usual reparametrization of (A.11), based on an auxiliary variable w_{it},

$$y_{it} = \mathbb{1}\{w_{it} > 0\}, \qquad w_{it} \mid \boldsymbol{\beta}_t \sim \mathcal{N}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t, 1\right), \qquad i = 1,\ldots,n,\ t = 1,\ldots,q.$$

The mean-field distribution corresponding to w_{it} is

$$q(w_{it}) = \left\{\frac{\mathbb{1}(w_{it} > 0)}{\Phi\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t^{(1)}\right)}\right\}^{y_{it}} \left\{\frac{\mathbb{1}(w_{it} \le 0)}{1 - \Phi\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t^{(1)}\right)}\right\}^{1-y_{it}} (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}\left(w_{it} - \boldsymbol{X}_i^T \boldsymbol{\beta}_t^{(1)}\right)^2\right\},$$

where cst does not depend on w_{it}, that is, the variational distribution of w_{it} is a truncated normal variable: for $\delta \in \{0,1\}$,

$$w_{it} \mid \boldsymbol{y} = \delta \sim \mathcal{TN}\left(\boldsymbol{X}_i^T \boldsymbol{\beta}_t^{(1)}, 1; \left\{0 < (-1)^{1-\delta} w_{it}\right\}\right).$$

The computations for corresponding variational updates and contribution to the variational objective function $\mathcal{L}(q)$ are similar to those of Appendix B.2.

A.5.4 Mixed linear-probit regression model

The model simply represents the binary responses using a probit link and the continuous responses using a linear link; the algorithm is therefore a straightforward extension of the algorithms described in Appendices A.2 and A.5.3.

B Appendix for Chapter 4

B.1 Derivation of the variational algorithm for the group sparsity model

B.1.1 Variational distributions

We detail the derivation of the variational algorithm for the group sparsity model presented in Section 4.2.1. Consider q centered responses, $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q)$, and p centered predictors, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, for n samples, and let $\boldsymbol{\nu} = (\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}, \boldsymbol{\tau}, \sigma^{-2})$. We have

$$\begin{aligned} p(\boldsymbol{y}, \boldsymbol{\nu}) &= \left\{ \prod_{t=1}^{q} p\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t\right) \right\} \left\{ \prod_{t=1}^{q} \prod_{g=1}^{G} p\left(\boldsymbol{\beta}_{gt} \mid \gamma_{gt}, \sigma^{-2}, \tau_t\right) \right\} \\ &\quad \times \left\{ \prod_{t=1}^{q} \prod_{g=1}^{G} p\left(\gamma_{gt} \mid \omega_g\right) \right\} \left\{ \prod_{g=1}^{G} p(\omega_g) \right\} \left\{ \prod_{t=1}^{q} p\left(\tau_t\right) \right\} p\left(\sigma^{-2}\right), \end{aligned}$$

with same distributions as for the reference hierarchical model, see Appendix A.2.1, except for the conditional distribution of $\boldsymbol{\beta}_{gt}$ which has

$$\boldsymbol{\beta}_{gt} \mid \gamma_{gt}, \tau_t, \sigma^{-2} \sim \gamma_{gt} \mathcal{N}_{|g|}\left(0, \sigma^2 \tau_t^{-1} \boldsymbol{I}_{|g|}\right) + (1 - \gamma_{gt}) \delta_0,$$

where $|g|$ is the cardinality of group g, and δ_0 is a point mass at $\mathbf{0} \in \mathbb{R}^{|g|}$.

Consider the following mean-field form for the variational approximation,

$$q(\boldsymbol{\nu}) = \left\{ \prod_{t=1}^{q} \prod_{g=1}^{G} q(\boldsymbol{\beta}_{gt}, \gamma_{gt}) \right\} \left\{ \prod_{g=1}^{G} q(\omega_g) \right\} \left\{ \prod_{t=1}^{q} q\left(\tau_t\right) \right\} q\left(\sigma^{-2}\right).$$

The variational updates are as follows. We have

$$\begin{aligned} q(\boldsymbol{\beta}_{gt}, \gamma_{gt}) &= \operatorname{cst} \left[(2\pi)^{-|g|/2} \exp\left\{ -\frac{1}{2} \tau_t^{(1)} \boldsymbol{\beta}_{gt}^T \left(\boldsymbol{X}_g^T \boldsymbol{X}_g + \left(\sigma^{-2}\right)^{(1)} \boldsymbol{I}_g \right) \boldsymbol{\beta}_{gt} + \tau_t^{(1)} \left(\boldsymbol{y}_t - \boldsymbol{X}_{-g} \boldsymbol{\beta}_{-gt}^{(1)} \right)^T \boldsymbol{X}_g \boldsymbol{\beta}_{gt} \right\} \right]^{\gamma_{gt}} \\ &\quad \times \exp\left\{ \frac{|g|}{2} \mathrm{E}_q\left(\log \sigma^{-2}\right) + \frac{|g|}{2} \mathrm{E}_q\left(\log \tau_t\right) + \mathrm{E}_q\left(\log \omega_g\right) \right\}^{\gamma_{gt}} \\ &\quad \times \left\{ \delta_0(\boldsymbol{\beta}_{gt}) \right\}^{1-\gamma_{gt}} \exp\left\{ \mathrm{E}_q \log\left(1 - \omega_g\right) \right\}^{1-\gamma_{gt}}, \end{aligned}$$

where

$$\boldsymbol{\beta}^{(1)}_{-gt} = \left(\boldsymbol{\beta}^{(1)}_{1t}, \ldots, \boldsymbol{\beta}^{(1)}_{(g-1)t}, \boldsymbol{\beta}^{(1)}_{(g+1)t}, \ldots, \boldsymbol{\beta}^{(1)}_{Gt}\right), \qquad \boldsymbol{\beta}^{(1)}_{gt} = \mathrm{E}_q\left(\boldsymbol{\beta}_{gt}\right) = \gamma_{gt}\boldsymbol{\mu}_{gt}.$$

Hence, we have

$$q(\boldsymbol{\beta}_{gt}, \gamma_{gt}) = q(\boldsymbol{\beta}_{gt} \mid \gamma_{gt})\, q(\gamma_{gt}),$$

with

$$\boldsymbol{\beta}_{gt} \mid \gamma_{gt} = 1, \boldsymbol{y} \sim \mathcal{N}_{|g|}\left(\boldsymbol{\mu}_{\beta,gt}, \boldsymbol{\Sigma}_{\beta,gt}\right), \quad \boldsymbol{\beta}_{gt} \mid \gamma_{gt} = 0, \boldsymbol{y} \sim \delta_0, \quad \gamma_{gt} \mid \boldsymbol{y} \sim \text{Bernoulli}\left(\gamma^{(1)}_{gt}\right),$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1}_{\beta,gt} &= \tau^{(1)}_t \left\{\boldsymbol{X}_g^T \boldsymbol{X}_g + \left(\sigma^{-2}\right)^{(1)} \boldsymbol{I}_g\right\}, && t = 1, \ldots, q, \\
\boldsymbol{\mu}_{\beta,gt} &= \tau^{(1)}_t \boldsymbol{\Sigma}_{\beta,gt} \boldsymbol{X}_g^T \left(\boldsymbol{y}_t - \boldsymbol{X}_{-g} \boldsymbol{\beta}^{(1)}_{-gt}\right), && g = 1, \ldots, G,
\end{aligned}$$

and

$$\begin{aligned}
\gamma^{(1)}_{gt} = \Big[& 1 + \det\left(\boldsymbol{\Sigma}_{\beta,gt}\right)^{-1/2} \\
& \times \exp\left\{\mathrm{E}_q \log\left(1 - \omega_g\right) - \mathrm{E}_q\left(\log \omega_g\right) - \frac{|g|}{2}\mathrm{E}_q\left(\log \sigma^{-2}\right) - \frac{|g|}{2}\mathrm{E}_q\left(\log \tau_t\right) - \frac{1}{2}\boldsymbol{\mu}^T_{\beta,gt}\boldsymbol{\Sigma}^{-1}_{\beta,gt}\boldsymbol{\mu}_{\beta,gt}\right\}\Big]^{-1}.
\end{aligned}$$

The updates for ω_g are the same as for the reference model, namely,

$$\omega_g \mid \boldsymbol{y} \sim \text{Beta}\left(a^*_g, b^*_g\right),$$

where

$$a^*_g = a_g + \sum_{t=1}^{q} \gamma^{(1)}_{gt}, \qquad b^*_g = b_g - \sum_{t=1}^{q} \gamma^{(1)}_{gt} + q.$$

Then, we have

$$\begin{aligned}
\log q\left(\sigma^{-2}\right) &= \frac{1}{2}\sum_{t=1}^{q}\sum_{g=1}^{G}\left\{\gamma^{(1)}_{gt}|g|\log\left(\sigma^{-2}\right) - \sigma^{-2}\tau^{(1)}_t \mathrm{E}_q\left(\boldsymbol{\beta}^T_{gt}\boldsymbol{\beta}_{gt}\right)\right\} + (\lambda - 1)\log\left(\sigma^{-2}\right) - \nu\sigma^{-2} + \text{cst} \\
&= \left(\frac{1}{2}\sum_{t=1}^{q}\sum_{g=1}^{G}|g|\gamma^{(1)}_{gt} + \lambda - 1\right)\log\sigma^{-2} - \left(\frac{1}{2}\sum_{t=1}^{q}\sum_{g=1}^{G}\mathrm{E}_q\left(\boldsymbol{\beta}^T_{gt}\boldsymbol{\beta}_{gt}\right)\tau^{(1)}_t + \nu\right)\sigma^{-2} + \text{cst},
\end{aligned}$$

where

$$\mathrm{E}_q\left(\boldsymbol{\beta}^T_{gt}\boldsymbol{\beta}_{gt}\right) = \mathrm{E}_q\left(\boldsymbol{\beta}_{gt}\right)^T \mathrm{E}_q\left(\boldsymbol{\beta}_{gt}\right) + \mathrm{tr}\left\{\mathrm{Var}\left(\boldsymbol{\beta}_{gt}\right)\right\}.$$

Similarly, we will use

$$\mathrm{E}_q\left(\boldsymbol{\beta}^T_{gt}\boldsymbol{X}^T_g\boldsymbol{X}_g\boldsymbol{\beta}_{gt}\right) = \mathrm{E}_q\left(\boldsymbol{\beta}_{gt}\right)^T \boldsymbol{X}^T_g\boldsymbol{X}_g\mathrm{E}_q\left(\boldsymbol{\beta}_{gt}\right) + \mathrm{tr}\left\{\boldsymbol{X}^T_g\boldsymbol{X}_g\mathrm{Var}\left(\boldsymbol{\beta}_{gt}\right)\right\},$$

where

$$\begin{aligned}
\mathrm{Var}\left(\boldsymbol{\beta}_{gt}\right) &= \mathrm{E}_q\left(\boldsymbol{\beta}_{gt}\boldsymbol{\beta}^T_{gt}\right) - \mathrm{E}_q\left(\boldsymbol{\beta}_{gt}\right)\mathrm{E}_q\left(\boldsymbol{\beta}_{gt}\right)^T \\
&= \mathrm{E}_q\left(\boldsymbol{\beta}_{gt}\boldsymbol{\beta}^T_{gt} \mid \gamma_{gt} = 1\right)\gamma^{(1)}_{gt} - \mathrm{E}_q\left(\boldsymbol{\beta}_{gt} \mid \gamma_{gt} = 1\right)\mathrm{E}_q\left(\boldsymbol{\beta}_{gt} \mid \gamma_{gt} = 1\right)^T\left(\gamma^{(1)}_{gt}\right)^2 \\
&= \gamma^{(1)}_{gt}\left\{\boldsymbol{\Sigma}_{gt} + \boldsymbol{\mu}_{\beta,gt}\boldsymbol{\mu}^T_{\beta,gt}\right\} - \left(\gamma^{(1)}_{gt}\right)^2\boldsymbol{\mu}_{\beta,gt}\boldsymbol{\mu}^T_{\beta,gt} = \gamma^{(1)}_{gt}\boldsymbol{\Sigma}_{gt} + \gamma^{(1)}_{gt}\left(1 - \gamma^{(1)}_{gt}\right)\boldsymbol{\mu}_{\beta,gt}\boldsymbol{\mu}^T_{\beta,gt}.
\end{aligned}$$

We identify a Gamma distribution

$$\sigma^{-2} \mid \boldsymbol{y} \sim \text{Gamma}\left(\lambda^*, \nu^*\right),$$

with

$$\lambda^* = \lambda + \frac{1}{2}\sum_{g=1}^{G}|g|\sum_{t=1}^{q}\gamma_{gt}^{(1)}, \qquad \nu^* = \nu + \frac{1}{2}\sum_{t=1}^{q}\sum_{g=1}^{G}\text{E}_q\left(\boldsymbol{\beta}_{gt}^T\boldsymbol{\beta}_{gt}\right)\tau_t^{(1)}.$$

Finally, we have

$$\begin{aligned}
\log q(\tau_t) &= \text{E}_{-\tau_t}\left\{\log p\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t\right)\right\} + \sum_{g=1}^{G}\text{E}_{-\tau_t}\left\{\log p\left(\boldsymbol{\beta}_{gt} \mid \gamma_{gt}, \tau_t, \sigma^{-2}\right)\right\} + \log p(\tau_t) + \text{cst} \\
&= \frac{n}{2}\log\tau_t - \frac{\tau_t}{2}\text{E}_q\left(\|\boldsymbol{y}_t - \boldsymbol{X}\boldsymbol{\beta}_t\|^2\right) + \frac{1}{2}\log\tau_t\sum_{g=1}^{G}|g|\gamma_{gt}^{(1)} - \frac{\tau_t}{2}\left(\sigma^{-2}\right)^{(1)}\sum_{g=1}^{G}\text{E}_q\left(\boldsymbol{\beta}_{gt}^T\boldsymbol{\beta}_{gt}\right) \\
&\quad + (\eta_t - 1)\log\tau_t - \kappa_t\tau_t + \text{cst} \\
&= \log\tau_t\left(\eta_t + \frac{n}{2} + \frac{1}{2}\sum_{g=1}^{G}|g|\gamma_{gt}^{(1)} - 1\right) - \tau_t\left[\kappa_t + \frac{1}{2}\|\boldsymbol{y}_t\|^2 - \boldsymbol{y}_t^T\sum_{g=1}^{G}\boldsymbol{X}_g\boldsymbol{\mu}_{\beta,gt}\gamma_{gt}^{(1)}\right. \\
&\quad + \sum_{g=1}^{G-1}\gamma_{gt}^{(1)}\boldsymbol{\mu}_{\beta,gt}\boldsymbol{X}_g^T\sum_{g'=g+1}^{G}\boldsymbol{X}_{g'}\boldsymbol{\mu}_{\beta,g't}\gamma_{g't}^{(1)} + \frac{1}{2}\sum_{g=1}^{G}\text{E}_q\left(\boldsymbol{\beta}_{gt}^T\boldsymbol{X}_g^T\boldsymbol{X}_g\boldsymbol{\beta}_{gt}\right) \\
&\quad \left. + \frac{1}{2}\sum_{g=1}^{G}\left(\sigma^{-2}\right)^{(1)}\text{E}_q\left(\boldsymbol{\beta}_{gt}^T\boldsymbol{\beta}_{gt}\right)\right] + \text{cst}.
\end{aligned}$$

Therefore

$$\tau_t \mid \boldsymbol{y} \sim \text{Gamma}\left(\eta_t^*, \kappa_t^*\right), \qquad \tau_t^{(1)} = \eta_t^*/\kappa_t^*,$$

where

$$\begin{aligned}
\eta_t^* &= \eta_t + \frac{n}{2} + \frac{1}{2}\sum_{g=1}^{G}|g|\gamma_{gt}^{(1)}, \\
\kappa_t^* &= \kappa_t + \frac{1}{2}\|\boldsymbol{y}_t\|^2 - \boldsymbol{y}_t^T\sum_{g=1}^{G}\boldsymbol{X}_g\boldsymbol{\mu}_{\beta,gt}\gamma_{gt}^{(1)} + \sum_{g=1}^{G-1}\gamma_{gt}^{(1)}\boldsymbol{\mu}_{\beta,gt}\boldsymbol{X}_g^T\sum_{g'=g+1}^{G}\boldsymbol{X}_{g'}\boldsymbol{\mu}_{\beta,g't}\gamma_{g't}^{(1)} \\
&\quad + \frac{1}{2}\sum_{g=1}^{G}\text{E}_q\left(\boldsymbol{\beta}_{gt}^T\boldsymbol{X}_g^T\boldsymbol{X}_g\boldsymbol{\beta}_{gt}\right) + \frac{1}{2}\sum_{g=1}^{G}\left(\sigma^{-2}\right)^{(1)}\text{E}_q\left(\boldsymbol{\beta}_{gt}^T\boldsymbol{\beta}_{gt}\right).
\end{aligned}$$

B.1.2 Variational lower bound

We provide the variational lower bound, $\mathcal{L}(q)$, of the marginal log-likelihood, $\log p(\boldsymbol{y})$:

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{\nu})\log\left\{\frac{p(\boldsymbol{y},\boldsymbol{\nu})}{q(\boldsymbol{\nu})}\right\}\text{d}\boldsymbol{\nu} \\
&= \sum_{t=1}^{q}\mathcal{L}_y\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \gamma_t, \tau_t\right) + \sum_{t=1}^{q}\sum_{g=1}^{G}\mathcal{L}_\beta\left(\boldsymbol{\beta}_{gt}, \gamma_{gt} \mid \sigma^{-2}, \tau_t\right) + \sum_{g=1}^{G}\mathcal{L}_\omega\left(\omega_g\right) + \mathcal{L}_\sigma\left(\sigma^{-2}\right) + \sum_{t=1}^{q}\mathcal{L}_\tau\left(\tau_t\right),
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{L}_y\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t\right) &= \mathrm{E}_q\left\{\log p(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t)\right\} \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\mathrm{E}_q\left(\log\tau_t\right) - \frac{1}{2}\tau_t^{(1)}\left[\|\boldsymbol{y}_t\|^2 - 2\boldsymbol{y}_t^T\sum_{g=1}^{G}\boldsymbol{X}_g\boldsymbol{\mu}_{\beta,gt}\gamma_{gt}^{(1)}\right. \\
&\quad \left. +2\sum_{g=1}^{G-1}\gamma_{gt}^{(1)}\boldsymbol{\mu}_{\beta,gt}\boldsymbol{X}_g^T\sum_{g'=g+1}^{G}\boldsymbol{X}_{g'}\boldsymbol{\mu}_{\beta,g't}\gamma_{g't}^{(1)} + \sum_{g=1}^{G}\mathrm{E}_q\left(\boldsymbol{\beta}_{gt}^T\boldsymbol{X}_g^T\boldsymbol{X}_g\boldsymbol{\beta}_{gt}\right)\right] \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\mathrm{E}_q\left(\log\tau_t\right) - \tau_t^{(1)}\left\{\kappa_t^* - \frac{1}{2}\sum_{g=1}^{G}\left(\sigma^{-2}\right)^{(1)}\mathrm{E}_q\left(\boldsymbol{\beta}_{gt}^T\boldsymbol{\beta}_{gt}\right) - \kappa_t\right\},
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}_\beta\left(\boldsymbol{\beta}_{gt}, \gamma_{gt} \mid \tau_t, \sigma^{-2}\right) &= \mathrm{E}_q\left\{\log p(\boldsymbol{\beta}_{gt} \mid \gamma_{gt}, \tau_t, \sigma^{-2})\right\} + \mathrm{E}_q\left\{\log p(\gamma_{gt} \mid \omega_g)\right\} - \mathrm{E}_q\left\{\log q(\boldsymbol{\beta}_{gt}, \gamma_{gt})\right\} \\
&= \frac{|g|}{2}\gamma_{gt}^{(1)}\left\{\mathrm{E}_q\left(\log\sigma^{-2}\right) + \mathrm{E}_q\left(\log\tau_t\right)\right\} - \frac{1}{2}\left(\sigma^{-2}\right)^{(1)}\tau_t^{(1)}\mathrm{E}_q\left(\boldsymbol{\beta}_{gt}^T\boldsymbol{\beta}_{gt}\right) \\
&\quad +\gamma_{gt}^{(1)}\mathrm{E}_q\left(\log\omega_g\right) + \left(1-\gamma_{gt}^{(1)}\right)\mathrm{E}_q\left\{\log(1-\omega_g)\right\} + \frac{1}{2}\gamma_{gt}^{(1)}\left\{\log\det\left(\boldsymbol{\Sigma}_{\beta,gt}\right) + |g|\right\} \\
&\quad -\gamma_{gt}^{(1)}\log\gamma_{gt}^{(1)} - \left(1-\gamma_{gt}^{(1)}\right)\log\left(1-\gamma_{gt}^{(1)}\right),
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}_\omega(\omega_g) &= \mathrm{E}_q\left\{\log p(\omega_g)\right\} - \mathrm{E}_q\left\{\log q(\omega_g)\right\} \\
&= \left(a_g - a_g^*\right)\mathrm{E}_q\left(\log\omega_g\right) + \left(b_g - b_g^*\right)\mathrm{E}_q\left\{\log(1-\omega_g)\right\} - \log B(a_g, b_g) + \log B\left(a_g^*, b_g^*\right),
\end{aligned}$$

$$\mathcal{L}_\sigma\left(\sigma^{-2}\right) = \left(\lambda-\lambda^*\right)\mathrm{E}_q\left(\log\sigma^{-2}\right) - \left(\nu-\nu^*\right)\left(\sigma^{-2}\right)^{(1)} + \lambda\log\nu - \lambda^*\log\nu^* - \log\Gamma(\lambda) + \log\Gamma\left(\lambda^*\right),$$

$$\mathcal{L}_\tau(\tau_t) = \left(\eta_t-\eta_t^*\right)\mathrm{E}_q\left(\log\tau_t\right) - \left(\kappa_t-\kappa_t^*\right)\tau_t^{(1)} + \eta_t\log\kappa_t - \eta_t^*\log\kappa_t^* - \log\Gamma(\eta_t) + \log\Gamma(\eta_t^*).$$

B.2 Derivation of the variational algorithm for the similarity sparsity model

B.2.1 Variational distributions

We provide the derivation of the variational algorithm for the similarity sparsity model presented in Section 4.2.2. Consider q centered responses, $\boldsymbol{y} = \left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q\right)$, and p centered predictors, $\boldsymbol{X} = \left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p\right)$, for n samples. We rewrite the model using the auxiliary variable z_{st}

$$\begin{aligned}
\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \boldsymbol{\tau}_t &\sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}_t, \tau_t^{-1}\boldsymbol{I}_n\right), & \tau_t &\sim \text{Gamma}(\eta_t, \kappa_t), & t &= 1, \ldots, q, \\
\beta_{st} \mid \gamma_{st}, \tau_t, \sigma^2 &\sim \gamma_{st}\mathcal{N}\left(0, \sigma^2\tau_t^{-1}\right) + (1-\gamma_{st})\,\delta_0, & \sigma^{-2} &\sim \text{Gamma}(\lambda, \nu), & s &= 1, \ldots, p, \\
\gamma_{st} &= \mathbb{1}\{z_{st} > 0\}, & z_{st} \mid \theta_s &\sim \mathcal{N}(\theta_s, 1), & \boldsymbol{\theta} &\sim \mathcal{N}_p(\boldsymbol{m}_0, \boldsymbol{\Sigma}_0).
\end{aligned}$$

Let $\boldsymbol{\nu} = \left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\tau}, \sigma^{-2}\right)$, we have

$$\begin{aligned} p(y, \boldsymbol{\nu}) &= \left\{\prod_{t=1}^{q} p\left(y_t \mid \beta_t, \tau_t\right)\right\}\left\{\prod_{t=1}^{q}\prod_{s=1}^{p} p\left(\beta_{st} \mid \gamma_{st}, \sigma^{-2}, \tau_t\right)\right\} \\ &\quad \times \left\{\prod_{t=1}^{q}\prod_{s=1}^{p} p\left(\gamma_{st} \mid z_{st}\right)\right\}\left\{\prod_{s=1}^{p} p(z_{st} \mid \theta_s)\right\} p(\boldsymbol{\theta})\left\{\prod_{t=1}^{q} p\left(\tau_t\right)\right\} p(\sigma^{-2}). \end{aligned}$$

Consider the following mean-field form for the variational approximation,

$$q\left(\boldsymbol{\nu}\right) = \left\{\prod_{t=1}^{q}\prod_{s=1}^{p} q(\beta_{st}, \gamma_{st}, z_{st})\right\} q(\boldsymbol{\theta})\left\{\prod_{t=1}^{q} p\left(\tau_t\right)\right\} q\left(\sigma^{-2}\right).$$

The variational updates are as follows. We have

$$\begin{aligned} q(\beta_{st}, \gamma_{st}, z_{st}) &= \text{cst} \times \exp\left(-\frac{\tau_t^{(1)}}{2} \mathrm{E}_{-(\beta_{st},\gamma_{st})} \|\boldsymbol{y}_t - \boldsymbol{X}\boldsymbol{\beta}_t\|^2\right) \exp\left\{\frac{1}{2}\mathrm{E}\left(\log\sigma^{-2}\right) + \frac{1}{2}\mathrm{E}\left(\log\tau_t\right)\right\}^{\gamma_{st}} \\ &\quad \times \left[(2\pi)^{-1/2} \exp\left\{-\frac{\left(\sigma^{-2}\right)^{(1)} \tau_t^{(1)}}{2}\beta_{st}^2\right\}\right]^{\gamma_{st}} \left\{\delta_0(\beta_{st})\right\}^{1-\gamma_{st}} \mathbb{1}\{z_{st} > 0\}^{\gamma_{st}} \\ &\quad \times \mathbb{1}\{z_{st} \le 0\}^{1-\gamma_{st}} \exp\left\{-\frac{1}{2}\mathrm{E}_{-z_{st}}\left(z_{st} - \theta_s\right)^2\right\}. \end{aligned}$$

We therefore observe that

$$q(\beta_{st}, \gamma_{st}, z_{st}) = q(\beta_{st} \mid z_{st}) q(z_{st} \mid \gamma_{st}) q(\gamma_{st}),$$

with

$$\begin{aligned} \beta_{st} \mid z_{st} > 0, \boldsymbol{y} &\sim \mathcal{N}\left(\mu_{\beta,st}, \sigma^2_{\beta,st}\right), \qquad \beta_{st} \mid z_{st} \le 0, \boldsymbol{y} \sim \delta_0, \\ z_{st} \mid \gamma_{st} = \delta, \boldsymbol{y} &\sim \mathcal{TN}\left(\theta_s^{(1)}, 1; \left\{0 < (-1)^{1-\delta} z_{st}\right\}\right), \\ \gamma_{st} \mid \boldsymbol{y} &\sim \text{Bernoulli}\left(\gamma_{st}^{(1)}\right), \end{aligned}$$

where $\mathcal{TN}\left(\mu, \sigma^2; \{a < x < b\}\right)$ denotes the truncated normal distribution, and where

$$\mu_{\beta,st} = \sigma^2_{\beta,st}\tau_t^{(1)} \boldsymbol{X}_s^T\left(\boldsymbol{y}_t - \sum_{j=1, j\neq s}^{p} \gamma_{jt}^{(1)} \mu_{\beta,jt} \boldsymbol{X}_j\right), \qquad \sigma^{-2}_{\beta,st} = \tau_t^{(1)}\left\{\|\boldsymbol{X}_s\|^2 + \left(\sigma^{-2}\right)^{(1)}\right\},$$

$$\gamma_{st}^{(1)} = \left[1 + \sigma^{-1}_{\beta,st} \frac{1 - \Phi\left(\theta_s^{(1)}\right)}{\Phi\left(\theta_s^{(1)}\right)} \exp\left\{-\frac{1}{2}\mathrm{E}\left(\log\sigma^{-2}\right) - \frac{1}{2}\mathrm{E}\left(\log\tau_t\right) - \frac{1}{2}\mu^2_{\beta,st}\sigma^{-2}_{\beta,st}\right\}\right]^{-1}.$$

The first moment of z_{st} is obtained by observing that

$$\mathrm{E}_q\left(z_{st} \mid \gamma_{st}\right) = \theta_s^{(1)} + \mathrm{M}\left(\theta_s^{(1)}, \gamma_{st}\right),$$

where

$$\mathrm{M}(u, \delta) = (-1)^{1-\delta} \frac{\varphi(u)}{\Phi(u)^{\delta}\left[1 - \Phi(u)\right]^{1-\delta}}, \qquad u \in \mathbb{R},\ \delta = 0, 1,$$

is the inverse Mills ratio (Mills, 1926), so

$$z_{st}^{(1)} = \gamma_{st}^{(1)} \left\{ \mathrm{M}(\theta_s^{(1)}, 1) - \mathrm{M}(\theta_s^{(1)}, 0) \right\} + \theta_s^{(1)} + \mathrm{M}(\theta_s^{(1)}, 0).$$

For the second moment, we observe that

$$\mathrm{E}_q \left(z_{st}^2 \mid \gamma_{st} \right) = 1 + \left\{ \theta_s^{(1)} \right\}^2 - \theta_s^{(1)} \mathrm{M}\left(\theta_s^{(1)}, \gamma_{st} \right) + 2\theta_s^{(1)} \mathrm{M}\left(\theta_s^{(1)}, \gamma_{st} \right) = 1 + \theta_s^{(1)} \mathrm{E}_q \left(z_{st} \mid \gamma_{st} \right),$$

so

$$\begin{aligned} z_{st}^{(2)} &= \gamma_{st}^{(1)} + \gamma_{st}^{(1)} \theta_s^{(1)} \mathrm{E}_q \left(z_{st} \mid \gamma_{st} = 1 \right) + (1 - \gamma_{st}^{(1)}) + (1 - \gamma_{st}^{(1)}) \theta_s^{(1)} \mathrm{E}_q \left(z_{st} \mid \gamma_{st} = 0 \right) \\ &= \theta_s^{(1)} z_{st}^{(1)} + 1. \end{aligned}$$

Finally, given γ_{st}, the entropy is

$$\mathrm{H}(z_{st} \mid \gamma_{st}) = \log \left[\sqrt{2\pi e}\, \Phi\left(\theta_s^{(1)} \right)^{\gamma_{st}} \left\{ 1 - \Phi\left(\theta_s^{(1)} \right) \right\}^{1-\gamma_{st}} \right] - \frac{1}{2} \theta_s^{(1)} \mathrm{M}\left(\theta_s^{(1)}, \gamma_{st} \right).$$

We also have

$$\log q\left(\boldsymbol{\theta} \right) = -\frac{1}{2} \left\{ q \boldsymbol{\theta}^T \boldsymbol{\theta} - 2 \boldsymbol{\theta}^T \boldsymbol{Z}^{(1)} \mathbb{1}_q + \boldsymbol{\theta}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\theta} - 2 \boldsymbol{\theta}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{m}_0 \right\} + \mathrm{cst}.$$

Therefore,

$$\boldsymbol{\theta} \sim \mathcal{N}_p \left(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta \right),$$

with

$$\boldsymbol{\mu}_\theta = \boldsymbol{\Sigma}_\theta \left(\boldsymbol{Z}^{(1)} \mathbb{1}_q + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{m}_0 \right), \qquad \boldsymbol{\Sigma}_\theta^{-1} = q \boldsymbol{I}_p + \boldsymbol{\Sigma}_0^{-1}.$$

The updates for σ^{-2} and τ are the same as for the reference model; see Appendix A.2.

B.2.2 Variational lower bound

We provide the variational lower bound, $\mathcal{L}(q)$, of the marginal log-likelihood, $\log p(\boldsymbol{y})$:

$$\begin{aligned} \mathcal{L}(q) &= \int q(\boldsymbol{\nu}) \log \left\{ \frac{p(\boldsymbol{y}, \boldsymbol{\nu})}{q(\boldsymbol{\nu})} \right\} \mathrm{d}\boldsymbol{\nu} \\ &= \sum_{t=1}^{q} \mathcal{L}_y \left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t, \tau_t \right) + \sum_{t=1}^{q} \sum_{s=1}^{p} \mathcal{L}_\beta \left(\beta_{st}, \gamma_{st}, z_{st} \mid \sigma^{-2}, \tau_t, \theta_s \right) + \sum_{t=1}^{q} \mathcal{L}_\tau \left(\tau_t \right) + \mathcal{L}_\sigma \left(\sigma^{-2} \right) + \mathcal{L}_\theta \left(\boldsymbol{\theta} \right), \end{aligned}$$

where $\mathcal{L}_y \left(y_t \mid \beta_t, \gamma_t, \tau_t \right)$, $\mathcal{L}_\tau \left(\tau_t \right)$ and $\mathcal{L}_\sigma \left(\sigma^{-2} \right)$ are the same as for the reference model, see Appendix A.2,

$$\begin{aligned} \mathcal{L}_\beta \left(\beta_{st}, \gamma_{st}, z_{st} \mid \sigma^{-2}, \tau_t, \theta_s \right) &= \mathrm{E}_q \log p \left(\beta_{st} \mid \gamma_{st}, \sigma^{-2}, \tau_t \right) + \mathrm{E}_q \log p \left(\gamma_{st} \mid z_{st} \right) + \mathrm{E}_q \log p \left(z_{st} \mid \theta_s \right) \\ &\quad - \mathrm{E}_q \log q \left(\beta_{st}, \gamma_{st}, z_{st} \right) \\ &= \mathcal{L}_{\beta,1} + \mathcal{L}_{\beta,2} + \mathcal{L}_{\beta,3}, \end{aligned}$$

$$\begin{aligned}\mathcal{L}_{\beta,1} &= \mathrm{E}_q \log p\left(\beta_{st} \mid \gamma_{st}, \sigma^{-2}, \tau_t\right) - \mathrm{E}_q \log q\left(\beta_{st} \mid z_{st}\right) \\ &= \frac{1}{2}\gamma_{st}^{(1)}\left\{\mathrm{E}_q\left(\log\sigma^{-2}\right) + \mathrm{E}_q\left(\log\tau_t\right) - \left(\mu_{\beta,st}^2 + \sigma_{\beta,st}^2\right)\left(\sigma^{-2}\right)^{(1)}\tau_t^{(1)}\right\} + \frac{1}{2}\gamma_{st}^{(1)}\left(\log\sigma_{\beta,st}^2 + 1\right),\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{\beta,2} &= \mathrm{E}_q \log p\left(\gamma_{st} \mid z_{st}\right) + \mathrm{E}_q \log p\left(z_{st}\right) - \mathrm{E}_q \log q\left(z_{st} \mid \gamma_{st}\right) \\ &= \left(1-\gamma_{st}^{(1)}\right)\log\left\{1-\Phi\left(\theta_s^{(1)}\right)\right\} + \gamma_{st}^{(1)}\log\Phi\left(\theta_s^{(1)}\right) - \mathrm{H}\left(z_{st} \mid \gamma_{st}=1\right)\gamma_{st}^{(1)} - \mathrm{H}\left(z_{st} \mid \gamma_{st}=0\right)\left(1-\gamma_{st}^{(1)}\right) \\ &\quad - \frac{1}{2}\sigma_{\theta,s}^2 + \mathrm{H}\left(z_{st} \mid \gamma_{st}=1\right)\gamma_{st}^{(1)} + \mathrm{H}\left(z_{st} \mid \gamma_{st}=0\right)\left(1-\gamma_{st}^{(1)}\right) \\ &= \left(1-\gamma_{st}^{(1)}\right)\log\left\{1-\Phi\left(\theta_s^{(1)}\right)\right\} + \gamma_{st}^{(1)}\log\Phi\left(\theta_s^{(1)}\right) - \frac{1}{2}\sigma_{\theta,s}^2,\end{aligned}$$

$$\mathcal{L}_{\beta,3} = -\gamma_{st}^{(1)}\log\gamma_{st}^{(1)} - \left(1-\gamma_{st}^{(1)}\right)\log\left(1-\gamma_{st}^{(1)}\right),$$

$$\begin{aligned}\mathcal{L}_\theta(\boldsymbol{\theta}) &= \mathrm{E}_q\left\{\log p(\boldsymbol{\theta})\right\} - \mathrm{E}_q\left\{\log q(\boldsymbol{\theta})\right\} \\ &= \frac{1}{2}\left\{-\log\det(\boldsymbol{\Sigma}_0) + \log\det(\boldsymbol{\Sigma}_\theta) - \left(\boldsymbol{\mu}_\theta - \boldsymbol{m}_0\right)^T\boldsymbol{\Sigma}_0^{-1}\left(\boldsymbol{\mu}_\theta - \boldsymbol{m}_0\right) - \mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_{\theta b}\right) + p\right\}.\end{aligned}$$

B.3 Hyperparameter specification for simulations of Section 4.2.2

We describe the top-level hyperparameter choices for the three models compared in the numerical experiments of Section 4.2.2. For model (3.1)–(3.2)–(3.3), we choose the hyperparameters a_s and b_s based on the multiplicity control procedure of Section 3.1.3 and using the simulated number of active SNPs as prior number of SNPs expected to be in the model, that is, $p^* = p_\gamma$. We then set the corresponding hyperparameters for remaining two models, so as to roughly match the reference hyperparameter setting: for model (4.1), we choose a_g and b_g as in Section 3.1.3, but replace the prior proportion of active predictors p^*/p, with a prior proportion of "active" groups, g^*/G, i.e., groups that have at least one predictor involved in associations. We choose the number of such groups as $g^* = p^*$, which corresponds to assuming that each active group contains a single active predictor. As model (4.3) doesn't place a Beta distribution on the probability parameter, we match the first moment of the marginal distribution of γ_{st},

$$\mathrm{pr}(\gamma_{st}=1) = \int \mathrm{pr}(\gamma_{st}=1 \mid \theta_s)\mathrm{pr}(\theta_s)\mathrm{d}\theta_s = \int \Phi(\theta_s)\,\varphi\left(\frac{\theta_s - m_{0s}}{s_{0s}}\right)\mathrm{d}\theta = \Phi\left(\frac{m_{0s}}{\sqrt{1+s_{0s}^2}}\right),$$

where s_{0s}^2 is the sth diagonal entry of $\boldsymbol{\Sigma}_0$, with that for the reference model, $\mathrm{pr}(\gamma_{st}=1) = \mathrm{E}(\omega_s) = a_s/(a_s+b_s)$.

We employ a block diagonal matrix $\boldsymbol{\Sigma}_0$ based on the autocorrelation block partition used to generate the data, namely, the b^{th} block of $\boldsymbol{\Sigma}_0$ is $\boldsymbol{\Sigma}_{0b} = \alpha \boldsymbol{X}_b^T\boldsymbol{X}_b$ where $\boldsymbol{X}_b$ corresponds to the $(b-1)\times 50+1$ to $b\times 50$ columns of $\boldsymbol{X}$. As discussed in Section 3.1.3, specifying the second moment of ω_s, ω_g or θ_s is difficult because we have no prior state of knowledge for it; this is addressed in Chapter 5. Here, we scaled the predictors, so $\boldsymbol{X}_b^T\boldsymbol{X}_b$ is the empirical correlation matrix of the SNPs in block b, and set $\alpha = 0.1$, to ensure that the prior distribution of the marginal probability parameter, $\Phi(\theta_s)$, concentrates on large negative values enforcing sufficient sparsity.

B.4 Derivation of the annealed variational algorithm

This section provides the annealed variational updates for the model and mean-field approximation described in Chapter 3; these are readily obtained by modifying the variational updates described in Appendix A.2. Using

$$\log q_c(v_j) = c\,\mathrm{E}_{-j}\{\log p(\boldsymbol{y},\boldsymbol{v})\} + \mathrm{cst}, \qquad j = 1,\ldots,J,$$

where $0 < c \leq 1$ is the inverse temperature parameter, $\mathrm{E}_{-j}(\cdot)$ is the expectation with respect to the distributions $q(v_k)$, for all the variables v_k $(k \neq j)$, and cst is constant with respect to v_j, we can express the annealed variational parameters $\boldsymbol{\alpha}(c)$ with respect to the classical variational parameters $\boldsymbol{\alpha}$. The annealed parameters for β_{st} and γ_{st} are

$$\mu_{\beta,st}(c) = c\mu_{\beta,st}, \qquad \sigma^{-2}_{\beta,st}(c) = c\sigma^{-2}_{\beta,st}, \qquad \frac{1}{\gamma^{(1)}_{st}(c)} = 1 + \left(\frac{1}{\gamma^{(1)}_{st}} - 1\right)^c.$$

For τ_t, we have

$$\eta^*_t(c) = c\left(\eta^*_t - 1\right) + 1, \qquad \kappa^*_t(c) = c\kappa^*_t.$$

Similarly, for σ^{-2}, we have

$$\lambda^*(c) = c\left(\lambda^* - 1\right) + 1, \qquad \nu^*(c) = c\nu^*.$$

Finally, for ω_s, we have

$$a^*_s(c) = c\,(a_s - 1) + 1, \qquad b^*_s(c) = c\,(b_s - 1) + 1.$$

The variational objective function, $\mathcal{L}(q)$, is evaluated only when the inverse temperature $c = 1$ has been reached, so is unchanged from Appendix A.2.

C Appendix for Chapter 5

C.1 Hyperparameter specification for top-level priors

We describe the hyperparameter settings for the prior distribution of ζ_t. We borrow ideas from Bottolo et al. (2011), in that we let the response-specific parameter $\zeta_t \overset{\text{iid}}{\sim} \mathcal{N}(n_0, t_0^2)$ control the sparsity level, i.e., the number of predictors associated with each response, and use parameter $\theta_s \overset{\text{iid}}{\sim} \mathcal{N}(m_0, s_0^2)$ as a predictor-specific modulator of this level.

We will rely on the following results: for $X \sim \mathcal{N}(\mu, \sigma^2)$,

$$\mathrm{E}\{\Phi(X)\} = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right), \tag{C.1}$$

$$\mathrm{E}\left\{\Phi(X)^2\right\} = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right) - 2\mathrm{T}\left(\frac{\mu}{\sqrt{1+\sigma^2}}, \frac{1}{\sqrt{1+2\sigma^2}}\right), \tag{C.2}$$

where

$$\mathrm{T}(h,a) = \varphi(h)\int_0^a \frac{\varphi(hx)}{1+x^2}\mathrm{d}x, \qquad a,h \in \mathbb{R},$$

is Owen's T function (Owen, 1956), with $\varphi(\cdot)$ the standard normal density function.

Equality (C.1) can be obtained as follows. Let $Z_1 \sim \mathcal{N}\left(-\sigma^{-1}\mu, \sigma^{-2}\right)$ and $Z_2 \sim \mathcal{N}(0,1)$ be independent and observe that

$$\mathrm{pr}(Z_1 \le Z_2 \mid Z_2 = z) = \mathrm{pr}(Z_1 \le z) = \Phi(\sigma z + \mu), \quad z \in \mathbb{R},$$

so that

$$\mathrm{pr}(Z_1 \le Z_2) = \int \Phi(\sigma z + \mu)\varphi(z)\mathrm{d}z,$$

which corresponds to the left hand-side of (C.1). But since $Z_1 - Z_2 \sim \mathcal{N}\left(-\sigma^{-1}\mu, \sigma^{-2}+1\right)$, we also have

$$\mathrm{pr}(Z_1 \le Z_2) = \mathrm{pr}(Z_1 - Z_2 \le 0) = \Phi\left(\frac{\mu}{\sqrt{1+\sigma^2}}\right),$$

which gives the result. Equality (C.2) can be obtained similarly.

Coming back to the hyperparameter setting, we need to proceed in two steps: we first restrict ourselves to a model with no predictor-specific modulation ($\theta_s = 0$) so that, given ζ_t, the prior probability of

association between predictor $\boldsymbol{X}_s$ and response $\boldsymbol{y}_t$ is

$$\mathrm{E}\left(\gamma_{st} \mid \theta_s = 0, \zeta_t\right) = \Phi(\zeta_t).$$

We then set n_0 and t_0^2 by specifying a prior expectation and a prior variance for the number p_γ of active predictors per response:

$$\begin{aligned}
\mathrm{E}\left(p_\gamma \mid \theta_s = 0\right) &= \mathrm{E}\left\{\mathrm{E}\left(p_\gamma \mid \theta_s = 0, \zeta_t\right)\right\} = p\,\mathrm{E}\left\{\Phi\left(\zeta_t\right)\right\}, \\
\mathrm{Var}\left(p_\gamma \mid \theta_s = 0\right) &= \mathrm{Var}\left\{\mathrm{E}\left(p_\gamma \mid \theta_s = 0, \zeta_t\right)\right\} + \mathrm{E}\left\{\mathrm{Var}\left(p_\gamma \mid \theta_s = 0, \zeta_t\right)\right\} \\
&= p(p-1)\mathrm{E}\left\{\Phi\left(\zeta_t\right)^2\right\} + p\mathrm{E}\left\{\Phi\left(\zeta_t\right)\right\}\left[1 - p\mathrm{E}\left\{\Phi\left(\zeta_t\right)\right\}\right],
\end{aligned}$$

in which we use (C.1) and (C.2) with $\mu = n_0$ and $\sigma^2 = t_0^2$. We then solve this system numerically to obtain n_0 and t_0^2.

We now reintroduce the predictor-specific effect θ_s and argue that

$$\mathrm{E}\left(p_\gamma\right) \approx \mathrm{E}\left(p_\gamma \mid \theta_s = 0\right),$$

i.e., the user-specified prior mean is roughly left unchanged after the reintroduction of the modulation parameter θ_s. To this end, we observe that

$$\mathrm{E}\left(p_\gamma\right) = p\mathrm{E}\left\{\Phi\left(\theta_s + \zeta_t\right)\right\},$$

where the expectation is with respect to $\theta_s + \zeta_t \sim \mathcal{N}\left(m_0 + n_0, s_0^2 + t_0^2\right)$. So using $\mu = m_0 + n_0$ and $\sigma^2 = s_0^2 + t_0^2$ in (C.1), we have

$$\begin{aligned}
\mathrm{E}\left(p_\gamma\right) = \mathrm{E}\left(p_\gamma \mid \theta_s = 0\right) &\iff \mathrm{E}\left\{\Phi\left(\theta_s + \zeta_t\right)\right\} = \mathrm{E}\left\{\Phi\left(\zeta_t\right)\right\} \\
&\iff \frac{m_0 + n_0}{\sqrt{1 + s_0^2 + t_0^2}} = \frac{n_0}{\sqrt{1 + t_0^2}},
\end{aligned} \tag{C.3}$$

giving

$$m_0 = n_0\left(\sqrt{1 + \frac{s_0^2}{1 + t_0^2}} - 1\right) \approx \frac{n_0 s_0^2}{2\left(1 + t_0^2\right)} - \frac{n_0 s_0^4}{8\left(1 + t_0^2\right)^2} + \ldots.$$

As in practice, $s_0^2 \ll 1$, $m_0 \approx 0$; we therefore take $m_0 = 0$ (which also simplifies the discussion in the horseshoe prior case).

Finally, assuming that (C.3) holds, we observe that the marginal variance is slightly inflated, as in Bottolo et al. (2011):

$$\begin{aligned}\mathrm{Var}\left(p_\gamma\right) - \mathrm{Var}\left(p_\gamma \mid \theta_s = 0\right) &= 2p(p-1)\left[\mathrm{T}\left(h, \frac{1}{\sqrt{1+2t_0^2}}\right)\right. \\ &\quad \left. -\mathrm{T}\left(h, \frac{1}{\sqrt{1+2\left(s_0^2+t_0^2\right)}}\right)\right] \\ &= 2p(p-1)\varphi(h)\int_{\left\{1+2\left(s_0^2+t_0^2\right)\right\}^{-1/2}}^{\left\{1+2t_0^2\right\}^{-1/2}} \frac{\varphi(hx)}{1+x^2}\mathrm{d}x \\ &> 0,\end{aligned}$$

with $h = (1+t_0^2)^{-1/2}\, n_0$.

C.2 Derivation of the annealed variational algorithm

C.2.1 Variational distributions

Recall the full model specification. For q centered responses, $\boldsymbol{y} = \left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q\right)$, and p centered predictors, $\boldsymbol{X} = \left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p\right)$, for n samples,

$$\begin{aligned}
\boldsymbol{y}_t &\mid \boldsymbol{\beta}_t, \tau_t \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}_t, \tau_t^{-1}\boldsymbol{I}_n\right), && \tau_t \overset{\text{ind}}{\sim} \text{Gamma}(\eta_t, \kappa_t), && t = 1, \ldots, q, \\
\beta_{st} &\mid \gamma_{st}, \tau_t, \sigma \sim \gamma_{st}\mathcal{N}\left(0, \sigma^2\tau_t^{-1}\right) + (1-\gamma_{st})\,\delta_0, && \sigma^{-2} \sim \text{Gamma}(\nu, \rho), && s = 1, \ldots, p, \\
\gamma_{st} &\mid \theta_s, \zeta_t \sim \text{Bernoulli}\left\{\Phi(\theta_s + \zeta_t)\right\}, && \zeta_t \overset{\text{iid}}{\sim} \mathcal{N}(n_0, t_0^2), && \\
\theta_s &\mid \lambda_s, \sigma_0 \sim \mathcal{N}\left(0, \lambda_s^2\sigma_0^2\right), && \lambda_s \overset{\text{iid}}{\sim} \mathrm{C}^+(0,1), && \sigma_0 \sim \mathrm{C}^+(0, q^{-1/2}),
\end{aligned} \tag{C.4}$$

where δ_0 is the Dirac distribution, $\Phi(\cdot)$ is the standard normal cumulative distribution function, and $\mathrm{C}^+(\cdot,\cdot)$ is a half-Cauchy distribution.

In order to obtain closed-form updates for our variational algorithm, we consider two data augmentations. We first rewrite the probit-link level using the classical representation

$$\gamma_{st} = \mathbb{1}\{z_{st} > 0\}, \qquad z_{st} \mid \theta_s, \zeta_t \sim \mathcal{N}\left(\theta_s + \zeta_t, 1\right),$$

where z_{st} is an auxiliary variable. We then consider the following formulation for the scale parameters of the horseshoe prior,

$$\sigma_0^{-2} \mid \xi \sim \text{Gamma}\left(\frac{1}{2}, \xi^{-1}\right), \qquad \xi^{-1} \sim \text{Gamma}\left(\frac{1}{2}, q\right), \qquad p\left(\lambda_s^{-2}\right) = \pi^{-1}\left(1+\lambda_s^2\right)^{-1}\lambda_s^3.$$

This parametrization introduces the auxiliary variable, ξ; it was first proposed by Neville et al. (2014). For completeness, we reformulate two lemmas which establish the equivalence with the original formulation in (C.4).

Lemma C.2.1. *If a is a random variable such that*

$$a \mid \xi \sim \text{Gamma}\left(\frac{1}{2}, \xi^{-1}\right), \qquad \xi^{-1} \sim \text{Gamma}\left(\frac{1}{2}, A^{-2}\right), \quad A > 0,$$

then $a^{-1/2} \sim \text{C}^+(0, A)$.

Lemma C.2.2. *If a is a random variable such that*

$$p(a) = \pi^{-1}(1+a)^{-1} a^{-1/2}, \quad a > 0,$$

then $a^{-1/2} \sim \text{C}^+(0, 1)$.

Proof. Both results are straightforward. □

We now provide the updates for all the heated variational parameters. Let $T > 1$ be the current temperature from the annealing schedule, let $\boldsymbol{\nu}$ be the parameter vector of interest, and let $q_T(\boldsymbol{\nu})$ be the mean-field variational approximation to the true posterior distribution, $p(\boldsymbol{\nu} \mid \boldsymbol{y})$. We maximize the lower bound on the marginal log-likelihood,

$$\mathcal{L}_T(q) = \int q_T(\boldsymbol{\nu}) \log p(\boldsymbol{\nu}, \boldsymbol{y}) \mathrm{d}\boldsymbol{\nu} - T \int q_T(\boldsymbol{\nu}) \log q_T(\boldsymbol{\nu}) \mathrm{d}\boldsymbol{\nu}.$$

Recall from Chapter 4 that the heated variational distributions $q_T(\nu_j)$ are given by

$$\log q_T(\nu_j) = T^{-1} \text{E}_{-j}\{\log p(\boldsymbol{y}, \boldsymbol{\nu})\} + \text{cst}, \qquad j = 1, \ldots, J, \tag{C.5}$$

where $\text{E}_{-j}(\cdot)$ is the expectation with respect to the distributions $q_T(\nu_k)$, for all the variables ν_k ($k \neq j$), and cst is constant with respect to ν_j. For ease of reading, we hereafter drop the subscript T in $q_T(\cdot)$, and write $c = T^{-1}$. We find that,

$$q(\beta_{st}, \gamma_{st}, z_{st}) = q(\beta_{st} \mid z_{st}) q(z_{st} \mid \gamma_{st}) q(\gamma_{st}), \qquad s = 1, \ldots, p, \; t = 1, \ldots, q$$

with

$$\begin{aligned}
\beta_{st} \mid z_{st} > 0, \boldsymbol{y} &\sim \mathcal{N}\left(\mu_{\beta,st}, \sigma^2_{\beta,st}\right), \quad \beta_{st} \mid z_{st} \leq 0, \boldsymbol{y} \sim \delta_0, \\
z_{st} \mid \gamma_{st} = \delta, \boldsymbol{y} &\sim \mathcal{TN}\left(\theta_s^{(1)} + \zeta_t^{(1)}, 1; \left\{0 < (-1)^{1-\delta} z_{st}\right\}\right), \\
\gamma_{st} \mid \boldsymbol{y} &\sim \text{Bernoulli}\left(\gamma_{st}^{(1)}\right),
\end{aligned}$$

where $X \sim \mathcal{TN}\left(\mu, \sigma^2; \{a < x < b\}\right)$ denotes a truncated normal variable,

$$\sigma^{-2}_{\beta,st} = c\,\tau_t^{(1)} \left\{\|\boldsymbol{X}_s\|^2 + \left(\sigma^{-2}\right)^{(1)}\right\}, \qquad \mu_{\beta,st} = c\,\sigma^2_{\beta,st} \tau_t^{(1)} \boldsymbol{X}_s^T \left(\boldsymbol{y}_t - \sum_{j=1, j\neq s}^{p} \gamma_{jt}^{(1)} \mu_{\beta,jt} \boldsymbol{X}_j\right),$$

and

$$\begin{aligned}
\frac{1}{\gamma_{st}^{(1)}} \;=\; 1 + \exp\Bigg[-c \Bigg\{ &\frac{1}{2}\left(\log \sigma^{-2}\right)^{(1)} + \frac{1}{2}\left(\log \tau_t\right)^{(1)} + \frac{1}{2}\mu^2_{\beta,st}\sigma^{-2}_{\beta,st} + \log \sigma_{\beta,st} \\
&- \log\left\{1 - \Phi\left(\theta_s^{(1)} + \zeta_t^{(1)}\right)\right\} + \log \Phi\left(\theta_s^{(1)} + \zeta_t^{(1)}\right) \Bigg\} \Bigg].
\end{aligned}$$

Writing $\alpha_{st} = \theta_s + \zeta_t$, the first moment of z_{st} given γ_{st} is

$$\mathrm{E}_q\left(z_{st} \mid \gamma_{st}\right) = \alpha_{st}^{(1)} + \mathrm{M}\left(\alpha_{st}^{(1)}, \gamma_{st}\right),$$

where

$$\mathrm{M}\left(u, \gamma\right) = (-1)^{1-\gamma} \frac{\varphi\left(u\right)}{\Phi\left(u\right)^{\gamma}\left[1-\Phi\left(u\right)\right]^{1-\gamma}}, \qquad u \in \mathbb{R},\ \gamma = 0, 1,$$

is the inverse Mills ratio. We therefore have

$$z_{st}^{(1)} \quad = \quad \gamma_{st}^{(1)}\left\{\mathrm{M}(\alpha_{st}^{(1)}, 1) - \mathrm{M}(\alpha_{st}^{(1)}, 0)\right\} + \alpha_{st}^{(1)} + \mathrm{M}(\alpha_{st}^{(1)}, 0).$$

The second moment of z_{st} given γ_{st} is

$$\mathrm{E}_q\left(z_{st}^2 \mid \gamma_{st}\right) = 1 + \left(\alpha_{st}^{(1)}\right)^2 - \alpha_{st}^{(1)}\mathrm{M}\left(\alpha_{st}^{(1)}, \gamma_{st}\right) + 2\alpha_{st}^{(1)}\mathrm{M}\left(\alpha_{st}^{(1)}, \gamma_{st}\right) = 1 + \alpha_{st}^{(1)}\mathrm{E}_q\left(z_{st} \mid \gamma_{st}\right),$$

which implies that

$$\begin{aligned} z_{st}^{(2)} &= \gamma_{st}^{(1)} + \gamma_{st}^{(1)}\alpha_{st}^{(1)}\mathrm{E}_q\left(z_{st} \mid \gamma_{st} = 1\right) + (1 - \gamma_{st}^{(1)}) + (1 - \gamma_{st}^{(1)})\alpha_{st}^{(1)}\mathrm{E}_q\left(z_{st} \mid \gamma_{st} = 0\right) \\ &= \alpha_{st}^{(1)} z_{st}^{(1)} + 1. \end{aligned}$$

Then, we find

$$\sigma^{-2} \mid \boldsymbol{y} \sim \mathrm{Gamma}\left(\nu_\sigma, \rho_\sigma\right), \qquad \left(\sigma^{-2}\right)^{(1)} = \nu_\sigma / \rho_\sigma,$$

with

$$\nu_\sigma = c\left(\nu + \frac{1}{2}\sum_{t=1}^{q}\sum_{s=1}^{p}\gamma_{st}^{(1)}\right) - c + 1, \qquad \rho_\sigma = c\left\{\rho + \frac{1}{2}\sum_{t=1}^{q}\tau_t^{(1)}\sum_{s=1}^{p}\gamma_{st}^{(1)}\left(\mu_{\beta,st}^2 + \sigma_{\beta,st}^2\right)\right\}.$$

The residual precision parameters have

$$\tau_t \mid \boldsymbol{y} \sim \mathrm{Gamma}\left(\eta_{\tau,t}, \kappa_{\tau,t}\right), \qquad \tau_t^{(1)} = \eta_{\tau,t} / \kappa_{\tau,t},$$

where

$$\begin{aligned} \eta_{\tau,t} \quad &= \quad c\left(\eta_t + \frac{n}{2} + \frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)}\right) - c + 1, \\ \kappa_{\tau,t} \quad &= \quad c\left[\kappa_t + \frac{1}{2}\|\boldsymbol{y}_t\|^2 - \boldsymbol{y}_t^T\sum_{s=1}^{p}\mu_{\beta,st}\gamma_{st}^{(1)}\boldsymbol{X}_s + \sum_{s=1}^{p-1}\mu_{\beta,st}\gamma_{st}^{(1)}\boldsymbol{X}_s^T\sum_{j=s+1}^{p}\mu_{\beta,jt}\gamma_{jt}^{(1)}\boldsymbol{X}_j\right. \\ &\qquad \left. + \frac{1}{2}\sum_{s=1}^{p}\gamma_{st}^{(1)}\left(\sigma_{\beta,st}^2 + \mu_{\beta,st}^2\right)\left\{\|\boldsymbol{X}_s\|^2 + \left(\sigma^{-2}\right)^{(1)}\right\}\right]. \end{aligned}$$

We then have

$$\zeta_t \mid \boldsymbol{y} \sim \mathcal{N}\left(\mu_{\zeta,t}, \sigma_{\zeta,t}^2\right),$$

with

$$\sigma_{\zeta,t}^{-2} = c\left(p + t_0^{-2}\right), \qquad \mu_{\zeta,t} = c\,\sigma_{\zeta,t}^2\left(\sum_{s=1}^{p} z_{st}^{(1)} - \sum_{s=1}^{p}\theta_s^{(1)} + t_0^{-2} n_0\right).$$

Similarly, we have

$$\theta_s \mid \boldsymbol{y} \sim \mathcal{N}\left(\mu_{\theta,s}, \sigma_{\theta,s}^2\right),$$

with

$$\sigma_{\theta,s}^{-2} = c\left\{q + \left(\sigma_0^{-2}\right)^{(1)}\left(\lambda_s^{-2}\right)^{(1)}\right\}, \qquad \mu_{\theta,s} = c\,\sigma_{\theta,s}^2\left(\sum_{t=1}^{q} z_{st}^{(1)} - \sum_{t=1}^{q}\zeta_t^{(1)}\right).$$

The global precision parameters have variational distributions

$$\sigma_0^{-2} \mid y \sim \text{Gamma}\left(\nu_{\sigma_0}, \rho_{\sigma_0}\right), \qquad \left(\sigma_0^{-2}\right)^{(1)} = \nu_{\sigma_0}/\rho_{\sigma_0},$$

with

$$\nu_{\sigma_0} = \frac{c}{2}\left(p-1\right)+1, \qquad \rho_{\sigma_0} = c\left\{\left(\xi^{-1}\right)^{(1)} + \frac{1}{2}\sum_{s=1}^{p}\left(\lambda_s^{-2}\right)^{(1)}\left(\mu_{\theta,s}^2 + \sigma_{\theta,s}^2\right)\right\},$$

and

$$\xi^{-1} \mid \boldsymbol{y} \sim \text{Gamma}\left(\nu_\xi, \rho_\xi\right), \qquad \left(\xi^{-1}\right)^{(1)} = \nu_\xi/\rho_\xi,$$

with

$$\nu_\xi = 1, \qquad \rho_\xi = c\left\{q + \left(\sigma_0^{-2}\right)^{(1)}\right\}.$$

Finally, the updates for the local precision parameters are given by the following lemma.

Lemma C.2.3. *Let* $0 < c \leq 1$. *Then*

$$\left(\lambda_s^{-2}\right)^{(1)} = \frac{\Gamma(-c+2, L_s)}{L_s\Gamma(-c+1, L_s)} - 1, \tag{C.6}$$

where

$$L_s = \frac{c}{2}\left(\sigma_0^{-2}\right)^{(1)}\left(\mu_{\theta,s}^2 + \sigma_{\theta,s}^2\right),$$

and $\Gamma(\cdot,\cdot)$ *is the incomplete Gamma function. For* $c = 1$, *(C.6) reduces to*

$$\left(\lambda_s^{-2}\right)^{(1)} = \frac{1}{L_s \exp(L_s)\mathrm{E}_1(L_s)} - 1,$$

where $E_1(\cdot)$ *is the exponential integral function of order 1.*

Proof. Write $a_s = \lambda_s^{-2}$ for simplicity. Using (C.5), one finds

$$q\left(a_s\right) \propto \left(1+a_s\right)^{-c}\exp\left(-L_s a_s\right), \quad a_s > 0.$$

One then needs to compute

$$a_s^{(1)} = \frac{\displaystyle\int_0^\infty a_s(1+a_s)^{-c}\exp\left(-L_s a_s\right)\mathrm{d}a_s}{\displaystyle\int_0^\infty (1+a_s)^{-c}\exp\left(-L_s a_s\right)\mathrm{d}a_s}.$$

The denominator is obtained as

$$\begin{aligned}\int_0^\infty (1+a_s)^{-c}\exp\left(-L_s a_s\right)\mathrm{d}a_s &= \exp\left(L_s\right)\int_0^\infty (1+a_s)^{-c}\exp\left\{-L_s(1+a_s)\right\}\mathrm{d}a_s \\ &= \exp\left(L_s\right)L_s^{c-1}\,\Gamma(-c+1, L_s)\end{aligned}$$

with $\Gamma(s,x) = \int_x^\infty t^{s-1}\mathrm{e}^{-t}\mathrm{d}t$. The numerator can be decomposed as

$$\int_0^\infty a_s(1+a_s)^{-c}\exp\left(-L_s a_s\right)\mathrm{d}a_s = \int_0^\infty (1+a_s)^{1-c}\exp\left\{-L_s a_s\right\}\mathrm{d}a_s - \int_0^\infty (1+a_s)^{-c}\exp\left\{-L_s a_s\right\}\mathrm{d}a_s$$

The second term is the denominator computed above (changing the sign). The first term can be computed in a similar fashion as

$$\int_0^\infty (1+a_s)^{1-c} \exp\{-L_s a_s\}\, \mathrm{d}a_s = \exp(L_s)\, L_s^{c-2}\, \Gamma(-c+2, L_s).$$

The first part of the lemma follows immediately. The second part is trivially obtained by noting that $\Gamma(1, L_s) = e^{-L_s}$ and $\Gamma(0, L_s) = \int_{L_s}^\infty t^{-1} e^{-t} \mathrm{d}t = \mathrm{E}_1(L_s)$. □

To avoid overflow/underflow issues and ensure numerical stability, we implemented these updates using the *log-sum-exp* formulation (Calafiore and El Ghaoui, 2014) where appropriate and we used an iterative scheme based on continued fractions for evaluating $\exp(x)\mathrm{E}_1(x)$, $x > 0$, similarly as described in Neville et al. (2014).

C.2.2 Variational lower bound

We now provide the variational lower bound, $\mathcal{L}(q)$, of the marginal log-likelihood, $\log p(\boldsymbol{y})$; $\mathcal{L}(q)$ is evaluated to monitor convergence at each iteration, once the final temperature $T = 1$ has been reached:

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{\nu}) \log\left\{\frac{p(\boldsymbol{y},\boldsymbol{\nu})}{q(\boldsymbol{\nu})}\right\} \mathrm{d}\boldsymbol{\nu} \\
&= \sum_{t=1}^q \mathcal{L}_y\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t, \tau_t\right) + \sum_{s=1}^p \sum_{t=1}^q \mathcal{L}_{\beta,\gamma}\left(\beta_{st}, \gamma_{st}, z_{st} \mid \sigma^{-2}, \tau_t, \theta_s, \zeta_t\right) + \sum_{t=1}^q \mathcal{L}_\zeta(\zeta_t) \\
&\quad + \sum_{s=1}^p \mathcal{L}_\theta(\theta_s) + \mathcal{L}_{\sigma_0}\left(\sigma_0^{-2}\right) + \mathcal{L}_\xi\left(\xi^{-1}\right) + \sum_{s=1}^p \mathcal{L}_\lambda\left(\lambda_s^{-2}\right) + \mathcal{L}_\sigma\left(\sigma^{-2}\right) + \sum_{t=1}^q \mathcal{L}_\tau(\tau_t),
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{L}_y\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t\right) &= \mathrm{E}_q\left\{\log p(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \tau_t)\right\} \\
&= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\mathrm{E}\left(\log\tau_t\right) - \tau_t^{(1)}\left\{\kappa_{\tau,t} - \frac{1}{2}\sum_{s=1}^p \gamma_{st}^{(1)}\left(\sigma^2_{\beta,st} + \mu^2_{\beta,st}\right)\left(\sigma^{-2}\right)^{(1)} - \kappa_t\right\},
\end{aligned}$$

$$\begin{aligned}
\mathcal{L}_{\beta,\gamma}\left(\beta_{st}, \gamma_{st}, z_{st} \mid \sigma^{-2}, \tau_t, \theta_s, \zeta_t\right) &= \mathrm{E}_q\left\{\log p\left(\beta_{st} \mid \gamma_{st}, \sigma^{-2}, \tau_t\right)\right\} + \mathrm{E}_q\left\{\log p\left(\gamma_{st} \mid z_{st}\right)\right\} \\
&\quad + \mathrm{E}_q\left\{\log p\left(z_{st} \mid \theta_s, \zeta_t\right)\right\} - \mathrm{E}_q\left\{\log q\left(\beta_{st}, \gamma_{st}, z_{st}\right)\right\} \\
&= \mathcal{L}_\beta\left(\beta_{st} \mid \gamma_{st}, z_{st}, \sigma^{-2}, \tau_t\right) + \mathcal{L}_\gamma\left(\gamma_{st}, z_{st} \mid \theta_s, \zeta_t\right),
\end{aligned}$$

with

$$\begin{aligned}
\mathcal{L}_\beta\left(\beta_{st} \mid \gamma_{st}, z_{st}, \sigma^{-2}, \tau_t\right) &= \mathrm{E}_q\left\{\log p\left(\beta_{st} \mid \gamma_{st}, \sigma^{-2}, \tau_t\right)\right\} - \mathrm{E}_q\left\{\log q\left(\beta_{st} \mid z_{st}\right)\right\} \\
&= \frac{1}{2}\gamma_{st}^{(1)}\left\{\mathrm{E}_q\left(\log\sigma^{-2}\right) + \mathrm{E}_q\left(\log\tau_t\right) - \left(\mu^2_{\beta,st} + \sigma^2_{\beta,st}\right)\left(\sigma^{-2}\right)^{(1)}\tau_t^{(1)}\right\} \\
&\quad + \frac{1}{2}\gamma_{st}^{(1)}\left(\log\sigma^2_{\beta,st} + 1\right),
\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}_\gamma\left(\gamma_{st}, z_{st} \mid \theta_s, \zeta_t\right) &= \mathrm{E}_q\left\{\log p\left(\gamma_{st} \mid z_{st}\right)\right\} + \mathrm{E}_q\left\{\log p\left(z_{st} \mid \theta_s, \zeta_t\right)\right\} - \mathrm{E}_q\left\{\log q\left(z_{st} \mid \gamma_{st}\right)\right\} - \mathrm{E}_q\left\{\log q\left(\gamma_{st}\right)\right\} \\ &= \left(1-\gamma_{st}^{(1)}\right)\log\left\{1-\Phi\left(\alpha_{st}^{(1)}\right)\right\} + \gamma_{st}^{(1)}\log\Phi\left(\alpha_{st}^{(1)}\right) - \frac{1}{2}\sigma_{\theta,s}^2 - \frac{1}{2}\sigma_{\zeta,t}^2 - \gamma_{st}^{(1)}\log\gamma_{st}^{(1)} \\ &\quad -\left(1-\gamma_{st}^{(1)}\right)\log\left(1-\gamma_{st}^{(1)}\right).\end{aligned}$$

We then find

$$\begin{aligned}\mathcal{L}_\zeta\left(\zeta_t\right) &= \mathrm{E}_q\left\{\log p(\zeta_t)\right\} - \mathrm{E}_q\left\{\log q(\zeta_t)\right\} \\ &= \frac{1}{2}\left\{-\log t_0^2 + \log\left(\sigma_{\zeta,t}^2\right) - t_0^{-2}\left(\mu_{\zeta,t} - n_0\right)^2 - t_0^{-2}\sigma_{\zeta,t}^2 + 1\right\},\end{aligned}$$

$$\begin{aligned}\mathcal{L}_\theta\left(\theta_s\right) &= \mathrm{E}_q\left\{\log p(\theta_s)\right\} - \mathrm{E}_q\left\{\log q(\theta_s)\right\} \\ &= \frac{1}{2}\left\{\left(\log\sigma_0^{-2}\right)^{(1)} + \left(\log\lambda_s^{-2}\right)^{(1)} + \log\left(\sigma_{\theta,s}^2\right) - \left(\sigma_0^{-2}\right)^{(1)}\left(\lambda_s^{-2}\right)^{(1)}\left(\mu_{\theta,s}^2 + \sigma_{\theta,s}^2\right) + 1\right\},\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{\sigma_0}\left(\sigma_0^{-2}\right) &= \mathrm{E}_q\left\{\log p(\sigma_0^{-2})\right\} - \mathrm{E}_q\left\{\log q(\sigma_0^{-2})\right\} \\ &= \left(\frac{1}{2} - \nu_{\sigma_0}\right)\left(\log\sigma_0^{-2}\right)^{(1)} - \left\{\left(\xi^{-1}\right)^{(1)} - \rho_{\sigma_0}\right\}\left(\sigma_0^{-2}\right)^{(1)} + \frac{1}{2}\left\{\log\xi^{-1}\right)^{(1)} - \nu_{\sigma_0}\log\rho_{\sigma_0} \\ &\quad -\frac{1}{2}\log\pi + \log\Gamma\left(\nu_{\sigma_0}\right),\end{aligned}$$

$$\begin{aligned}\mathcal{L}_\xi\left(\xi^{-1}\right) &= \mathrm{E}_q\left\{\log p(\xi^{-1})\right\} - \mathrm{E}_q\left\{\log q(\xi^{-1})\right\} \\ &= \left(\frac{1}{2} - \nu_\xi\right)\left(\log\xi^{-1}\right)^{(1)} - \left(q - \rho_\xi\right)\left(\xi^{-1}\right)^{(1)} - \nu_\xi\log\rho_\xi - \frac{1}{2}\log\pi + \log\Gamma\left(\nu_\xi\right) + \frac{1}{2}\log(q),\end{aligned}$$

$$\begin{aligned}\mathcal{L}_\lambda\left(\lambda_s^{-2}\right) &= \mathrm{E}_q\left\{\log p(\lambda_s^{-2})\right\} - \mathrm{E}_q\left\{\log q(\lambda_s^{-2})\right\} \\ &= -\log\pi - \frac{1}{2}\left(\log\lambda_s^{-2}\right)^{(1)} + L_s\left\{\left(\lambda_s^{-2}\right)^{(1)} + 1\right\} + \log\mathrm{E}_1\left(L_s\right),\end{aligned}$$

$$\begin{aligned}\mathcal{L}_\sigma\left(\sigma^{-2}\right) &= \mathrm{E}_q\left\{\log p\left(\sigma^{-2}\right)\right\} - \mathrm{E}_q\left\{\log q\left(\sigma^{-2}\right)\right\} \\ &= \left(\nu - \nu_\sigma\right)\left(\log\sigma^{-2}\right)^{(1)} - \left(\rho - \rho_\sigma\right)\left(\sigma^{-2}\right)^{(1)} + \nu\log\rho - \nu_\sigma\log\rho_\sigma - \log\Gamma\left(\nu\right) + \log\Gamma\left(\nu_\sigma\right),\end{aligned}$$

$$\begin{aligned}\mathcal{L}_\tau\left(\tau_t\right) &= \mathrm{E}_q\left\{\log p(\tau_t)\right\} - \mathrm{E}_q\left\{\log q(\tau_t)\right\} \\ &= \left(\eta_t - \eta_{\tau,t}\right)\left(\log\tau_t\right)^{(1)} - \left(\kappa_t - \kappa_{\tau,t}\right)\tau_t^{(1)} + \eta_t\log\kappa_t - \eta_{\tau,t}\log\kappa_{\tau,t} - \log\Gamma(\eta_t) + \log\Gamma(\eta_{\tau,t}).\end{aligned}$$

C.3 Student-t modification for the horseshoe local scales

C.3.1 Variational algorithm

We implemented a variant of the algorithm of Appendix C.2, which assigns half-t prior distributions to the local scales λ_s,

$$\theta_s \mid \lambda_s, \sigma_0 \sim \mathcal{N}\left(0, \lambda_s^2 \sigma_0^2\right), \qquad \lambda_s \sim \mathrm{t}_\nu^+(0,1), \tag{C.7}$$

where $\mathrm{t}_\nu^+(0,1)$ with degrees of freedom $\nu > 1$.

As for the half-Cauchy scales, we use a reparametrization of (C.7); it is based on the following generalization of Lemma C.2.1.

Lemma C.3.1. *If a is a random variable such that*

$$a \mid \xi \sim \mathrm{Gamma}\left(\frac{\nu}{2}, \nu\xi^{-1}\right), \qquad \xi^{-1} \sim \mathrm{Gamma}\left(\frac{1}{2}, A^{-2}\right), \quad A > 0, \nu \geq 1,$$

then $a^{-1/2} \sim \mathrm{t}_\nu^+(0, A)$.

With this lemma one finds, writing $a_s = \lambda_s^{-2}$,

$$p(a_s) = \pi^{-1/2}\Gamma\left(\frac{\nu}{2}\right)^{-1} \nu^{\nu/2}\, a_s^{\nu/2-1} \left(\frac{\nu-1}{2}\right)! \,(\nu a_s + 1)^{-(\nu+1)/2}, \quad a_s > 0, \nu = 2k+1, k \in \mathbb{N};$$

the updates derived hereafter are only valid for ν odd.

Using (C.5), we get

$$a_s^{(1)} = \frac{\displaystyle\int_0^\infty a_s^{c(\nu-1)/2+1}(1+\nu a_s)^{-c(\nu+1)/2}\exp\left(-L_s a_s\right)\mathrm{d}a_s}{\displaystyle\int_0^\infty a_s^{c(\nu-1)/2}(1+\nu a_s)^{-c(\nu+1)/2}\exp\left(-L_s a_s\right)\mathrm{d}a_s}.$$

Using formula 3.383(5) of Gradshteyn and Ryzhik (1994), we find, for the denominator,

$$D(\nu, c) = \nu^{-c(\nu-1)/2-1}\,\Gamma\left\{\frac{c}{2}(\nu-1)+1\right\}\Psi\left\{\frac{c}{2}(\nu-1)+1, 2-c, \frac{L_s}{\nu}\right\}, \tag{C.8}$$

and for the numerator,

$$N(\nu, c) = \nu^{-c(\nu-1)/2-2}\,\Gamma\left\{\frac{c}{2}(\nu-1)+2\right\}\Psi\left\{\frac{c}{2}(\nu-1)+2, 3-c, \frac{L_s}{\nu}\right\},$$

where

$$\Psi(\alpha, \beta; z) = \frac{\Gamma(1-\beta)}{\Gamma(\alpha-\beta+1)}\Phi(\alpha, \beta; z) + \frac{\Gamma(\beta-1)}{\Gamma(\alpha)} z^{1-\beta}\Phi(\alpha-\beta+1, 2-\beta; z),$$

and $\Phi(\alpha, \beta; z)$ is the confluent hypergeometric function (Gradshteyn and Ryzhik, 1994, formula 9.210(1)). For $c = 1$, a direct evaluation of these expressions is numerically unstable, even when using classical algorithmic techniques such as *log-sum-exp*, so we instead employ an recursive approach: we observe that

$$\begin{aligned}\int_0^\infty x^n(1+\alpha x)^{-m}\exp(-\beta x)\mathrm{d}x \quad &= \quad \frac{1}{\alpha}\left\{\int_0^\infty x^{n-1}(1+\alpha x)^{-m+1}\exp(-\beta x)\mathrm{d}x\right. \\ &\qquad \left. -\int_0^\infty x^{n-1}(1+\alpha x)^{-m}\exp(-\beta x)\mathrm{d}x\right\},\end{aligned}$$

for $m \geq n > 0$, $n, m \in \mathbb{N}$, $\alpha \in \mathbb{R}\setminus\{0\}$, and $\beta > 0$, and apply Formulas 3.3.52(4) and 3.353(2) of Gradshteyn and Ryzhik (1994). The resulting expression involves multiple evaluations of the exponential integral function $\mathrm{E}_1(\cdot)$; we experienced no numerical issue for ν up to 7, thanks to the continued fraction iterative implementation mentioned in Appendix C.2.1. The corresponding contribution to the variational lower bound is

$$\begin{aligned}\mathcal{L}_\lambda\left(\lambda_s^{-2}\right) &= \mathrm{E}_q\left\{\log p(\lambda_s^{-2})\right\} - \mathrm{E}_q\left\{\log q(\lambda_s^{-2})\right\} \\ &= -\frac{1}{2}\log\pi - \log\Gamma\left(\frac{\nu}{2}\right) + \frac{\nu}{2}\log\nu - \frac{1}{2}\left(\log\lambda_s^{-2}\right)^{(1)} + \log D(\nu,1) + L_s\left(\lambda_s^{-2}\right)^{(1)} + \log\left(\frac{\nu-1}{2}!\right),\end{aligned}$$

where $D(\nu,1)$ is given by (C.8).

C.3.2 Experiments on real data

We applied both the algorithm of Appendix C.2 and the half-t modification of Appendix C.3.1 to the eQTL data analyzed in Section 5.5. We considered the first 2,079 SNPs of chromosome one (so as to cover complete haplotypes) and all 24,461 unstimulated transcript expression levels, for all 413 individuals. Using simulated annealing with initial temperature of $T = 20$ produced unexpectedly large values for a few half-Cauchy local scales. But the half-t_7 variant did not solve the problem, rather, it was transferred to the global scale σ_0 which grew unreasonably large, owing to its half-Cauchy scale, $\mathrm{C}^+(0, q^{-1/2})$. Artefacts were only removed after fixing $\sigma_0 = q^{-1/2}$. We compared inferences with this model using initial temperature $T = 20$, to that of the horseshoe algorithm of Appendix C.2 using lower initial temperature, $T = 5$ (which did not trigger any artefact). Figure C.1 shows the largest posterior probabilities of inclusion, which are very similar for both inferences; some values are slightly larger for the half-t_7 variants. The maximum hotspot size obtained by thresholding these probabilities at 0.8 is 18 in both the half-Cauchy and the half-t cases.

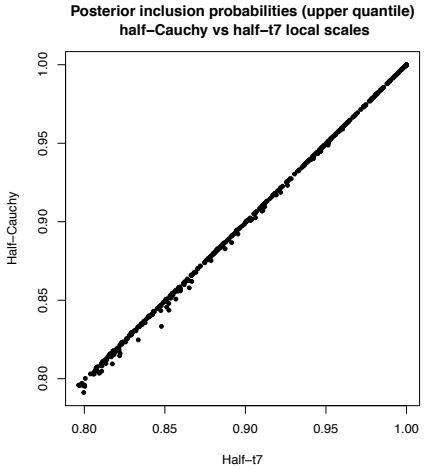

Figure C.1 – Posterior probabilities of inclusion above the $1-\alpha = 1-10^{-5}$ quantile of their empirical distribution, in the half-Cauchy or the half-t_7 case. The data are the first $p = 2{,}079$ SNPs from chromosome one, and the 24,461 unstimulated transcript expression levels, for all 413 individuals from the eQTL study of Section 5.5.

C.3.3 Simulation study

We compared the two variants (half-Cauchy local scales with $T = 5$ and half-t_7 local scales with $T = 20$) on two simulated datasets, each with 10 hotspots: in the first, hotspots have average size ≈ 20 and in the second, ≈ 70. The data involve $1,000$ predictors and $5,000$ responses with block correlation structure corresponding as in the experiments of Section 5.4.2. Figure C.2 indicates that both algorithms have same selection performance in the small-hotspot case, but the half-Cauchy local scale variant is more powerful in the large-hotspot case; the t_7 tails are too light to fetch the large hotspots.

This experiment and that of Appendix C.3.2 suggest using our original horseshoe algorithm of Appendix C.2, with weak annealing (e.g., $T = 5$).

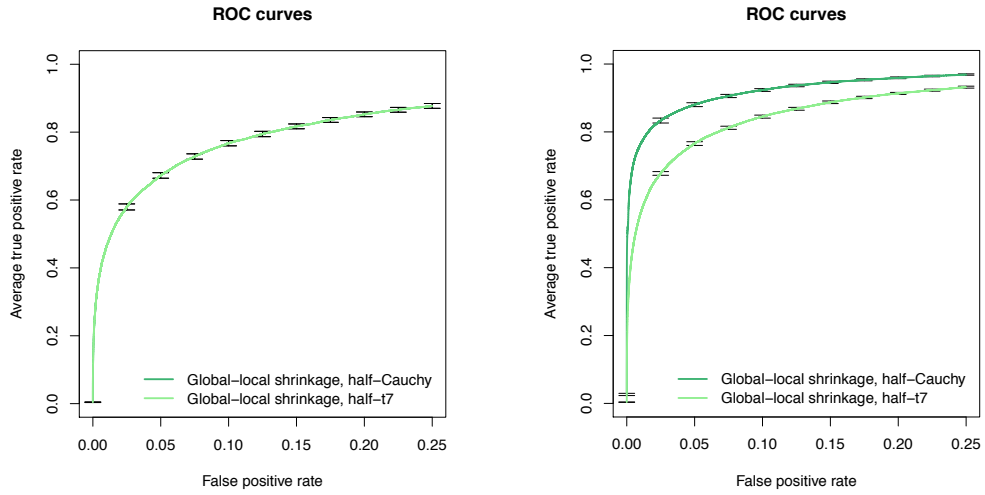

Figure C.2 – Truncated average ROC curves for predictor-response selection performance with half-t_7 and half-Cauchy local scales. Problem with $p = 1,000$ predictors, of which 10 are hotspots of average size ≈ 20 (left) and ≈ 70 (right), $q = 5,000$ responses and $n = 300$ samples. The block-autocorrelation coefficients for predictors were drawn from the interval $(0.75, 0.95)$, and the residual block-equicorrelation coefficients for the responses, from the interval $(0, 0.25)$. Hotspots explain at most 10% of each response variance. The expected number of SNPs associated per transcript is based on the same base rate as in Section 5.4.2, $\mathrm{E}_p = 0.002 \times p = 2$, and the variance for this number was set to $\mathrm{V}_p = 10$. The two curves overlap in the left panel.

C.4 Complements to simulation experiments

We next provide additional performance illustrations for the simulation experiments presented in Section 5.4.

C.4.1 Simulation study 1: performance with global-local modelling

Figure C.3 compares hotspot selection performance for the five models discussed in Section 5.4.2 on the reference scenario. It also compares the cumulated posterior probabilities for the small simulated hotspots for the fixed-variance model with largest variance and our global-local proposal.

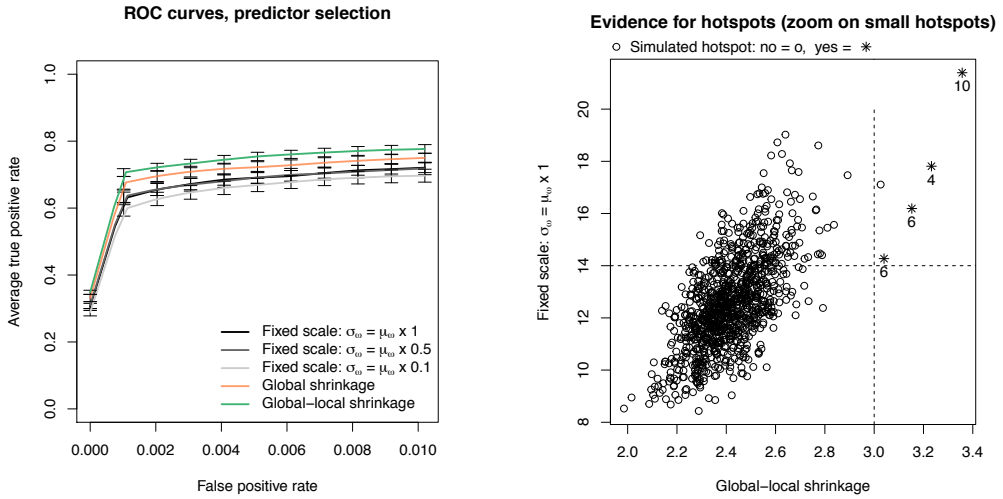

Figure C.3 – Performance of five hotspot modelling approaches. Left: truncated average ROC curves for hotspot selection with 95% confidence intervals obtained from 64 replications. Right: evidence for hotspots computed as, for each candidate predictor, the sum of its posterior probabilities of associations with all responses; average over 16 replications. Zoom on the noise level and four smallest hotspots, with simulated sizes 4, 6, 6 and 10. The dashed lines highlight how the global-local shrinkage proposal is better than the fixed-scale model with $\sigma_\omega = \mu_\omega$ in discriminating weak hotspot signals from the noise. The data comprise $p = 1,000$ simulated SNPs with 20 hotspots, and $q = 20,000$ responses, of which 200 are associated with at least one hotspot, leaving the rest of the responses unassociated. The block-autocorrelation coefficients for SNPs were drawn from the interval $(0.75, 0.95)$, and the residual block-equicorrelation coefficients for responses, from the interval $(0, 0.25)$. At most 25% of each response variance is explained by the hotspots. For the fixed-variance models, we used a base-rate of $\mu_\omega = 0.002$, and scales of $\sigma_\omega = \mu_\omega \times \{1, 0.5, 0.1\}$.

C.4.2 Simulation study 2: performance with and without simulated annealing

Figure C.4 compares the performances of classical and annealed variational inferences on the data of Section 5.4.4.

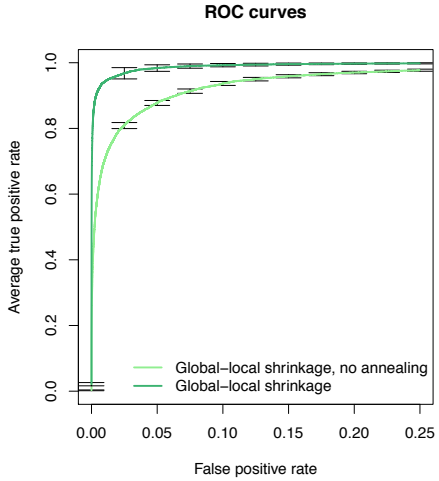

Figure C.4 – Pairwise selection performance by classical and annealed variational algorithms. Truncated average ROC curves for hotspot selection with 95% confidence intervals obtained from 16 replications.

C.5 Stimulated eQTL analysis: overlap of transcripts associated with hotspot rs6581889 across conditions

Figure C.5 shows the overlap of transcripts found associated with hotspot rs6581889 in the application to monocyte eQTL data presented in Section 5.5. Most associations are shared between the unstimulated and IFN-gamma conditions, although 107 associations are specific to the latter condition. Stimulations by LPS of 2 and 24 hours have no triggering effect.

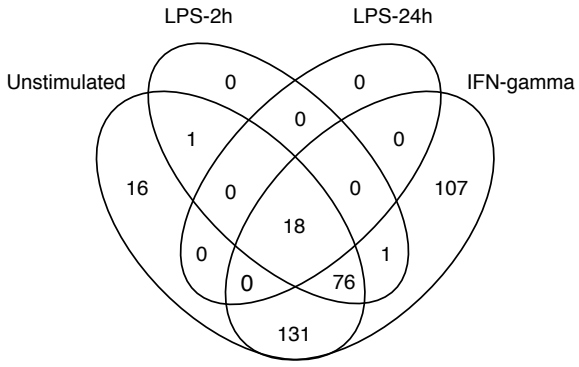

Figure C.5 – Venn diagram for transcripts associated with rs6581889 across conditions.

D Appendix for Chapter 6

D.1 Derivation of the variational-EM algorithm

D.1.1 Variational distributions

We provide the detailed derivation of the variational-EM algorithm for our two-stage hierarchical model encoding predictor-level covariates presented in Section 6.2. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q)$ be q centered responses, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ be p centered predictors, for n samples. We rewrite the model using the auxiliary variable z_{st} as

$$
\begin{array}{rcll}
\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \boldsymbol{\tau}_t & \sim & \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}_t, \tau_t^{-1}\boldsymbol{I}_n\right), & \tau_t \sim \text{Gamma}(\eta_t, \kappa_t), \qquad t = 1, \ldots, q, \\
\beta_{st} \mid \gamma_{st}, \tau_t, \sigma^2 & \sim & \gamma_{st}\mathcal{N}\left(0, \sigma^2\tau_t^{-1}\right) + (1-\gamma_{st})\,\delta_0, & \sigma^{-2} \sim \text{Gamma}(\lambda, \nu), \qquad s = 1, \ldots, p, \\
\gamma_{st} & = & \mathbb{1}\{z_{st} > 0\}, & z_{st} \mid \theta_{b,s}, \zeta_t, \boldsymbol{\xi}_b \sim \mathcal{N}\left(\theta_{b,s} + \zeta_t + \boldsymbol{V}_s^T\boldsymbol{\xi}_b, 1\right), \\
\xi_{b,l} \mid \rho_{b,l} & \sim & \rho_{b,l}\mathcal{N}\left(0, s_b^2\right) + (1-\rho_{b,l})\,\delta_0, & \theta_{b,s} \sim \mathcal{N}(0, s_{0b}^2), \qquad \zeta_t \sim \mathcal{N}(n_0, t_0^2), \\
\rho_{b,l} & \sim & \text{Bernoulli}\left(\omega_{b,l}\right), & l = 1, \ldots, r,
\end{array}
$$

where $\boldsymbol{V} = (V_1, \ldots V_r)$ is the $p \times r$ matrix of centered predictor-level covariates, $b \in \mathscr{P}$ is a block of predictor and response variables, with $\mathscr{P}$ a partition of $\{1, \ldots, p\} \times \{1, \ldots, q\}$, $b \ni (s,t)$.

Let $\boldsymbol{\nu} = \left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\zeta}, \boldsymbol{\xi}, \boldsymbol{\rho}, \boldsymbol{\tau}, \sigma^{-2}\right)$ be the parameter vector, $\boldsymbol{\eta}_b = \left(\boldsymbol{\omega}_b, s_{0b}^2, s_b^2\right)$ be the hyperparameter vector for the second-stage model for block b, and $\boldsymbol{\eta} = \left(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_B\right)$. We have

$$
\begin{aligned}
p(\boldsymbol{y}, \boldsymbol{\nu} \mid \boldsymbol{\eta}) \quad = \quad & \left\{\prod_{t=1}^{q} p\left(\boldsymbol{y}_t \mid \beta_t, \tau_t\right)\right\}\left\{\prod_{t=1}^{q}\prod_{s=1}^{p} p\left(\beta_{st} \mid \gamma_{st}, \sigma^{-2}, \tau_t\right)\right\}\left\{\prod_{t=1}^{q} p\left(\tau_t\right)\right\} p(\sigma^{-2}) \\
& \times \left\{\prod_{t=1}^{q}\prod_{s=1}^{p} p\left(\gamma_{st} \mid z_{st}\right) p(z_{st} \mid \theta_{b,s}, \zeta_t, \boldsymbol{\xi}_b)\right\}\left\{\prod_{b=1}^{B}\prod_{s\in b} p(\theta_{b,s} \mid s_{0b}^2)\right\} \\
& \times \left\{\prod_{t=1}^{q} p(\zeta_t)\right\}\left\{\prod_{b=1}^{B}\prod_{l=1}^{r} p(\xi_{b,l} \mid \rho_{b,l}, s_b^2) p(\rho_{b,l} \mid \omega_{b,l})\right\},
\end{aligned}
$$

where $s \in b$ is a shortening meaning s is a predictor index from the predictor-response block b. We wrote the second conditional distribution of the second line, $p(z_{st} \mid \theta_{b,s}, \zeta_t, \boldsymbol{\xi}_b)$, where b implicitly corresponds to the block containing (s,t); we may use such tacit notation hereafter for brevity.

We use the following mean-field form for the variational approximation,

$$\begin{aligned} q(\boldsymbol{v}) &= \left\{\prod_{t=1}^{q}\prod_{s=1}^{p} q(\beta_{st},\gamma_{st},z_{st})\right\}\left\{\prod_{t=1}^{q} q(\tau_t)\right\} q\left(\sigma^{-2}\right)\left\{\prod_{b=1}^{B}\prod_{s\in b} q(\theta_{b,s})\right\}\left\{\prod_{t=1}^{q} q(\zeta_t)\right\} \\ &\quad \times\left\{\prod_{b=1}^{B}\prod_{l=1}^{r} q(\xi_{b,l},\rho_{b,l})\right\}. \end{aligned}$$

The annealed variational updates for $\boldsymbol{\tau}$, σ^{-2}, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{z}$ are the same as those for the global-local model, replacing $\theta_s^{(1)}+\zeta_t^{(1)}$ with $\theta_{b,s}^{(1)}+\zeta_t^{(1)}+\boldsymbol{V}_s^T\boldsymbol{\xi}_b^{(1)}$ in Appendix C.2. For $\theta_{b,s}$, we have

$$\theta_{b,s} \mid \boldsymbol{y} \sim \mathcal{N}\left(\mu_{\theta,b,s},\sigma^2_{\theta,b,s}\right),$$

with

$$\sigma^{-2}_{\theta,b,s} = c\left(q_b + s_{0b}^{-2}\right), \qquad \mu_{\theta,b,s} = c\,\sigma^2_{\theta,b,s}\left\{\sum_{t\in b}\left(z_{st}^{(1)}-\zeta_t^{(1)}\right) - \boldsymbol{V}_s^T\boldsymbol{\xi}_b^{(1)}\right\},$$

where q_b is the number of responses in block b, $t\in b$ means t is a response index from block b, and $0<c\leq 1$ is the annealing inverse temperature parameter. For ζ_t, we find

$$\zeta_t \mid \boldsymbol{y} \sim \mathcal{N}\left(\mu_{\zeta,t},\sigma^2_{\zeta,t}\right),$$

with

$$\sigma^{-2}_{\zeta,t} = c\left(p + t_0^{-2}\right), \qquad \mu_{\zeta,t} = c\,\sigma^2_{\zeta,t}\left\{\sum_{b:t\in b}\sum_{s\in b}\left(z_{st}^{(1)}-\theta_{b,s}^{(1)}-\boldsymbol{V}_s^T\boldsymbol{\xi}_b^{(1)}\right) + t_0^{-2}n_0\right\}.$$

The variational distribution for the annotation effects is

$$q(\xi_{b,l},\rho_{b,l}) = q(\xi_{b,l}\mid\rho_{b,l})q(\rho_{b,l}),$$

with

$$\xi_{b,l}\mid\rho_{b,l}=1,\boldsymbol{y}\sim\mathcal{N}\left(\mu_{\xi,b,l},\sigma^2_{\xi,b,l}\right), \quad \xi_{b,l}\mid\rho_{b,l}=0,\boldsymbol{y}\sim\delta_0, \quad \rho_{b,l}\mid\boldsymbol{y}\sim\text{Bernoulli}\left(\rho_{b,l}^{(1)}\right),$$

where

$$\sigma^{-2}_{\xi,b,l} = c\left\{q_b\sum_{s\in b}V_{sl}^2 + s_b^{-2}\right\},$$

$$\mu_{\xi,b,l} = c\sigma^2_{\xi,b,l}\sum_{s\in b}V_{sl}\left\{\sum_{t\in b}z_{st} - q_b\mu_{\theta,b,s} - \sum_{t\in b}\mu_{\zeta,t} - q_b\sum_{j=1,j\neq l}^{r}\rho_{b,j}^{(1)}\mu_{\xi,b,j}V_{sj}\right\},$$

and

$$\frac{1}{\rho_{b,l}^{(1)}} = 1+\exp\left[-c\left\{\log\omega_{b,l}+\frac{1}{2}\mu^2_{\xi,b,l}\sigma^{-2}_{\xi,b,l}-\frac{1}{2}\log s_b^2-\log\left(1-\omega_{b,l}\right)+\log\sigma_{\xi,b,l}\right\}\right].$$

D.1.2 Variational lower bound

We provide the computational details for the lower bound, $\mathcal{L}(q;\boldsymbol{\eta})$, of the marginal log-likelihood, $\log p(\boldsymbol{y} \mid \boldsymbol{\eta})$:

$$\begin{aligned}\mathcal{L}(q;\boldsymbol{\eta}) &= \int q(\boldsymbol{v}) \log\left\{\frac{p(\boldsymbol{y},\boldsymbol{v})}{q(\boldsymbol{v})}\right\} \mathrm{d}\boldsymbol{v} \\ &= \sum_{t=1}^{q} \mathcal{L}_y\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t, \tau_t\right) + \sum_{t=1}^{q}\sum_{s=1}^{p} \mathcal{L}_\beta\left(\beta_{st}, \gamma_{st}, z_{st} \mid \sigma^{-2}, \tau_t, \theta_{b,s}, \zeta_t, \boldsymbol{\xi}\right) + \sum_{t=1}^{q} \mathcal{L}_\tau\left(\tau_t\right) \\ &\quad + \mathcal{L}_\sigma\left(\sigma^{-2}\right) + \sum_{b=1}^{B}\sum_{s\in b} \mathcal{L}_\theta\left(\theta_{b,s} \mid s_{0b}^2\right) + \sum_{t=1}^{q} \mathcal{L}_\zeta\left(\zeta_t\right) + \sum_{b=1}^{B}\sum_{l=1}^{r} \mathcal{L}_{\xi,\rho}\left(\xi_{b,l}, \rho_{b,l} \mid s_b^2, \omega_{b,l}\right),\end{aligned} \tag{D.1}$$

where $\mathcal{L}_y\left(\boldsymbol{y}_t \mid \boldsymbol{\beta}_t, \boldsymbol{\gamma}_t, \tau_t\right)$, $\mathcal{L}_\tau\left(\tau_t\right)$, $\mathcal{L}_\sigma\left(\sigma^{-2}\right)$ and $\mathcal{L}_\zeta(\zeta_t)$ are the same as for the model of Chapter 5 (see Appendix C.2), and we have

$$\begin{aligned}\mathcal{L}_\beta\left(\beta_{st}, \gamma_{st}, z_{st} \mid \sigma^{-2}, \tau_t, \theta_{b,s}, \zeta_t, \boldsymbol{\xi}_b\right) &= \frac{1}{2}\gamma_{st}^{(1)}\left\{\mathrm{E}_q\left(\log\sigma^{-2}\right) + \mathrm{E}_q\left(\log\tau_t\right) - \left(\mu_{\beta,st}^2 + \sigma_{\beta,st}^2\right)\left(\sigma^{-2}\right)^{(1)}\tau_t^{(1)}\right\} \\ &\quad + \frac{1}{2}\gamma_{st}^{(1)}\left(\log\sigma_{\beta,st}^2 + 1\right) + \gamma_{st}^{(1)}\log\Phi\left(\theta_{b,s}^{(1)} + \zeta_t^{(1)} + \boldsymbol{V}_s^T\boldsymbol{\xi}_b^{(1)}\right) \\ &\quad + \left(1-\gamma_{st}^{(1)}\right)\log\left\{1 - \Phi\left(\theta_{b,s}^{(1)} + \zeta_t^{(1)} + \boldsymbol{V}_s^T\boldsymbol{\xi}_b^{(1)}\right)\right\} \\ &\quad - \frac{1}{2}\sigma_{\theta,b,s}^2 - \frac{1}{2}\sigma_{\zeta,t}^2 - \frac{1}{2}\sum_{l=1}^{r} V_{sl}^2\left\{\sigma_{\xi,l}^2 + \mu_{\xi,l}^2(1-\rho_l^{(1)})\right\}\rho_l^{(1)} \\ &\quad - \gamma_{st}^{(1)}\log\gamma_{st}^{(1)} - \left(1-\gamma_{st}^{(1)}\right)\log\left(1-\gamma_{st}^{(1)}\right),\end{aligned}$$

for $(s,t) \ni b$. Then,

$$\mathcal{L}_\theta(\theta_{b,s} \mid s_{0b}^2) = \frac{1}{2}\left\{-\log s_{0b}^2 + \log\sigma_{\theta,b,s}^2 - s_{0b}^{-2}\left(\mu_{\theta,b,s}^2 + \sigma_{\theta,b,s}^2\right) + 1\right\},$$

and

$$\begin{aligned}\mathcal{L}_{\xi,\rho}\left(\xi_{b,l}, \rho_{b,l} \mid s_b^2, \omega_{b,l}\right) &= -\frac{1}{2}\rho_{b,l}^{(1)}\log s_b^2 + \frac{1}{2}\rho_{b,l}^{(1)}\left(\log\sigma_{\xi,b,l}^2 + 1\right) - \rho_{b,l}^{(1)}\log\rho_{b,l}^{(1)} - (1-\rho_{b,l}^{(1)})\log\left(1-\rho_{b,l}^{(1)}\right) \\ &\quad - \frac{1}{2s_b^2}\rho_{b,l}^{(1)}\left(\sigma_{\xi,b,l}^2 + \mu_{\xi,b,l}^2\right) + \rho_{b,l}^{(1)}\log\omega_{b,l} + \left(1-\rho_{b,l}^{(1)}\right)\log\left(1-\omega_{b,l}\right).\end{aligned}$$

D.1.3 EM hyperparameter updates

The M-step updates of the variational-EM algorithm are obtained by taking the first derivative of (D.1) with respect to the hyperparameters. We have

$$\frac{\partial}{\partial s_{0b}^2}\mathcal{L}(q;\boldsymbol{\eta}) = \sum_{s\in b}\frac{\partial}{\partial s_{0b}^2}\mathcal{L}_\theta(\theta_{b,s} \mid s_{0b}^2) = -\frac{p_b}{2}s_{0b}^{-2} + \frac{1}{2}s_{0b}^{-4}\sum_{s\in b}\left(\mu_{\theta,b,s}^2 + \sigma_{\theta,b,s}^2\right),$$

where p_b is the number of predictors in block b, so

$$s_{0b}^2 = \frac{1}{p_b}\sum_{s\in b}\left(\mu_{\theta,b,s}^2 + \sigma_{\theta,b,s}^2\right),$$

$$\frac{\partial}{\partial s_b^2}\mathcal{L}(q;\boldsymbol{\eta}) = \frac{\partial}{\partial s_b^2}\sum_{l=1}^{r}\mathcal{L}_{\xi,\rho}(\xi_{b,l},\rho_{b,l}\mid s_b^2,\omega_{b,l}) = -\frac{1}{2}s_b^{-2}\sum_{l=1}^{r}\rho_{b,l}^{(1)} + \frac{1}{2}s_{0b}^{-4}\sum_{l=1}^{r}\rho_{b,l}^{(1)}\left(\mu_{\xi,b,l}^2+\sigma_{\xi,b,l}^2\right),$$

so

$$s_b^2 = \frac{\sum_{l=1}^{r}\rho_{b,l}^{(1)}\left(\mu_{\xi,b,l}^2+\sigma_{\xi,b,l}^2\right)}{\sum_{l=1}^{r}\rho_{b,l}^{(1)}},$$

and finally,

$$\frac{\partial}{\partial\omega_{b,l}}\mathcal{L}(q;\boldsymbol{\eta}) = \frac{\partial}{\partial\omega_{b,l}}\mathcal{L}_{\xi,\rho}(\xi_{b,l},\rho_{b,l}\mid s_b^2,\omega_{b,l}) = \frac{\rho_{b,l}^{(1)}}{\omega_{b,l}} - \frac{1-\rho_{b,l}^{(1)}}{1-\omega_{b,l}},$$

so

$$\omega_{b,l} = \rho_{b,l}^{(1)}.$$

E Appendix for Chapter 7

E.1 Methods

E.1.1 Ethics

The study was approved by the local human research ethic committees. Participants provided informed written consent, and all procedures were conducted in accordance with the Declaration of Helsinki.

E.1.2 Study Samples

The *Ottawa* study was a medically supervised program set up by the Weight Management Clinic of Ottawa (Dent et al., 2002). Participants followed a low caloric diet at 900 kcal/day, for six to twelve weeks using a meal-replacement product (Optifast900, Nestlé Health Science, Switzerland). Subjects under medication known to affect the rate of weight loss, glucose homeostasis or thyroid indices were excluded from all analyses, and subjects who were not under fasting conditions at plasma sample collection were excluded from the proteomic analyses.

The *DiOGenes* study was an interventional, multi-center pan-European program (Larsen et al., 2010). Eight partner states participated to the study: Bulgaria, the Czech Republic, Denmark, Germany, Greece, the Netherlands, Spain and United Kingdom. Participants followed an eight-week low calorie diet at 800 kcal/day, using Modifast (Nutrition et Santé, France).

The main clinical characteristics of both cohorts are given in Supplementary Table S4.

E.1.3 Proteomic data

Plasma protein expression data were obtained using two types of technologies: mass-spectrometry (MS) and a multiplexed aptamer-based assay developed by SomaLogic (Kraemer et al., 2011). Samples were randomized, ensuring that the plate numbers were not associated with age, gender, ethnicity, weight-related measures, glycemic indices, measures of chemical biochemistry, and, for the DiOGenes samples, collection centers.

The MS proteomic quantification used plasma samples spiked with protein standard lactoglobulin (LACB). Samples were immuno-depleted, reduced, digested, isobarically 6-plex labeled and purified.

They were analysed in duplicates on two separate but identical systems using linear ion trap with Orbitrap Elite analyzer and Ultimate 3000 RSLCnano System (Thermo Scientific). Protein identification was done with the UniProtKB/Swiss-Prot database (Boutet et al., 2007), using Mascot 2.4.0 (Matrix Sciences) and Scaffold 4.2.1 (Proteome Software). Both peptide and protein false discovery rates were set to 1%, with a criterion of two unique peptides. The relative quantitative protein values corresponded to the $\log_2$-transformation of the protein ratio fold changes with respect to their measurements in the biological plasma reference sample. The sample preparation and all other manipulations relative to the MS measurements are detailed further in Cominetti et al. (2015).

The SomaLogic protein measurements were characterized using the SOMAscan assay (Kraemer et al., 2011), which relies on fluorescent labelling of poly-nucleotide aptamers targeting specific protein epitopes. Protein measurements were obtained in relative fluorescence unit and were then $\log_2$-transformed.

We discarded MS-based proteins if their measurements were missing for more than 5% of the samples, leaving 210 proteins in the Ottawa cohort, and 136, in the DiOGenes cohort; we restricted all downstream analyses to the 133 proteins available for both cohorts. The SomaLogic measurements had no missing values. A total of 1,100 proteins were assayed in the Ottawa cohort, and 1,129 in the DiOGenes cohort. All our analyses focus on the 1,096 proteins quantified for both cohorts. The overlap of between the MS and SomaLogic panels is of 72 proteins only.

We excluded samples with extreme expression values in more than 5% of the proteins, i.e., values beyond the outer fences of the empirical distribution ($q_{0.25} - 3 \times \text{IQR}$, $q_{0.75} + 3 \times \text{IQR}$, where $q_{0.25}$, $q_{0.75}$ are the lower and upper quartiles, and IQR, the interquartile range). After this quality control procedure, approximately 10 to 20 samples were removed from each of the four datasets; $n = 577$ and 428 Canadian samples remained in the MS and SomaLogic datasets, respectively, and $n = 481$ and 563 DiOGenes samples, in the MS and SomaLogic datasets.

E.1.4 Genotyping

Genotypes were generated using HumanCoreExome-12 v1.1 Illumina SNP arrays, according to their manufacturer's instructions (Steemers and Gunderson, 2005), and were called with the GenomeStudio Software (provided by Illumina). We discarded SNPs with call rate < 95%, violating Hardy–Weinberg equilibrium (false discovery rate < 20%), and we discarded subjects with low call rate (< 95%), abnormally high autosomal heterozygosity (false discovery rate < 1%), an XXY karyotype, or gender inconsistencies between genotype data and clinical records. For subjects with high identity-by-state (IBS> 95%), we kept only the one with the highest call rate. The subjects from both cohorts were of European ancestry and the two cohorts had similar genetic structure. We used principal component analyses separately on each cohort to exclude subjects that were extremely heterogeneous genetically. We performed genotype imputation using SHAPEIT (Delaneau et al., 2008) and IMPUTE2 (Howie et al., 2009), based on the European reference panel from the 1000 Genomes Project Consortium (2012, March 2012 release, phase 1 version 3). We then discarded SNPs with INFO score < 0.8, which left 4.9M imputed SNPs in both datasets. In order to avoid near-collinearity, which may render multivariate analyses unstable, we applied a light linkage disequilibrium (LD) pruning with PLINK (Purcell et al., 2007) using pairwise r^2 threshold 0.95. We also applied a minor allele frequency threshold of 5%, after having restricted the genotype data to the subjects with available proteomic data.

The above steps were performed separately for the Ottawa and the DiOGenes cohorts, so in order to define a common set of SNPs for discovery and replication, we restricted each dataset to the SNPs available for both cohorts. After all genetic quality controls, and in both cohorts, $p = 275,485$ SNPs remained for the SomaLogic analysis, and $p = 275,297$ for the MS analysis. In the Ottawa cohort there were $n = 376$ subjects having both genotype and MS proteomic data, and $n = 394$ subjects having both genotype and MS proteomic data. In the DiOGenes cohort, these numbers were $n = 400$ and 548.

E.1.5 Clinical data

Both cohorts had records on age, gender, anthropometric traits (weight and BMI), glycemic variables (fasting glucose, fasting insulin, HOMA-IR), and total lipid levels obtained from blood biochemistry (total cholesterol, triglycerides, HDL). We derived LDL values using the Friedewald formula (Friedewald et al., 1972), and obtained gender-specific *visceral adiposity index* (VAI) values using the formula proposed by Amato et al. (2010). In each cohort and for each clinical variable, we removed samples with extreme measurements, similarly as for the proteomic data quality control.

E.1.6 Overview of LOCUS

LOCUS is an efficient Bayesian approach for estimating QTL associations jointly from $p = 10^5 - 10^6$ genetic variants, typically single nucleotide polymorphisms (SNPs), and $q = 10^2 - 10^4$ expression outcomes, for $n = 10^2 - 10^4$ individuals (see Figure 7.1a and Ruffieux et al., 2017). It is based on a hierarchical sparse regression model that involves a collection of q high-dimensional regressions, each having all p SNPs as candidate predictors. These regressions are linked hierarchically via parameters controlling the propensity of SNPs to be associated with several outcomes (Figure 7.1b), which allows the leveraging of shared association patterns across all molecular variables, and enhances the estimation of weak *trans* and pleiotropic QTL effects. The model enforces sparsity on the regression coefficients, so LOCUS identifies just one or few markers per relevant locus, even in regions of high LD. Moreover, LOCUS estimates interpretable posterior probabilities of association for all SNP-outcome pairs (Figure 7.1c), from which Bayesian false discovery rate (FDR) are easily calculated (Newton et al., 2004).

Inference on high-dimensional Bayesian models is both computationally and statistically difficult. Previous joint QTL approaches (Jia and Xu, 2007; Richardson et al., 2010; Bottolo et al., 2011; Scott-Boyer et al., 2012) are based on sampling procedures, such as Markov Chain Monte Carlo (MCMC) algorithms, and require prohibitive computational times on data with more than few hundreds of SNPs or outcomes. LOCUS instead relies on deterministic inference: it implements a fast variational inference algorithm, which scales to the typical sizes of QTL problems. Ruffieux et al. (2017) extensively compared the performance of LOCUS with existing QTL methods, whether sampling-based or deterministic, univariate or multivariate. We recently augmented our algorithm with a simulated annealing procedure (Ueda and Nakano, 1998; Rose et al., 1990) to enhance exploration of multimodal parameter spaces such as induced by strong LD structures. LOCUS is tailored to genomic, proteomic, lipidomic and methylation QTL analyses; it can also be used for genome-wide association with several clinical endpoints. Details and extensive performance studies are in Ruffieux et al. (2017); see also Appendix E.2 for simulations based on the Ottawa pQTL data.

The applicability of a fully multivariate method to large molecular QTL data also hinges on the effective computational implementation of its algorithmic procedure. The annealed variational updates of

LOCUS are analytical and performed by batches of variables. The software is written in R with C++ subroutines; it is publicly available at https://github.com/hruffieux/locus.

E.1.7 Proteomic quantitative trait locus analyses

We performed proteomic quantitative trait locus (pQTL) analyses separately for each platform, i.e., a first analysis for the MS proteomic dataset, and a second for the SomaLogic proteomic dataset. Each analysis consisted of two stages: a discovery stage using the Ottawa cohort, and a replication stage based on the DiOGenes cohort.

For discovery, we used the multivariate Bayesian method LOCUS on both the MS and the SomaLogic datasets, with an annealing schedule of 50 geometrically-spaced temperatures, and set the initial temperature to 20; pilot experiments indicated that estimation was not sensitive to these choices. We used a convergence tolerance of 10^{-3} on the changes in the objective function as the stopping criterion. The algorithm can handle missing data in the response matrix, so no imputation was necessary for the MS proteomic data.

We adjusted all analyses for age, gender, and BMI at baseline. No important stratification was observed in the genotype data; the first ten principal components together explained little of the total variance ($< 4\%$), so we did not include them as covariates. We derived FDR values from the posterior probabilities of association obtained between each SNP and each protein, and reported pQTL associations using an FDR threshold of 5%. Both LOCUS runs completed within hours; convergence was reached after 2 hours (79 iterations) for the MS dataset, and after 10 hours and 20 minutes (72 iterations) for the SomaLogic dataset, on an Intel Xeon CPU, 2.60 GHz.

We performed a validation study of the discovered pQTLs in the DiOGenes cohort using the GEMMA method (Zhou and Stephens, 2012), with centered relatedness matrix (default). We then obtained adjusted p-values using Benjamini–Hochberg false discovery rates, and validated our hits using an FDR threshold of 5%.

E.1.8 pQTL annotation

We used the Ensembl database (GRCh37, release 94, Zerbino et al., 2018) to retrieve the list of genes within 2 Mb of each sentinel SNP (i.e., involved in the pQTL associations identified by LOCUS), and the SNAP database (Johnson et al., 2008) to retrieve the SNPs in LD ($r^2 > 0.8$), limiting the search to 500 Kb upstream and downstream of the sentinel SNP position. We called *cis pQTLs* all sentinel SNPs located within ± 1 Mb of the gene encoding for the controlled protein, and *trans pQTLs*, all other pQTLs.

We also evaluated the overlap between our pQTL associations and previously reported pQTL signals with the PhenoScanner database (Staley et al., 2016), using the default p-value threshold $p < 1 \times 10^{-5}$, and an LD proxy search ($r^2 > 0.8$).

E.1.9 Epigenomic annotation

We retrieved epigenomic annotations of 1,000 Genomes Project (release 20110521) from Pickrell (2014). The data covered 450 annotation features, each was binary-coded according to the presence or absence of overlap with the SNPs. The features corresponded to DNase-I hypersensitivity, chromatin state,

SNP consequences (coding, non-coding, 5'UTR, 3'UTR, etc), synonymous and nonsynonymous status and histone modification marks. We obtained distances to the closest transcription start site from the UCSC genome browser (Karolchik et al., 2003). Ninety-seven of our 104 validated sentinel SNPs had annotation data; to evaluate their functional enrichment, we resampled SNP sets of size 97 from our initial SNP panel, and, for each set, we computed the cumulated number of annotation. We did the same for the distances to transcription start site. We repeated this 10^5 times to derive empirical p-values.

E.1.10 Colocalization with known eQTLs and with GWAS risk loci

We evaluated the overlap of our pQTLs with the eQTL variants reported by the GTEx Consortium (2015, release 7) at q-value < 0.05. We used a one-sided Fisher exact test to assess the enrichment for the 104 validated sentinel pQTLs. We considered all 49 tissues listed by GTEx but eQTL SNPs for several tissues were counted only once. We made both general queries and queries asking whether a pQTL uncovered by LOCUS was also an eQTL for the gene coding for the controlled protein.

We retrieved known associations between the validated sentinel pQTLs and diseases or clinical traits, based on the GWAS catalog (v1.0 release e92, Welter et al., 2014), also using an LD proxy search ($r^2 > 0.8$).

E.1.11 Associations with clinical variables

We tested associations between the proteins under genetic control and clinical parameters separately in each cohort. For the DiOGenes data, we used linear mixed-effect models, adjusting for age, gender as fixed effects, and center as a random effect. For Ottawa data, we used linear models, adjusting for age and gender. Except when testing associations with anthropomorphic traits, all analyses were also adjusted for BMI. For the clinical variables available in the two cohorts (total cholesterol, HDL, LDL, fasting glucose, fasting insulin, HOMA-IR, triglycerides and VAI), we performed meta-analyses using the R package metafor (Viechtbauer, 2010). We used random-effect models to account for inter-study variability, which may in part result from geographical differences. We did not interpret the results if between-study heterogeneity estimates were high ($I^2 > 80\%$), and evaluated the directional consistency of the effects between Ottawa and DiOGenes. We adjusted for multiplicity using Benjamini–Hochberg correction across all tests, i.e., involving the 88 tested proteins and the two proteomic technologies, and reported associations using a 5% threshold.

E.1.12 Data availability

The MS proteomic data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (Deutsch et al., 2016) with the dataset identifiers PXD005216 for DiOGenes and PXD009350 for Ottawa (Username: reviewer63509@ebi.ac.uk; Password: YXuqutMH).

E.1.13 Code availability

All statistical analyses were performed using the R environment (version 3.3.2; R Core Team, 2018). LOCUS and ECHOSEQ are freely available under GPL-2 licenses from Github. Annotation queries used the Ensembl, GTEx, GWAS Catalog, JASPAR, PhenoScanner, SNP2TFBS, UCSC, UniProt websites.

E.1.14 URLs

ECHOSEQ: https://github.com/hruffieux/echoseq
LOCUS: https://github.com/hruffieux/locus
GEMMA: http://www.xzlab.org/software.htmlGTExhttps://GTExportal.org/home
Ensembl: http://grch37.ensembl.org/index.html
GTEx: https://gtexportal.org/home
GWAS Catalog: https://www.ebi.ac.uk/gwas
JASPAR: http://jaspar.genereg.net
PhenoScanner: http://www.phenoscanner.medschl.cam.ac.uk
SNP2TFBS: https://ccg.vital-it.ch/cgi-bin/snp2tfbs/snpviewer_form_parser.cgi
UCSC: https://genome.ucsc.edu
UniProt: https://www.uniprot.org

E.2 Statistical and computational performance of LOCUS

We evaluated the expected performance of LOCUS on our data by conducting two simulation studies. We compared its statistical power to detect pQTL associations with that of GEMMA, a univariate linear mixed model approach by Zhou and Stephens (2014). We used the R package `echoseq` to generate synthetic data that emulate real data conditions.

We ran the LOCUS and GEMMA on the SNPs of all $n = 376$ Ottawa subjects, and on simulated expression outcomes with residual dependence replicating that of the $q = 133$ mass-spectrometry proteomic expression levels. We first used the SNPs from chromosome one ($p = 20,900$), and generated associations between 20 SNPs and 25 proteins chosen randomly, leaving the remaining variables unassociated. Some proteins were under pleiotropic control; we drew the degree of pleiotropy of the 20 SNPs from a positively-skewed Beta distribution so only a few SNPs were hotspots, i.e., were associated with many proteins. We generated associations under an additive dose-effect scheme and drew the proportions of outcome variance explained by a given SNP from a Beta$(2,5)$ distribution to give more weight to smaller effect sizes. We then rescaled these proportions so that the variance of each protein attributable to genetic variation was below 35%. These choices led to an inverse relationship between minor allele frequencies and effect sizes, which is to be expected under natural selection. We generated 50 replicates, re-drawing the protein expression levels and effect sizes for each.

The ROC curves of Figure E.1a show a net gain in power for selections with LOCUS compared to GEMMA. The average standardized partial areas under the curve (pAUC) with 95% confidence intervals are 0.926 ± 0.005 for LOCUS and 0.840 ± 0.005 for GEMMA, using a false positive threshold of 25%.

In the second simulation, we re-assessed the performance of LOCUS for a grid of data generation scenarios. We considered a wide range of sparsity levels (numbers of proteins under genetic control) and effect sizes (proportions of outcome variance explained by the genetic variants). Given the large number of configurations (247), and in order to limit the computational burden, we used the first $p = 2,000$ SNPs, and ran LOCUS and GEMMA on 20 replicates for each configuration. Figure E.1b demonstrates that the superiority of LOCUS over GEMMA generalizes to all data generation scenarios, as the average standardized pAUC is everywhere greater for LOCUS than for GEMMA.

The performance of LOCUS is largely attributable to the multivariate modelling of all the SNPs and proteomic outcomes, which allows sharing of information across and within loci, as well as across

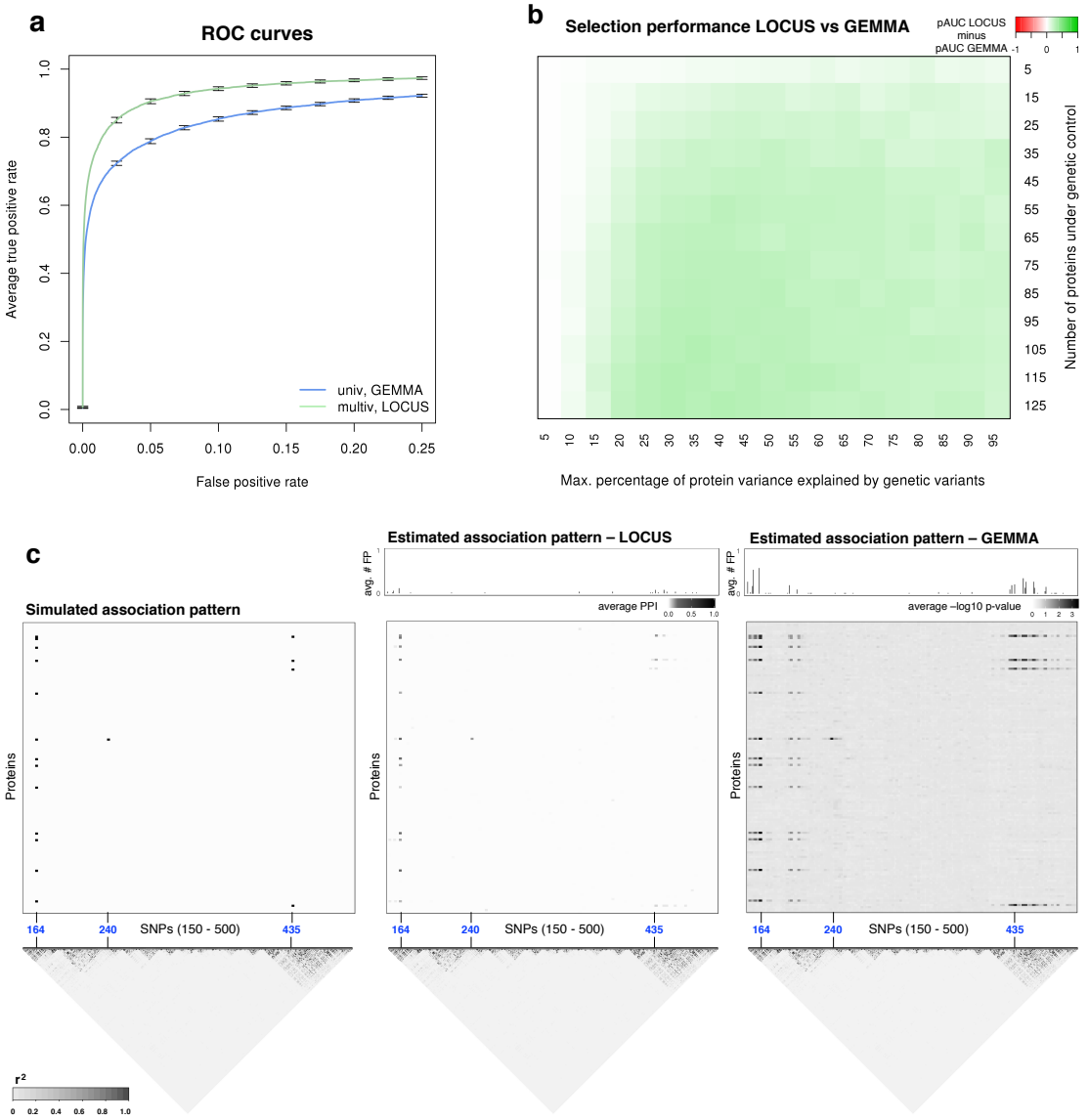

Figure E.1 – Selection performances of LOCUS and GEMMA. (a) Truncated average ROC curves with 95% confidence intervals, obtained from 50 replications, for identification of SNP-trait associations. (b) Difference of average standardized pAUC of LOCUS and GEMMA for a grid of effect sizes (x-axis) and signal sparsity (y-axis), using 20 replications for each scenario. (c) Simulated pattern, and patterns recovered by LOCUS and GEMMA, averaged over the 50 replications. The plots display a window of 350 SNPs (x-axis) containing the first three SNPs having simulated associations (blue labels), along with their linkage disequilibrium (LD) pattern. The top panels of the middle and right plots display the average number of false positives when selecting SNPs at FDR 25%. GEMMA indicates many false positive associations in regions of high LD, while LOCUS usually pinpoints the relevant SNPs.

different proteins under common genetic regulation. By design, univariate screening approaches do not exploit association patterns common to multiple outcomes or markers; they analyze the outcomes one by one, and do not account for LD structures, thereby increasing false discoveries at loci with strong LD structures (Figure E.1c). At a given FDR, such spurious associations hamper the detection of weak but genuine signals. Owing to its simulated annealing procedure that improves exploration at loci

with strong LD, LOCUS better discriminates truly associated SNPs from their correlated neighbours (Figure E.1c).

The runtime of LOCUS was of the same order as that of the univariate GEMMA analysis. On average, for one replicate, LOCUS took 5 minutes and 26 seconds to complete, while GEMMA took 7 minutes and 4 seconds, running in parallel on four cores of an Intel Xeon CPU, 2.60 GHz.

Bibliography

1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491:56, 2012.

M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: A System for Large-Scale Machine Learning. In K. Keeton and T. Roscoe, editors, *Operating Systems Design and Implementation*, pages 265–283, Savannah, United States, 2016. USENIX.

J. H. Albert and S. Chib. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679, 1993.

K. G. M. M. Alberti, R. H. Eckel, S. M. Grundy, P. Z. Zimmet, J. I. Cleeman, K. A. Donato, J. Fruchart, W. P. T. James, C. M. Loria, and S. C. Smith. Harmonizing the metabolic syndrome: a joint interim statement of the International Diabetes Federation Task Force on Epidemiology and Prevention; National Heart, Lung, and Blood Institute; American Heart Association; World Heart Federation; International Atherosclerosis Society; and International Association for the Study of Obesity. *Circulation*, 120: 1640–1645, 2009.

S.-I. Amari. *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics. Springer-Verlag, New York, United States, 1985.

S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10:251–276, 1998.

M. C. Amato, C. Giordano, M. Galia, A. Criscimanna, S. Vitabile, M. Midiri, A. Galluzzo, and AlkaMeSy Study Group. Visceral adiposity index (VAI): a reliable indicator of visceral fat function associated with cardiometabolic risk. *Diabetes Care*, 2010.

J. R. Anderson and C. Peterson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

C. Andrieu and E. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *The Annals of Applied Probability*, 16:1462–1505, 2006.

E. Angelino, M. J. Johnson, and R. P. Adams. Patterns of scalable Bayesian inference. *Foundations and Trends in Machine Learning*, 9:119–247, 2016.

A W Armstrong, C T Harskamp, and E J. Armstrong. The association between psoriasis and obesity: a systematic review and meta-analysis of observational studies. *Nutrition & Diabetes*, 2:e54, 2012.

D. J. Balding, M. Bishop, and C. Cannings. *Handbook of Statistical Genetics*. John Wiley & Sons, Chichester, United Kingdom, 2008.

L. Balles, J. Romero, and P. Hennig. Coupling adaptive batch sizes with learning rates. *arXiv preprint arXiv:1612.05086*, 2016.

M. Banerjee and M. Saxena. Interleukin-1 (IL-1) family of cytokines: role in type 2 diabetes. *International Journal of Clinical Chemistry*, 413:1163–1170, 2012.

D. Barber and W. Wiegerinck. Tractable variational structures for approximating graphical models. In M. S. Kearns, M. J. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 183–189, Cambridge, Massachusetts, 1999. MIT Press.

M. M. Barbieri and J. O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32:870–897, 2004.

R. Bardenet, A. Doucet, and C. C. Holmes. Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, pages 405–413, Bejing, China, 2014. Proceedings of Machine Learning Research.

R. Bardenet, A. Doucet, and C. C. Holmes. On Markov chain Monte Carlo methods for tall data. *The Journal of Machine Learning Research*, 18:1515–1557, 2017.

J. C Barrett. Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harbor Protocols*, 2009:pdb–ip71, 2009.

L. Barreyro, B. Will, B. Bartholdy, L. Zhou, T. I. Todorova, R. F. Stanley, S. Ben-Neriah, C. Montagna, S. Parekh, and A. Pellagatti. Overexpression of IL-1 receptor accessory protein in stem and progenitor cells and outcome correlation in AML and MDS. *Blood*, 120:1290–1298, 2012.

A. Barron, M. J Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27:536–561, 1999.

N. M. Bass, E. M. Brunt, J. M. Clark, A. M. Diehl, J. H. Hoofnagle, D. E. Kleiner, K. V. Kowdley, A. J. McCullough, B. A. Neuschwander-Tetri, and P. R. Robuck. Clinical, laboratory and histological associations in adults with nonalcoholic fatty liver disease. *Hepatology (Baltimore, Md.)*, 52:913–924, 2010.

M. J. Bayarri, J. O. Berger, A. Forte, and G. García-Donato. Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40:1550–1577, 2012.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57:289–300, 1995.

Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29:1165–1188, 2001.

J. Berger. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics*, 8:716–761, 1980.

T. Berisa and J. K. Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32:283, 2016.

M. Betancourt and M. Girolami. Hamiltonian Monte Carlo for hierarchical models. *Current Trends in Bayesian Methodology with Applications*, 79:30, 2015.

A. Bhadra and B. K. Mallick. Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69:447–457, 2013.

A. Bhadra, J. Datta, N. G. Polson, and B. Willard. Default Bayesian analysis with global-local shrinkage priors. *Biometrika*, 103:955–969, 2016.

A. Bhadra, J. Datta, N. G. Polson, and B. Willard. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12:1105–1131, 2017a.

A. Bhadra, J. Datta, N. G. Polson, and B. T. Willard. Lasso meets horseshoe. *arXiv preprint arXiv:1706.10179*, 2017b.

A. Bhattacharya and D. B. Dunson. Nonparametric Bayesian density estimation on manifolds with applications to planar shapes. *Biometrika*, 97:851–865, 2010.

A. Bhattacharya and D. B. Dunson. Sparse Bayesian infinite factor models. *Biometrika*, 98:291–306, 2011.

A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson. Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110:1479–1490, 2015.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, United States, 2006.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112:859–877, 2017.

M. J. Bonder, R. Luijk, D. V. Zhernakova, M. Moed, P. Deelen, M. Vermaat, M. van Iterson, F. van Dijk, M. van Galen, and J. Bot. Disease variants alter transcription factor levels and methylation of their binding sites. *Nature Genetics*, 49:131–138, 2017.

D. Bontemps. Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. *The Annals of Statistics*, 39:2557–2584, 2011.

L. Bottolo and S. Richardson. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5:583–618, 2010.

L. Bottolo, E. Petretto, S. Blankenberg, F. Cambien, S. A. Cook, L. Tiret, and S. Richardson. Bayesian detection of expression quantitative trait loci hot spots. *Genetics*, 189:1449–1459, 2011.

E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, and A. Bairoch. Uniprotkb/Swiss-Prot. In *Plant Bioinformatics*, pages 89–112. Springer, 2007.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, United States, 2004.

K. Bozaoglu, C. Attard, H. Kulkarni, N. Cummings, V. P. Diego, M. A. Carless, K. A. Shields, M. P. Johnson, S. Kowlessur, and T. D. Dyer. Plasma Levels of Soluble Interleukin 1 Receptor Accessory Protein Are Reduced in Obesity. *The Journal of Clinical Endocrinology and Metabolism*, 99:3435–3443, 2014.

L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

A. T. Brunger and L. M. Rice. Crystallographic refinement by simulated annealing: methods and applications. *Methods in Enzymology*, 277:243–268, 1997.

B. Brynedal, J. Choi, T. Raj, R. Bjornson, B. E. Stranger, B. M. Neale, B. F. Voight, and C. Cotsapas. Large-scale trans-eQTLs affect hundreds of transcripts and mediate patterns of transcriptional co-regulation. *The American Journal of Human Genetics*, 100:581–591, 2017.

P. Bühlmann and S. van de Geer. *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer, Berlin, Germany, 2011.

R. H. Byrd, G. M. Chin, J. Nocedal, and Y. Wu. Sample size selection in optimization methods for machine learning. *Mathematical Programming*, 134:127–155, 2012.

M. Böni-Schnetzler, S. P. Häuselmann, E. Dalmas, D. T. Meier, C. Thienel, S. Traub, F. Schulze, L. Steiger, E. Dror, and P. Martin. Beta Cell-Specific Deletion of the IL-1 Receptor Antagonist Impairs beta Cell Proliferation and Insulin Secretion. *Cell Reports*, 22:1774–1786, 2018.

G. C. Calafiore and L. El Ghaoui. *Optimization Models*. Cambridge University Press, Cambridge, United Kingdom, 2014.

J. Carayol, C. Chabert, A. Di Cara, C. Armenise, G. Lefebvre, D. Langin, N. Viguerie, S. Metairon, W. H. M. Saris, and A. Astrup. Protein quantitative trait locus study in obesity during weight-loss identifies a leptin regulator. *Nature Communications*, 8:2084, 2017.

P. Carbonetto and M. Stephens. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis*, 7:73–108, 2012.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 73–80, Clearwater Beach, United States, 2009. Proceedings of Machine Learning Research.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480, 2010.

G. Casella. Empirical Bayes Gibbs sampling. *Biostatistics*, 2:485–500, 2001.

I. Castillo and A. W. van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40:2069–2101, 2012.

L. Chen, B. Ge, F. P. Casale, L. Vasquez, T. Kwan, D. Garrido-Martín, S. Watt, Y. Yan, K. Kundu, and S. Ecker. Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, 167:1398–1414.e24, 2016.

W. Chen, B. R. Larrabee, I. G. Ovsyannikova, R. B. Kennedy, I. H. Haralambieva, G. A. Poland, and D. J. Schaid. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. *Genetics*, page 115, 2015.

H. Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24: 17–36, 1996.

J. H. Chung. The role of DNA-PK in aging and energy metabolism. *The FEBS Journal*, 285:1959–1972, 2018.

A. Cichocki and S.-I. Amari. Families of alpha-beta-and gamma-divergences: flexible and robust measures of similarities. *Entropy*, 12:1532–1568, 2010.

M. Clyde. *BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging*, 2016. R package version 1.0.9.

M. Clyde, H. Desimone, and G. Parmigiani. Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91:1197–1208, 1996.

M. A. Clyde. Bayesian model averaging and model search strategies (with discussion). In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics*, volume 6, pages 157–185. Oxford University Press, New York, United States, 1999.

P. M. Coan, M. Barrier, N. Alfazema, R. N. Carter, S. Marion de Procé, X. C. Dopico, A. Garcia Diaz, A. Thomson, L. H. Jackson-Jones, and B. Moyon. Complement Factor B Is a Determinant of Both Metabolic and Cardiovascular Features of Metabolic Syndrome. *Hypertension*, 2017.

O. Cominetti, A. Nunez Galindo, J. Corthesy, S. Oller Moreno, I. Irincheeva, A. Valsesia, A. Astrup, W. H. M. Saris, J. Hager, and M. Kussmann. Proteomic biomarker discovery in 1000 human plasma samples with mass spectrometry. *Journal of Proteome Research*, 15:389–399, 2015.

International HapMap Consortium, D. Altshuler, and P. Donnelly. A haplotype map of the human genome. *Nature*, 437:1299, 2005.

D. R. Cox, E. Spjøtvoll, S. Johansen, W. R. van Zwet, J. F. Bithell, O. Barndorff-Nielsen, and M. Keuls. The role of significance tests (with discussion). *Scandinavian Journal of Statistics*, pages 49–70, 1977.

J. Datta and J. K. Ghosh. Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, 8:111–132, 2013.

A. P. Dawid. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68:265–274, 1981.

S. De, A. Yadav, D. Jacobs, and T. Goldstein. Automated inference with adaptive batches. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1504–1513, Fort Lauderdale, United States, 2017. Proceedings of Machine Learning Research.

N. de Freitas, P. Højen-Sørensen, M. I. Jordan, and S. Russell. Variational MCMC. In J. Breese and D. Koller, editors, *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence*, pages 120–127, San Francisco, United States, 2001. Morgan Kaufmann Publishers Inc.

J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51:107–113, 2008.

O. Delaneau, C. Coulonges, and J. Zagury. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics*, 9:540, 2008.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society. Series B (methodological)*, 39: 1–38, 1977.

R. M. Dent, R. M. Penwarden, N. Harris, and S. B. Hotz. Development and evaluation of patient-centered software for a weight-management clinic. *Obesity Research*, 10:651–656, 2002.

E. W. Deutsch, A. Csordas, Z. Sun, A. Jarnuczak, Y. Perez-Riverol, T. Ternent, D. S. Campbell, M. Bernal-Llinares, S. Okuda, and S. Kawano. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Research*, page gkw936, 2016.

P. Donnelly. Progress and challenges in genome-wide association studies in humans. *Nature*, 456:728, 2008.

D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 41–81, 1992.

J. L. Doob. Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, pages 23–27, Paris, 1949. Colloques Internationaux du Centre National de la Recherche Scientifique.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

B. Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102:93–103, 2007.

B. Efron. Microarrays, empirical Bayes and the two-groups model. *Statistical Science*, 23:1–22, 2008.

B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press, Cambridge, United Kingdom, 2010.

B. Efron and R. J. Tibshirani. Empirical Bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23:70–86, 2002.

B. Efron, R. J. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96:1151–1160, 2001.

B. E. Engelhardt and R. P. Adams. Bayesian structured sparsity from Gaussian fields. *arXiv preprint arXiv:1407.2235*, 2014.

L. Er, S. Wu, L. Hsu, M. Teng, Y. Sun, and Y. Ko. Pleiotropic Associations of RARRES2 Gene Variants and Circulating Chemerin Levels: Potential Roles of Chemerin Involved in the Metabolic and Inflammation-Related Diseases. *Mediators of Inflammation*, 2018, 2018.

B. P. Fairfax, S. Makino, J. Radhakrishnan, K. Plant, S. Leslie, A. Dilthey, P. Ellis, C. Langford, F. O. Vannberg, and J. C. Knight. Genetics of gene expression in primary immune cells identifies cell type–specific master regulators and roles of HLA alleles. *Nature Genetics*, 44:502, 2012.

B. P. Fairfax, P. Humburg, S. Makino, V. Naranbhai, D. Wong, E. Lau, L. Jostins, K. Plant, R. Andrews, and C. McGee. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science*, 343:1246949, 2014.

J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70:849–911, 2008.

M. Favennec, B. Hennart, R. Caiazzo, A. Leloire, L. Yengo, M. Verbanck, A. Arredouani, M. Marre, M. Pigeyre, and A. Bessede. The kynurenine pathway is activated in human obesity and shifted toward kynurenine monooxygenase activation. *Obesity*, 23:2066–2074, 2015.

T. Flutre, X. Wen, J. Pritchard, and M. Stephens. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genetics*, 9:e1003486, 2013.

L. Folkersen, E. Fauman, M. Sabater-Lleal, R. J. Strawbridge, M. Frånberg, B. Sennblad, D. Baldassarre, F. Veglia, S. E. Humphries, and R. Rauramaa. Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genetics*, 13:e1006706, 2017.

A. Franke, D. P. B. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, and R. Roberts. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nature Genetics*, 42:1118–1125, 2010.

W. T. Friedewald, R. I Levy, and D. S. Fredrickson. Estimation of the concentration of low-density lipoprotein cholesterol in plasma without use of the preparative ultracentrifuge. *Clinical Chemistry*, 18:499–502, 1972.

J. H. Friedman, T. J. Hastie, and R. J. Tibshirani. *glmnet: Lasso and elastic-net regularized generalized linear models*, 2009. R package version 2.0.2.

J. H. Friedman, T. J. Hastie, and R. J. Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

D. J. Gaffney, J.-B. Veyrieras, J. F. Degner, R. Pique-Regi, A. A. Pai, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard. Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology*, 13:R7, 2012.

C. B. Garrison, K. J. Lastwika, Y. Zhang, C. I. Li, and P. D. Lampe. Proteomic analysis, immune dysregulation, and pathway interconnections with obesity. *Journal of Proteome Research*, 16:274–287, 2016.

A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1:515–534, 2006.

A. Gelman, A. Jakulin, M. G. Pittau, and Y. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2:1360–1383, 2008.

A. Gelman, J. Hill, and M. Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5:189–211, 2012.

A. Gelman, H. S. Stern, J. B. Carlin, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, United States, 3rd edition, 2013.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

C. R. Genovese and L. Wasserman. Rates of convergence for the Gaussian mixture sieve. *The Annals of Statistics*, 28:1105–1127, 2000.

E. I. George and D. P. Foster. Calibration and empirical Bayes variable selection. *Biometrika*, 87:731–747, 2000.

E. I. George and R. E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889, 1993.

E. I. George and R. E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7: 339–373, 1997.

C. J. Geyer and E. A. Thompson. Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90:909–920, 1995.

B. Ghorbani, H. Javadi, and A. Montanari. An instability in variational inference for topic models. *arXiv preprint arXiv:1802.00568*, 2018.

S. Ghosal. Asymptotic normality of posterior distributions in high-dimensional linear models. *Bernoulli*, 5:315–331, 1999.

S. Ghosal and A. W. van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35:192–223, 2007.

S. Ghosal and A. W. van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, Cambridge, United Kingdom, 2017.

S. Ghosal, J. K. Ghosh, and R. V. Ramamoorthi. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27:143–158, 1999.

S. Ghosal, J. K. Ghosh, and A. W. van der Vaart. Convergence rates of posterior distributions. *The Annals of Statistics*, 28:500–531, 2000.

J. Ghosh and A. E. Ghattas. Bayesian variable selection under collinearity. *The American Statistician*, 69: 165–173, 2015.

J. Ghosh, Y. Li, and R. Mitra. On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, 13:359–383, 2017.

P. Ghosh and A. Chakrabarti. Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems. *Bayesian Analysis*, 12:1133–1161, 2017.

P. Ghosh, X. Tang, M. Ghosh, and A. Chakrabarti. Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Analysis*, 11:753–796, 2016.

Y. Gilad, S. A. Rifkin, and J. K. Pritchard. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics*, 24:408–415, 2008.

J. Gill. Is partial-dimension convergence a problem for inferences from MCMC algorithms? *Political Analysis*, 16:153–178, 2008.

M. E. Goddard, K. E. Kemper, I. M. MacLeod, A. J. Chamberlain, and B. J. Hayes. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. *Proceedings of the Royal Society of London B: Biological Sciences*, 283, 2016.

E. L. Goode. Linkage disequilibrium. In M. Schwab, editor, *Encyclopedia of Cancer*, pages 2043–2048. Springer, Berlin, Heidelberg, Germany, 2011.

I. I. Gottesman and T. D. Gould. The endophenotype concept in psychiatry: etymology and strategic intentions. *American Journal of Psychiatry*, 160:636–645, 2003.

A. Goustin and A. B. Abou-Samra. The "thrifty" gene encoding Ahsg/Fetuin-A meets the insulin receptor: Insights into the mechanism of insulin resistance. *Cellular Signalling*, 23:980–990, 2011.

I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products*. Academic Press, Burlington, United States, 5th edition, 1994.

R. Gramacy, R. Samworth, and R. King. Importance tempering. *Statistics and Computing*, 20:1–7, 2010.

S. Greenland and J. M. Robins. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology*, pages 244–251, 1991.

J. Griffin and P. Brown. Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, 12: 135–159, 2017.

GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*, 348:648–660, 2015.

Y. Guan and M. Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *Annals of Applied Statistics*, 5:1780–1815, 2011.

G. Guennebaud and B. Jacob. Eigen v3. http://eigen.tuxfamily.org, 2010.

R. Guhaniyogi, S. Qamar, and D. B. Dunson. Bayesian conditional density filtering. *Journal of Computational and Graphical Statistics*, 2018.

F. Guo, X. Wang, K. Fan, T. Broderick, and D. B. Dunson. Boosting variational inference. *arXiv preprint arXiv:1611.05559*, 2016.

H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242, 2001.

E. Halperin and D. A. Stephan. SNP imputation in association studies. *Nature Biotechnology*, 27:349, 2009.

J. L. Harden, S. M. Lewis, S. R. Lish, M. Suárez-Fariñas, D. Gareau, T. Lentini, L. M. Johnson-Huang, J. G. Krueger, and M. A. Lowes. The tryptophan metabolism enzyme L-kynureninase is a novel inflammatory factor in psoriasis and other inflammatory diseases. *The Journal of Allergy and Clinical Immunology*, 137:1830–1840, 2016.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, pages 97–109, 1970.

A. L. Hess, J. Carayol, T. Blædel, J. Hager, A. Di Cara, A. Astrup, W. H. M. Saris, L. H. Larsen, and A. Valsesia. Analysis of circulating angiopoietin-like protein 3 and genetic variants in lipid metabolism and liver health: the DiOGenes study. *Genes & Nutrition*, 13:7, 2018.

N. J. Higham. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22:329–343, 2002.

G. E. Hinton and D. van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In L. Pitt, editor, *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pages 5–13, New York, United States, 1993. ACM.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14:1303–1347, 2013.

C. C. Holmes, D. G. T. Denison, and B. K. Mallick. Accounting for model uncertainty in seemingly unrelated regressions. *Journal of Computational and Graphical Statistics*, 11:533–551, 2002.

A. Honkela, M. Tornio, T. Raiko, and J. Karhunen. Natural conjugate gradient in variational inference. In M. Ishikawa, Kenji Doya, H. Miyamoto, and T. Yamakawa, editors, *Neural Information Processing: 14th International Conference, ICONIP 2007, Kitakyushu, Japan, November 13-16, 2007, Revised Selected Papers, Part II*, pages 305–314. Springer, Berlin, Germany, 2008.

B. N. Howie, P. Donnelly, and J. Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, 5:e1000529, 2009.

J. M. Howson, W. Zhao, D. R. Barnes, W. Ho, R. Young, D. S. Paul, L. L. Waite, D. F. Freitag, E. B. Fauman, and E. L. Salfati. Fifteen new risk loci for coronary artery disease highlight arterial wall-specific mechanisms. *Nature Genetics*, 49:1113–1119, 2017.

X. Huang, J. Wang, and F. Liang. A variational algorithm for Bayesian variable selection. *arXiv preprint arXiv:1602.07640*, 2016.

J. Ingraham and D. Marks. Variational inference for sparse and undirected models. *arXiv preprint arXiv:1602.03807*, 2016.

International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437:1299, 2005.

H. Ishwaran and J S. Rao. Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association*, 98:438–455, 2003.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics*, 33:730–773, 2005.

H. Ishwaran and J. S. Rao. Consistency of spike and slab regression. *Statistics & Probability Letters*, 81: 1920–1928, 2011.

S. K. Iyengar, J. R. Sedor, B. I. Freedman, W. L. Kao, M. Kretzler, B. J. Keller, H. E. Abboud, S. G. Adler, L. G. Best, and D. W. Bowden. Genome-wide association and trans-ethnic meta-analysis for advanced diabetic kidney disease: family investigation of nephropathy and diabetes (FIND). *PLoS Genetics*, 11: e1005352, 2015.

T. Jaakkola, L. K. Saul, and M. I. Jordan. Fast learning by bounding likelihoods in sigmoid type belief networks. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 528–534, Cambridge, Massachusetts, 1996. MIT Press.

T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37, 2000.

K. R. Jacobs, G. Castellano-González, G. J. Guillemin, and D. B. Lovejoy. Major Developments in the Design of Inhibitors along the Kynurenine Pathway. *Current Medicinal Chemistry*, 24:2471–2495, 2017.

W. James and C. M. Stein. Estimation with quadratic loss. In J. Neyman, editor, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability 1*, pages 361–379, Berkeley, United States, 1961. University of California Press.

H. Jeffreys. Theory of probability, 1961.

Z. Jia and S. Xu. Mapping quantitative trait loci for expression abundance. *Genetics*, 176:611–623, 2007.

B. Jiang and J. S. Liu. Bayesian partition models for identifying expression quantitative trait loci. *Journal of the American Statistical Association*, 110:1350–1361, 2015.

J. E. Johndrow, J. C. Mattingly, S. Mukherjee, and D. Dunson. Approximations of Markov Chains and high-dimensional Bayesian inference. *arXiv preprint*, 2015.

A. D. Johnson, R. E. Handsaker, S. L. Pulit, M. M. Nizzari, C. J. O'Donnell, and P. W. De Bakker. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24: 2938–2939, 2008.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 315–323, Cambridge, Massachusetts, 2013. Curran Associates, Inc.

V. E. Johnson and D. Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:143–170, 2010.

V. E. Johnson and D. Rossell. Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107:649–660, 2012.

I. M. Johnstone. On minimax estimation of a sparse normal mean vector. *The Annals of Statistics*, pages 271–289, 1994.

I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32:1594–1649, 2004.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

K. J. Karczewski, J. T. Dudley, K. R. Kukurba, R. Chen, A. J. Butte, S. B. Montgomery, and M. Snyder. Systematic functional regulatory assessment of disease-associated variants. *Proceedings of the National Academy of Sciences*, 110:9607–9612, 2013.

D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, and D. J. Thomas. The UCSC genome browser database. *Nucleic Acids Research*, 31:51–54, 2003.

K. Katahira, K. Watanabe, and M. Okada. Deterministic annealing variant of variational Bayes method. In K. Hukushima, Y. Kabashima, H. Nishimori, and T. Tanaka, editors, *Journal of Physics: Conference Series*, page 012015, Kyoto, Japan, 2008. IOP Publishing.

A. Khan, O. Fornes, A. Stigliani, M. Gheorghe, J. A. Castro-Mondragon, R. van der Lee, A. Bessy, J. Chèneby, S. R. Kulkarni, and G. Tan. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research*, 46:D260–D266, 2018. ISSN 1362-4962.

G. Kichaev, W.-Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft, and B. Pasaniuc. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genetics*, 10:e1004722, 2014.

S. Kim, J. Becker, M. Bechheim, V. Kaiser, M. Noursadeghi, N. Fricker, E. Beier, S. Klaschik, P. Boor, and T. Hess. Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes. *Nature Communications*, 5:5236, 2014.

D. P. Kingma and J. Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220: 671–680, 1983.

H. Kitano. Systems biology: a brief overview. *Science*, 295:1662–1664, 2002.

A. Korattikara, Y. Chen, and M. Welling. Austerity in MCMC land: cutting the Metropolis-Hastings budget. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, pages 181–189, Bejing, China, 2014. Proceedings of Machine Learning Research.

S. Kraemer, J. D. Vaught, C. Bock, L. Gold, E. Katilius, T. R. Keeney, N. Kim, N. A Saccomano, S. K. Wilcox, and D. Zichi. From SOMAmer-based biomarker discovery to diagnostic and clinical applications: a SOMAmer-based, streamlined multiplex proteomic assay. *PloS one*, 6:e26332, 2011.

D. H. Krantz. The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, 94:1372–1381, 1999.

M. Kulis and M. Esteller. DNA methylation and cancer. In Z. Herceg and T. Ushijima, editors, *Advances in Genetics*, volume 70, pages 27–56. Elsevier, San Diego, United States, 2010.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22: 79–86, 1951.

S. Kumar, G. Ambrosini, and P. Bucher. SNP2tfbs - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Research*, 45, 2017.

D. Kwon, M. T. Landi, M. Vannucci, H. J Issaq, D. Prieto, and R. M. Pfeiffer. An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. *Computational Statistics & Data Analysis*, 55:2807–2818, 2011.

M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5:369–411, 2010.

J. C. Lambert, C. A. Ibrahim-Verbaas, D. Harold, A. C. Naj, R. Sims, C. Bellenguez, A. L. DeStafano, J. C. Bis, G. W. Beecham, and B. Grenier-Boley. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45:1452–1458, 2013.

E. S. Lander and N. J. Schork. Genetic dissection of complex traits. *Science*, 265:2037–2048, 1994.

P. S. Laplace. Memoir on the probability of the causes of events. *Statistical Science*, 1:364–378, 1986.

T. M. Larsen, S. Dalskov, M. van Baak, S. A. Jebb, A. Kafatos, A. F. H. Pfeiffer, J. A. Martinez, T. Handjieva-Darlenska, M. Kunesova, and C. Holst. The Diet, Obesity and Genes (Diogenes) Dietary Study in eight European countries—a comprehensive design for long-term intervention. *Obesity Reviews*, 11: 76–91, 2010.

M. N. Lee, C. Ye, A.-C. Villani, T. Raj, W. Li, T. M. Eisenhaure, S. H. Imboywa, P. I. Chipendo, F. A. Ran, and K. Slowikowski. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, 343:1246980, 2014.

J. T. Leek and J. D. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3:e161, 2007.

R. N. Lemaitre, I. B. King, E. K. Kabagambe, J. H. Y. Wu, B. McKnight, A. Manichaikul, W. Guan, Q. Sun, D. I. Chasman, and M. Foy. Genetic loci associated with circulating levels of very long-chain saturated fatty acids. *Journal of Lipid Research*, 56:176–184, 2015.

A. Lewin, H. Saadi, J. E. Peters, A. Moreno-Moral, J. C. Lee, K. G. C. Smith, E. Petretto, L. Bottolo, and S. Richardson. MT-HESS: an efficient Bayesian approach for simultaneous association detection in OMICS datasets, with application to eQTL mapping in multiple tissues. *Bioinformatics*, 32:523–532, 2015.

B. Lewin, J. Krebs, S. T. Kilpatrick, and E. S. Goldstein. *Lewin's Genes*, volume 10. Jones & Bartlett, Sudbury, United States, 2011.

F. Li and N. R. Zhang. Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association*, 105:1202–1214, 2010.

H. Li and D. Pati. Variable selection using shrinkage priors. *Computational Statistics & Data Analysis*, 107:107–119, 2017.

Y. Li and M. Kellis. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Research*, 44:e144–e144, 2016.

Y. Li and R. E. Turner. Rényi divergence variational inference. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 1073–1081, Cambridge, Massachusetts, 2016. Curran Associates, Inc.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423, 2008.

F. Liang, C. Liu, and R. Carroll. *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. John Wiley & Sons, New York, United States, 2011.

J. C. Lin, W. Ho, A. Gurney, and A. Rosenthal. The netrin-G1 ligand NGL-1 promotes the outgrowth of thalamocortical axons. *Nature Neuroscience*, 6:1270–1276, 2003.

J. Z. Liu, S. van Sommeren, H. Huang, S. C. Ng, R. Alberts, A. Takahashi, S. Ripke, J. C. Lee, L. Jostins, and T. Shah. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, 47:979–986, 2015.

R. J. F. Loos and G. S. H. Yeo. The bigger picture of FTO—the first GWAS-identified obesity gene. *Nature Reviews Endocrinology*, 10:51, 2014.

E. López-Villar, G. A. Martos-Moreno, J. A. Chowen, S. Okada, J. J. Kopchick, and J. Argente. A proteomic approach to obesity and type 2 diabetes. *Journal of Cellular and Molecular Medicine*, 19:1455–1470, 2015.

J. Lucas, C. Carvalho, Q. Wang, A. Bild, J. R. Nevins, and M. West. Sparse statistical modelling in gene expression genomics. In *Bayesian Inference for Gene Expression and Proteomics*, volume 1, pages 1–25. Cambridge University Press, New York, United States, 2006.

A. J. Lusis. A thematic review series: systems biology approaches to metabolic and cardiovascular disorders. *Journal of Lipid Research*, 47:1887–1890, 2006.

T. F. C. Mackay, E. A. Stone, and J. F. Ayroles. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10:565, 2009.

D. Maclaurin and R. P. Adams. Firefly Monte Carlo: exact MCMC with subsets of data. In N. Zhang and J. Tian, editors, *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 543–552, Arlington, United States, 2014. AUAI Press.

S. Mandt, J. McInerney, F. Abrol, R. Ranganath, and D. Blei. Variational tempering. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 704–712, Cadiz, Spain, 2016. Proceedings of Machine Learning Research.

J. Marchini and B. Howie. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11:499, 2010.

E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters*, 19: 451, 1992.

Y. Maruyama and E. I. George. Fully Bayes factors with a generalized g-prior. *The Annals of Statistics*, 39:2740–2765, 2011.

H. Matsunaga, M. Iwashita, T. Shinjo, A. Yamashita, M. Tsuruta, S. Nagasaka, A. Taniguchi, M. Fukushima, N. Watanabe, and F. Nishimura. Adipose tissue complement factor B promotes adipocyte maturation. *Biochemical and Biophysical Research Communications*, 495:740–748, 2018.

M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, and J. Brody. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, page 1222794, 2012.

R. Mcdonald, M. Mohri, N. Silberman, D. Walker, and G. S. Mann. Efficient large-scale distributed training of conditional maximum entropy models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1231–1239, Cambridge, Massachusetts, 2009. Curran Associates, Inc.

J. B. Meigs, F. B. Hu, N. Rifai, and J. E. Manson. Biomarkers of endothelial dysfunction and risk of type 2 diabetes mellitus. *JAMA*, 291:1978–1986, 2004.

N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44:335–341, 1949.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092, 1953.

M. Mézard, G. Parisi, and M. Virasoro. *Spin Glass Theory and Beyond: an Introduction to the Replica Method and its Applications*. World Scientific Publishing Company, Singapore, Singapore, 1987.

J. P. Mills. Table of the ratio: area to bounding ordinate, for any portion of normal curve. *Biometrika*, pages 395–400, 1926.

T. Minka. Divergence measures and message passing. Technical report, Microsoft Research, 2005.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In J. Breese and D. Koller, editors, *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 362–369, San Francisco, United States, 2001. Morgan Kaufmann Publishers Inc.

S. Minsker, S. Srivastava, L. Lin, and D. B. Dunson. Robust and scalable Bayes via a median of subset posterior measures. *The Journal of Machine Learning Research*, 18:4488–4527, 2017.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032, 1988.

O. Monestier and V. Blanquet. WFIKKN1 and WFIKKN2: "Companion" proteins regulating TGFB activity. *Cytokine & Growth Factor Reviews*, 32:75–84, 2016.

S. Monni and M. G. Tadesse. A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Analysis*, 4:413–436, 2009.

I. E. Monroy-Muñoz, J. Angeles-Martinez, R. Posadas-Sánchez, T. Villarreal-Molina, E. Alvarez-León, C. Flores-Dominguez, G. Cardoso-Saldaña, A. Medina-Urrutia, J. G. Juárez-Rojas, and C. Posadas-Romero. PLA2g2a polymorphisms are associated with metabolic syndrome and type 2 diabetes mellitus. Results from the genetics of atherosclerotic disease Mexican study. *Immunobiology*, 222: 967–972, 2017.

J. M. Moreno-Navarrete, R. Martínez-Barricarte, V. Catalán, M. Sabater, J. Gómez-Ambrosi, F. J. Ortega, W. Ricart, M. Blüher, G. Frühbeck, and S. Rodríguez de Cordoba. Complement factor H is expressed in adipose tissue in association with insulin resistance. *Diabetes*, 59:200–209, 2010.

R. D. Morey and J. N. Rouder. *BayesFactor: computation of Bayes factors for common designs*, 2015. R package version 0.9.12-2.

D. Morgensztern, S. Devarakonda, T. Mitsudomi, C. Maher, and R. Govindan. Mutational events in lung cancer: present and developing technologies. In H. I. Pass, D. Ball, and G. V. Scagliotti, editors, *IASLC Thoracic Oncology*, pages 95–103. Elsevier, Philadelphia, United States, 2nd edition, 2018.

D. Mozaffarian, E. K. Kabagambe, C. O. Johnson, R. N. Lemaitre, A. Manichaikul, Q. Sun, M. Foy, L. Wang, H. Wiener, and M. R. Irvin. Genetic loci associated with circulating phospholipid trans fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *The American Journal of Clinical Nutrition*, 101:398–406, 2015.

J. S. Murray, D. B. Dunson, L. Carin, and J. E. Lucas. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108:656–665, 2013.

K. Müssig, H. Staiger, F. Machicao, C. Thamer, J. Machann, F. Schick, C. D. Claussen, N. Stefan, A. Fritsche, and H. Häring. RARRES2, encoding the novel adipokine chemerin, is a genetic determinant of disproportionate regional body fat distribution: a comparative magnetic resonance imaging study. *Metabolism*, 58:519–524, 2009.

A. C. Naj, G. Jun, G. W. Beecham, L. Wang, B. N. Vardarajan, J. Buros, P. J. Gallins, J. D. Buxbaum, G. P. Jarvik, and P. K. Crane. Common variants at MS4a4/MS4a6e, CD2ap, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nature Genetics*, 43:436–441, 2011.

N. N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42:789–817, 2014.

R. M. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

S. E. Neville, J. T. Ormerod, and M. P. Wand. Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics*, 8:1113–1151, 2014.

M. A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5:155–176, 2004.

T. Nic Suibhne, T. C. Raftery, O. McMahon, C. Walsh, C. O'Morain, and M. O'Sullivan. High prevalence of overweight and obesity in adults with Crohn's disease: Associations with disease and lifestyle factors. *Journal of Crohn's and Colitis*, 7:e241–e248, 2013.

A. C. Nica and E. T. Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transaction of the Royal Society B*, 368:20120362, 2013.

S. M. O'Brien and D. B. Dunson. Bayesian multivariate logistic regression. *Biometrics*, 60:739–746, 2004.

M. Opper and D. Saad. *Advanced Mean Field Methods: Theory and Practice*. MIT Press, Cambridge, Massachusetts, 2001.

P. F. O'Reilly, C. J. Hoggart, Y. Pomyen, F. C. F. Calboli, P. Elliott, M.-R. Jarvelin, and L. J. Coin. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One*, 7:e34861, 2012.

T. Ormerod, C. You, and S. Müller. A variational Bayes approach to variable selection. *Preprint*, 2014.

D. B. Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27:1075–1090, 1956.

J. Paisley, D. Blei, and M. Jordan. Variational Bayesian inference with stochastic search. *arXiv preprint arXiv:1206.6430*, 2012.

O. Papaspiliopoulos, G. O. Roberts, and M. Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, pages 59–73, 2007.

J.-H. Park, M. H. Gail, C. R. Weinberg, R. J. Carroll, C. C. Chung, Z. Wang, S. J. Chanock, J. F. Fraumeni, and N. Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108: 18026–18031, 2011.

S. Park, O. Gavrilova, A. L. Brown, J. E. Soto, S. Bremner, J. Kim, X. Xu, S. Yang, J. Um, and L. G. Koch. DNA-PK promotes the mitochondrial, metabolic and physical decline that occurs during aging. *Cell Metabolism*, 25:1135–1146.e7, 2017.

D. Perekrestenko, V. Cevher, and M. Jaggi. Faster coordinate descent via adaptive importance sampling. *arXiv preprint arXiv:1703.02518*, 2017.

C. B. Peterson, M. Bogomolov, Y. Benjamini, and C. Sabatti. TreeQTL: hierarchical error control for eQTL findings. *Bioinformatics*, 32:2556–2558, 2016.

E. Petretto, L. Bottolo, S. R. Langley, M. Heinig, C. McDermott-Roe, R. Sarwar, M. Pravenec, N. Hübner, T. J. Aitman, and S. A. Cook. New insights into the genetic control of gene expression using a Bayesian multi-tissue approach. *PLoS Computational Biology*, 6:e1000737, 2010.

C. Pfleger, H. B. Mortensen, L. Hansen, C. Herder, B. O. Roep, H. Hoey, H. Aanstoot, M. Kocova, and N. C. Schloot. Association of IL-1ra and adiponectin with C-peptide and remission in patients with type 1 diabetes. *Diabetes*, 2008.

J. Phieler, R. Garcia-Martin, J. D. Lambris, and T. Chavakis. The role of the complement system in metabolic organs and metabolic diseases. In *Seminars in Immunology*, volume 25, pages 47–53. Elsevier, 2013.

J. K. Pickrell. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94:559–573, 2014.

J. K. Pickrell, T. Berisa, J. Z. Liu, L. Ségurel, J. Y. Tung, and D. A. Hinds. Detection and interpretation of shared genetic influences on 42 human traits. *Nature Genetics*, 48:709, 2016.

J. Piironen and A. Vehtari. Projection predictive variable selection using Stan+ R. *arXiv preprint arXiv:1508.02502*, 2015.

J. Piironen and A. Vehtari. On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *arXiv preprint arXiv:1610.05559*, 2016.

J. Piironen and A. Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11:5018–5051, 2017.

M. Plummer, N. Best, K. Cowles, and K. Vines. *CODA: Convergence Diagnosis and Output Analysis for MCMC*, 2006. R package version 0.18.1.

N. G. Polson and J. G. Scott. Alternative global–local shrinkage rules using hypergeometric–Beta mixtures. Technical report, Duke University Department of Statistical Science, 2009.

N. G. Polson and J. G. Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics*, volume 9, pages 501–538. Oxford University Press, New York, United States, 2010.

N. G. Polson and J. G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7: 887–902, 2012.

N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association*, 108:1339–1349, 2013.

C. Poole. Multiple comparisons? No problem! *Epidemiology*, 2:241–243, 1991.

S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. De Bakker, and M. J. Daly. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81:559–575, 2007.

M. A. Quintana and D. V. Conti. Integrative variable selection via Bayesian model uncertainty. *Statistics in Medicine*, 32:4938–4953, 2013.

M. Quiroz, R. Kohn, M. Villani, and M.-N. Tran. Speeding up MCMC by efficient data subsampling. *Journal of the American Statistical Association*, pages 1–35, 2018.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

V. K. Ramanan, S. L. Risacher, K. Nho, S. Kim, L. Shen, B. C. McDonald, K. K. Yoder, G. D. Hutchins, J. D. West, and E. F. Tallman. GWAS of longitudinal amyloid accumulation on 18F-florbetapir PET in Alzheimer's disease implicates microglial activation gene IL1rap. *Brain: A Journal of Neurology*, 138: 3076–3088, 2015.

B. Ramkhelawon, E. J. Hennessy, M. Ménager, T. D. Ray, F. J. Sheedy, S. Hutchison, A. Wanschel, S. Oldebeken, M. Geoffrion, and W. Spiro. Netrin-1 promotes adipose tissue macrophage retention and insulin resistance in obesity. *Nature Medicine*, 20:377–384, 2014.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In S. Kaski and J. Corander, editors, *Artificial Intelligence and Statistics*, pages 814–822, Reykjavik, Iceland, 2014. Proceedings of Machine Learning Research.

R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In M. F. Balcan and K. Q. Weinberger, editors, *International Conference on Machine Learning*, pages 324–333, New York, United States, 2016. Proceedings of Machine Learning Research.

M. Rebhan, V. Chalifa-Caspi, J. Prilusky, and D. Lancet. GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. *Bioinformatics*, 14: 656–664, 1998.

E. W. Rebnord, E. Strand, Oi. Midttun, G. F. T. Svingen, M. H. E. Christensen, P. M. Ueland, G. Mellgren, P. R. Njolstad, G. S. Tell, and O. K. Nygard. The kynurenine:tryptophan ratio as a predictor of incident type 2 diabetes mellitus in individuals with coronary artery disease. *Diabetologia*, 60:1712–1721, 2017.

A. P. Reiner, C. L. Carty, N. S. Jenny, C. Nievergelt, M. Cushman, D. J. Stearn-Kurosawa, S. Kurosawa, L. H. Kuller, and L. A. Lange. PROC, PROCR, and PROS1 polymorphisms, plasma anticoagulant phenotypes, and risk of cardiovascular disease and mortality in older adults: the Cardiovascular Health Study. *Journal of Thrombosis and Haemostasis*, 6:1625–1632, 2008.

A. Rényi. On measures of entropy and information. Technical report, Hungarian Academy of Sciences Budapest Hungary, 1961.

S. Richardson, L. Bottolo, and J. S. Rosenthal. Bayesian models for sparse regression analysis of high-dimensional data. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, editors, *Bayesian Statistics*, volume 9, pages 539–569. Oxford University Press, New York, United States, 2010.

D. Ricklin, G. Hajishengallis, K. Yang, and J. D. Lambris. Complement: a key system for immune surveillance and homeostasis. *Nature Immunology*, 11:785–797, 2010.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951.

C. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, United States, 2013.

A. J. Rogers and S. T. Weiss. Epidemiologic and population genetic studies. In Robertson D. and G. H. Williams, editors, *Clinical and Translational Science*, pages 313–326. Elsevier, 2nd edition, 2017.

K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11:589–594, 1990.

J. S. Rosenthal and G. O. Roberts. Coupling and ergodicity of adaptive MCMC. *Journal of Applied Probablity*, 44:458–475, 2007.

M. Rotival, T. Zeller, P. S. Wild, S. Maouche, S. Szymczak, A. Schillert, R. Castagné, A. Deiseroth, C. Proust, and J. Brocheton. Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. *PLoS Genetics*, 7:e1002367, 2011.

H. Ruffieux, A. C. Davison, J. Hager, and I. Irincheeva. Efficient inference for genetic association studies with multiple outcomes. *Biostatistics*, 18:618–636, 2017.

H. Ruffieux, A. C. Davison, J. Hager, J. Inshaw, B. Fairfax, S. Richardson, and L. Bottolo. A global-local approach for detecting hotspots in multiple response regression. submitted, 2018a.

H. Ruffieux, W. H. M. Saris, A. Astrup, M. E. Harper, R. Dent, A. C. Davison, J. Hager, and A. Valsesia. A pqtl study sheds light on the genetic architecture of obesity. in preparation, 2018b.

T. Salimans, D. Kingma, and M. Welling. Markov Chain Monte Carlo and variational inference: bridging the gap. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, pages 1218–1226, Lille, France, 2015. Proceedings of Machine Learning Research.

M.-A. Sato. Online model selection based on the variational Bayes. *Neural Computation*, 13:1649–1681, 2001.

L. K. Saul and M. I. Jordan. Exploiting tractable substructures in intractable networks. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 486–492, Cambridge, Massachusetts, 1996. MIT Press.

L. K. Saul, T. Jaakkola, and M. I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.

L. Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4: 10–26, 1965.

J. G. Scott and J. O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38:2587–2619, 2010.

S. L. Scott, A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch. Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11:78–88, 2016.

M. P. Scott-Boyer, G. C. Imholte, A. Tayeb, A. Labbe, C. F. Deschepper, and R. Gottardo. An integrated hierarchical Bayesian model for multivariate eQTL mapping. *Statistical Applications in Genetics and Molecular Biology*, 11:1515–1544, 2012.

A. A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28: 1353–1358, 2012.

R. D. Sheldon, K. M. Kanosky, K. D. Wells, L Miles, J. W. Perfield, S. Xanthakos, T. H Inge, and R. S. Rector. Transcriptomic differences in intra-abdominal adipose tissue in extremely obese adolescents with different stages of NAFLD. *Physiological Genomics*, 48:897–911, 2016.

X. Shen and L. Wasserman. Rates of convergence of posterior distributions. *The Annals of Statistics*, 29: 687–714, 2001.

R. Sheth and R. Khardon. Excess risk bounds for the Bayes risk using variational inference in latent gaussian models. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5151–5161, Cambridge, Massachusetts, 2017. Curran Associates, Inc.

M. A. Shogren-Knaak and C. L. Peterson. Chromatin and Chromatin Remodeling Enzymes. In C. Wu and C. Allis, editors, *Methods in Enzymology*, volume 375, pages 62–76. Elsevier, San Diego, United States, 2003.

N. Simon, J. H. Friedman, and T. J. Hastie. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529*, 2013.

R. Sims, S. J. van der Lee, A. C. Naj, C. Bellenguez, N. Badarinarayan, J. Jakobsdottir, B. W. Kunkle, A. Boland, R. Raybould, and J. C. Bis. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nature Genetics*, 49:1373–1384, 2017.

S. Singh, P. S. Dulai, A. Zarrinpar, S. Ramamoorthy, and W. J. Sandborn. Obesity in IBD: epidemiology, pathogenesis, disease course and treatment outcomes. *Nature Reviews Gastroenterology & Hepatology*, 14:110–121, 2017.

S. Sivakumaran, F. Agakov, E. Theodoratou, J. Prendergast, L. Zgaga, T. Manolio, I. Rudan, P. McKeigue, J. Wilson, and H. Campbell. Abundant Pleiotropy in Human Complex Diseases and Traits. *American Journal of Human Genetics*, 89, 2011.

S. Smemo, J. J. Tena, K.-H. Kim, E. R. Gamazon, N. J. Sakabe, C. Gómez-Marín, I. Aneas, F. L. Credidio, D. R. Sobreira, and N. F. Wasserman. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507:371, 2014.

N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14:483–495, 2013.

P. Song, T. Ramprasath, H. Wang, and M. Zou. Abnormal kynurenine pathway of tryptophan catabolism in cardiovascular diseases. *Cellular and molecular life sciences: CMLS*, 74:2899–2916, 2017.

Q. Song and F. Liang. A split-and-merge Bayesian variable selection approach for ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:947–972, 2015.

Y. Song, J. E. Manson, L. Tinker, N. Rifai, N. R. Cook, F. B. Hu, G. S. Hotamisligil, P. M. Ridker, B. L. Rodriguez, and K. L. Margolis. Circulating levels of endothelial adhesion molecules and risk of diabetes in an ethnically diverse cohort of women. *Diabetes*, 56:1898–1904, 2007.

D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks. *BUGS 0.5: Bayesian inference using Gibbs sampling*, 1996. MRC Biostatistics Unit, Cambridge, United Kingdom.

D. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. *OpenBUGS user manual, version 3.0. 2*, 2007. MRC Biostatistics Unit, Cambridge, United Kingdom.

J. R. Staley, J. Blackshaw, M. A. Kamat, S. Ellis, P. Surendran, B. B. Sun, D. S. Paul, D. Freitag, S. Burgess, and J. Danesh. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics*, 32:3207–3209, 2016.

F. J. Steemers and K. L. Gunderson. Illumina, Inc. *Future Medicine*, 2005.

C. M. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University, United States, 1956.

C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, pages 1135–1151, 1981.

M. Stephens. False discovery rates: a new deal. *Biostatistics*, 18:275–294, 2016.

M. Stephens and D. J. Balding. Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10:681, 2009.

F. C. Stingo, Y. A. Chen, M. G. Tadesse, and M. Vannucci. Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *The Annals of Applied Statistics*, 5:1978–2002, 2011.

T. W. Stone, M. McPherson, and L. Gail Darlington. Obesity and Cancer: Existing and New Hypotheses for a Causal Connection. *EBioMedicine*, 30:14–28, 2018.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:479–498, 2002.

J. D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31:2013–2035, 2003.

J. D. Storey and R. J. Tibshirani. Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences*, 100:9440–9445, 2003.

R. L. Strausberg. Talkin' Omics. *Disease Markers*, 17:39, 2001.

W. E Strawderman. Proper Bayes minimax estimators of the multivariate normal mean. *The Annals of Mathematical Statistics*, 42:385–388, 1971.

K. Suhre, M. Arnold, A. M. Bhagwat, R. J. Cotton, R. Engelke, J. Raffler, H. Sarwath, G. Thareja, A. Wahl, and R. K. DeLisle. Connecting genetic risk to disease end points through the human blood plasma proteome. *Nature Communications*, 8:14357, 2017.

B. B. Sun, J. C. Maranville, J. E. Peters, D. Stacey, J. R. Staley, J. Blackshaw, S. Burgess, T. Jiang, E. Paige, and P. Surendran. Genomic atlas of the human plasma proteome. *Nature*, 558:73–79, 2018.

C. J. Tack, R. Stienstra, L. A. B. Joosten, and M. G. Netea. Inflammation links excess fat to insulin resistance: the role of the interleukin-1 family. *Immunological Reviews*, 249:239–252, 2012.

Y. G. Tak and P. J. Farnham. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin*, 8:57, 2015.

O. Tavana, N. Puebla-Osorio, J. Kim, M. Sang, S. Jang, and C. Zhu. Ku70 functions in addition to nonhomologous end joining in pancreatic beta-cells: a connection to beta-catenin regulation. *Diabetes*, 62:2429–2438, 2013.

D. E. te Beest, S. W. Mes, S. M. Wilting, R. H. Brakenhoff, and M. A. van de Wiel. Improved high-dimensional prediction with Random Forests by the use of co-data. *BMC Bioinformatics*, 18:584, 2017.

A. B. Thrush, G. Antoun, M. Nikpay, D. A. Patten, C. DeVlugt, J. F. Mauger, B. L. Beauchamp, P. Lau, R. Reshke, E. Doucet, P. Imbeault, R. Boushel, D. Gibbings, J. Hager, A. Valsesia, R. S. Slack, O. Y. Al-Dirbashi, R. Dent, R. McPherson, and M. E. Harper. Diet-resistant obesity is characterized by a distinct plasma proteomic signature and impaired muscle fiber metabolism. *International Journal of Obesity*, 42:353, 2018.

R. J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

D. Tran, R. Ranganath, and D. M. Blei. The variational Gaussian process. *arXiv preprint arXiv:1511.06499*, 2015.

G. Trynka, C. Sandor, B. Han, H. Xu, B. E. Stranger, X. S. Liu, and S. Raychaudhuri. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nature Genetics*, 45:124, 2013.

C. Tsallis. Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, 52: 479–487, 1988.

J. W. Tukey. The philosophy of multiple comparisons. *Statistical Science*, pages 100–116, 1991.

E. Turro, N. Bochkina, A.-M. K Hein, and S. Richardson. BGX: a Bioconductor package for the Bayesian integrated analysis of Affymetrix GeneChips. *BMC Bioinformatics*, 8:439, 2007.

N. Ueda and R. Nakano. Deterministic annealing EM algorithm. *Neural Networks*, 11:271–282, 1998.

M. A. van de Wiel, T. G. Lien, W. Verlaat, W. N. van Wieringen, and S. M. Wilting. Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine*, 35:368–381, 2016.

M. A. van de Wiel, D. E. Te Beest, and M. M. Münch. Learning from a lot: Empirical Bayes for high-dimensional model-based prediction. *Scandinavian Journal of Statistics*, 2018.

P. van der Harst and N. Verweij. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circulation Research*, 122:433–443, 2018.

S. van der Pas, B. T. Szabó, and A. W. van der Vaart. Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics*, 11:3196–3225, 2017.

S. L. van der Pas, B. J. K. Kleijn, and A. W. van der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8:2585–2618, 2014.

S. L. van der Pas, J.-B. Salomond, and J. Schmidt-Hieber. Conditions for posterior contraction in the sparse normal means problem. *Electronic Journal of Statistics*, 10:976–1000, 2016.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, United Kingdom, 2000.

J. C. van Wolfswinkel and R. F. Ketting. The role of small non-coding RNAs in genome stability and chromatin organization. *Journal of Cell Science*, 123:1825–1839, 2010.

W. Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 2010.

J. Von Neumann and S. Ulam. Monte Carlo method. *National Bureau of Standards Applied Mathematics Series*, 12:36, 1951.

B. Wang and D. M. Titterington. Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values. In M. Chickering and J. Halpern, editors, *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 577–584, Arlington, United States, 2004. AUAI Press.

B. Wang and D. M. Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. In R. G. Cowell and Z. Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, Barbados, 2005. Society for Artificial Intelligence and Statistics.

C. Wang, X. Chen, A. J. Smola, and E. P. Xing. Variance reduction for stochastic gradient optimization. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 181–189, Cambridge, Massachusetts, 2013. Curran Associates, Inc.

X. Wang and D. B. Dunson. Parallelizing MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

X. Wang, F. Guo, K. A. Heller, and D. B. Dunson. Parallelizing MCMC with random partition trees. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 451–459, Cambridge, Massachusetts, 2015. Curran Associates, Inc.

X. Wang, D. B. Dunson, and C. Leng. DECOrrelated feature space partitioning for distributed sparse regression. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 802–810, Cambridge, Massachusetts, 2016. Curran Associates, Inc.

Y. Wang and D. M. Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 0:1–85, 2018.

L. D Ward and M. Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology*, 30:1095, 2012.

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In L. Getoor and T. Scheffer, editors, *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688, New York, United States, 2011. ACM.

D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, and L. Hindorff. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42:D1001–D1006, 2014.

H.-J. Westra, M. J. Peters, T. Esko, H. Yaghootkar, C. Schurmann, J. Kettunen, M. W. Christiansen, B. P. Fairfax, K. Schramm, and J. E. Powell. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature Genetics*, 45:1238, 2013.

R. H. F. Wong, I. Chang, C. S. S. Hudak, S. Hyun, H. Kwan, and H. S. Sul. A role of DNA-PK for the metabolic gene regulation in response to insulin. *Cell*, 136:1056–1072, 2009.

W. L. Xu, A. R. Atti, M. Gatz, N. L. Pedersen, B. Johansson, and L. Fratiglioni. Midlife overweight and obesity increase late-life dementia risk: A population-based twin study. *Neurology*, 76:1568–1574, 2011.

J. Yang, T. Ferreira, A. P. Morris, S. E. Medland, P. A. F. Madden, A. C. Heath, N. G. Martin, G. W. Montgomery, M. N. Weedon, and R. J. Loos. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics*, 44:369, 2012.

J. Yang, L. G. Fritsche, X. Zhou, G. Abecasis, and International Age-Related Macular Degeneration Genomics Consortium. A scalable Bayesian method for integrating functional information in genome-wide association studies. *The American Journal of Human Genetics*, 101:404–416, 2017.

C. Yao, R. Joehanes, A. D. Johnson, T. Huan, C. Liu, J. E. Freedman, P. J. Munson, D. E. Hill, M. Vidal, and D. Levy. Dynamic role of trans regulation of gene expression in relation to complex traits. *The American Journal of Human Genetics*, 100:571–580, 2017.

C. Yao, G. Chen, C. Song, J. Keefe, M. Mendelson, T. Huan, B. B. Sun, A. Laser, J. C. Maranville, and H. Wu. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature Communications*, 9:3268, 2018a.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Yes, but Did It Work?: Evaluating Variational Inference. *arXiv preprint arXiv:1802.02538*, 2018b.

C. You, J. T. Ormerod, and S. Mueller. On variational Bayes estimation and variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistics*, 56:73–87, 2014.

Y. Yu and X.-L. Meng. To center or not to center: That is not the question—an Ancillarity–Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, 20:531–570, 2011.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:49–67, 2006.

M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster computing with working sets. In E. Nahum and D. Xu, editors, *HotCloud '10*, page 95, Boston, United States, 2010. USENIX.

M. D. Zeiler. ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Studies in Bayesian Econometrics*, volume 6, pages 233–243. Elsevier, New York, United States, 1986. P. K. Goel and A. Zellner, editors.

D. R Zerbino, P. Achuthan, W. Akanni, M. Amode, D. Barrell, J. Bhai, K. Billis, C. Cummins, A. Gall, and C. G. Girón. Ensembl 2018. *Nucleic Acids Research*, 46:D754–D761, 2018.

C. Zhang, J. Butepage, H. Kjellstrom, and S. Mandt. Advances in variational inference. *arXiv preprint arXiv:1711.05597*, 2017.

F. Zhang and C. Gao. Convergence rates of variational posterior distributions. *arXiv preprint arXiv:1712.02519*, 2017.

W. Zhang, J. Zhu, E. E. Schadt, and J. S. Liu. A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules. *PLoS Computational Biology*, 6:e1000642, 2010.

P. Zhao and T. Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In F. Bach and D. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, pages 1–9, Lille, France, 2015. Proceedings of Machine Learning Research.

J. Zheng, A. M. Erzurumluoglu, B. L. Elsworth, J. P. Kemp, L. Howe, P. C. Haycock, G. Hemani, K. Tansey, C. Laurin, and B. S. Pourcain. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, 33:272–279, 2017.

X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44:821, 2012.

X. Zhou and M. Stephens. Efficient algorithms for multivariate linear mixed models in genome-wide association studies. *Nature Methods*, 11:407, 2014.

X. Zhou, P. Carbonetto, and M. Stephens. Polygenic modeling with Bayesian sparse linear mixed models. *PLoS Genetics*, 9:e1003264, 2013.

Hélène RUFFIEUX

helene.ruffieux@alumni.epfl.ch
ch.linkedin.com/in/hruffieux
github.com/hruffieux

EDUCATION

2014 – 2019	PhD in Statistics Large-scale Bayesian Inference for Genetic Association with Multiple Outcomes	Ecole Polytechnique Fédérale de Lausanne (EPFL) and Nestlé Institute of Health Sciences (NIHS), Lausanne, Switzerland
Sep – Dec 2017	Research visit at MRC Biostatistics Unit	University of Cambridge, United Kindgom
2011 – 2013	Master of Science MSc in Applied Mathematics	EPFL
Feb – Jul 2013	Master's Thesis in Numerical Analysis Multiscale finite element method for advection-diffusion problems in convection-dominated regimes	Ecole Nationale des Ponts et Chaussées, Paris, France
2008 – 2011	Bachelor of Science BSc in Mathematics	EPFL
2003 – 2007	High School Diploma Applied Mathematics and Physics Option	Collège St-Michel, Fribourg, Switzerland

PROFESSIONAL ACTIVITIES

2011 – 2019	Teaching assistant of Bachelor's and Master's degree courses Statistics for Genomic Data Analysis, Time Series, Linear Algebra, Probability and Statistics Supervision of a Master's thesis	EPFL
Jan – Nov 2014	Junior data analyst Genomic statistical analyses to improve risk stratification of cancer patients via the identification of prognostic markers	Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
2012 – 2013	Internship Characterization of circulating tumor cells by optimized microfluidic chip	SIB

PEER-REVIEWED PUBLICATIONS AND PREPRINTS

H. Ruffieux, A. C. Davison, J. Hager, and I. Irincheeva. Efficient inference for genetic association studies with multiple outcomes, *Biostatistics,* 18:618–636, 2017, doi: 10.1093/biostatistics/kxx007.

H. Ruffieux, A. C. Davison, J. Hager, J. Inshaw, B. Fairfax, S. Richardson, and L. Bottolo. A global-local approach for detecting hotspots in multiple response regression. *Submitted,* 2018.

H. Ruffieux, J. Carayol, M. E. Harper, R. Dent, W. H. M. Saris, A. Astrup, A. C. Davison, J. Hager, and A. Valsesia. A large-scale multivariate pQTL study sheds light on the genetic architecture of obesity. *In preparation,* 2018.

CONFERENCES AND PRESENTATIONS

27 June 2018	International Society for Bayesian Analysis (ISBA) Conference Contributed poster presentation : A global-local approach for detecting hotspots in multiple-response regression	Edinburgh, United Kingdom
23 August 2017	Molecular modeling group Invited seminar : Large-scale Bayesian inference for integrating multiple sources of molecular data	SIB, Lausanne, Switzerland
11 July 2017	International Society for Clinical Biostatistics (ISCB) Contributed presentation : Joint variational inference for genetic association with multiple outcomes	Vigo, Spain
20 June 2017	MRC Biostatistics Unit Invited seminar : Joint variational inference for genetic association with multiple outcomes	University of Cambridge, United Kingdom
23 April 2017	International Biometric Society (IBS), Channel Network Conference Contributed presentation : Joint variational inference for genetic association with multiple outcomes	Hasselt, Belgium
2 August 2016	Joint Statistical Meetings (JSM) Invited presentation : Efficient inference for genetic association with multiple outcomes	Chicago, United States
16 June 2016	Conférence Universitaire de Suisse Occidentale (CUSO) Young Researcher Conference Contributed presentation : Variational inference for hierarchical regression models with multiple responses	EPFL, Lausanne, Switzerland
6 January 2016	Workshop Chair of Statistics Contributed presentation : Bayesian inference on genetic association using multiple genomic data types	Davos, Switzerland
6 January 2015	Workshop Chair of Statistics Contributed presentation : Modelling heterogeneous high-dimensional data	Davos, Switzerland
2014 – 2017	Internal presentations for Nestlé Science Advisory Boards, Project Reviews, Statistical meetings, Training Sessions	Multiple sites of Nestlé, Switzerland

PROGRAMMING AND IT

R, Matlab, Mathematica, knowledge of C++ and Python
Unix and high performance computing environments
Git, LateX, Office

LANGUAGES

French: native, English: advanced, German: basic

HONOURS AND AWARDS

Undergraduate: Rank 1/19 Bachelor, Grade 5.87/6 Master
High-School: two prices for top grades in Applied Mathematics and Physics,
one price for best diploma thesis: On the mechanisms of retroactions in climate change.