9

Multivariate linear models for GWAS

Research in Genetics in the first decade of the 21^{st} century has been dominated by the attempt to characterize common variation in the human genome and its impact on complex phenotypes. The decade opened with the announcement of the completion of the first draft(s) of the human genome (Lander et al., 2001; Venter et al., 2001), which provided one reference sequence. An international effort (The HapMap, 2003), analogous to the one that had facilitated this first achievement, was then devoted to the characterization of common variants in different human populations (originally focusing on trios to represent European, Yoruba, Beijing Chinese and Japanese populations). By 2007, commercial enterprises had developed technologies that allowed hundreds of thousands of single nucleotide polymorphisms (SNPs) to be genotyped in thousands of individuals at reasonable costs: genome-wide association studies (GWAS), first described in Risch and Merikangas (1996), became possible and popular. These studies aim to identify genetic loci that influence complex phenotypes: that is, traits whose genetic underpinning is not ascribable to one, or even a handful, of genes. When very many loci influence a trait, it is reasonable to assume that the effect of any of these might be quite modest, requiring a large sample size for detection. Genome wide association studies, that recruit individuals from a population, without need to study relatives, represent a convincing design in this context, and indeed, they have become the method of choice for many groups. In May 2012, the NHGRI catalog of published genome-wide association studies (`http://www.genome.gov/gwastudies/`) included 1261 publications and 6411 SNPs, with a lot of research along these lines still under progress (13 of the 19 articles in the May 2012 of *Nature Genetics* rely on GWAS). In this chapter, we survey the standard method of analysis for GWAS, compare it to the underlying genetic model, and describe sta-

tistical approaches that have attempted to bridge the gap between the theoretical model and the methods of analysis, with particular emphasis on Bayesian methods. The chapter is organized as follows. Section 1 reviews the polygenic model originally proposed by R.A. Fisher; Section 2 describes standard GWAS approaches, their results, and the data driven indication that additional loci remain to be identified; Section 3 outlines the challenges encountered when attempting to use multivariate models in GWAS; and Sections 4 and 5 describe how penalized methods and Bayesian approaches address them.

9.1 The polygenic model

Throughout the chapter, we will limit our discussion to quantitative traits. This is only in the interest of simplicity: the arguments we will describe apply equally to complex binary traits, and the extension of statistical methodology to these is conceptually clear, even if it might require substantial complication at the level of data analysis and algorithms. Quantitative traits were the first type of complex traits whose genetic underpinning were investigated systematically. The impact and importance of Fisher's 1918 paper should be understood in the context of the scientific debate of the time: on the one hand, biometrists, like Galton, collected data that clearly underscored the existence of a genetic, heritable component of continuous traits like height; on the other hand, the understanding of the molecular mechanisms of inheritance was very rudimentary, with the most advanced conceptualization being Mendel's model, of a two allelic system resulting in a discrete phenotype. Fisher (1918) postulates that continuous traits have both an environmental and genetic component, with the latter the result of many Mendelian factors. The simplest version of such model assumes no interaction between genes and environment or between the genes themselves. Formally, if y_i indicates the phenotype value for subject i,

$$y_i = \sum_{k=1}^{K} Z_{ik} + \epsilon_i, \tag{9.1}$$

with the independent environmental variable $\epsilon_i \perp Z_{ik}$ and Z_{ik} representing the effect of Mendelian factor k on subject i. In Fisher's time, it was impossible to gather any observations on Z_{ik}, but model (14.1) could be used to explain (and predict) the levels of correlation between relatives, which in turn was used to define the heritability of a trait. Let

Y be a vector of phenotype levels for related individuals. Then, under the assumption of independence between the genetic and environmental factors, as well as across the genetic signals,

$$\text{cov}(Y) = \sigma_a^2 \frac{1}{2}\Phi_+ \sigma_d^2 \Delta + \sigma_\epsilon^2 I, \tag{9.2}$$

where Φ and Δ are two matrices defined by the relationships between subjects and σ_a^2 and σ_d^2 are two components of the genetic variance. Specifically, σ_a^2 is the variability of the phenotype that can be explained by additive effects of the Mendelian factors that each subjects inherits from her/his two parents, while σ_d^2 is the variability that can be explained by interactions between these Mendelian factors (dominance). The heritability of any trait is defined as the ratio σ_a^2/σ^2.

In contrast, the technological advances of the last decades made it possible to observe genetic variation at hundreds of thousands (millions) of locations across the genome. Specifically, the genotypes at biallelic markers are recorded into a matrix $X = \{x_{ij}\}$ with $x_{ij} \in \{0, 1, 2\}$ representing the number of alleles of type A observed in individual i at location k. The loci are chosen sufficiently close to each other that $\text{cov}(X_k, X_{k+1}) \neq 0$ and the collection of variables $(X_1, \ldots, X_p)$ can serve as a proxy for other unobserved locations across the genome. This is possible by exploiting what geneticists call "linkage disequilibrium," a local dependency between alleles at neighboring loci that derives from the fact that we inherit DNA from our parents in consecutive chunks. On the basis of this data it becomes possible to try to uncover the genome location of Fisher's Mendelian factors fitting a model like

$$y_i = \sum_{k=1}^{p} \beta_k X_{ki} + e_i, \tag{9.3}$$

where $\beta_i \neq 0$ correspond to locations in the genome harboring one factor relevant for the trait of interest.

9.2 Analysis of GWAS

While the implicit genetic model for complex traits is polygenic, the practice of statistical analysis in GWAS has consisted substantially in testing for association between the trait and each genetic locus, one locus at a time, with models like

$$y_i = \alpha + \beta_k X_{ki} + \eta_i. \tag{9.4}$$

There are a number of reasons behind this choice. Relying on univariate regression is computationally simple, allows one to use all available observations at each marker without having to rely on imputation, and avoids the problem of choosing among correlated explanatory variables, when they all represent equally good proxies of the same unobserved Mendelian factor. Moreover, univariate regression results in a SNP-specific p-value that is easily comparable and combinable across different studies— a relevant point given the considerable importance of meta-analysis in present-day genetic investigations.

Despite the attractiveness of its simplicity, model (14.4) is clearly misspecified with respect to our understanding of the underlying genetic mechanism, with a variety of consequences. Firstly, while typically subjects in GWAS studies are recruited among unrelated individuals with similar ethnic background, it is impossible to exactly determine the degree of relatedness among individuals and their ethnic background prior to genotyping. Invariably, the sample under study will include individuals with different degrees of similarity at the genetic level. When analyzing a polygenic trait with model (14.4), the effects of all genetic factors that are not approximated by X_{ik} become part of the error term. As soon as the degree of genetic similarity across individuals varies, this will result in a correlation structure in the error that, if ignored, leads to wrong evaluation of significance for each of the β_ks values. This has been recognized and it has been shown that fitting models like (14.4), but accounting, even rather crudely, for the covariance structure induced by genetic similarity leads to more coherent inference (Kang et al. (2010)).

Secondly, excluding relevant regressors from the model as in (14.4) leads to biased estimates of the β_k coefficients. Given that the goal of genetic studies is to identify loci that influence the phenotype, rather than exactly estimate their influence, downward bias is particularly serious as it results in possibly missing relevant loci. Indeed, there has been an accumulation of evidence that the genomewide significant loci identified with strategy (14.3) do not provide a complete picture of the genetic underpinning of the studied traits and that more information can be potentially extracted from the GWAS datasets. We now review some of these results.

Based on Fisher's model (14.2), over the years, a set of estimates of heritabilities for various traits has been obtained. Once a set of loci $\mathcal{D}$ has been identified with GWAS, one can evaluate the proportion of phenotypic variance explained by them: $\sigma^2_{a,\mathcal{D}} = \sum_{k\in\mathcal{D}} \beta_k^2 \sigma^2_{X_k}$. For most traits for which this calculation has been possible, σ^2_a, as estimated from

family-based studies, results to be quite substantially larger than $\sigma^2_{a,\mathcal{D}}$. This observation, that the SNPs identified via univariate regression explain only a small portion of the trait variance, has led geneticist to speculate on the reasons of this discrepancy, in the debate that is often referred to as "missing heritability" (Manolio et al. (2009)). Multiple hypotheses have been advanced, and most likely, the truth lies behind many of them: (1) the effect of the discovered loci is under-estimated by the coefficient of X_k, which only imperfectly tracks the signal due to the true causative variant, typically not directly genotyped; (2) there might be a substantial contribution to phenotypic variation by rare variants, whose effect are difficult to detect with a GWAS design; (3) the traditional estimates of σ^2_a might be biased up-wards, as they are derived making a series of assumptions (as absence of interaction between genetic factors) that might not be correct; (4) there are more genetic loci, acting both additively and interacting, whose effects have not been discovered by univariate analysis. A number of papers provide empirical data to support (4), as we discuss in the following.

Sabatti et al. (2009), mapping cardiovascular traits, noted how the analysis of the p-values from univariate model (14.4) underscores the presence of residual association signals, beyond those strong enough to be identified with genomewide significance. An application of the Higher Criticism statistics to test the global null of no association between any of the genomic loci and the phenotypes resulted in clear rejections of all the null hypotheses, even for those phenotypes for which we were unable to identify any significant locus. Switching the emphasis, from model selection and loci identification to prediction, can be a useful way of quantifying the amount of signal present in the data. This point is made clear in the investigation on schizophrenia and bipolar disorder in Purcell et al. (2009), and in the survey of the Welcome trust data in Quevedo et al. (2011), where it becomes apparent that it is possible to predict disease risk with much higher accuracy when using large number of SNPs genotyped in GWAS studies, rather than when relying only on the "genomewide significant associations." Approaching the same question from yet a slightly different view point, Visscher and collaborators (Yang et al., 2010), in a series of papers, suggest reconstructing σ^2_a directly from GWAS data, estimating a decomposition of variance similar to that in Kang et al. (2010), and argue that this should be used as a measure of what proportion of the phenotypic variance can be explained on the basis of the common genetic variation captured by SNPs. This estimate of the phenotypic variance explainable by common genetic vari-

ation is considerably higher than $\sigma^2_{a,\mathcal{D}}$, again underscoring that there is more information on the traits in the SNPs than what is captured in the mapped loci. In this context, then, it seems that attempting to fit multivariate models as (14.3) might increase the power to identify all relevant loci and the rest of the chapter is devoted to a review of the methodologies developed with this goal.

9.3 Challenges in multivariate linear models for GWAS

Many different statistical approaches can be collected under the umbrella of "multivariate analysis of GWAS," including haplotype-based methods, pathway analysis, as well as methodology specifically designed to target interactions. We will not consider all of these, but restrict our attention to the simplest model that consider single SNPs as explanatory variables, but try to reconstruct their possible effects in the context of a multivariate regression (14.3). This is for need of focus, but also because the challenges posed by this relatively simple step are sufficiently interesting and deserving discussion. Other chapters (15, 16) complement this, providing the reader with examples of applications of multivariate methods to GWAS data as well as other association studies.

Once the decision has been made to try to identify a multivariate model (14.3) to explain the trait of interest, researchers are faced with a number of challenges. Even a 'small' GWAS relies on genotypes at 500,000 SNPs, putting the number of possible regressor p up in the hundreds of thousands. By contrast, in each single study, the number of observations n is in the thousands. The fact that $p >> n$ makes it clearly impossible to consider all possible models and hope to estimate their parameters. However, our expectations with regard to the genetic architecture of the studied traits translate in a structure on the vector β of regressor coefficients that makes its estimation possible. While the polygenic model suggests that possibly a large number of genetic loci might be influencing traits, this should still represent a very small portion of the entire genome. In other words, we expect the vector β to include a majority of zero values (or to be sparse, using current terminology)– which makes the problem tractable.

Researchers attempting to fit a multivariate regression as in (14.3) are presented with a number of challenges. We single out four major difficulties. The **first (1) challenge** is presented by the **size of the**

model space. Even capping the number of regressors with non-zero coefficients, the number of possible models is astronomical: clearly exhaustive enumeration is not an option and computationally clever strategies are needed.

Intimately connected, is the need to appropriately **(2)** account for the **effect of search** in selecting a model. The statistics literature has paid increasing attention to problems like GWAS where the number of potential regressors p is much larger than the number of observations n. The pioneering work of Foster and George (1994), for example, discusses many related issues and proves that, under some conditions, model selection can be carried out comparing models on the basis of

$$\|y - X\beta\|_2^2 + \lambda\sigma^2\|\beta\|_0, \tag{9.5}$$

where the ℓ_0 'norm' indicates the number of non zero coefficients in the vector β and $\lambda \approx 2\log p$ (rather than $\log n$ or 2, as suggested by BIC or AIC respectively), to account for the effect of search. While some asymptotic optimality properties are provable for this criteria (known as RIC, risk inflation criterion), its applicability is limited by the fact that it requires searching the entire model space, and its performance under a greedy exploration is not clear (Birgé and Massart (2001)). It is interesting to note that, in the context of GWAS studies, Bogdan et al. (2011) and Frommlet et al. (2012) propose and explore the effectiveness of a related criterion, inspired by Bayesian models and which we will discuss in a later section. Simulation studies seem to suggest that a penalization of order $\log p$ might be appropriate also coupled with greedy search.

A third **(3) difficulty** arises when a model has been chosen, and some **confidence statements on the relevance of the selected predictors** are needed. One thing is to state that a selection procedure is consistent asymptotically, another to evaluate its validity for the given sample size. Moreover, scientific interpretability in gene mapping really calls for evaluation of the confidence with which every single predictor belongs to the model. This topic of inference after selection has received increased attention in the literature, but it is still by and large an open area of research. It is worth mentioning the work being carried out at the University of Pennsylvania (Berk et al. (2011)) and at the university of Tel Aviv (Benjamini et al. (2009); Benjamini and Gavrilov (2009)), with the first group focusing on simultaneous inference, and the latter on selective one.

Finally **(4)**, in the context of GWAS, it is important to consider that

it is not clear that the "true" model is part of the search space: the genetic factors influencing the trait of interest are likely not typed directly, and multiple neighboring SNPs might serve as reasonable proxies.

The extent of the difficulties **(1)-(4)** might explain the reticence so far to use model selection strategies to directly build multivariate models as (14.3) in gene mapping. A brief analysis of previous experience in this field underscores how researchers are aware of the challenges.

Unlike many other areas of scientific investigation, where the need for reproducibility of results obtained by high-dimensional searches emerged only recently, the structure of the gene mapping problem is such that from early on, researchers are acutely aware of the necessity to account for the large model space. Even when only a handful of markers were available for typing, geneticists became accustomed (Morton (1955)) to very stringent significance thresholds, to account for the size of the genome and are reluctant to accept results of selection strategies that do not come with a clear statement on the confidence of the predictors.

While computationally the problem was possibly not as challenging, linkage studies also were faced with the possibility of searching for multiple loci simultaneously, or sequentially (conditionally), or marginally (locus-by-locus). Without doing justice to the rich literature that this question inspired, it is useful to recall at least the paper by Dupuis et al. (1995), which concludes that the multiple comparison correction necessary for simultaneous search was such to make this not definitely preferable to the single locus search, while conditional search appeared rather problematic. Recently, Wu and Zhao (2009) have revisited a similar question in the context of GWAS, comparing the effectiveness of marginal, conditional, and joint search for a trait under the influence of two loci, again concluding that single locus search is–perhaps surprisingly–well powered. Perhaps a slightly different message comes from animal research, where multivariate models are routinely constructed and accepted, even if researchers have concluded that criteria like BIC tend to result in overparametrized models (Broman and Speed (2002)).

After having argued why it might be interesting to explore more complex statistical models for GWAS as well as outlined the difficulties that such approaches might present, we are going to summarize the two substantial efforts we are aware of to address this challenge: Lasso models and Bayesian model selection approaches. Both have been described in detail and quite clearly in the literature: therefore, rather than dwelling on implementation and technicalities, we will continue with the bird-eye

view of this chapter that we hope will motivate the reader to dig deeper on the numerous problems. Both Lasso and Bayesian model selection tackle a much wider set of problems than GWAS, but we will focus on their application and successes in this area.

9.4 Lasso approaches to GWAS

There are many ways of introducing the Lasso model (Tibshirani, 1996), but for the purpose of this exposition consider a reformulation of the selection criteria (14.5) where the ℓ_0 penalty is substituted by an ℓ_1 penalty (Chen et al., 1998). The implied minimization problem changes from combinatorial search to convex optimization, resulting in substantial computational savings. Indeed, the Lasso presents a very effective solution to challenge (1). As interest in the use of Lasso has increased in the last years, it has also become apparent that it is quite possible to obtain solutions to the minimization problem with high computational efficiency. A thorough review of progress in this area is beyond our goals, but it is important to mention at least two features that have found direct application in GWAS.

Coordinate descent methods have been proven very effective in obtaining the Lasso solution (Wu and Lange, 2008; Friedman et al., 2007) and have been implemented in software that takes input files that are particularly convenient for genetic researchers (Wu et al., 2009). Coordinate descent, that is updating β_k values one k at the time using marginal optimization, is simple and attractive because of the modest memory requirements. In sparse settings, where most of the β_ks are equal to zero and an algorithm that starts from $\beta = 0$ already has most of the coefficients at their optimal value, coordinate descent is also very efficient in that it converges to the right solution in a limited time.

Furthermore, it has become apparent that it is possible to pre-select variables, so that not only few coordinate descent cycles are needed, but these do not need to involve updates of all β_k. Wu et al. (2009) describes this property first, and in the context of GWAS for binary traits. Looking at the Karush–Kuhn–Tucker (KKT) conditions for the solution, it is possible to rank predictors in terms of their "likelihood" to have non-zero coefficient. It is an acceptable strategy only to update the most promising predictors and, once convergence has been achieved for their parameters, check if the KKT conditions are satisfied for the entire vector β. If yes, the current β values represent the solution for the

original problem; if not, the variables for which KKT are not satisfied can be added to the set to be updated and the process iterated. Generally, this "swindle" can result in substantial computational savings. Since the paper by Wu et al. (2009), other rules have been proposed to reduce computation times with predictor elimination as in El Ghaoui et al. (2010) and in Tibshirani et al. (2012).

Lasso methodology, then, generally speaking, effectively addresses challenge (1) (dealing with the large model space). How about the other difficulties? There are by now a number of results (see Candès and Plan (2009) and references therein) to suggest that, under some conditions, the Lasso strategy finds the correct model, so that Lasso is not simply a convenient approach, but a principled one (challenge 2). These results are based on properties of the predictors X, the β coefficients, and the choice of the penalization parameter λ. With regard to this latter, it is interesting to note that the value $\lambda \propto \sqrt{\log p}$, corresponding to the suggestion for the risk inflation criterion in Foster and George (1994), appears to result in good model selection properties for the Lasso (Candès and Plan, 2009). The properties required on X have different degree of verificability and generality, but are typically such that one cannot assume they are satisfied in GWAS, where neighboring SNPs tend to be correlated. In some contexts, it is suggested that a value of λ appropriate for the data at hand may be identified via cross validation. However, this typically results in a λ value that guarantees a Lasso solution with good prediction properties, rather than model selection ones—which is the goal in gene mapping studies. Additionally, it has been reported that cross validation leads to noisy results in genomic contexts, where the effect size of important variables is quite small (see Martinez et al. (2011)).

If the answer to challenge 2 is only partially understood in the context of the Lasso, challenge 3 is certainly the one that requires more progress. How to provide the user with confidence statements on each of the predictors? Wu et al. (2009) profess ignorance and provide the users of their program with p-values from univariate regression and t-statistics from the multivariate model including only the selected predictors—cautioning against interpreting these as an appropriate measure of significance that account for selection. Benjamini and Gavrilov (2009), not specifically in the context of GWAS, suggest using an FDR-based penalty to re-screen the Lasso solution. Meinshausen and Bühlmann (2010) proposed Stability Selection as a way to obtain some type of confidence statements on variable inclusion. In their bootstrap-like procedure, the

sample size is repeatedly halved and on each of these subsamples, the Lasso optimization is solved for different values of λ. For each predictor, one keeps track of the proportion π_λ^k of subsamples in which it was selected as a function of λ, and calculates the max of this proportion $\pi_*^k = \max_\lambda \pi_\lambda^k$ as the regularization parameter varies. Meinshausen and Bühlmann (2010) suggest selecting the predictors for which $\pi_*^k > 1/2$ and prove that, under some rather unrealistic conditions, this can be used to control the Family Wise Error Rate (FWER) among the predictors. Alexander and Lange (2011) have implemented a version of this stability selection for GWAS, but encountered only modest success: on the one hand, it is quite clear that the typical GWAS regressors do not satisfy the exchangeability condition of Meinshausen and Bühlmann (2010), one the other hand, a procedure that aims at controlling FWER in this framework appears to be severely underpowered. Alexander and Lange (2011) also discuss challenge 4, noting that the selection probabilities of SNPs belonging to the same region should really be considered jointly, as the Lasso tends to select randomly one of multiple predictors that contribute to the model equally well and hence inclusion probabilities for each single variable tend to be diluted when there is correlation among the predictors.

To conclude this review of implementation of the Lasso for GWAS, we want to point the reader's attention to two other contributions that use Lasso approaches in hierarchical fashion. He and Lin (2011) implement Fan and Lv (2008)'s results on Sure Independence Screening, suggesting that marginal tests can be used to exclude regressors. Note that this, while similar to the variable elimination rules we described based on KKT, represents a different approach: the goal is not simply to solve the Lasso optimization problem, but to find the correct model, and once irrelevant variables are excluded, multiple model selection strategies could be implemented. Wu et al. (2010) also have studied the power of marginal regression as well as explored how to address challenges 2 and 3 while relying on the computational efficiency of the Lasso. In their proposal, *Screen and clean*, they advocate for the use of separate datasets to select models and try to evaluate their validity.

9.5 Bayesian approaches to GWAS

We now turn to the description of Bayesian approaches, whose specific appeal and challenges can be best understood by comparing them

with the approaches described this far. If Lasso-like approaches excel in tackling the computational challenge (1), Bayesian approaches are particularly attractive for their ability to provide the user with inferential results on (2) model selection and (3) individual variables inclusion, accounting for the search on the vast model space.

In a Bayesian framework, it is required that prior probabilities be specified on the different models M under consideration, as well as on the parameters of interest β and σ (the variance of the errors). For definiteness, we will consider for most of the chapter the framework introduced by George and McCulloch (1993), which is the most commonly used in GWAS. For each possible regressor X_k, let γ_k be an indicator variable equal to 1 when variable k belongs to the linear model for the mean of the phenotype, and 0 otherwise. Prior probability on the models M can be specified in terms of prior probabilities on the vector γ and prior probabilities on β conditional on γ are generally easier to describe. The posterior distribution will be defined on the parameters, as well as on γ, enabling the researcher to summarize evidence in favor of different model choices in a coherent and easily interpretable fashion.

Such statements can often be made when comparing Bayesian and frequentists approaches: the former are said to be more intuitive and interpretable, and it is often stressed how a posterior distribution, once defined, provides a coherent framework within which much of the challenges faced by frequentist inference become mute. While these statements might lead to cheering among the Bayesian lines, they generally do little to convince the skeptics of the opportunity to subscribe to a way of making inference that requires explicit input of prior information, in forms that often might seem unacceptable and inadequate for scientific investigation, where the evidence needs to be accepted and uniformly interpreted by an entire community. Therefore, rather than dwelling on general statements, we would like to make **two remarks**. **(a)** Often, to study what the inference would be adopting a Bayesian approach leads to guidelines and procedures that also have validity in a frequentist framework. **(b)** For the Bayesian results to be interesting, the specification of priors cannot be haphazard, but needs to be honest and intelligent. In large scale problems, a 'good' prior can often be based on very little 'arbitrariness', and mainly reflect implications of the high dimensional nature of the parameter space.

As an example of **(a)**, consider BIC: Schwarz (1978) derived it as an approximation of the posterior distribution of different models as the number of observations $n \to \infty$, and the evidence from the data dom-

inates over priors. Still, it has proven to be a useful model selection criterion in a frequentist framework as well—where one can show that it leads to consistent model choices as $n \to \infty$. Nevertheless, the setting under which BIC was derived does not apply to GWAS. Swartz's approximation is valid as $n \to \infty$, while the number of models is fixed. In GWAS, p, the number of possible regressors, is much larger than n, and the number of models certainly cannot be considered as dominated by n. It is no wonder, then, that the application of of BIC to such high dimensional settings leads to inappropriate model choices. Over the years, a number of modifications of BIC have been proposed: Bogdan et al. (2004) and Frommlet et al. (2012) have recently described one presumed to be appropriate for GWAS. The authors argue that using the traditional BIC approximation in this context, where n cannot be considered so large as to outweigh any prior on the models, is equivalent to give any model equal probability. This might seem harmless enough, but it results in an expected number of $p/2$ regressors to be included in the linear model—which does not correspond to the scientists understanding of the traits' architecture and naturally leads to the choice of too large models (see remark **(b)**). The authors, then, suggest to work with a different and more realistic prior specification, such that the number of regressors in the model has a priori a Binomial probability of parameters (ω, p), and with ω small enough to adequately represent the underlying notion of sparsity. This leads to a modified BIC criteria:

$$\|y - X\beta\|_2^2 + \sigma^2 \|\beta\|_0 (\log n + \log p + d),$$

with $d = -2 \log \omega p$; more recently, the authors propose

$$\|y - X\beta\|_2^2 + \sigma^2 \|\beta\|_0 (\log n + \log p + d) - 2 \log (\|\beta\|_0!),$$

where the last term introduces a penalty that depends on the factorial of the number of non zero coefficients. Beyond the exact form of these criteria, we want to point the readers' attention to the fact that they both bring a $\log p$ term in the penalization parameter, similarly to the results in Foster and George (1994), but with considerably less effort or the need to introduce a new inferential paradigm. Furthermore, the second version of the criteria (Frommlet et al., 2012), under some conditions, controls for FDR across the selected regressors: this Bayesian-derived procedure leads to a selection strategy with properties that can be understood entirely in a frequentist context.

Let us now consider how the high dimensional nature of the parameter space can be taken into account to define useful prior distribu-

tions (as in our remark (b)). To appreciate how a Bayesian model selection approach tackles challenge (3), we need to look more specifically at the prior/posterior distribution of the indicator vector γ. One common choice (that corresponds to the prior on model size just described), is to assume γ_k independent and identically distributed as Bernoulli(ω). The assumption of independence a priori does not reflect the shared knowledge that when the true causal variant is not among the genotyped regressors, multiple neighboring variants could equally well represent it. The hyper-parameter ω represents the sparsity of the model, which in GWAS one expects to be substantial. Typically, researchers feel more comfortable estimating this from the data, rather than specifying an arbitrary value, so that ω is often given a prior distribution. However, substantial care has to be exercised at this stage. While setting $\omega = 0.0001$, for example, might appear arbitrary, it leads to more honest inference than giving ω a uniform "non informative" prior, with the rather 'informative' consequence that the a-priori expected number of significant regressors is $p/2$ (remark (b)). Not only this does not reflect our understanding of the genetic architecture of the trait, but using such priors would hardly result in a posterior that leads to inference immune from the problems of multiple comparison and that accurately reflects selection. To avoid such problems, two recent implementations propose the following. Guan and Stephens (2011) suggest $\log \omega \sim U[a, b]$, with $a = \log 1/p$ and $b = \log V/p$, V being the upper-limit of the number of variables in the model. Wilson et al. (2010) uses $\omega \sim \text{Beta}(1, p/\phi)$, which implies that the expected model size is ϕ, a value that can be chosen appropriately.

Given γ, for σ and β, one typically uses some version of the conjugate priors. For example, in Guan and Stephens (2011), β_k are $iid\, \mathcal{N}(0, \sigma_a^2)$. These authors use a novel prior on the hyper-parameter σ_a^2, to reflect current knowledge on heritabilty. To insure that models with larger number of parameters do not coincide a priori with models with higher heritability, the authors suggest making the prior on σ_a^2 dependent on the total number of non zero elements of γ.

If the prior distributions have been carefully specified, the probabilities $P(\gamma_k|Data)$, obtained averaging out all other parameters from the posterior distribution, provide an intuitive and effective instrument to evaluate the importance of each regressor. $P(\gamma_k|Data)$ summarizes the evidence in favor of X_k, not taken out of context, but accounting for the effects of all other variables; not within the framework of a specific model, but averaging over all possible models. Wilson et al. (2010) provide an

eloquent and detailed discussion of these properties, as well as those of the associated Bayes factors (see also chapter 15). These authors also suggest how to approximate the Bayes factors relative to the inclusion in the model of a single regressor using Laplace approximation. This last remark leads us to discussing the most challenging aspect of the Bayesian approach: computation of relevant quantities of the posterior distribution, which needed to select models, estimate parameters, and provide confidence statements. While these task are often difficulty, the challenge here is exacerbated by the size of the model space.

With enumeration of all possible models not a possibility, Monte Carlo sampling of models according to their probability represents a reasonable approach. Markov chain Monte Carlo (MCMC) methods have been very successful across Bayesian statistics, and they have also been used in the context of GWAS. Guan and Stephens (2011) describe their experience with an MCMC scheme that relies on sampling γ from its posterior, unconditional on β; usage of Gibbs Sampler and Metropolis updates, based on a combination of local and long range moves; and 'smart' proposals for new values of γ. When considering to add a new variable to the model (changing one of the γ_k from 0 to 1), these authors focus on covariates with the strongest marginal association by giving them higher probability in the proposal distribution (ranking of variables is obtained via the Bayes factor comparing the model with no covariates with the model that includes only the variant in question). Wilson et al. (2010) use a combination of approximation and stochastic search algorithms. To calculate the SNP-specific Bayes factors, they use Laplace approximations. They then eliminate from consideration all SNPs with Bayes factor smaller than 1. The distribution of all possible models constructed with the surviving variables is explored using an evolutionary Monte Carlo algorithm. The authors of both papers report satisfactory results, but the time requirements for MCMC and EMC are not comparable to those of Lasso in GWAS setting. In addition, in absence of precise theoretical result on the time to convergence, it is hard to produce software with reliable performance when in use by researchers whose skills set does not include MCMC.

Possibly in reaction to this difficulty, Hoggart et al. (2008) suggest using a Bayesian version of the Lasso. Departing from the model selection framework that we described (which relies on the γ indicators) the authors assign independent shrinkage priors to each of the βs parameters with sharp peaks at zero and are only interested in the mode of the posterior distribution. When using double exponential priors, this results

exactly in the Lasso optimization; the authors consider also the more general normal–exponential–gamma prior and use a coordinate descent algorithm to achieve the solution. Clearly, this approach leads to a computational performance similar to that of the Lasso, but it seems to this writer that most if not all the advantages of a Bayesian models selection have been lost. Li et al. (2011) also propose a Bayesian Lasso for GWAS. Their contribution differs from that of Hoggart et al. (2008) as they a) add a preconditioning step; b) explore the entire posterior distribution using MCMC; c) they show how the amount of the penalty defined by the parameter λ can be estimated from the data. All of these generalizations are reasonable and natural extensions: a) variable elimination steps make sense also in Bayesian settings; b) given the effectiveness of coordinate descent methods, one can presume that Gibbs sampler iterations might work well; c) estimating λ adaptively is really the one advantage of a Bayesian take on the Lasso. Still, the adoption of continuous priors on β_ks results in a posterior that is not easily interpretable from a model selection standpoint. Laplace-type priors induce a sparse parameter estimate only if this is based on the mode of the posterior distribution—which, instead, does not recognize zero as a likely value for the parameters and cannot be used to evaluate the importance of every variable in the model.

Most recently, there has been an effort to put to use variational methods in Bayesian approaches to GWAS. Variational Bayes methods have become popular in the data mining/ learning literature as a computationally attractive approach to approximations of the posterior distributions. Rather than discussing them in general, we focus on their application to GWAS, as explored in Logsdon et al. (2010), Carbonetto and Stephens (2011), and Logsdon et al. (2012). In variational methods, one seeks an approximation of the posterior from within a class of distributions that are easy to manipulate. Consider again the Bayesian model selection setting with the vector β of regression coefficients and γ the associated 0/1 values indicating if each of the explanatory variables belongs to the model or not. Let $P(\beta, \gamma; \Phi)$ be the posterior distribution of the parameters (β, γ) given the data and all other remaining parameters or hyperparameters, collected in the vector Φ. An approximation for $P(\beta, \gamma; \Phi)$ is sought among the class Q of distributions of the form $Q(\beta, \gamma; \phi) = \prod_{k=1}^{p} q_k(\beta_k, \gamma_k; \phi_k)$, with each of the independent

components $q_k(\beta_k, \gamma_k; \phi_k)$ a mixture of Gaussian and mass at zero:

$$q_k(\beta_k, \gamma_k : \phi_k) = \begin{cases} \alpha_k \varphi(\beta_k; \mu_k, s_k^2) & \text{if } \gamma_k = 1 \\ (1 - \alpha_k)\delta_0(\beta_k) & \text{otherwise} \end{cases} .$$

Note that a commonly used prior distribution for β, γ has precisely this form: γ_k are assumed to be independent Bernoulli, and the β_k, conditional on γ, are independent, and Gaussian if $\gamma_k \neq 0$. It is well known that, unless the regressors X are orthogonal, assuming such prior will not result in a posterior in the same functional class. However, the approach discussed in Logsdon et al. (2010) and subsequent papers, is based on finding an approximation for the posterior within this class. The values of (α_k, μ_k, s_k^2) are selected to minimize the Kullback-Leibler distance between $P(\beta, \gamma; \Phi)$ and $Q(\beta, \gamma; \phi)$: a solution to this problem can be generally cast in the framework of calculus of variation, leading to the name "variational methods". In the specific context at hand, a close form solution does not exist, but optimal values for (α_k, μ_k, s_k^2) can be obtained with a few iterative steps. In fact, coordinate descent methods are available, resulting in formulas with clear interpretation (see Carbonetto and Stephens (2011) for an explicit discussion). Because of the functional form selected for $Q(\beta, \gamma : \phi)$, the approximation of the posterior lends itself naturally to model selection.

Contrasting variational Bayes with MCMC is useful to fully understand the advantage/limitation of the two approaches. Generally speaking, MCMC aims to evaluate integrals of the form

$$\mathrm{E}_P(g(\beta, \gamma)) = \int g(\beta, \gamma) P(\beta, \gamma; \Phi) d\beta d\gamma,$$

where $P(\beta, \gamma; \Phi)$ is the true, intractable posterior. The resulting estimators $\hat{\mathrm{E}}_m(g(\beta, \gamma))$ are unbiased and consistent for $\mathrm{E}_P(g(\beta, \gamma))$, that is, as the number of MCMC samples m tends to infinity, the estimation error tends to zero (we are, of course, assuming that the Markov chain indeed converges to the target distribution, and ignoring the fact that this might require an unreasonable number of steps). In variational methods, instead, a few iterative steps are used to obtain an exact solution to an approximation problem resulting in the definition of $Q(\beta, \gamma; \phi)$, the closest distribution to $P(\beta, \gamma; \Phi)$, within a certain class. $Q(\beta, \gamma; \phi)$ is selected to be tractable, and $\mathrm{E}_Q(g(\beta, \gamma))$ is evaluated instead of $\mathrm{E}_P(g(\beta, \gamma))$. Typically, the function $g(\beta, \gamma)$ does not influence the choice of the approximation class: while Q might be the closest to P in KL distance within $\mathcal{Q}$, one does not know how precise this approximation is or what are

the effects of approximating $\mathrm{E}_P(g(\beta,\gamma))$ with $\mathrm{E}_Q(g(\beta,\gamma))$ for a specific g. Therefore, a great deal of caution has to be used in interpreting the results of variational inference. The problem seems particularly acute for our GWAS application. On the one hand, proponents of variational methods insist that it is important to use $Q(\beta,\gamma;\phi)$ as an approximation of the entire posterior $P(\beta,\gamma;\Phi)$, rather than viewing $q_k(\beta_k,\gamma_k;\phi_k)$ as an approximation of the marginal $p_k(\beta_k,\gamma_k;\Phi)$. And yet, in the GWAS model selection, we would precisely want valid marginal statements of the probability with which each of the β_k is different from zero, and these would be based on q_k. On the other hand, it is often noted how approximations as Q tend to be "more compact than the true distribution" (MacKay (2003)); and yet we are especially interested in the amount of variability implied by the posterior–rather than its mode, which can think of exploring with tools as the Lasso.

While Logsdon et al. (2010) pioneered variational methods in GWAS, Carbonetto and Stephens (2011) carry out a more systematic evaluation of the implication of the variational strategy described. In their hands, the posterior inclusion probabilities for the single parameter β_k can be incorrect, even though accurate information on hyper-paramters with global effects on the posterior can be obtained. This is perhaps not surprising, given that inference on the hyper-parameters is based on the entire approximation Q, rather than on the marginal q_k alone. Carbonetto and Stephens (2011) also note that variational methods work well when the regressors are orthogonal: while this result is also not unexpected, it seems that a multivariate method that works well in this case alone is not of great interest–as marginal regression would also perform satisfactorily. More recently, Logsdon et al. (2012) use a Bayesian set-up, variational approximation, and model averaging to propose a statistics to test, in a frequentist framework, the null hypothesis $H_0 : \beta_k = 0$. Simulations suggest that the statistics has good properties, and indeed, the combination of variational approximation and model averaging should address some of the difficulties outlined above. Nevertheless, it is difficult to provide a precise analysis of the latest suggestion, given the complexity of the framework adopted.

9.6 Conclusion

A few years have now passed since the first GWAS were published in 2007 and there has been an effort in the statistical genetics community

to explore the potential of the use of multivariate models, of which we provided a succinct account.

From the statistical perspective, this has been a very stimulating endeavour: it has motivated the development of more efficient algorithms to tackle multivariate regression when $p >> n$, and it has contributed to sharpen the focus on the shortcoming of currently available methodology.

From the genetic perspective, results are perhaps less exciting. While it has become apparent that using multivariate models one better accounts for the overall genetic determinant of the traits, this author would be hard press to state that this has resulted in the identification of novel loci–which is possibly the most coveted measure of success in this area of research. We have cited, in Section 2 of these chapter, a number of publications that underscore how multivariate models based on a large number of SNP improve prediction, and this is certainly the case. Each of the methodology papers we have surveyed also illustrate how, attempting to construct a multivariate model, brings to the attention of the researchers SNPs that would be otherwise overlooked; however, the authors all show a clear restrain in over-interpreting these. Bolder is the recent paper by Yang et al. (2012) who claim that, by using a conditional and joint analysis of summary statistics, additional variants influencing complex traits are identified. Still, as it turns out, the additional variants are all in regions already known to be associated to height (the trait in question) by marginal analysis. Brodsky (2011), a PhD thesis devoted to the analysis of multivariate linear models for GWAS, carefully illustrates how using multiple regressors leads to residuals with better statistical properties and how a combination of lasso and stability selection has the power to identify some location that are overlooked by marginal regression—but possibly at the cost of a false discovery rate that the community is not willing to accept.

This modest applied success might motivate further developments in the theory of statistics. Genovese et al. (2009) have circulated a provocative manuscript advocating to "revisit marginal regression" as having comparable performance to that of the lasso when $p >> n$. It is an interesting question to try to establish under which conditions multivariate linear models outperform marginal regression when model selection is the main objective and $p >> n$.

From the point of view of gene mapping, it remains the question of how to further explore the genetic basis of complex traits. The challenges presented by multivariate analysis will only be exacerbated if sequencing

substitutes genotyping. It is possible that methods of analysis and study design that incorporate more information and reflect genetic structure might help. This author believes that substantial progress in our understanding of the genetic architecture of the traits would come if the very numerous loci identified by GWAS were to be followed up carefully, so that the genes influencing the traits be singled out, together with all their relevant variant. This would clarify what portion of heritability is really explained by the identified loci, and pinpoint to biological pathways, which would inform successive studies aiming to detect subtler effects or effects due to interaction.

References

Alexander, D. H., and Lange, K. 2011. Stability selection for genome-wide association. *Genet. Epidemiol.*, **35**(7), 722–728.

Benjamini, Yoav, and Gavrilov, Yulia. 2009. A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Stat.*, **3**(1), 179–198.

Benjamini, Yoav, Heller, Ruth, and Yekutieli, Daniel. 2009. Selective inference in complex research. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, **367**(1906), 4255–4271.

Berk, R., Brown, L. D., Buja, A., Zhang, K., and Zhao, L. 2011. Valid Post-Selection Inference. *Submitted*.

Birgé, L., and Massart, P. 2001. Gaussian Model Selection. *Journal of the European Mathematical Society*, **3**(3), 203–268.

Bogdan, M., Ghosh, J. K., and Doerge, R. W. 2004. Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, **167**(2), 989–999.

Bogdan, Małgorzata, Chakrabarti, Arijit, Frommlet, Florian, and Ghosh, Jayanta K. 2011. Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.*, **39**(3), 1551–1579.

Brodsky, Janice. 2011. *Block, Pass, Score: A Multivariate Methodology for Genome-wide Association Studies*. PhD thesis, Statistics Department, UCLA.

Broman, Karl W., and Speed, Terence P. 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **64**(4), 641–656.

Candès, Emmanuel J., and Plan, Yaniv. 2009. Near-ideal model selection by ℓ_1 minimization. *Ann. Statist.*, **37**(5A), 2145–2177.

Carbonetto, P., and Stephens, M. 2011. Scalable variationa invference for Bayesian variable selection, and its accuracy in genetic association studies. *Bayesian Analysis*, **6**, 1–42.

Chen, Scott Shaobing, Donoho, David L., and Saunders, Michael A. 1998. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20**(1), 33–61.

Dupuis, J., Brown, P. O., and Siegmund, D. 1995. Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. *Genetics*, **140**(2), 843–856. 7498758.

El Ghaoui, L., Viallon, V., and Rabbani, T. 2010. Safe feature elimination in sparse supervised learning. *Technical Report UC/EECS*, **126**.

Fan, J., and Lv, J. 2008. Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, **70**, 849–911.

Fisher, R. A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edinb.*, **52**, 399–433.

Foster, Dean P., and George, Edward I. 1994. The risk inflation criterion for multiple regression. *Ann. Statist.*, **22**(4), 1947–1975.

Friedman, Jerome, Hastie, Trevor, Höfling, Holger, and Tibshirani, Robert. 2007. Pathwise coordinate optimization. *Ann. Appl. Stat.*, **1**(2), 302–332.

Frommlet, Florian, Ruhaltinger, Felix, Twaróg, Piot, and Bogdan, Małgorzata. 2012. Modified versions of Bayesian Information Criterion for genome-wide association studies. *Computational Statistics and Data Analysis*, **56**(5), 1038–1051.

Genovese, C., Jin, J., and Wasserman, L. 2009. Revisiting Marginal Regression. *ArXiv e-prints*, Nov.

George, E., and McCulloch, R. 1993. Variable Selection via Gibbs Sampling. *JASA*, **88**, 881–889.

Guan, Y., and Stephens, M. 2011. Bayesian variable selection regression for genome-wide association studies and other large scale problems. *Annals of Applied Statistics*, **5**, 1780–1815.

He, Q., and Lin, D. Y. 2011. A variable selection method for genome-wide association studies. *Bioinformatics*, **27**(1), 1–8.

Hoggart, C J, Whittaker, J C, De Iorio, M, and Balding, D J. 2008. Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet*, **4**(7). 18654633.

Kang, H M, Sul, J H, Service, S K, Zaitlen, N A, Kong, S Y, Freimer, N B, Sabatti, C, and Eskin, E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, **42**(4), 348–354. 20208533.

Lander, E. S., Linton, L. M., Birren, B., and Nusbaum, C. *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822), 860–921. 11237011.

Li, J., Das, K., Fu, G., Li, R., and Wu, R. 2011. The Bayesian lasso for genome-wide association studies. *Bioinformatics*, **27**(4), 516–523.

Logsdon, B. A., Hoffman, G. E., and Mezey, J. G. 2010. A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, **11**, 58.

Logsdon, B. A., Carty, C. L., Reiner, A. P., Dai, J. Y., and Kooperberg, C. 2012. A novel variational Bayes multiple locus Z-statistic for genome-wide association studies with Bayesian model averaging. *Bioinformatics*, May.

MacKay, David. 2003. *Information theory, inference, and learning algorithms*. Cambridge University Press.

Manolio, T. A., Collins, F. S., Cox, N. J., and Goldstein, D. B. *et al.* 2009. Finding the missing heritability of complex diseases. *Nature*, **461**(Oct), 747–753. 19812666.

Martinez, J., Carrol, R., Müller, S., Sampson, J., and Chatterjee, N. 2011. Empirical Performance of Cross-Validation with Oracle Methods in a Genomic Context. *The American Statistician*, **65**(4), 223–228.

Meinshausen, N., and Bühlmann, P. 2010. Stability Selection (with discussion). *Journal of the Royal Statistical Society: Series B*, **72**, 417–473.

Morton, N. 1955. Sequential tests for the detection of linkage. *Amer. J. Hum. Genet.*, **7**(3), 277–318. 13258560.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., and Sklar, P. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, **460**(7256), 748–752. 19571811.

Quevedo, J. R., Bahamonde, A., Perez-Enciso, M., and Luaces, O. 2011. Disease Liability Prediction from Large Scale Genotyping Data Using Classifiers with a Reject Option. *IEEE/ACM Trans Comput Biol Bioinform*, Mar. 21383414.

Risch, N., and Merikangas, K. 1996. The future of genetic studies of complex human diseases. *Science*, **273**(5281), 1516–1517. 8801636.

Sabatti, C, Service, S K, Hartikainen, A L, and *et al.*, Pouta. 2009. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*, **41**(1), 35–46. 19060910.

Schwarz, Gideon. 1978. Estimating the dimension of a model. *Ann. Statist.*, **6**(2), 461–464.

The HapMap, Consortium. 2003. The International HapMap Project. *Nature*, **426**(6968), 789–796. 14685227.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., and Taylor, J. amd Tibshirani, R. 2012. Strong Rules for Discarding Predictors in Lasso-Type Problems. *J. R. Stat. Soc. Ser. B*, **74**(2), 245–266.

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, **58**(1), 267–288.

Venter, J. C., Adams, M. D., Myers, E. W., and Li, P. W. *et al.* 2001. The sequence of the human genome. *Science*, **291**(5507), 1304–1351. 11181995.

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C., and Schildkraut, J. M. 2010. Bayesian model search and multilevel inference for SNP association studies. *Ann Appl Stat*, **4**(3), 1342–1364.

Wu, Jing, Devlin, Bernie, Ringquist, Steven, Trucco, Massimo, and Roeder, Kathryn. 2010. Screen and Clean: a tool for Identifying Interactions in Genome-wide Association Studies. *Genet Epidemiol*, **34**, 275–285. 20088021.

Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. 2009. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**(6), 714–721.

Wu, Tong Tong, and Lange, Kenneth. 2008. Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*, **2**(1), 224–244.

Wu, Z., and Zhao, H. 2009. Statistical power of model selection strategies for genome-wide association studies. *PLoS Genet.*, **5**(7), e1000582. 19649321.

Yang, J., Benyamin, B., McEvoy, B. P., and Gordon, S. *et al.* 2010. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, **42**(Jul), 565–569. 20562875.

Yang, J., Ferreira, T., Morris, A. P., and Medland, S. E. *et al.* 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, **44**(4), 369–375.