Bayesian QTL Mapping

Fei Zou
Department of Biostatistics
Gillings School of Global Public Health
University of North Carolina-Chapel Hill
fzou@bios.unc.edu

Bayesian QTL Mapping

- One gene one trait
  - very unlikely
- Most traits have a significant environmental exposure component
- The vast majority of biological traits are caused by complex polygenic interactions
  - also context dependent

Bayesian QTL Mapping

- Single QTL Mapping
  - Single marker analysis (Sax, 1923 Genetics)
  - Interval mapping: Lander & Botstein (1989, Genetics)

  Multiple QTL mapping
  - Composite interval mapping (Zeng 1993 PNAS, 1994 Genetics; Jansen & Stam, 1994 Genetics)
  - Multiple interval mapping (Kao et al., 1999 Genetics)
  - Bayesian analysis (Satagopan et al., 1997 Genetics)

Bayesian QTL Mapping

- Most complicated traits are caused by multiple (potentially interacting) genes, which also interact with environment stimuli
- Single QTL interval mapping
  - Ghost QTL (Lander & Botstein 1989)
  - Low power

Bayesian QTL Mapping

- Bayesian methods (Stephens and Fisch 1998 Biometrics; Sillanpaa and Arjas 1998 Genetics; Yi and Xu 2002 Genetic Research, and Yi et al. 2003 Genetics): treat the number of QTLs as a parameter by using reversible jump Markov chain Monte Carlo (MCMC) of Green (1995 Biometrika)
    - change of dimensionality, the acceptance probability for such dimension change, which in practice, may not be handled correctly (Ven 2004 Genetics)

Bayesian QTL Mapping

- Alternative, multiple QTL mapping can be viewed as a variable selection problem
  - Forward and step-wise selection procedures (Broman and Speed 2002 JRSSB)
  - LASSO, etc
  - Bayesian QTL mapping
    - Xu (2003 Genetics), Wang et al (2005 Genetics) Huang et al (2007 Genetics): Bayesian shrinkage
    - Yi et al (2003 Genetics): stochastic search variable selection (SSVS) of George and McCulloch (1993 JASA)
    - Yi (2004 Genetics): composite model space of Godsill (2001 J. Comp. Graph. Stat)
    - Software: R/qtlbim by Yi's group

Bayesian QTL Mapping

- Limitations of existing QTL mapping methods
  - do not model covariates at all or only model covariate effect linearly
  - do not model interactions at all or model only lower order interactions, such as two way interactions

Bayesian QTL Mapping

- The multiple QTL mapping is a very large variable selection problem: for p potential genes, with p being in the hundreds or thousands, there are 2^p possible main effect models, $2^{\binom{p}{2}}$ possible two-way interactions and $2^{\binom{p}{k}}$ possible higher order ($k > 2$) interactions.

Semiparmetric QTL Mapping: Non-linear Covariate Effect

- Goal: allow arbitrary covariate effect in QTL Mapping model
- Semiparametric model:

  $y_i = \sum_j \beta_j x_{ij} + \eta(t_i) + e_i, \quad i = 1, \cdots n \quad$ with $e_i \sim N(0, \sigma^2)$
- Funtion η is unspecified and t_i is a scalar or a vector of covariates
- Deifne $\tilde{x}_j = (x_{1j}, \cdots, x_{nj})'$ and $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})'$

Existing Semi/non-parametric Methods

- Dirichlet process (Muller et al. 1996)
- Splines (Smith and Kohn 1996; Denison et al. 1998 and DiMatteo et al. 2001)
- Wavelets (Abramovich et al. 1998 JRSSB)
- Kernel models (Feng et al 2007)
- Gaussian process (Neal 1997; 1996)
    - Gaussian process priors have a large support in the space of all smooth functions through an appropriate choice of covariance kernel.
    - Gaussian process is flexible for curve estimation because of their flexible sample path shapes
    - Gaussian process related to smoothing spline somehow (Wahba 1978 JRSSB)

Semiparmetric QTL Mapping: Non-linear Covariate Effect

- Model

  $$y_i = \sum_j \beta_j x_{ij} + \eta(t_i) + e_i, \quad i = 1, \cdots n \quad \text{with } e_i \sim N(0, \sigma^2)$$

- The unobserved variables are: function η, error variane σ^2 and the QTL effects $\boldsymbol{\beta} = \{\beta_j\}$

Gaussian Process Prior on η

- A Gaussian process where all possible finite dimensional distributions $\boldsymbol{\eta} = (\eta(t_1), \ldots, \eta(t_n))'$ follow a multivariate normal with $E(\eta(t_i)) = \mu(t_i)$ and $cov(\eta(t_i), \eta(t_j)) = \sigma(t_i, t_j)$ where
  $\mu(t; \alpha) = \alpha_1 f_1(t) + \cdots + \alpha_l f_l(t)$ and
  $\sigma(t_i, t_j) = \frac{1}{\tau} \exp(-\rho(t_i - t_j)^2)$
- τ controls the smoothness of η: when $\tau \to 0$, the posterior mean of η almost interpolates the data while centered around the prior mean function if $\tau \to \infty$.

Prior Specifications

- On $\boldsymbol{\beta} = \{\beta_j\}$: follow SSVS idea such that
  $P(\gamma_j = 1) = 1 - P(\gamma_j = 0) = p_j$
  $\beta_i \mid \gamma_j \sim (1-\gamma_j)N(0, \sigma_0^2) + r_j N(0,{}_j^2\, \sigma_0^2)$
- we also put hyper priors on τ and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_l)$ as
  $\tau \sim$ inverse-Gamma(a_τ, b_τ) and $\alpha \sim N(\alpha_0, \Gamma)$

Conditional Probability

- Define $\mathbf{\Sigma} = \tau Var(\boldsymbol{\eta})$, i.e., prior of $\boldsymbol{\eta}$ is $N(\boldsymbol{\mu}, \mathbf{\Sigma}/\tau)$.
- Define $\mathbf{M}$ as a $k \times l$ matrix with the (i,j)th element equal to $\mu_j(t_i)$.
- Then the conditional posterior distributions are
  - $\beta_j \mid \mathbf{y}, \boldsymbol{\theta}_{-\beta_j} \sim N(\hat{\beta}_j, \frac{\sigma^2}{\mathbf{x}_j' \mathbf{x}_j + \sigma^2/\sigma_j^2})$ [**exercise**: find $\hat{\beta}_j$]
  - $p(\gamma_j = 1 \mid \theta_{-\gamma_j}) = \frac{c_j}{c_j + \frac{p_j}{1-p_j} exp\left\{ \frac{\beta_j^2}{2\sigma_1^2} \left(1 - \frac{1}{c_j^2}\right) \right\}}$
  - $\boldsymbol{\eta} \mid \mathbf{y}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \tau \sim N(\boldsymbol{\mu}^\star, \mathbf{\Sigma}^\star)$
  - ......

  where $\boldsymbol{\theta}$ is the vector of all unknown parameters, $\sigma_j = c_j^{\gamma_j} \sigma_0$, $\boldsymbol{\mu}^\star = \mathbf{\Sigma}^\star \mathbf{D}(\mathbf{U} - \boldsymbol{\mu}) + \boldsymbol{\mu}$, $\mathbf{\Sigma}^\star = (\mathbf{D} + \mathbf{\Sigma}^{-1})^{-1}$, $\mathbf{D} = diag(d_1, \ldots, d_k)/\sigma^2$ with $d_j = \#$ of subjects that have covariates equal to t_j and $\mathbf{U} = \{u_j\}$ with $u_j =$ average of $y_i - \sum_j \beta_j x_{ij}$ for those samples whose covariate equals t_j.

Bayesian Model: Simulation Results

- See lecture4 ppt file

Semiparmetric QTL Mapping II: Multiple Interacting QTLs

- Goal: map multiple potentially interacting QTLs without specifically model all potential main and higher order interaction effects
- Semiparametric model:
  $y_i = \eta(x_{i1}, \cdots, x_{ip}) + e_i, \qquad i = 1, \cdots, n$ with $e_i \sim N(0, \sigma^2)$
- Again function η is unspecified
  - very flexible
  - $\eta(x_{i1}, \cdots, x_{ip}) = x_{i1} + x_{i3}$, or $x_{i1}x_{i3}$ or $x_{i1} + x_{i4}x_{i5}x_{i6}$ ......

Semiparmetric QTL Mapping II: Multiple Interacting QTLs

- Again Gaussian process prior is placed on η function such that
  - $E(\eta_i) = 0$
  - $cov(\eta_i, \eta_k) = \frac{1}{\tau} \exp[-\sum_j \rho_j (x_{ij} - x_{kj})^2]$ where $\eta_i = \eta(x_{i1}, \cdots, \eta_{ip})$ and let $\boldsymbol{\eta} = \{\eta_i\}$.
- Hyperparameters ρ_j related to length scales $\frac{1}{\sqrt{\rho_j}}$ which characterize the distance in that particular direction over which η is expected to vary significantly.
- When $\rho_j = 0$, η is expected to be an essentially constant function of that input variable j, which is therefore deemed irrelevant (Mackay 1998).

Semiparmetric QTL Mapping II: Multiple Interacting QTLs

- The original papers on the Gaussian process (Mackay 1998; Neal 1997) did not view this method as an approach for variable selection and imposed a Gamma prior on the ρ_j parameters. However, ρ_j does provide information about the relevance of any QTL with value near zero indicating an irrelevant QTL.
- For variable selection purpose, we can impose the following mixture priors on ρ_j based on latent variable γ_j:
  - $P(\gamma_j = 1) = p_j$
  - $\tau_j (= 1/\rho_j) \sim (1 - \gamma_j) Ga(\frac{\alpha_0}{2}, \frac{\alpha_0}{2\mu_0}) + \gamma_j Ga(\frac{\alpha_1}{2}, \frac{\alpha_1}{2\mu_1})$

Semiparmetric QTL Mapping II: Multiple Interacting QTLs

- No closed posterior form for τ_js and we resort to Metropolis-Hastings algorithm
  - Direct use of MH is not very efficient for our model and it would explore region of hight probability by an very inefficient random walk
  - hybrid MC method was proposed (Neal 1993,1996: Rasmussen 1996; Barber et al 1997) and we adopt this approach
  - hybird approach merges the MH algorithm with sampling techniques called dynamic simulation based on a "energy" function
- Not computationally feasible for GWAS data where millions of genotypes available on thousands of samples
  - deterministic algorithms to replace MCMC sampling, such as conjugate gradient optimization technique for maximum-a-posterior estimates?

Semiparmetric QTL Mapping: Extensions

- A) non-genetic factors, $\mathbf{z}_i = (z_{i1}, \cdots, z_{iq})$ can be also inluded into η
  $y_i = \eta(x_{i1}, \cdots, x_{ip}, z_{i1}, \cdots, z_{iq}) + e_i, \qquad i = 1, \cdots, n$ with
  $e_i \sim N(0, \sigma^2)$
- B) longitudinal data
  $y_{ij} = \eta(x_{i1}, \cdots, x_{ip}, t_{ij}) + e_{ij}$ with
  $\mathbf{e} = (e_{i1}, \cdots, e_{i,k_i}) \sim N(0, \boldsymbol{\Sigma}_i)$. We have considered cases where
  - $\boldsymbol{\Sigma}_i$ is known up to certain parameters
  - $\boldsymbol{\Sigma}_l$ is unknown and modelled vai the deomposition method of Chen and Dunson (2003, Biometrics)