Bayesian Statistics in the Big Data Era, CIRM, Marseille, 27.11.2018

Bayesian non-parametric regression for analyzing time course quantitative genetic data

Zitong Li

Melbourne Integrative Genomics

and School of Mathematics and Statistics

University of Melbourne

Email: zitong.li1@unimelb.edu.au

Introduction to quantitative genetics

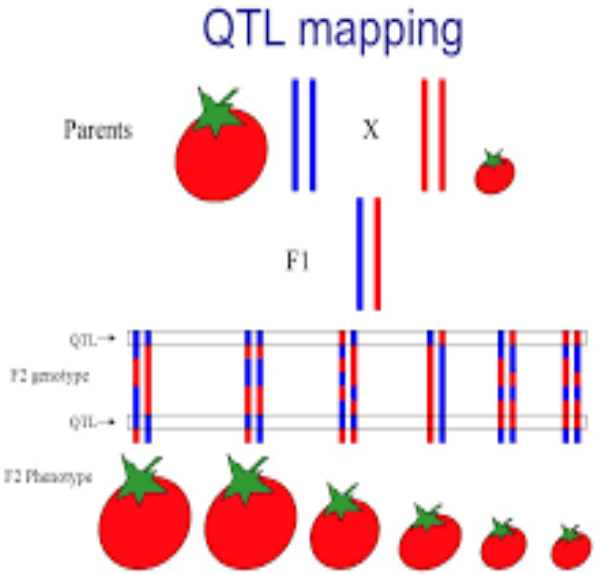
- Use statistical and computational methods to study the genotype-phenotype relationship

- Identify **causal molecular markers** which contributes to the phenotypic variation

- Estimate the **heritability**: $\frac{\mathrm{Var}(G)}{Var(P)} = \frac{Var(G)}{Var(G) + Var(E)}$

- Applications: human health and medicine, animal and plant breeding

Quantitative Trait Locus mapping and linear models



Genotype at a marker:
AA=1, AB=0, BB=-1

Phenotype

$Y = X\beta + E$

Additive genetic effect

The environmental factor~ $N(0,\sigma^2)$

The locus is claimed as a putative QTL, when the null hypothesis is rejected

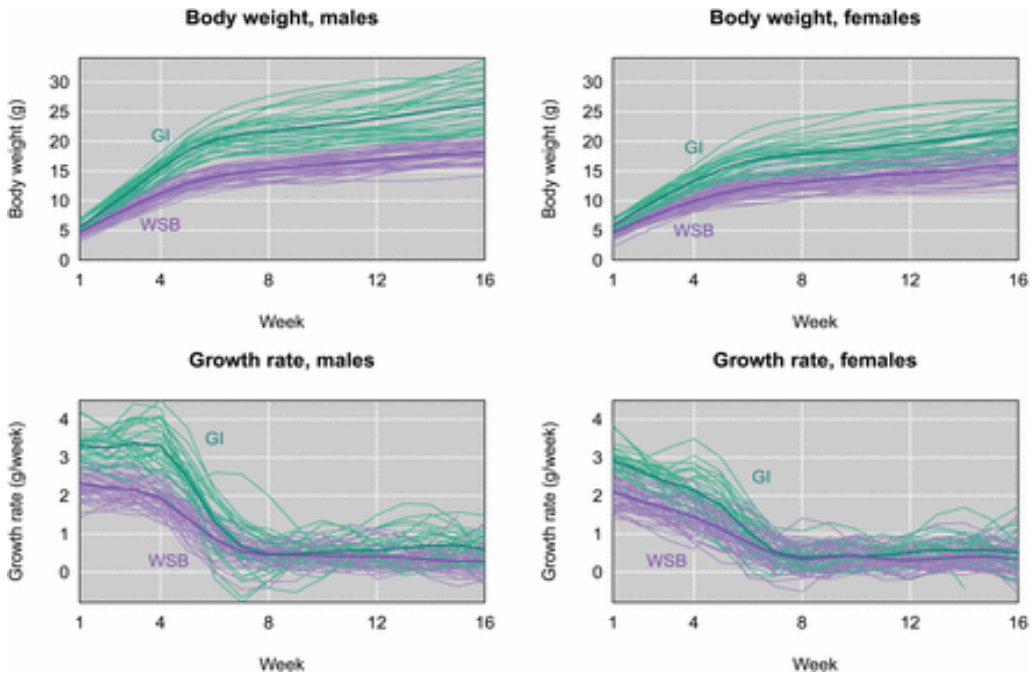
Hypothesis test: β=0 VS β ≠0

Time course data

Many biological processes are not static, but can change over time

Mice body size evolution on islands

- Study causes of the dramatic change in body size that accompanies island colonization (Gray et al. 2015, *Genetics*)
- Study target: Gough island (located in South Atlantic) mice, are twice as the mass of wild mice in UK
- Genetic explanations for the body size differentiation?



Varying-coefficient models

- The standard linear regression for single time point data (individuals *i*=1,…,*n*, markers *j*=1,..,*p*)

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \qquad e_i \overset{i.i.d.}{\sim} N(0, \sigma_0^2)$$

- Now for each individual, assume we have *m* repeated measurements over time. Naturally, we could extend the model as

$$y_i(t_r) = \beta_0(t_r) + \sum_{j=1}^{p} x_{ij}\beta_j(t_r) + e_i(t_r), \qquad \mathbf{e}_i = [e_i(t_1), ..., e_i(t_m)] \overset{i.i.d.}{\sim} \mathrm{MVN}(\mathbf{0}, \mathbf{\Sigma}_0)$$

- In the VC model, the regression coefficient for each explanatory variable is not assumed to be constant, but are allowed to change over time (Hastie and Tibshirani 1993, J. R. Stat. Soc.).

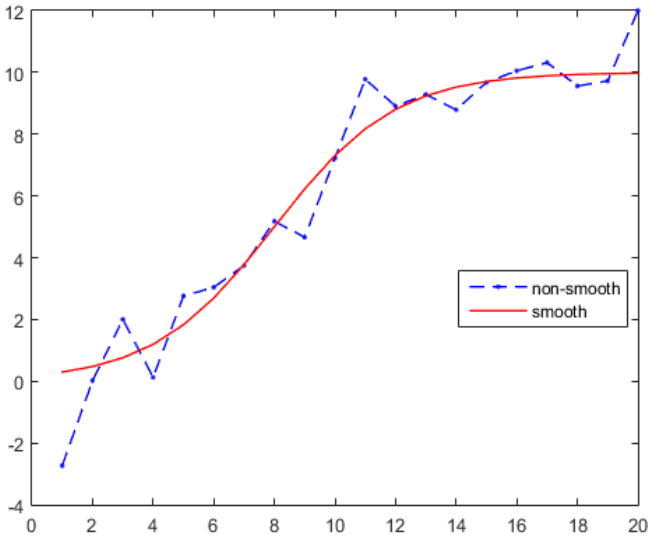
Objectives

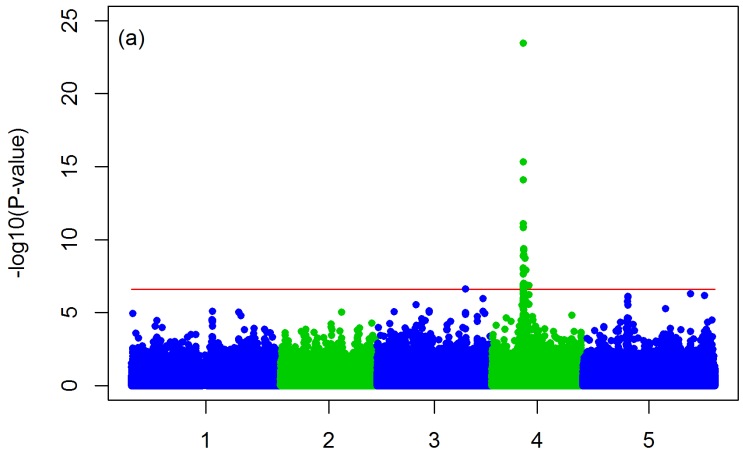
- **Smoothing (across time points):**

-the two measurements at nearby time points should be more close to each other than the measurements at two further distances.

- **Variable selection (across genetic features):**

-Among thousands of molecular markers, choose a few markers that are most associated with the phenotypes





B-splines

•B-Splines: truncated power series defined in the data domain. Knots: the breakpoint

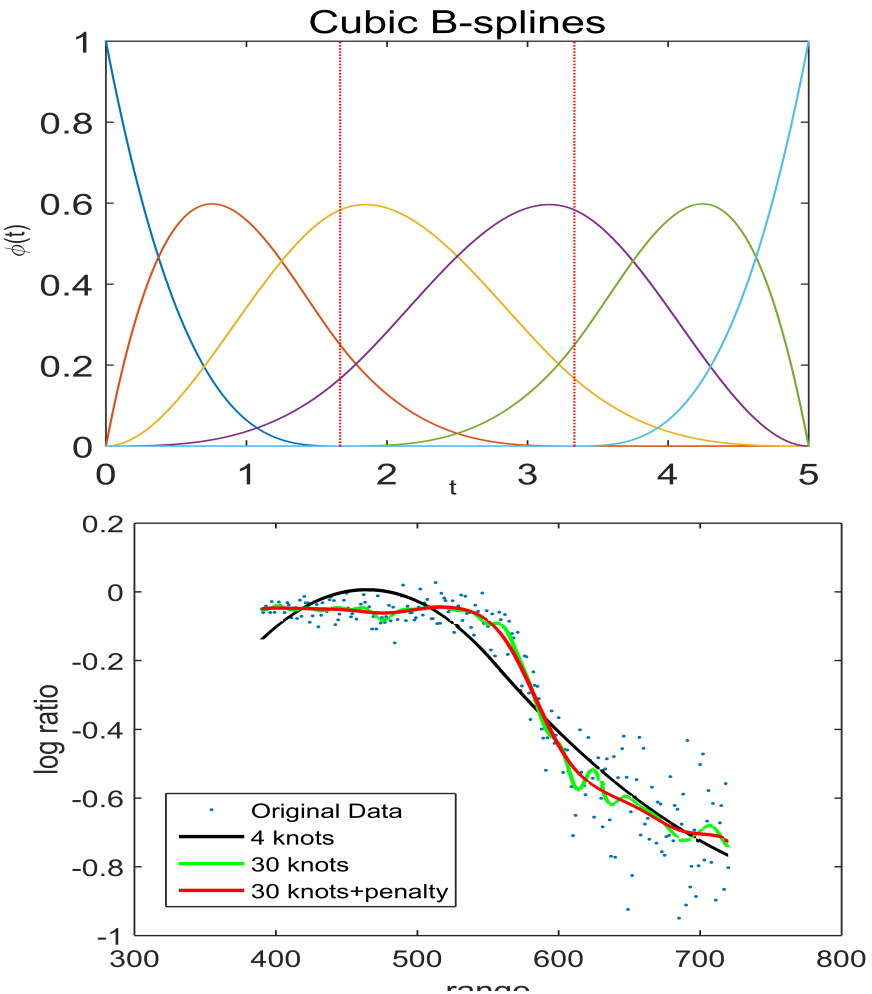
$$\beta(t) = \sum_{l=1}^{k} \Phi_l(t)\alpha_l$$

-Choosing an appropriate number of knots is crucial

•P-spline (Eilers and Marx 1996, Stat. Sci.):

-add a **difference penalty** to the likelihood function to avoid overfitting:

$$\lambda \sum_{l=1}^{k-1} (\alpha_l - \alpha_{l-1})^2$$



Bayesian P-splines prior

- Bayesian statistics: combine objective data with subjective prior knowledge posterior probability $\propto$ Likelihood × Prior

  Likelihood: Splines model

  Prior: difference penalty (to induce smoothness)

- Bayesian interpretation of the difference penalty $\lambda_j \sum_{l=1}^{k-1} (\alpha_{jl} - \alpha_{jl-1})^2$:

-Random walk prior:

$$p(\boldsymbol{\alpha}_j \mid \sigma_j^2) = \mathrm{MVN}(\boldsymbol{\alpha}_j \mid \mathbf{0}, \sigma_j^2 \mathbf{K}^{-1})$$

Bayesian VC model

$$y_i(t_r) = \beta_0(t_r) + \sum_{j=1}^{p} x_{ij}\beta_j(t_r) + e_i(t_r), \qquad \mathbf{e}_i = [e_i(t_1), \ldots, e_i(t_m)] \overset{i.i.d.}{\sim} \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_0)$$

$$\prod_{i=1}^{n} \mathrm{MVN}(\mathbf{y}_i \mid \boldsymbol{\beta}_0 + \sum_{i=1}^{p} x_{ij}\boldsymbol{\beta}_j, \boldsymbol{\Sigma}),$$

Regression parameters: $\boldsymbol{\beta}_j = \boldsymbol{\Psi}\boldsymbol{\alpha}_j$ $\Rightarrow$ $p(\boldsymbol{\alpha}_j \mid \sigma_j^2) = \mathrm{MVN}(\boldsymbol{\alpha}_j \mid \mathbf{0}, \sigma_j^2\mathbf{K}^{-1})$

$$p(\sigma_j^2) = \mathrm{IG}(0.0001, 0.0001)$$

Residuals: $\boldsymbol{\Sigma} = \sigma_\varepsilon^2\mathbf{I}, \quad \mathrm{p}(\sigma_\varepsilon^2) \propto \mathrm{IG}(0.0001, 0.0001)$

Posterior was evaluated using a Variational Bayes algorithm (Li and Sillanpää 2013, Genetics**)**

Alternative model structure: Gaussian process

- Definition:

$$\beta(t) \sim GP(0, \mathrm{Cov}(\beta(t), \beta(t'))) \equiv \mathrm{MVN}(0, \mathbf{C})$$

- Covariance functions fully determines the properties of the process

- The covariance function can induce a certain degree of smoothness in the model

Link the P-splines to Gaussian process

- Recall the B-splines $\beta(t) = \sum_{l=1}^{k} \Phi_l(t)\alpha_l$ or $\boldsymbol{\beta} = \boldsymbol{\Psi}\boldsymbol{\alpha}$ (1)
- and random walk prior $\boldsymbol{\alpha} \sim \mathrm{MVN}(0, \sigma^2 \mathbf{K})$ (2)
- If in (1) and (2) we marginalize the weight parameter $\boldsymbol{\alpha}$, we have

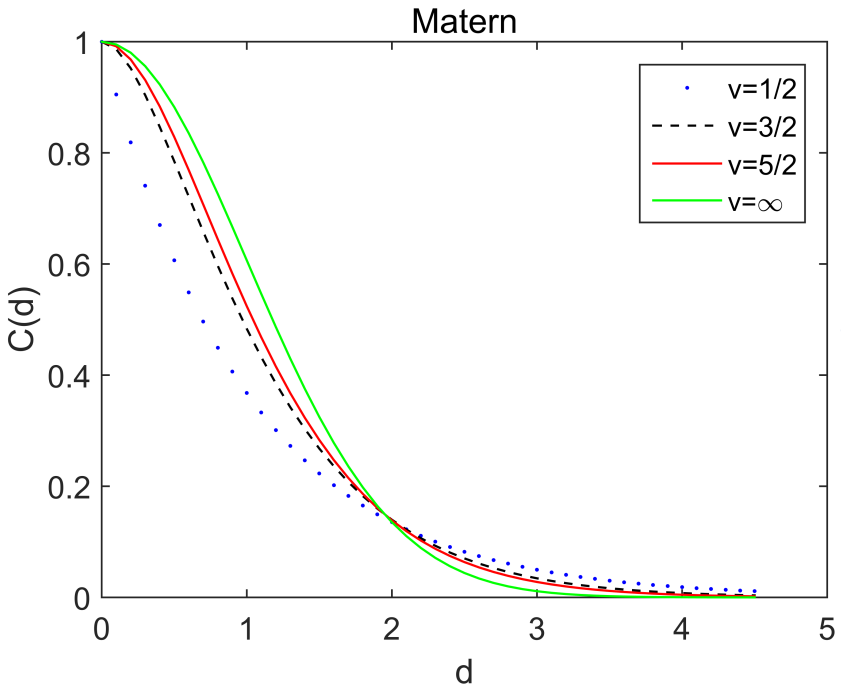
$$\boldsymbol{\beta} \sim \mathrm{MVN}(0, \sigma^2 \boldsymbol{\Psi}\mathbf{K}\boldsymbol{\Psi}^T)$$

- Define $\mathrm{Cov}(\beta(t), \beta(t')) = \sigma^2 \boldsymbol{\Psi}(t)\mathbf{K}\boldsymbol{\Psi}(t)^T$

- Therefore, P-Splines model is a special case of GP

Mátern covariance function

$$C_v(d) = \sigma^2 \frac{2}{\Gamma(v)} (\sqrt{2v}\frac{d}{\rho})^v K_v(\sqrt{2v}\frac{d}{\rho})$$



v=1/2, Laplace

v=**5/2** common choice in Spatial Statistics and Machine Learning fields

v= ∞, squared exponential

- Distance measure: $d = |t - t'|$
- K_v is a modified Bessel function
- v degree of freedom
- Hyper-parameters $\boldsymbol{\theta} = (\sigma^2, \rho)$, and σ_ε^2 assigned with non-informative hyper-priors

Posterior estimation: Empirical Bayes

(1) Hyper-parameter (such as σ^2 and ρ) inference by first analytically integrating out β:

$$\hat{\boldsymbol{\theta}} = \arg\max\, p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta}, \sigma_\varepsilon^2)\, p(\boldsymbol{\theta}, \sigma_\varepsilon^2)$$

(2) Posterior inference of β after fixing hyper-parameters:

$$\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \hat{\boldsymbol{\theta}} \sim \mathrm{MVN}(m_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$$

$$\text{where } m_{\boldsymbol{\beta}} = \mathbf{C}_{\boldsymbol{\beta}} \mathbf{X}^T (\mathbf{X}\mathbf{C}_{\boldsymbol{\beta}} \mathbf{X}^T + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{y}, \;\; \boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \mathbf{C}_{\boldsymbol{\beta}} - \mathbf{C}_{\boldsymbol{\beta}} \mathbf{X}^T (\mathbf{X}\mathbf{C}_{\boldsymbol{\beta}} \mathbf{X}^T + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{X}\mathbf{C}_{\boldsymbol{\beta}}$$

Variable selection

- Aim: detect a pasimony model M={a subset of markers have strong association with the phenotypes}.

- **Model posterior Marginal likelihood**

  P(M|y,X) $\propto$ P(y|X, M) × p(M)

- **Marginal likelihood P(y|X, M)** can be estimated by numerical integration：

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{M}) = \int p(\mathbf{y} \mid \mathbf{X}, \mathbf{M}, \boldsymbol{\theta}, \sigma_\varepsilon^2) p(\sigma_\varepsilon^2) d\sigma_\varepsilon^2$$

$$\approx \sum_{l=1}^{N} p(\mathbf{X}, \mathbf{M}, \boldsymbol{\theta}, \sigma_{\varepsilon,l}^2) p(\sigma_{\varepsilon,l}^2) \Delta_{\sigma_\varepsilon^2}$$

- **The model Prior:** $p(M \mid \pi) = \pi^{qm}(1-\pi)^{pm-qm}$ (p: total number of markers, q: number of selected markers, m: number of time points)

$\pi < 0.5$ ： in favour of small number of variables

Stepwise selection

- The goal is to find an optimal model satisfying

$$\hat{M} = \max[p(\mathbf{y} \mid \mathbf{X}, \mathbf{M}) + \ln p(\mathbf{M})]$$

- We use stepwise regression, to seek $\hat{M}$.

(i) starting from null model

(ii) add one variable into the model which improve the model most

(iii) repeat (ii) until no variable can improve the model posterior anymore

Computational issue

- Evaluating the marginal likelihood **P(y|X, M)** can be expensive, because it involves an inversion of an $nm \times nm$ matrix (*n*=NO. of individuals; *m*=NO. of time points):
- Complexity: **O((nm)^3)**

- Applying Woodbury-Sherman-Morrison lemma:

  -Computational complexity now becomes **O(m^3 q)**

  -q: number of markers selected into the model

Practical implementation

- On the basis of Matlab toolbox: **GPstuff** (https://research.cs.aalto.fi/pml/software/gpstuff/)

- To learn more about the software, please refer to:

-Vanhatalo et al. (2013) Gpstuff: Bayesian modeling with Gaussian Processes. Journal of Machine Learning Research 14: 1175-1179

- We also have a plan to develop a complementary R package for this method in near future

Case study: simulation

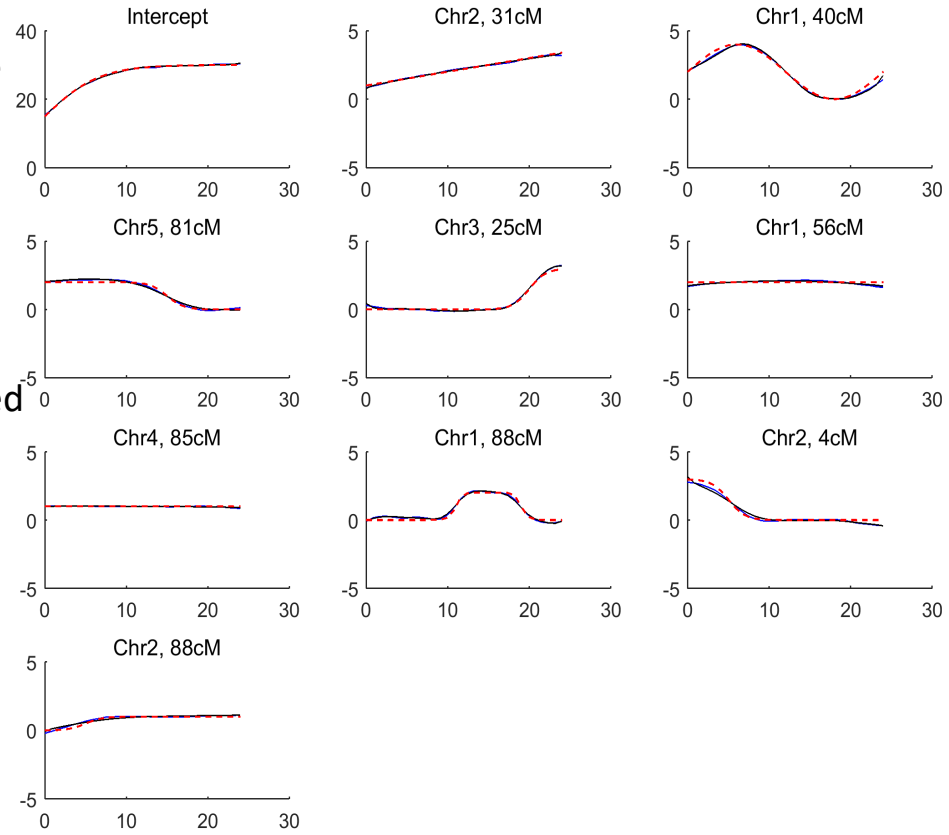
- A data with 1000 individuals, and 453 covariates with 9 QTL simulated
- The study was replicated for 50 times

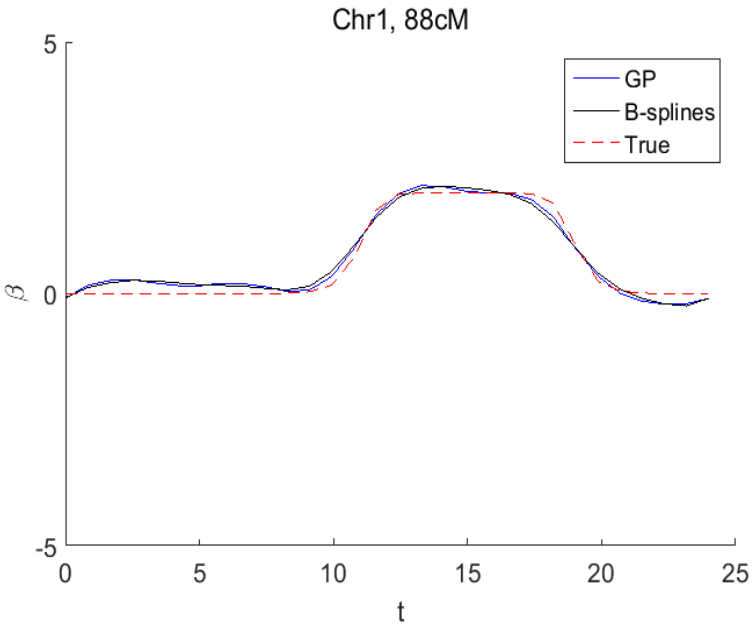
Descriptions	Trend functions
The intercept with a logistic growth curve	$\beta_0 = \frac{30}{1+\exp(-0.3t)}$
Loci 52 and 358 have constant effects over time	$\beta_{52} = 2$ $\beta_{358} = 1$
Locus 118 has linear growth effect over time	$\beta_{118} = 0.1t + 1$
Loci 174 and 216 are active only at the early stage	$\beta_{174} = \frac{1}{1+\exp(-t+5)}$ $\beta_{216} = \frac{3}{1+\exp(-t+20)}$
Loci 98 and 433 are active only at the late stage	$\beta_{98} = \frac{3}{1+\exp(t-5)}$ $\beta_{433} = \frac{2}{1+\exp(t-15)}$
Locus 78 is active only at the middle stage	$\beta_{78} = \frac{2}{1+(\frac{t-15}{4})^{10}}$
Locus 35 has the periodic effects over time	$\beta_{35} = 2 + 2\sin(\frac{\pi t}{12})$

Evaluation of parameter estimation

Red curve: the true simulated effect

Blue curve: estimated effect





Evaluation of variable selection

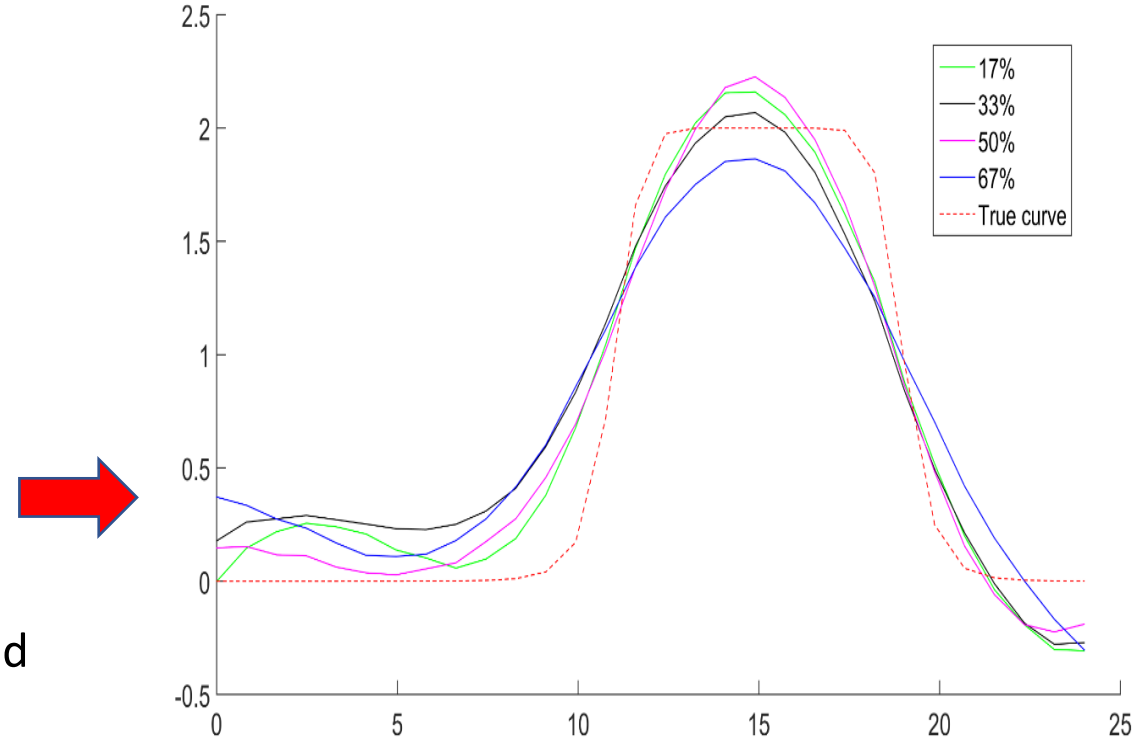
Table A simulation study of 50 replicates: the average performance of GP approach with different setting of model priors (with the choice of the model inclusion probability to be pi=0.5, equivalent as using marginal likelihood, pi=0.02 and pi=0.05) on data sets with number of time points k=10 and k=30, respectively.

Simulated QTL	Frequency of QTL detected by GP and Bspline					
	n=500, m=10			n=500, m=30		
	ML	pi=0.2	pi=0.01	ML	pi=0.2	pi=0.01
35 (Chr1, 40cM)	1.00	1.00	1.00	1.00	1.00	1.00
52 (Chr1, 56cM)	1.00	1.00	0.72	1.00	1.00	1.00
78 (Chr1, 88cM)	1.00	1.00	0.30	0.98	0.88	0.76
98 (Chr2, 3.6cM)	1.00	0.98	0.10	1	0.70	0.50
118 (Chr2, 31cM)	1.00	1.00	1.00	1	1.00	1.00
174 (Chr2, 88cM)	0.50	0.44	0.00	0.98	0.84	0.00
216 (Chr3, 25cM)	1.00	1.00	0.88	1	1.00	0.96
358 (Chr4, 85cM)	0.78	0.74	0.30	0.98	0.90	0.86
433 (Chr5, 81cM)	0.78	0.76	0.32	0.90	0.86	0.78
No. of false positives	1.78	1.08	0.38	1.7	0.36	0.24

Performance of GP on incomplete phenotype data

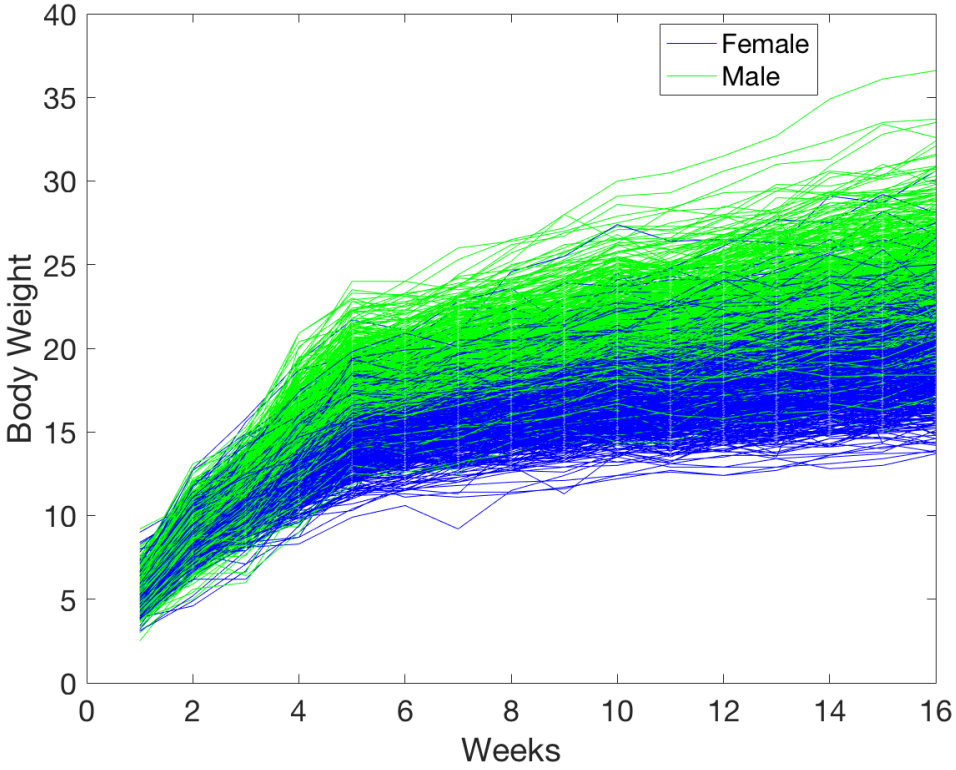
- In GP regression, missing observations in the phenotype data can simply be marginalized out of the model.

- The comparison of the additive genetic effects estimates under missing data scenario: with missing observations generated at 5, 10, 15 and 20 out of 30 time points in all the 500 individuals, which corresponds to 17%, 33%, 50%, and 67% incomplete data rates.



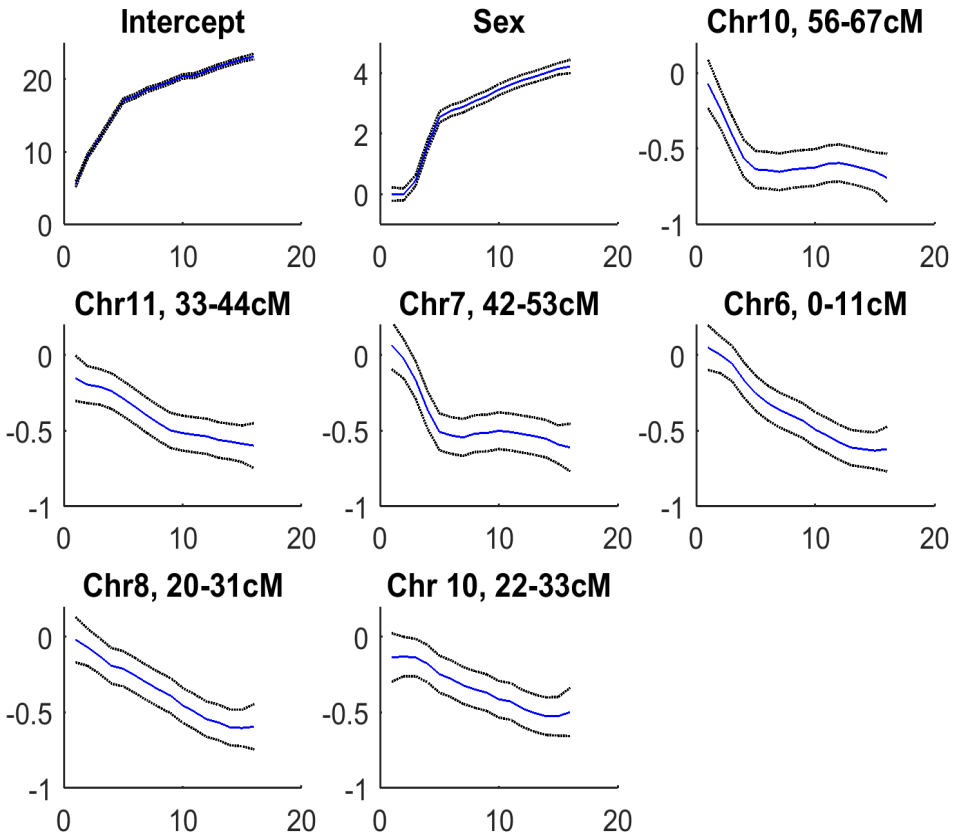
Mouse body size study

- A F2 cross generated from two divergent lines (a Gough island mouse and a Mainland mouse)
- About 1100 individuals. 12000 markers over 19 chromosomes
- The body weight was measured repeated for 16 weeks since born until mature

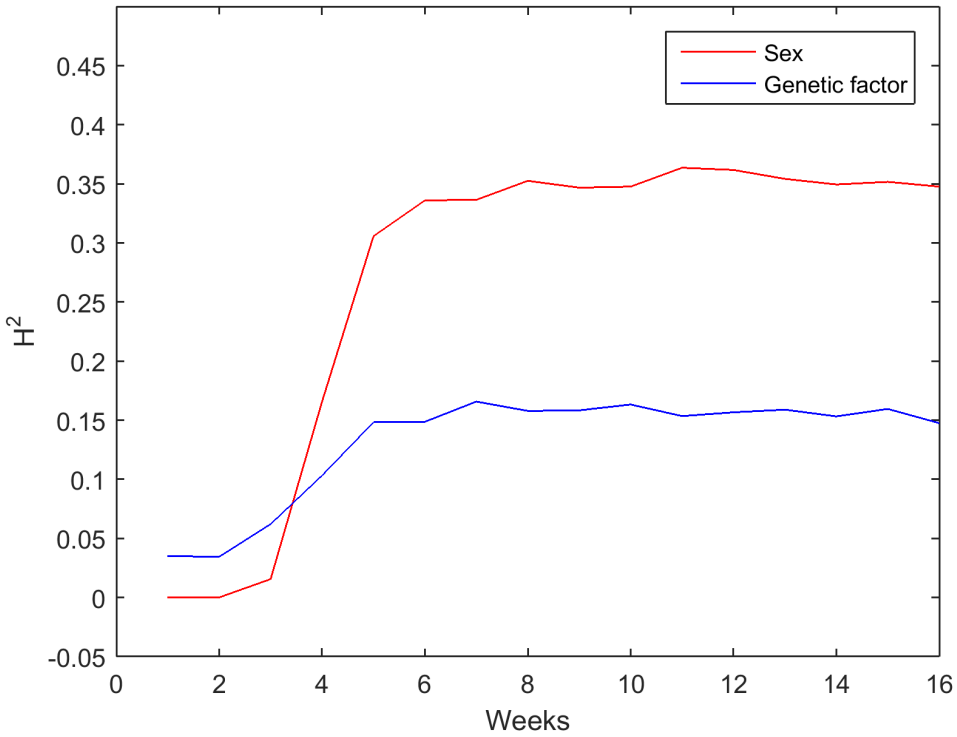


Results of Mouse body size study

Additive genetic effects



Heritability: proportion of phenotype variance explained by genetic factors



Manuscript

-Jarno Vanhatalo[1]*, Zitong Li[2]*, Mikko Sillanpää[3] (2018) A Gaussian process model for mapping quantitative trait loci in functional valued traits (under revision in *Bioinformatics*).

1 University of Helsinki

2 University of Melbourne

3 University of Oulu

*equal contribution

Conclusion

- We proposed efficient and non-parametric Bayesian inference for analyzing time course quantitative genetic data.

- Smoothing and variable selection were simultaneously achieved.

- Feasible to analyze high dimensional data sets of thousands of variables and hundreds of time points:

-n=1000, m=30, p=10 000: 10 hours (can speed up by parallel computing)

-n=200, m=250, p=200: 20 minutes

(MAC, I5 CPU, RAM=16Gb)

Things need to improve

- The posterior uncertainty (such as credible bands) might be under-estimated due to the use of an EP algorithm.
- Stepwise regression is greedy.
- Variable selection is sensitive to the choice of model priors, when the sample size is small.

Thank you very much!