Hierarchical Evolutionary Stochastic Search with Adaptation

Leonardo Bottolo[1] Sylvia Richardson[2] Enrico Petretto[3]

[1] Institute of Mathematical Sciences, Imperial College, London UK

[2] Centre for Biostatistics, Imperial College, London UK

[3] Division of Clinical Sciences and Division of Epidemiology, Public Health and Primary Care, Imperial College, London UK

Warwick, 17 March 2009

Introduction

- Searching for sparse structure in high dimensional data sets is one of the key challenges for statisticians today
- Variable selection in regression models is another fundamental approach to finding sparse structure
- It has become a research focus in view of the large genetic/genomic data sets that have become available
- In this context, different objectives can be sought:
  - Improving prediction, in particular by using model averaging
  - Better understanding of underlying process

Our statistical objective

- Building parsimonious regression models for high dimensional data sets to facilitate interpretation
- Analyse jointly large number of covariates and multiple outcomes
- Capture adequately the uncertainty related to the role of each feature $\Rightarrow$ Full Bayesian inference
- Avoid arbitrary (influential) tuning parameters in priors
- Provide efficient family of algorithms

Our biological motivation

- Combined application of genome-wide expression profiling with linkage enables the mapping of expression quantitative trait loci (eQTLs), i.e. genetic control points for gene expression

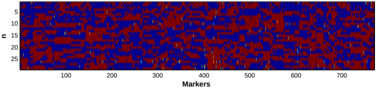
- *Cis*-acting (marker and transcript on same chromosome, typically with large effects) or *trans*-acting (different chromosome, with low effects) master regulators of gene expression are key control points in gene networks

- *Trans*-regulated genes are of primarily interest since they appears to be more complex, i.e. under polygenic control

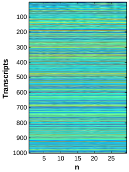
- Mining of eQTL data has led to new insights into gene functions and regulatory pathways

Motivating example: RI rat strains ($n = 29$)

$Y_{29\times1000}$
transcripts from
Adrenal tissue

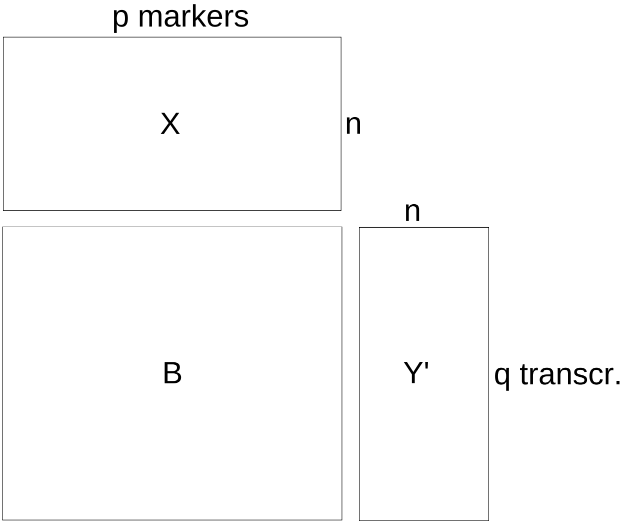


$X_{29\times770}$
informative
microsatellites



$\sim 7.7 \times 10^5$ tests
for association

Motivating example: generic set-up



p markers

X

n

n

B

Y'

q transcr.

Modelling strategy

Two possible approaches

- Use Multivariate Gaussian distributions for formulating a multiple response model of $Y(n \times q)$ on $X(n \times p)$
  - Imposes a strong assumption that all q outcomes are associated to same j^{th} covariate
  - Suitable for small q, e.g. transcripts in multiple tissues, preselected small group of transcripts, ...
- Link q separate regressions for each outcome $Y(n \times 1)$ through a flexible hierarchical structure on the selection indicators

Outline

- Bayesian variable selection set-up for hierarchically linked regressions ($1 \leq k \leq q$)
  - Priors specifications
  - Posterior inference
- MCMC Sampler
  - Evolutionary Monte Carlo: Local and Global moves (given k)
  - Updating global parameters
  - Adaptive Exploration Relevant Outcomes
- Illustration and demonstration of performance
  - Simulated example
  - Evidence for polygenic control and hot spot in the real data

Regression set-up and likelihood

- For every response, $k = 1, \ldots, q$, Gaussian linear regression:

  $$y_k = X\beta_k + \epsilon_k, \quad \epsilon_k \sim N\left(0, \sigma_k^2\right)$$

  with $X_{n\times p}$, centred

- Let $B_{q\times p} = (\beta_1, \ldots, \beta_k, \ldots, \beta_q)^T$ matrix of regression coefficients with $\beta_k = \left(\beta_{k1}, \ldots, \beta_{kj}, \ldots \beta_{kp}\right)$

- Let $\underline{\sigma}^2 = \left(\sigma_1^2, \ldots, \sigma_k^2, \ldots, \sigma_q^2\right)^T$

Then likelihood:

$$Y \Big| X, B, \underline{\sigma}^2 \sim \prod_{k=1}^{q} N\left(X\beta_k, \sigma_k^2\right)$$

Variable selection on X

Introduce prior structure on βs through latent binary matrix

- Latent binary matrix: $\Gamma = (\gamma_1, \ldots, \gamma_k, \ldots, \gamma_q)^T$
    - with $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kj}, \ldots, \gamma_{kp})$ the usual binary vector indicating which of the j^{th} covariates are included in the k^{th} regression
    - and $\gamma_{kj} = \{0, 1\}$
- Likelihood: $Y \,\big|\, X, B_\Gamma, \underline{\sigma}^2, \Gamma \sim \prod_{k=1}^{q} N\left(X_{\gamma_k}\beta_{\gamma_k}, \sigma_k^2\right)$

Prior specification for β_{γ_k}

Traditionally, two classes of priors have been considered for the variances of the regression coefficients

- $\beta_{\gamma_k} \,|g, \sigma_k^2, \gamma_k \sim N\left(0, g\sigma_k^2 \left(X_{\gamma_k}^T X_{\gamma_k}\right)^{-1}\right)$: g-prior structure
- Alternatively, replace $\left(X_{\gamma_k}^T X_{\gamma_k}\right)^{-1}$ by identity matrix : Independence prior

Prior specification

- $\beta_{\gamma_k} \left| g, \sigma_k^2, \gamma_k \right. \sim N\left(0, g\sigma_k^2 \left(X_{\gamma_k}^T X_{\gamma_k}^T\right)^{-1}\right)$: g-prior structure
- $g \sim InvGam\left(1/2, n/2\right)$ leading to Zellner-Siow priors

$$p\left(\beta_{\gamma_k} \left| \gamma_k, \sigma_k^2 \right.\right) \propto \int N\left(0, \sigma_k^2 g \left(X_{\gamma_k}^T X_{\gamma_k}\right)^{-1}\right) p\left(g\right) dg$$

- $\sigma_k^2 \sim InvGam\left(a_\sigma, b_\sigma\right)$
- $p\left(\gamma_{kj} \left| \omega_{kj} \right.\right) = \omega_{kj}^{\gamma_{kj}} \left(1 - \omega_{kj}\right)^{1-\gamma_{kj}}$, so
  $\gamma_{kj} \left| \omega_{kj} \right. \sim Bern\left(\omega_{kj}\right), 1 \leq k \leq q, 1 \leq j \leq p$

Prior structure for selection probabilities

Several possible structures might be appropriate

Most natural biologically: borrow information along columns to enhance the estimation of the hot spots

- Let $\Omega = \left(\omega_{kj}\right)_{k=1,\dots,q;j=1,\dots,p}$, then

  $$\omega_{kj} = \omega_j,$$

  where ω_j is the *a priori* column effect ("hot spot")

- Alternatively, could add a row effect (with a constraint)

  $$\omega_{kj} = \omega_j + \omega_k; \omega_j + \omega_k \leq 1$$

- $\omega_j, \omega_k \sim \mathit{Beta}\,(0.5, 0.5)$ or $\mathit{Beta}\,(a_\omega, b_\omega)$

Posterior inference

- Integrate out B_Γ and $\underline{\sigma}^2$ with marginal likelihood:

$$p\left(Y\,|X,g,\Gamma\right)\propto\int p\left(Y\,\middle|X,g,B_\Gamma,g,\underline{\sigma}^2,\Gamma\right)p\left(B_\Gamma\,\middle|g,\underline{\sigma}^2,\Gamma\right)p\left(\underline{\sigma}^2\right)dB_\Gamma d\underline{\sigma}$$

$$= \prod\nolimits_{k=1}^{q}(1+g)^{-p_{\gamma_k}/2}\left(2b_\sigma+S\left(\gamma_k\right)\right)^{-(2a_\sigma+n-1)/2}$$

$S\left(\gamma_k\right)=\left(y_k\right)^T\left(y_k\right)-\frac{g}{1+g}\left(y_k\right)^T X_{\gamma_k}\left(X_{\gamma_k}^T X_{\gamma_k}\right)^{-1}X_{\gamma_k}^T\left(y_k\right)$ where $y_k(n\times 1)$ is centred.

- Posterior estimates of g,Γ and Ω based on alternate sampling from their full conditionals

MCMC strategy

After integrating out variances and coefficients, left with sampling from full conditionals

1. $p\,(\Gamma\,|\cdots)\propto p\,(\,Y\,|X,g,\Gamma\,)\,p\,(\Gamma\,|\Omega\,)$

   This is particularly challenging as model space is huge: $dim(\Gamma)=q\times 2^p$. We use Evolutionary Monte Carlo (EMC)

2. $p\,(\Omega\,|\cdots)\propto p\,(\Gamma\,|\Omega\,)\,p\,(\Omega)$

   We use adaptive Metropolis-within-Gibbs (Roberts and Rosenthal, 2008) to adapt the tuning of the proposal for ω_{jk} on the logit scale

3. $p\,(g\,|\cdots)\propto p\,(\,Y\,|X,g,\Gamma\,)\,p\,(g)$

   To avoid tuning of the proposal, we also use adaptive MwG for g

Evolutionary Monte Carlo for Γ

We reduce stochastic search complexity by sampling Γ at each sweep:

- separately from each k with probability α_k

Given k, we use a tempered population of Markov Chains:

- Temperature reduces the influence of likelihood such that subsets of covariates can come in out during exploration
- Temperature reduces the dependence between γ_k and g
- Population based MCMC allows simultaneous exploration of different parts of the model space, each chain exchanging information with the others
- We retain just the non heated chain, while the other chains are used as “good proposals” for the indicator vector γ_k

Evolutionary MC for γ_k

How do we use it? At each sweep, a set of moves is attempted:

- “Local (mutation)” moves within each chain
- “Global moves”, a combination of:
  - Selection
  - Exchange
  - Crossover

The moves are tuned to improve efficiency. In particular:

- Local moves use restricted Gibbs sampling
- Selection move for “Exchange” operator based on joint information on all pairs of chains (Calvo 2005)

Adaptive Exploration Relevant Outcomes

- Not all outcomes are equally important, for some of them $\gamma_k = \emptyset$
- Idea is to spend more time on responses where there is more "action", i.e. $p_\gamma \gg 0$
- We propose to modify α_k, i.e. the probability of selecting the j^{th} full conditional

$$[p(\gamma_k | \cdots)]^{1/t} \propto [p(y_k | X_{\gamma_k}, \gamma_k, g) \, p(\gamma_k | \Omega_k)]^{1/t}$$

where $\Omega_k = (\omega_{kj})_{j=1,\dots,p}$ in an adaptive way

Adaptive Exploration Relevant Outcomes (continued)

- Optimising random scan Gibbs samplers has been proposed by Levin and Casella (2006): they adaptively updates the selection coefficients α_k based on the precision of the estimators of interest
- We propose a quasi-finite adaptation for α_k: $\tilde{\alpha}_k(b) = (1-\varepsilon)\, r_k(b) + \varepsilon$ with

$$\varepsilon = \begin{cases} 1 & \text{if } b \leq 2 \\ \frac{\sqrt{2}}{\sqrt{b}} + 10^{-3} & \text{otherwise} \end{cases}$$

$r_k(b) = \frac{\bar{p}_{\gamma_k}(b)}{\sum_k \bar{p}_{\gamma_k}(b)}$ and $\alpha_k(b) = \frac{\tilde{\alpha}_k(b)}{\sum_k \tilde{\alpha}_k(b)}$

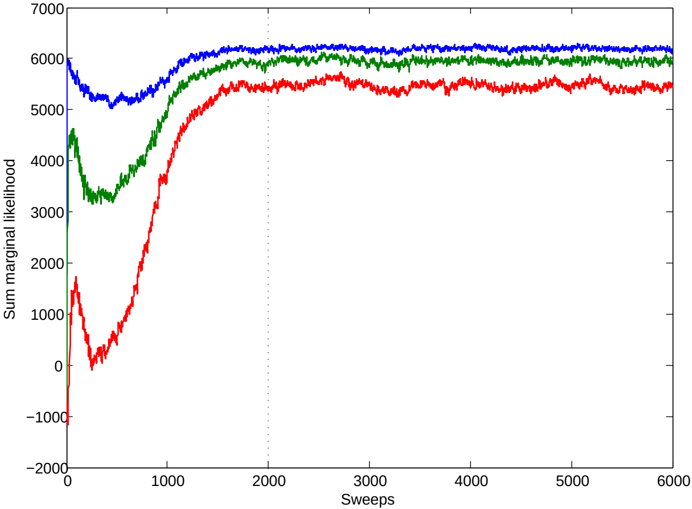
- After B batches, we “freeze” $\alpha_k(B) \rightarrow \alpha_k$

Simulated example

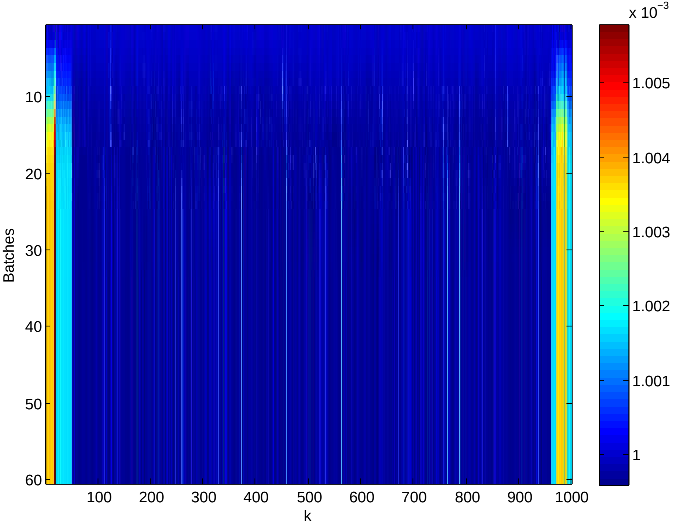
- Toy example where $q = 1000$, $p = 10$ and $n = 50$
- X_1, X_3, X_6 and X_{10}, associated with different outcomes in a complicated way
- Goal: find how many outcomes are associated with each predictor (“hot spot”)
- Here, we focus on illustrating the hierarchically related regression results under the model

$$\omega_{kj} = \omega_j, \forall k, 1 \leq j \leq p$$

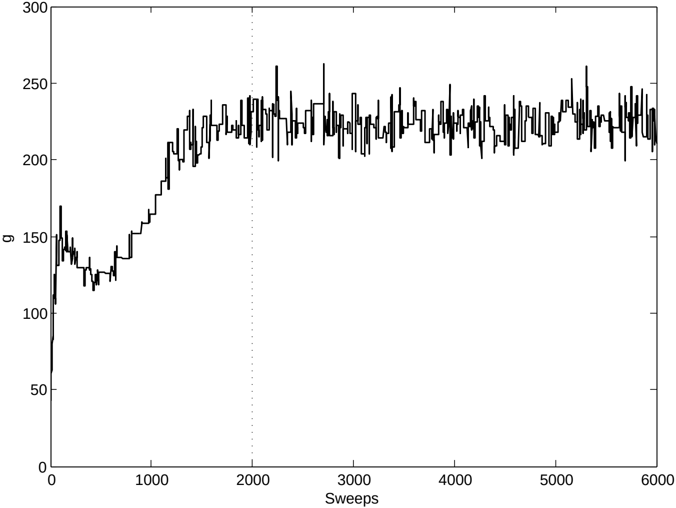
Simulated example: chains, temperature tuning and exploration model space



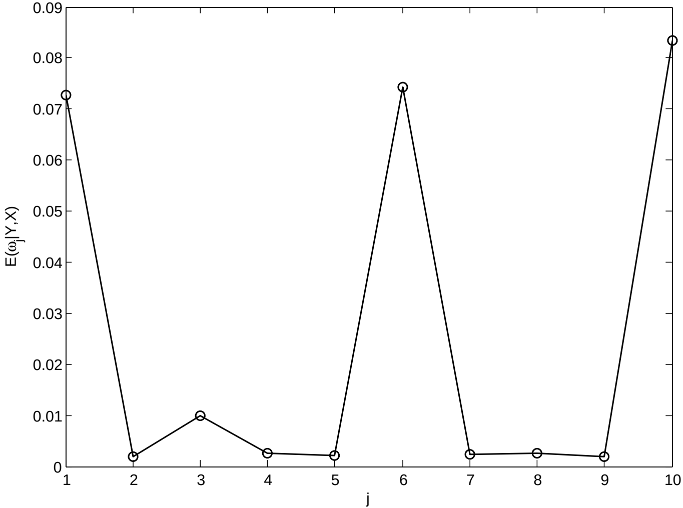
Simulated example: quasi-finite adaptation



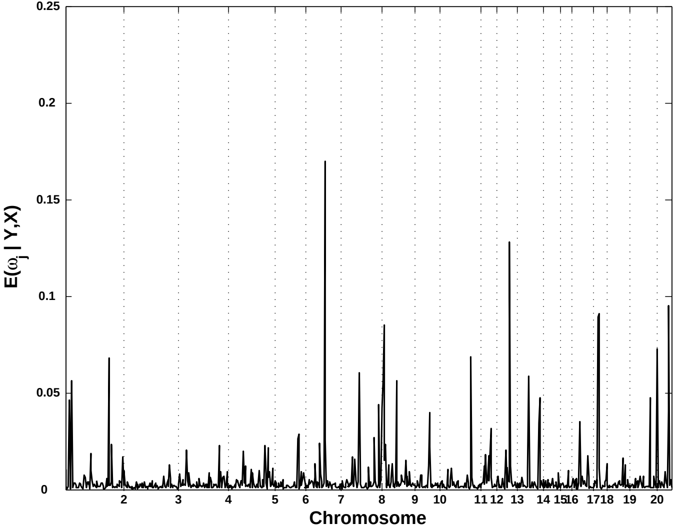
Simulated example: updating selection coefficient g



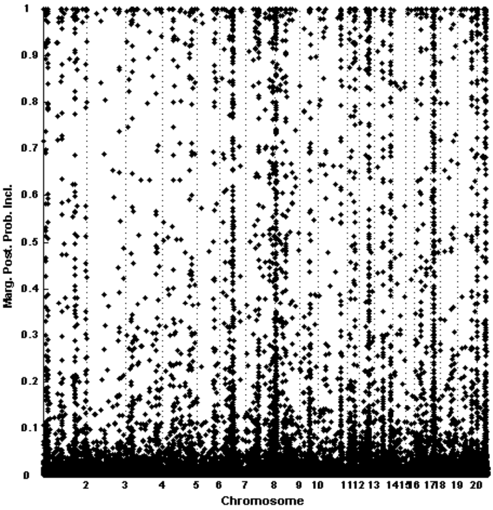
Simulated example: Hot spot evidence



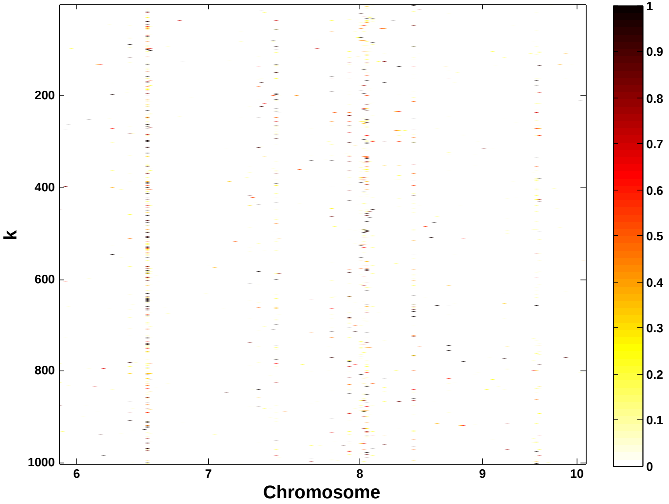
Real data example: Hot spot evidence



Real data example: marginal posterior probability of inclusion



Real data example: heat map



Conclusions

- Built a class of models suitable for joint analysis of genomic data sets, in particular for investigating link between genetic markers and multiple phenotypes
- For the huge dimensional space, we sample using Evolutionary Monte Carlo
- For global hyper-parameters, we sample using adaptive MwG with diminishing condition and bounded convergence conditions easy to check
- We implemented quasi-finite adaptation, but work in progress for a full adaptation in the spirit of Roberts and Rosenthal, 2008.