Nonparametric Bayesian Variable Selection with Applications to Multiple Quantitative Trait Loci Mapping with Epistasis and Gene-Environment Interaction

Fei Zou\textsuperscript{1}, Hanwen Huang\textsuperscript{1}, Seunggeun Lee\textsuperscript{1}, and Ina Hoeschele\textsuperscript{2}

\textsuperscript{1} Department of Biostatistics
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27599, U.S.A.

\textsuperscript{2} Virginia Bioinformatics Institute and Department of Statistics, Virginia Tech
Blacksburg, Virginia, 24061, U.S.A.
Bayesian Variable Selection for QTL Mapping with Epistasis

Keywords: Gaussian Process; Higher order interaction; MCMC; Multiple QTL.

Fei Zou
Department of Biostatistics
The University of North Carolina at Chapel Hill
4115D McGavran-Greenberg Hall, CB# 7420
Chapel Hill, NC 27599
(Tel) 919-843-4822
(Fax) 919-966-3840
(e-mail) fzou@bios.unc.edu
Abstract

The joint action of multiple genes is an important source of variation for complex traits and human diseases. However, mapping genes with epistatic effects and gene-environment interactions is a difficult problem because of relatively small sample sizes and very large parameter spaces for quantitative trait locus models that include such interactions. Here we present a nonparametric Bayesian method to map multiple quantitative trait loci (QTL) by considering epistatic and gene-environment interactions. The proposed method is not restricted to pair-wise interactions among genes, as is typically done in parametric QTL analysis. Rather than modeling each main and interaction term explicitly, our nonparametric Bayesian method measures the importance of each QTL, irrespective of whether it is mostly due to a main effect or due to some interaction effect(s), via an unspecified function of the genotypes at all candidate QTL. A Gaussian process prior is assigned to this unknown function. In addition to the candidate QTL, non-genetic factors and covariates, such as age, gender, and environmental conditions, can also be included in the unspecified function. The importance of each genetic factor (QTL) and each non-genetic factor/covariate included in the function is estimated by a single hyperparameter, which enters the covariance function and captures any main or interaction effect associated with a given factor/covariate. An initial evaluation of the performance of the proposed method is obtained via analysis of simulated and real data.

keywords Gaussian Process; Higher order interaction; MCMC; Multiple QTL.

1 Introduction

Traits showing continuous variation are called quantitative traits and are typically controlled by multiple genetic and non-genetic factors, which tend to have relatively small effects individually. Crosses between inbred lines produce suitable populations for QTL mapping and are available for agricultural plants and for animal (e.g., mouse) models of human diseases. Such crosses are often used to detect QTL. For these inbred line crosses, uniform genetic backgrounds, controlled breeding schemes and controlled environment ensure that there is little or no confounding of uncontrolled sources of variability with genetic effects. The potential for such confounding complicates and limits the analysis and interpretation of human data. Because of the homology between humans and rodents, rodent models can be extremely useful in advancing our understanding of certain human diseases. In the past two decades, various statistical approaches have been developed to identify QTL in inbred line crosses (see for example Doerge (1997) for review). To perform QTL mapping (identification), a large number of candidate positions (candidate QTL) along the genome are selected. These candidate QTL may all be located at genetic markers (positions of sequence variants in the genome where the genotypes of all individuals in a mapping population can be measured), or also in between markers if the marker density is not high. QTL mapping may then be performed by considering one candidate QTL at a time or multiple candidate QTL simultaneously. For inbred line crosses with low marker density and considering a single candidate QTL at a time, the interval mapping method was proposed by Lander and Botstein (1989). However, these authors showed that interval mapping tends to identify a “ghost” QTL located in between two actual linked QTL if there exist two or more closely linked QTL. This problem can be reduced or eliminated in two ways: (1) By using composite interval mapping (Zeng, 1994 and Jansen and Stam, 1994) which
still performs a one-dimensional QTL search but conditional on the genotypes at a pair of markers flanking the marker interval containing the current QTL, to absorb the effects of background (non-target QTL) outside of the target interval. (2) By performing multiple QTL mapping, where two or more QTL are mapped simultaneously. Furthermore, if several QTL affect a quantitative trait mostly through their interactions (epistasis) while having non-existent or weak main effects, then interval mapping or single marker analysis will fail to detect such QTL. QTL interactions may not be limited to pair-wise interactions. Marchini et al. (2005) have shown by simulation that searching for three loci jointly in the presence of a three-way interaction is more powerful than searching for a single or a pair of QTL. There are various different implementations of multiple QTL mapping. Most methods still only perform pairwise searches, with and without epistasis. The most recent methods are based on Bayesian variable selection and consider a group of candidate QTL or all candidate QTL in the genome simultaneously (e.g., Yi et al., 2007). These methods are typically still limited to pairwise interactions among QTL and do not consider gene-environment interactions.

The identification of QTL can be viewed as a very large variable selection problem: for \( p \) candidate QTL, with \( p \) typically in the hundreds or thousands and sample size in the low hundreds, there are \( 2^p \) possible main effect models, \( \binom{p}{2} \) possible two-way interactions and \( \binom{p}{k} \) possible higher order (\( k > 2 \)) interactions. For inbred line crosses, where multiple QTL mapping models can be represented as multiple linear regression models, classical variable selection methods such as forward and step-wise selection (Broman and Speed, 2002) have been used in searching for main and two-way interaction effects. Bayesian analysis implemented by Markov chain Monte Carlo (MCMC) and based on the composite model space framework (Godsill, 2001, 2003) has been introduced to genetic mapping (Yi, 2004). Well-known Bayesian variable selection methods such as reversible jump MCMC (Green, 1995) and stochastic search variable selection (SSVS) (George and McCulloch, 1993) are special cases. SSVS and similar methods employ mixture priors for the regression coefficients, which specify different distributions for the coefficients under the null (effect negligible) and alternative (effect non-negligible) hypotheses. The marginal posterior probabilities of the alternative hypotheses can be used to identify a subset of important parameters based on Bayesian multiple comparison rules, including the median probability model (with a threshold of 0.5) and Bayesian False Discovery Rate control (e.g., Müller, et al., 2006).

An alternative to variable selection with mixture priors is classical and Bayesian shrinkage-or penalty-based inference. For the classical approach of penalized regression, while an \( L_2 \) based shrinkage method (ridge regression) cannot perform variable selection, other methods, in particular the \( L_1 \) based lasso of Tibshirani (1996) and later lasso extensions are capable of performing variable selection by reducing the effects of unimportant variables effectively to zero. The lasso has been applied to parametric, regression based QTL mapping (Yi and Xu 2008). The penalized regression methods can be interpreted as Bayesian regression models with particular sparsity priors imposed on the regression coefficients (Park and Casella 2008).

Regression methods are also used for association mapping in human populations. Recently, Kwee et al. (2008) proposed a semiparametric regression based approach for candidate regions in human association mapping, where a quantitative trait is regressed on a nonparametric function of the tagSNP genotypes within a region. They analyzed a (small) subset of the genome and tested for the joint significance of the subset. Their method potentially can be used to model interactions
among SNPs and covariates. However, Kwee et al. (2008) fit their model using least-squares kernel machines, a dimension reducing technique that is identical to an analysis based on a specific linear mixed model. Model selection for different types of kernels and different sets of variables is performed using criteria such as AIC (Akaike, 1974) and BIC (Schwartz, 1978), which may not be appropriate or feasible in large-scale, sparse variable selection situations.

We (Huang et al., 2010) have recently developed a Bayesian semi-parametric QTL mapping method, where non-genetic covariate effects are modeled non-parametrically. This method was implemented via MCMC, and a Gaussian process prior (O’Hagan, 1978; Neal, 1996, 1997) was placed on the unknown covariate function. The Gaussian process is particularly well suited for curve estimation due to its flexible sample path shapes. This method allows one or more non-genetic covariates to have an arbitrary (nonlinear) relationship with the phenotype. Another strong advantage of the Gaussian process is its ability to deal with high dimensional data compared to other non-parametric techniques such as spline regression (Wahba, 1984; Heckman, 1986; Chen, 1988; Speckman, 1988; Cuzick, 1992; Hastie and Loader, 1993 and Mays, 1995). There has been a growing interest in using Gaussian processes as a unifying framework for studying multivariate regression (Rasmussen, 1996), pattern classification (Williams and Barber, 1998), and hierarchical modeling (Menzefricke, 2000). In this paper, we build on this work and propose a non-parametric Bayesian method for multiple QTL mapping by including not only non-genetic covariates but also all candidate QTL in the unknown function. A Gaussian process prior (GPP) is again placed on the unknown function, and a variable selection approach is implemented for the hyperparameters of the GPP (one for each QTL and non-genetic covariate). Here, we rely on mixture priors and MCMC implementation, and we focus on linkage mapping in inbred line crosses, while in ongoing and future work we are considering shrinkage priors, deterministic algorithms, and association mapping. Our application of the GPP differs from ”standard” applications in that the QTL covariates included in the unknown function are discrete, not continuous, with a small number (2 or 3) of possible values (the genotype codes). The goal of using a GPP here is not curve or response surface modeling but rather high-dimensional variable selection (QTL and non-genetic covariates) with a method requiring only a single parameter for each variable while accounting for any multi-way interactions among the candidate variables.

To improve current methods for linkage mapping in inbred line crosses and for association analysis of human populations, we need to be able to detect QTL irrespective of whether they act mostly through main effects, interactions with other QTL, or interactions with environment. Fitting a parametric model including all these potential effects for a genome-wide search would substantially increase the multiple testing problem, in addition to being computationally extremely demanding. Here we offer an alternative. We will show that our non-parametric Bayesian method can identify QTL irrespective of whether they act through main effects, interactions with other QTL, or with environmental factors. This method cannot identify the source(s) of a QTL’s importance (main or interaction effects involving this QTL). Therefore, once a small number of important QTL have been identified in a genome-wide scan, then these QTL can be further analyzed with detailed parametric models to determine the source(s) of their importance.

The remainder of the paper is organized as follows. We first present the nonparametric multiple QTL model and outline the MCMC sampler in Section 2. Simulation results and the analysis of a real data set are presented in Section 3. We end the paper with a discussion and conclusions in
2 Nonparametric Regression with Gaussian Process

2.1 Model and prior

For the $i$th individual, we observe (i) the genotype codes $x_i = \{x_{ik}\}_{k=1}^p$ at $p$ markers, where $x_{ik}$ is the genotype code at the $k$th marker; (ii) $q$ non-genetic covariates or factors $t_i = \{t_{ij}\}_{j=1}^q$, where $t_{ij}$ is the value of non-genetic covariate $j$; and (iii) the phenotype or trait value $y_i$. The primary goal of the analysis is to map QTL (also loosely referred to as "genes") associated with the phenotype. Assuming for simplicity of presentation that the set of candidate QTL is the set of markers, the problem reduces to identifying which markers influence the phenotype through their genotypes. We have considered the following semiparametric QTL mapping model in Huang et al. (2010):

$$y_i = \sum_{k=1}^{p} x_{ik} \beta_k + \eta(t_{i1}, \cdots, t_{iq}) + e_i, \quad i = 1, \cdots, n,$$

where $\eta$ is an unknown function of non-genetic covariates which we modeled via a Gaussian process prior; $\beta_k$ is the partial regression coefficient associated with the $k$th marker; and $e_i$ is a random error with distribution $N(0, \sigma_e^2)$. Model (1) only considers main QTL effects, which of course can be extended to pair-wise interactions by including the terms $\sum_{j \neq k} x_{ij} x_{ik} \beta_{jk}$, into (1), and similarly to higher-order interactions. The explicit modeling of interactions among genes causes an increase in the number of parameters in (1) which is exponential in the order of the interactions considered. Consequences are computational difficulties and poor inferences due to small sample sizes. To overcome this problem, we can alternatively consider the following model:

$$y_i = \eta(x_{i1}, \cdots, x_{ip}, t_{i1}, \cdots, t_{iq}) + e_i, \quad i = 1, \cdots, n. \tag{2}$$

This model is flexible and considers all interactions among genes and gene-environment interactions non-explicitly. For example, if the unknown function $\eta(x_{i1}, \cdots, x_{ip}, t_{i1}, \cdots, t_{iq})$ equals $x_{i1}x_{i2} + x_{i3}t_{i1}$, then equation (2) non-explicitly models the two way interaction between genes 1 and 2 and the gene-environmental interaction between gene 3 and environmental covariate 1. Let $\eta_i = \eta(x_{i1}, \cdots, x_{ip}, t_{i1}, \cdots, t_{iq})$ and define $\eta = \{\eta_i\}_{i=1}^n$.

To estimate $\eta$, we again assume that $\eta$ has a Gaussian Process prior (as in Huang et al., 2010) with mean 0 and with a covariance matrix $\Sigma$ whose element $ii'$ $(i \neq i')$ associated with individuals $i$ and $i'$ is

$$\Sigma_{ii'} = \text{Cov}[\eta_i, \eta_{i'}] = \xi \exp \left( -\sum_{k=1}^{p} \rho_{xk}^2 (x_{ik} - x_{i'k})^2 - \sum_{j=1}^{q} \rho_{tj}^2 (t_{ij} - t_{i'j})^2 \right), \tag{3}$$

where $\xi$, the $\rho_{xk}^2$s and the $\rho_{tj}^2$s are hyperparameters. This is the most commonly employed stationary covariance function for a Gaussian process (a detailed presentation of Gaussian processes with many valid covariance functions is in Abrahamsen (1997); see also Mackay (1998)). Hyperparameter $\xi$ defines the vertical scale of variations, i.e., controls the magnitude of the exponential part.
Hyperparameters $\rho_{x_k}^2$ and $\rho_{t_j}^2$ are related to length scales which characterize the distance in that particular direction over which $y$ is expected to vary significantly. When for example, $\rho_{x_k}^2 = 0$, then $\eta$ is expected to be essentially a constant function of variable (gene) $x_k$, which is therefore deemed irrelevant (Mackay, 1998). When $\rho_{x_k}^2$ is large, then the resulting function has a short characteristic length and will vary rapidly along the corresponding axis of $x_k$, indicating that variable $x_k$ is of high importance. Similarly, $\rho_{t_j}^2$ indicates the importance of non-genetic covariate $j$ in combination with the genetic factors and other non-genetic covariates.

The original papers on the Gaussian process (Mackay, 1998; Neal, 1997) did not view this method as an approach for variable selection and imposed an inverse Gamma prior on the $\rho^2$ parameters. Though $\rho_{x_k}^2$ does provide information about the relevance of any QTL $k$ with values near zero indicating an irrelevant QTL (similar to the parameters $\beta_k$ in the parametric linear QTL model (1)), determining which $\rho_{x_k}^2$s are significantly non-zero is challenging. It is convenient to represent the hyperparameters $\rho_{x_k}^2$ in terms of their reciprocals, defined to be $\tau_{x_k} = 1/\rho_{x_k}^2$ and $\tau_{t_j} = 1/\rho_{t_j}^2$. We perform Bayesian variable selection by imposing a Gamma mixture priors on the parameters $\tau_{x_k}$ and $\tau_{t_j}$. We introduce the latent variables $\gamma_{x_k}$ ($\gamma_{x_k} = 0$ or 1) and $\gamma_{t_j}$($\gamma_{t_j} = 0$ or 1). Then the Gamma mixture priors for the QTL associated parameters are represented as

$$p_{x_k} \sim Be(p_{x_k}|a_{x_0}, b_{x_0}),$$
$$P(\gamma_{x_k} = 1) = 1 - P(\gamma_{x_k} = 0) = p_{x_k},$$
$$\tau_{x_k} \sim (1 - \gamma_{x_k})Ga \left( \frac{\alpha_{x_0}}{2}, \frac{\alpha_{x_0}}{2\mu_{x_0}} \right) + \gamma_{x_k}Ga \left( \tau_{x_k} \left| \frac{\alpha_{x_1}}{2}, \frac{\alpha_{x_1}}{2\mu_{x_1}} \right. \right). \tag{4}$$

Here $Be(p|a, b)$ represents the Beta density $p^{a-1}(1-p)^{b-1}/B(a, b)$, $Ga(\tau|a, b)$ represents the Gamma density $\tau^{a-1}\exp(-\beta\tau)b^a/\Gamma(a)$, with $E(\tau) = \mu$ and $\beta = \frac{\alpha}{2\mu}$, and $(\alpha_{x_0}, \mu_{x_0}, \alpha_{x_1}, \mu_{x_1}, a_{x_0}, b_{x_0})$ are hyperparameters to be specified or inferred. Similarly, for the non-genetic covariate associated parameters, we assume the mixture priors

$$p_{t_j} \sim Be(p_{t_j}|a_t, b_t),$$
$$P(\gamma_{t_j} = 1) = 1 - P(\gamma_{t_j} = 0) = p_{t_j},$$
$$\tau_{t_j} \sim (1 - \gamma_{t_j})Ga \left( \tau_{t_j} \left| \frac{\alpha_{t_0}}{2}, \frac{\alpha_{t_0}}{2\mu_{t_0}} \right. \right) + \gamma_{t_j}Ga \left( \tau_{t_j} \left| \frac{\alpha_{t_1}}{2}, \frac{\alpha_{t_1}}{2\mu_{t_1}} \right. \right). \tag{5}$$

Note that here $\mu_{x_0}$, $\mu_{t_0}$, $\mu_{t_1}$, and $\mu_{t_1}$ are the means of the two Gamma distributions in (4) and in (5), respectively. Setting $\mu_{x_0}$ (as well as $\mu_{t_0}$) to a large value ensures that if $\gamma_{x_k} = 0$, then $\rho_{x_k}$ will take on very small values, and thus the corresponding variable is irrelevant. In contrast, setting $\mu_{x_1}$ (as well as $\mu_{t_1}$) to a smaller value ensures that if $\gamma_{x_k} = 1$, then a non-zero value of $\rho_{x_k}$ will be included in the model.

Define $\tau_\xi = 1/\xi^2$, $\tau_e = 1/\sigma_e^2$ and let the prior distributions of the two parameters be Gamma and given by

$$p(\tau_\xi) = \frac{(\alpha_\xi/2\mu_\xi)^{-\alpha_\xi/2}}{\Gamma(\alpha_\xi/2)}(\tau_\xi)^{-\alpha_\xi/2-1}\exp(-\tau_\xi\alpha_\xi/2\mu_\xi) \tag{6}$$
$$p(\tau_e) = \frac{(\alpha_e/2\mu_e)^{-\alpha_e/2}}{\Gamma(\alpha_e/2)}(\tau_e)^{-\alpha_e/2-1}\exp(-\tau_e\alpha_e/2\mu_e). \tag{7}$$
Values for the parameters \((\alpha_{\xi}, \mu_{\xi}, \alpha_{e}, \mu_{e})\) are chosen prior to analysis.

The Gaussian process was originally proposed for modeling curves with continuous covariates, where the smoothness assumption on Gaussian process guarantees the smoothness of the estimated curves. However, in QTL mapping and other similar genetics analysis (Kwee et al. 2008), the primary goal is to map genes. The violation of the continuity assumption may be highly influential on the QTL effect estimation, but since the QTL effect estimation is only a secondary task in QTL mapping, the discreteness of genetic variables is less of a concern. As one extreme example, when only one gene is included, the Gaussian process model is equivalent to the random effect model where the genetic effect is treated as random with a normal distribution.

2.2 MCMC algorithm for posterior computation.

Inference is based on the joint posterior distribution of the unknown parameters \((\tau_{x1}, \cdots, \tau_{xp}, \tau_{t1}, \cdots, \tau_{tq}, \tau_{\xi}, \tau_{e})\) and the unknown function vector \(\eta\), conditional on the phenotype \(y\), covariate \(t\) and marker \(x\) data. One potential problem in working with this joint posterior arises due to the discrete nature of the marker data: When the number of significant markers (i.e. markers with distinctly non-zero \(\rho_{xk}\)) is small, then the covariance matrix of \(\eta, \Sigma\), may become (nearly) singular, because multiple individuals will share the same genotype configuration at these few markers. In this case the performance of the method deteriorates, and we therefore prefer to work with the joint posterior of the unknown parameters, or the joint posterior marginalized with respect to \(\eta\). Because of the normal prior on \(\eta\), this marginalization is equivalent to substituting the likelihood function of \(y\) conditional on \(\eta\) by the unconditional likelihood of \(y\), or

\[
y \sim N(0, \Sigma_y) \text{ where } \Sigma_y = \Sigma + \sigma_e^2 I. \tag{8}
\]

where \(\Sigma_y\) is non-singular even if \(\Sigma\) is singular. We compared inferences based on the joint posterior of the unknown parameters and \(\eta\) versus the joint posterior of the parameters (using the same simulated data as presented below) and found the latter to provide clearly superior results. Therefore, from here on we only consider the marginalized posterior.

Most of the posterior computation is quite straightforward, and details can be found in the Supplement section. Below we describe an efficient sampling scheme, hybrid MCMC, that is essential for dealing with the large-scale QTL mapping data.

Let \(\upsilon = (\log(\tau_{x1}), \cdots, \log(\tau_{xp}), \log(\tau_{t1}), \cdots, \log(\tau_{tq}), \log(\tau_{\xi}), \log(\tau_{e}))\). Due to the complexity of the covariance form (3), one cannot sample from the full conditional posterior distributions of \(\upsilon\) directly. The Metropolis-Hastings algorithm could be used with some proposal distribution, but it would explore the region of high probability by an inefficient random walk. To overcome this problem, the hybrid Monte Carlo method was proposed for sampling the hyperparameters in Gaussian process regression (Neal, 1993, 1996; Rasmussen, 1996; Barber and Williams, 1997), and we adopt this approach here. The hybrid Monte Carlo method merges the Metropolis-Hastings algorithm with sampling techniques based on dynamics simulation.

In order to sample the \(p+q+2\) elements of vector \(\upsilon\) from their posterior distribution \(p(\upsilon|y, \theta_{-\upsilon})\), we consider a physical system including \(p+q+2\) particles with the coordinate of the \(i\)th particle being \(v_i\). The potential energy of this system is defined in such a way that \(E(\upsilon) = -\log p(\upsilon|y, \theta_{-\upsilon})\). To
allow the use of the dynamical method, we introduce a "momentum" variable, \( \phi \), which has \( p+q+2 \) real-valued components, \( \phi_i \), in one-to-one correspondence with the components of \( \nu \). The kinetic energy of this system is defined as \( K(\phi) = \frac{1}{2} \sum_{i=1}^{p+q+2} \phi_i^2 \). Therefore, sampling \( \nu \) from \( p(\nu) = e^{-E(\nu)} \) is equivalent to sampling \( (\nu, \phi) \) from \( p(\nu, \phi) = e^{-E(\nu) - K(\phi)} \) by simply ignoring the momentum \( \phi \). The canonical distribution over \( (\nu, \phi) \) is defined to be \( p(\nu, \phi) = e^{-\mathcal{H}(\nu, \phi)} \), where \( \mathcal{H}(\nu, \phi) = E(\nu) + K(\phi) \) is the "Hamiltonian" function, which gives the total energy of the system. It is well known in physics that the evolutions of such a canonical dynamical system through fictitious time \( s \) are governed by the following differential equations:

\[
\frac{d\nu_i}{ds} = \frac{\partial \mathcal{H}}{\partial \phi_i} = \phi_i, \quad \frac{d\phi_i}{ds} = -\frac{\partial \mathcal{H}}{\partial \nu_i} = -\frac{\partial E}{\partial \nu_i}.
\]

(9)

By simulating this dynamical system, the transitions of the Markov chain in the hybrid Monte Carlo method take place as follows:

a). Starting from the current state \( (\nu(s), \phi(s)) \), perform \( L \) steps based on the discretized equation (9) with step size \( \epsilon \), resulting in the state \( (\nu^*, \phi^*) = (\nu(s + L\epsilon), \phi(s + L\epsilon)) \). A single step from \( s \) to \( s + \epsilon \) can be explicitly written as

\[
\phi_i(s + \frac{\epsilon}{2}) = \phi_i(s) - \frac{\epsilon}{2} \frac{\partial E}{\partial \nu_i}(\nu(s)),
\]

(10)

\[
v_i(s + \epsilon) = v_i(s) + \epsilon \phi_i(s + \frac{\epsilon}{2}),
\]

(11)

\[
\phi_i(s + \epsilon) = \phi_i(s + \frac{\epsilon}{2}) - \frac{\epsilon}{2} \frac{\partial E}{\partial \nu_i}(\nu(s + \epsilon)).
\]

(12)

b). With probability \( \min(1, \exp[\mathcal{H}(\nu, \phi) - \mathcal{H}(\nu^*, \phi^*)]) \), accept the new state \( (\nu, \phi) = (\nu^*, \phi^*) \); otherwise reject the new state and retain the old state with negated momentum \( (\nu, \phi) = (\nu, -\phi) \).

c). Update the total energy of the system by perturbing the momenta according to \( \phi_i = c\phi_i + z_i\sqrt{1 - c^2} \) for all \( i \), where \( z_i \) is drawn randomly from the standard normal distribution. The momentum causes the particle to continue in a consistent direction until a region of high energy (low probability) is encountered. Following Rasmussen (1996), we set \( \epsilon = 0.5n^{-1/2} \) and \( c = 0.95 \).

3 Simulation and Real Data Analysis

3.1 Simulation of multiple QTL models with or without higher order interactions.

We simulated a backcross population with 200 individuals and a single chromosome with 151 evenly spaced markers at 5cM intervals. To investigate the ability of the nonparametric Bayesian multiple QTL analysis based on (2) to map higher order interacting QTL that have no main effects, we simulated four interacting QTL without main effects and without lower order interactions. The four simulated QTL are located at markers 9, 39, 69 and 99, respectively. The simulated function \( \eta \) equals \( \eta(x_{i1}, \cdots, x_{151}) = x_{i9}x_{i39}x_{i69}x_{i99} \), where the \( x_{ik}, \ k = 9, 39, 69, 99 \), are the genotype codes (1 and -1) of the four simulated QTL of individual \( i \) and \( \sigma_c^2 = 1 \). The total heritability of this model is 50%.
For the analysis, we set $\alpha_0 = \alpha_t = \alpha_x = \alpha_c = 0.5$, $C = 100$, and $\mu_\xi = \mu_c = 400$. We also set $a_{x,\gamma} = a_{t,\gamma} = 0.95$ and $b_{x,\gamma} = b_{t,\gamma} = 0.05$, so that the prior probabilities that each variable (QTL or non-genetic covariate) is relevant or irrelevant for the phenotype are 0.05 and 0.95, respectively. The left panel of Figure 1 provides a plot of the posterior mean estimate of the hyperparameter $\gamma_{xk}$ for each marker $k$ versus the marker position from the general model (2). As we hoped, the estimates of the hyperparameters associated with the true QTL markers are much larger than the estimates of the hyperparameters associated with the irrelevant markers, and all four, purely interacting QTL were identified based on the marginal posterior probability of inclusion $> 0.5$. Selecting all variables with marginal posterior probability of inclusion $> 0.5$ produces the median probability model which is known to frequently correspond to the optimal predictive model while often differing from the highest probability model.

For comparison, we also ran R/qtlbim (www.qtlbim.org/), a popular software for Bayesian multiple QTL mapping developed by Yandell et al. (2007). R/qtlbim is an extensible, interactive environment for parametric Bayesian analysis of multiple interacting QTL models for experimental crosses (limited to two-way interactions). The results are summarized in the right panel of Figure 1. In the R/qtlbim manual (Banerjee et al 2008), the following criteria are suggested for judging the significance of QTL: weak support if the Bayes Factor (BF) falls between 3 and 10; moderate support if the BF falls between 10 and 30; strong support if the BF $> 30$, and no support if BF $< 3$. According to these criteria, R/qtlbim fails to detect any QTL simulated.

To further test the method, we then simulated data sets containing QTL that have only main effects ($\eta_i(x_i, \cdots, x_i) = 0.25(x_{i21} + x_{i51} + x_{i81} + x_{i111})$) or main and two-way interaction effects ($\eta_i(x_i, \cdots, x_i) = 0.25 \cdot (x_{i21} + x_{i31} + x_{i21}x_{i51} + x_{i81}x_{i111})$). These are situations which R/qtlbim was specifically designed for. All other simulation parameters remained the same as in the previous simulation. Both models have a total heritability of 20% ($\sim 5\%$ heritability for each QTL). We used the same priors as in the previous simulation for the nonparametric method, and as before we used the default priors of R/qtlbim. Figure 2 summarizes the results for the additive model. Our nonparametric method detects three of the four QTL based on the marginal posterior probability of inclusion $> 0.5$ and misses one QTL. Similarly, R/qtlbim detects the same three QTL with weak support ($3 < BF < 10$). For the model with the two-way interactions, results were very similar and are therefore not shown.

Our method and R/qtlbim use different criteria (median inclusion probability versus BF based selection) for the selection of a relevant subset of QTL. This difference is confounded with the comparison between the nonparametric and the linear parametric method in terms of their ability to detect existing QTL correctly. To overcome this problem, we varied the cut-offs imposed on the inclusion probability and BF respectively, for declaring the significance of QTL, and we generated receiver operating characteristic (ROC) curves. For each scenario (four-way interaction, additive, additive plus two-way interaction models as above), we ran 100 simulations. Instead of fixing the positions of the four simulated QTL, we uniformly generated their positions subject to the restriction that any pair of QTL had to be at least 10cM apart. We divided the whole genome into non-overlapping 10cM wide intervals. For a given cut-off (on inclusion probability or BF), a significant interval was defined as an interval that contains at least one marker whose significance measure exceeds the cut-off. A significant interval is defined as a true positive if it includes one of the simulated QTL. Otherwise, it is called a false positive. We defined: True positive rate = (# of
significant, true intervals) / (# of significant intervals); false positive rate = (# of significant, false intervals) / (# of significant intervals). The ROC curves up to a false positive rate of 0.1 are given in Figure 3 for all three models simulated. For the four way interaction model, our nonparametric method performed much better than R/qtlbim which essentially failed to detect any QTL. It is interesting to see that our method appears to perform essentially as well as R/qtlbim for the model with both main and two-way interactions. It is even more interesting to find out that our method is superior to R/qtlbim for the main effects model. This is because we ran R/qtlbim by searching for both main effects and two-way interactions simultaneously, even when analyzing the data generated under the pure main effects (additive) model.

3.2 Real data analysis

In addition to the simulation, we tested our method on a real mouse study on obesity, a major risk factor for type II diabetes. To genetically dissect a polygenic mouse model of obesity-driven type II diabetes, Reifsnyder et al. (2000) outcrossed the obese, diabetes-prone, NZO (New Zealand Obese)/HILt strain to the relatively lean NON (Nonobese Nondiabetic)/Lt strain, and then reciprocally backcrossed obese F1 mice to the lean NON/Lt parental strain. They have measured the body weights of 187 backcross males. In addition, inguinal, gonadal, retroperitoneal and mesenteric fat pad weights have also been measured. Stylianou et al. (2006) have studied the fat pad weights using F2 progeny between the SM/J and NZB/BINJ inbred mouse strains. They have identified several QTL associated with the gonadal fat pad weight after adjusting for the total lean body weight (LBWT). Following Stylianou et al. (2006), we first calculated the total fat pad weight as the mesenteric fat pad weight plus twice the sum of the inguinal, gonadal, and retroperitoneal fat pad weights. Then the lean body weight (LBWT) was obtained as the difference between the total body weight and the total weight of the fat pads. We applied our nonparametric Bayesian variable selection method to the Reifsnyder et al. (2000) data. The results are presented in Figure 4. Clearly, two among the 86 predictors (85 markers plus the continuous covariate LBWT) are selected. The first ranked predictor is the covariate LBWT, and the second ranked predictor is marker D4Mit311 located on chromosome 4. Figures 4 strongly indicates a QTL on chromosome 4 while other QTL in the genome are (much) less likely. Further studies based on this observation can be done by investigating the relationship between the phenotype and these two variables in more detail. For each genotype of the QTL identified on chromosome 4, we estimated the weight curve function on LBWT, and the results are reported in Figure 5. From the two estimated curves, there is no clear evidence for an interaction between the QTL and LBWT.

4 Discussion

In this paper, we have proposed a novel nonparametric QTL mapping method where both the genetic as well as non-genetic factors are modeled via a function $\eta$, whose form is unspecified. The advantage of our approach is that it models all potential genetic and non-genetic effects, including main effects and all interaction effects of any order, non-explicitly. It only determines which of the genetic and non-genetic factors are important, on their own through main effects and/or in
combination with other factors. This was achieved by combining the Gaussian Process prior for the unknown function with variable selection. Although in this paper we have assumed that all putative QTL are located at the marker positions, it is straightforward to extend the method to consider any candidate QTL in between marker positions as in Wang et al. (2005) and Huang et al. (2010). A similar nonparametric variable selection procedure has been proposed for computer experiments by Linkletter et al. (2006). These authors mainly focused on identifying active factors having non-linear relationships with the response variable. However, mapping multiple interacting QTL is our main purpose, and our paper appears to be the first one to propose modeling the joint action of multiple QTL with an unknown function having a Gaussian process prior, which accommodates any multi-way interactions. Moreover, Linkletter et al. (2006) only consider a relatively small (< 50) number of continuous covariates while in our paper and in QTL linkage and association mapping in general, there is a large number of discrete marker covariates (hundreds or thousands) in addition to a small number of environmental, continuous covariates or discrete factors. Therefore, an efficient sampling scheme, such as the hybrid MCMC described in the paper, is essential for dealing with these large-scale datasets.

While the linear parametric method in R/qtlbim may have little or no power to detect QTL acting through higher order interactions, computationally it is fast, and it can handle large numbers of individuals and markers. We do not recommend to replace the linear parametric analysis with the nonparametric method, but rather to use it as an additional or preliminary tool to screen the genome for QTL acting through higher-order interactions, which existing QTL mapping methods fail to detect. Once important factors have been identified with the nonparametric method, they then can and should be further analyzed with a detailed parametric model to elucidate the mode of action of the identified QTL (and environmental factors). Application of a detailed parametric method on a genome-wide scale to search for all possible main and interaction effects would dramatically increase the multiple testing problem, in addition to the computational burden, while the nonparametric method can identify all these effects with a single parameter per candidate QTL (and environmental factor).

Our current research focuses on further improving the computational feasibility of our nonparametric method. Our current implementation of the Bayesian Gaussian process prior method, with the mixture priors on the variable selection parameters ($\rho$'s) and using the hybrid Monte Carlo method, allows us to analyze datasets with up to several hundred individuals and several hundred markers, in hours rather than in minutes as with R/qtlbim. A major reason for this increase in computing time is the need to compute the inverse of an $n \times n$ matrix in each MCMC cycle to sample $\nu$. This is particularly a problem for genome-wide association studies (GWAS), for which our nonparametric method is also potentially useful. GWAS typically require larger sample size than linkage studies (in the order of thousands or ten thousands) and several hundred thousand markers (tag SNPs). Further, in the paper, we propose a simple Gibbs sampler for the latent binary variables that code for inclusion of a marker in the covariance function. For QTL mapping with only hundreds of markers, this algorithm works well. For very large $p$, it is likely that the algorithm may not properly mix over the huge sample space, a legitimate concern when we apply the method directly to GWAS data where hundreds of thousands of SNPs are available. Berger and Molina (2005) propose an approach to search for important models through large model space without visiting every model. Their approach provides a nice alternative, which deserves further investigation for GWAS data. Besides alternative sampling schemes we are currently investigating shrinkage priors to
replace the mixture priors and increase computational efficiency, and we are exploring deterministic algorithms to replace MCMC sampling, in particular a conjugate gradient optimization technique to compute the maximum-a-posteriori estimates of the parameters (Rasmussen, 1996). A genome-wide dataset may first be analyzed with the deterministic implementation to screen out many variables (predictors) that are clearly irrelevant. Then the selected, promising subset of predictors (markers, genomic regions) may be re-analyzed by full MCMC which provides much more information than a deterministic mode-finding algorithm. With an initial implementation of shrinkage priors and the conjugate gradient optimization technique we have been able to analyze a data set in a candidate gene association study with approximately 900 participants and 2500 tag SNPs.

Selection of a subset of QTL can be performed based on the estimated marginal posterior probabilities of inclusion with cut-off determined using the median probability model or Bayesian FDR. Alternatively, we may add pseudo-null variable(s) into the model and use the posterior distribution of their $\gamma$'s to guide the variable selection. Linkletter et al. (2006) suggested adding a single pseudo-null variable but running the analysis many times (say 100). For computational reasons, this approach works for their smaller size problems but is computationally very demanding or infeasible in the QTL mapping context. Furthermore, adding a single pseudo-null variable would not work (well) for QTL mapping because marker (null) variables are correlated due to linkage. Wu et al. (2007) proposed a similar idea for variable selection in linear regression models using a set of pseudo-null variables. Their method requires no additional repeated analysis as in Linkletter et al (2006) and can also incorporate the linkage structure of the observed markers into the generation of the pseudo-null variables. We are planning to extend the method of Wu et al (2007) to our Gaussian process based QTL selection methodology.

Much work has been done recently on sparse signal detection in (generalized) linear regression models, where there are two groups of sparsity priors, shrinkage or one-group priors, and mixture or two (multiple) group priors. Here we have employed a mixture prior for the parameters related to variable selection. Our current and future work focuses on further studies and modifications of this mixture prior and of alternative shrinkage priors. The goal of our present paper was to convincingly demonstrate that the nonparametric Bayesian analysis based on the Gaussian process prior is indeed able to detect QTL irrespectively of whether they act on the trait of interest through main effects, any order of interaction among QTL, or interactions of QTL with environmental factors.

5 Acknowledgements

The authors wish to thank the Editor, Associate editor and Referees for their helpful comments and suggestions, which have led to a great improvement of this article. This research was partially supported by NIH grant GM074175.
Literature Cited


Yi, N., Shriner, D., Banerjee, S., Mehta, T., Pomp, D. and Yandell, B.S. 2007 An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. Genetics, 176: 1865-1877


Figure 1: Four way interaction model: (a) Average marginal posterior probabilities of being included in the model for each marker. (b) Estimated marginal Bayes factor for each marker from R/qtlbim. The dashed vertical lines indicate the true QTL positions. The dotted horizontal line is the threshold suggested by R/qtlbim for weak support of significance.
Figure 2: Additive (main effects only) model: see Figure 1 for figure legends.
Figure 3: ROC curves estimated for the three models simulated. The solid lines represent the nonparametric method, and the dotted lines the linear parametric method in R/qtlbim. (a) Additive (main effects) model; b) main plus two-way interaction effects model; c) four-way interaction model.
Figure 4: Nonparametric Bayesian analysis of the backcross mesenteric fat pad data.
Figure 5: Estimated $\eta$ function separated for the QTL genotypes of marker D4Mit311 on Chromosome 4.
Supplement

MCMC algorithm for posterior computation. Let $\theta$ be the vector of all unknown quantities in the model, including the $\tau_{xk}$s, the $\tau_{tj}$s, the $\gamma_{zk}$s, the $\gamma_{tj}$s, $\tau_c$, $\tau_e$ and the latent variables, the $\eta$'s. Furthermore, we define $\theta_{-z}$ as the remaining sub-vector of $\theta$ after removing a parameter or parameter subset $z$ from $\theta$. Below we present the MCMC algorithm for the posterior computation.

For given values of the hyperparameters, which include $\alpha_{x0}, \alpha_{t0}, \alpha_{x1}, \alpha_{t1}, \alpha_{e}, C, \mu_{x}, \mu_{e}, a_{x\gamma}, a_{t\gamma}, b_{x\gamma}$ and $b_{t\gamma}$, we first sample the $\tau_{xk}$s, the $\tau_{tj}$s and $\tau_e$ from their prior distributions. Then we perform the following updating steps many times:

Step 1. Sample $\eta$ directly from its conditional distribution, which is the multivariate normal

$$
\eta | y, \theta_{-\eta} \sim N_n(\mu^*, \Sigma^*),
$$

with the covariance matrix $\Sigma^* = \left( \frac{1}{\sigma^2} I_n + \Sigma^{-1} \right)^{-1}$ and mean vector $\mu^* = \frac{1}{\sigma^2} \Sigma^* y$. For the posterior distribution marginalized with respect to $\eta$ we skip this step and directly go to Step 2.

Step 2: Sample the $\gamma_{zk}$s and the $\gamma_{tj}$s directly from their conditional posterior distributions. The conditional distribution of $\gamma_{zk}$ does not depend on $y$ but rather on $\rho_{zk}$ and several hyper-parameters and is of the form

$$
p(\gamma_{zk} = 1 | \theta_{-\gamma_{zk}}) = \frac{b_{x\gamma} \left( \frac{\alpha_{x1}}{2\mu_{x1}} \right)^{\frac{\alpha_{x1}}{2\rho_{zk}}} \rho_{zk}^{-\frac{\alpha_{x1}}{2}} e^{-\frac{\alpha_{x1}}{2\rho_{zk}}} \frac{\alpha_{x1}}{2\rho_{zk}}}{a_{x\gamma} \left( \frac{\alpha_{x0}}{2\mu_{x0}} + \frac{\alpha_{x1}}{2\mu_{x1}} \right)^{\frac{\alpha_{x0}}{2\rho_{zk}}} \rho_{zk}^{-\frac{\alpha_{x0}}{2}} e^{-\frac{\alpha_{x0}}{2\rho_{zk}}}}.
$$

Similarly, we have

$$
p(\gamma_{tj} = 1 | \theta_{-\gamma_{tj}}) = \frac{b_{t\gamma} \left( \frac{\alpha_{t1}}{2\mu_{t1}} \right)^{\frac{\alpha_{t1}}{2\rho_{tj}}} \rho_{tj}^{-\frac{\alpha_{t1}}{2}} e^{-\frac{\alpha_{t1}}{2\rho_{tj}}} \frac{\alpha_{t1}}{2\rho_{tj}}}{a_{t\gamma} \left( \frac{\alpha_{t0}}{2\mu_{t0}} + \frac{\alpha_{t1}}{2\mu_{t1}} \right)^{\frac{\alpha_{t0}}{2\rho_{tj}}} \rho_{tj}^{-\frac{\alpha_{t0}}{2}} e^{-\frac{\alpha_{t0}}{2\rho_{tj}}}}.
$$

Step 3: Sample the $\tau_{xk}$s, the $\tau_{tj}$s, $\tau_e$ and $\tau_c$ from their posterior distribution, which is proportional to

$$
\frac{1}{|\Sigma^{\frac{1}{2}}|} \exp \left\{ -\frac{1}{2} \eta' \Sigma \eta \right\} \prod_k p(\tau_{zk} | \gamma_{zk}) \prod_j p(\tau_{tj} | \gamma_{tj}) p(\tau_c) \left[ \tau_c^{n/2} \exp \left\{ -\frac{1}{2} \tau_c (y - \eta)' (y - \eta) \right\} p(\tau_e) \right].
$$

See Section 2.2 for the implementation details.

For the posterior distribution marginalized with respect to $\eta$, Equation 16 is changed to

$$
\frac{1}{|\Sigma^{\frac{1}{2}}|} \exp \left\{ -\frac{1}{2} y' \Sigma^{-1} y \right\} \prod_k p(\tau_{zk} | \gamma_{zk}) \prod_j p(\tau_{tj} | \gamma_{tj}) p(\tau_c) p(\tau_e).
$$

One iteration or cycle of our MCMC sampler consists of steps 1 to 3 for the joint posterior of the parameters and $\eta$, or steps 2 and 3 for the marginal posterior of the parameters. When the chain converges to its stationary distribution, the sampled values of all parameters are from their joint posterior distribution. Likewise, the samples of any single parameter represent the marginal posterior distribution of this parameter.