Bayesian regularized regression methods for quantitative genetics with focus on longitudinal data

Zitong Li

Department of Mathematics and Statistics
University of Helsinki, Finland

ACADEMIC DISSERTATION

To be presented with the permission of the Faculty of Science of the University of Helsinki, for public examination in lecture hall A129, Chemicum (A.I. Virtasen aukio 1) on 3rd of October 2014, at 12 o'clock noon.

Helsinki 2014

Supervisor: **Professor Mikko J. Sillanpää**
Department of Mathematical Sciences
and Department of Biology
University of Oulu
Biocenter Oulu
Finland

Reviewers: **Adjunct Professor Aki Vehtari**
Department of Biomedical Engineering
and Computational Science
Aalto University
Finland

Professor Shizhong Xu
Department of Botany and Plant Sciences
Department of Statistics
University of California, Riverside
USA

Opponent: **Associate Professor Ole Winther**
Department of Applied Mathematics
and Computer Science
Technical University of Denmark
Denmark

ISBN 978-951-51-0075-7 (paperback)
ISBN 978-951-51-0076-4 (PDF, http://ethesis.helsinki.fi)
Unigrafia Oy
Helsinki 2014

List of original publications

This thesis is based on the following publications, which are referred to in the text by their Roman numerals:

I. **Li, Z.** and M. J. Sillanpää 2012. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. Theoretical and Applied Genetics 125: 419-435.

II. **Li, Z.** and M. J. Sillanpää 2012. Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. Genetics 190: 231-249.

III. **Li, Z.** and M. J. Sillanpää 2013. A Bayesian nonparametric approach for mapping dynamic quantitative traits. Genetics 194: 997-1016.

IV. **Li, Z.**, H. R. Hallingbäck, S. Abrahamsson, A. Fries, B. Andersson, M. J. Sillanpää and M. R. García-Gil 2014. Functional QTL-analysis of wood properties in a full-sib family in Scots pine (*Pinus sylvestris* L.). (submitted)

Author contributions

I Both authors were involved in the conception and design of the study. ZL performed the example analyses, and drafted the manuscript. Both authors participated in the interpretation of results and critically revised the manuscript.

II, III Both authors were involved in the conception and design of the study. ZL developed and implemented the method, and drafted the manuscript. Both authors participated in the interpretation of results and critically revised the manuscript.

IV ZL, HH, MS and RGG were involved in the conception and design of the study. HH performed the preprocessing of the data. ZL developed and implemented the method, and performed the statistical analyses. ZL and HH jointly drafted the manuscript. ZL, HH, MS and RGG participated in the interpretation of results. All the authors critically revised the manuscript.

Contents

Foreword

Quantitative trait loci (QTL) /association mapping aims to identify the genomic loci associated with the complex traits. From a statistical perspective, multiple linear regression is often used to model, estimate and test the effects of genetic markers on a trait. With genotype data derived from contemporary genomics techniques, however, the number of markers typically exceed the number of individuals, and it is therefore necessary to perform some sort of variable selection or parameter regularization to provide reliable estimates of model parameters. In addition, many quantitative traits are dynamic in nature. Accordingly, a longitudinal study that jointly maps the repeated measurements of the phenotype over time may increase the statistical power to identify QTLs, compared with the single trait analysis. This thesis focuses on the Bayesian modeling and variable selection/regularization of QTL data derived from longitudinal studies. First, we review the principal frequentist regularization methods for analyzing a single trait. In the second work, we move to the Bayesian regularization methods, we consider a fast variational Bayes algorithm for parameter estimation, and we compare it to the classic Markov chain Monte Carlo method. In the third work, a non-parametric Bayesian varying coefficient method for analyzing longitudinal data with a large number of time points is developed. In the fourth work, we apply another two possible longitudinal models: (1) a multilevel model and (2) a mixed effect model to map a wood properties data set that is characterized by a small number of time points. An important perspective throughout this thesis is multiple hypothesis testing, which is applied to formally judge the statistical significance of the QTLs and reduce the number of false positive loci. Several existing frequentist and Bayesian procedures for multiple testing have been evaluated in the thesis.

1 Introduction to linear regression

Regression is a statistical technique aiming to model and estimate the associative relationship between two or more variables. This section provides a brief introduction to linear regression, which is the simplest and perhaps the most widely used regression technique.

A simple linear regression model with two variables involved can be formally defined as

$$y_i = \beta_0 + x_i\beta_1 + e_i, \tag{1.1}$$

where y_i $(i = 1, \ldots, n)$ is the ith observable value of the response variable (or dependent variable) $\boldsymbol{y}$, and x_i is the ith observable value of the explanatory variable $\boldsymbol{x}$. The unknown regression coefficients are β_0 and β. The intercept term β_0 describes the population mean, and the slope term β describes how strongly the explanatory variable $\boldsymbol{x}$ is related to the response variable $\boldsymbol{y}$. Furthermore, the error terms e_i are assumed to follow a normal distribution $N(0, \sigma_0^2)$ with mean zero and variance σ_e^2 independently for $i = 1, \ldots, n$.

In practice, the response variables are often influenced by more than one explanatory variable. In the case we have p $(p > 1)$ explanatory variables, it is natural to extend the equation (1.1) to the form of multiple regression, which is

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + e_i, \tag{1.2}$$

and the model can be further written in the matrix form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1.3}$$

where $\boldsymbol{y} = [y_1, ..., y_n]^T$, $\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ & & & \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$, $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_p]^T$, and $\boldsymbol{e} = [e_1, ..., e_n]^T$.

1.1 Least squares and uncertainties

Under the condition $n > p + 1$, by minimizing the sum of squared errors (SSE) function $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ with respect to the unknown regression coefficients $\boldsymbol{\beta}$, the following ordinary least square (OLS) estimates are obtained:

$$\hat{\boldsymbol{\beta}}_{\text{ols}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{1.4}$$

The estimation can be done from a maximum likelihood (ML) point of view. The likelihood function of model (1.3) is

$$p(\boldsymbol{y}|\boldsymbol{\beta},\sigma_0^2) = (2\pi)^{-\frac{n}{2}}(\sigma_0^2)^{-\frac{n}{2}}\exp[-\frac{1}{2\sigma_0^2}(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})]. \tag{1.5}$$

By maximizing the likelihood function with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\sigma_0^2}$, we obtain the ML estimates: $\hat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, and $\hat{\sigma}^2_{0\mathrm{ML}} = \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}$. In fact, the ML estimates $\hat{\boldsymbol{\beta}}_{\mathrm{ML}}$ are exactly the same as $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}$. Thus, below we discuss these matters mainly from the SSE and OLS estimation point of view. The ML estimates of σ_0^2 is known to be biased, and an unbiased alternative is a restricted maximum likelihood (REML) (Patterson and Thompson 1971) estimate $\hat{\sigma}_0^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-p-1}$.

The OLS method directly provides point estimates of the regression parameters, but not the uncertainty estimates. In order to evaluate the uncertainty quantities such as the variances or confidence intervals, we need to consider the sampling distribution of $\boldsymbol{\beta}$, which is constructed by repeated sampling with the levels of explanatory variable $\boldsymbol{X}$ held constant. The sampling distribution of $\boldsymbol{\beta}$ is a multivariate normal distribution $\mathrm{MVN}(E[\hat{\boldsymbol{\beta}}], \mathrm{COV}(\hat{\boldsymbol{\beta}}))$. The mean $E[\hat{\boldsymbol{\beta}}]$ is equivalent to $\boldsymbol{\beta}$ indicating that the OLS estimators are unbiased. The covariance of $\hat{\boldsymbol{\beta}}$ is $\mathrm{COV}(\hat{\boldsymbol{\beta}}) = \sigma_0^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Typically, σ_0^2 is unknown, and its unbiased estimate can be applied here. So we obtain an estimated covariance matrix as $\widehat{\mathrm{COV}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}_0^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

For a single regression coefficient β_j $(j = 0, 1, ..., p)$, we have

$$\frac{\hat{\beta}_j - \beta_j}{s[\hat{\beta}_j]} \sim t(n-p), \tag{1.6}$$

where $s[\hat{\beta}_j] = \sqrt{\mathrm{COV}(\hat{\boldsymbol{\beta}})_{jj}}$ is the standard error, and $t(n-p)$ is a Student-t distribution with $n-p$ degree of freedom. The density function of the Student-t distribution with ν degree of freedom is

$$p(x|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}(1+\frac{x^2}{\nu})^{-\frac{\nu+1}{2}}. \tag{1.7}$$

On the basis of (1.6), we can construct the confidence interval with $1-\alpha$ confidence level as $\mathrm{CI}(1-\alpha;\beta_j) = [\hat{\beta}_j - Q_t(1-\frac{\alpha}{2}|n-p), \hat{\beta}_j + Q_t(1-\frac{\alpha}{2}|n-p)]$, where $Q_t(\bullet)$ represents the Student-t distribution. The confidence interval can be used as a criterion to determine whether the parameter β_j is zero or not (i.e. $\beta_j = 0$ or $\beta_j \neq 0$). If $0 \notin \mathrm{CI}(1-\alpha;\beta_j)$, we conclude that the variable j is significant at the $1-\alpha$ confidence level. The typical choice of the significance level α is 0.01, 0.05 or 0.1.

Hypothesis testing is an alternative way of assessing significance. We name the event H_0: $\beta_j = 0$

as a null hypothesis, and H_1: $\beta_j \neq 0$ as an alternative hypothesis. For linear regression, we often consider the t test. The test statistic is $t_j^* = \frac{\beta_j}{s[\beta_j]}$. If $t_j^* > Q_t(1-\frac{\alpha}{2}|n-p)$, we reject null hypothesis H_0 and accept H_1. Another interpretation can be given from the perspective of p-value. The p-value is defined as "the probability of obtaining a test statistic at least as extreme as the one that was actually observed given that the null hypothesis H_0 is true" (Goodman 1999). If the p-value is smaller than the significance level α (say α=0.05), we conclude that the alternative hypothesis H_1 is accepted, with the risk of making a wrong decision controlled at α. In fact, the t test is equivalent to the above mentioned confidence interval approach. Other possibilities for assessing statistical confidence are F test and Wald test (Kutner et al. 2004), which are not covered here.

1.2 Model fitting and prediction

After obtaining the OLS estimates $\hat{\boldsymbol{\beta}}$ from the data $(\boldsymbol{X}, \boldsymbol{y})$, we could evaluate how well the estimated linear model fits the original data by calculating the mean square error

$$\text{MSE} = \frac{1}{n}(\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}), \tag{1.8}$$

where $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. Note that here the same data are used to first estimate the unknown parameters, and second "predict" their own response values.

In practice, a more interesting task is to predict the response values for some new explanatory data $\boldsymbol{X}_{\text{new}}$ by calculating $\hat{\boldsymbol{y}}_{\text{new}} = \boldsymbol{X}_{\text{new}}\hat{\boldsymbol{\beta}}$. If it happens that we obtain the true response values $\boldsymbol{y}_{\text{new}}$ later, we could use a similar mean square error type of criterion to evaluate the predictive ability:

$$\text{PE} = \frac{1}{n}(\boldsymbol{y}_{\text{new}} - \hat{\boldsymbol{y}}_{\text{new}})^T(\boldsymbol{y}_{\text{new}} - \hat{\boldsymbol{y}}_{\text{new}}), \tag{1.9}$$

Note that we name this metric as "PE" (representing the prediction error), in order to distinguish it from the "MSE" defined in the equation (1.8).

In general, MSE often provides over-optimistic estimation of model fitting compared with PE, because MSE involves model estimation and prediction using the same data. This is known as the "over-fitting" phenomenon, which will be discussed in **Section 4**.

Linear regression has been widely applied in diverse fields such as biology, economics and social science. This work is mainly concerned with the application of linear regression techniques in quantitative genetics. Briefly, a typical data set combines, phenotype data, which reflects certain observable characteristics or traits such as human height or barley kernel density, and genotype data that comprises information about the DNA sequence. A linear regression can be used to (i) identify segments of DNA sequence which are highly associated with the traits, and (ii) predict

the phenotype values based on genotype data. A more comprehensive description of the genetics applications will be given in **Section 6**.

2 Beyond the linear model

The standard linear regression model is mainly based on the assumptions that (i) the relationship between response and explanatory variables is roughly (additively) linear, (ii) the response variables are Gaussian distributed continuous variables and (iii) the responses are i.i.d. distributed. Fitting data that violate these assumptions into a standard linear model may not be efficient for either identifying significant explanatory variables or for making predictions. More specific regression techniques are needed in order to fit such data. For example, data with binary response variables such as disease status, can be fitted using logistic regression where a logistic link function is applied to transform the binary response into continuous space; in this way, a connection between discrete responses and (continuous) explanatory variables can be built. The logistic regression belongs to the generalized linear model (GLM) family, which aims to release the assumption (ii) of Gaussian residual error structures in the linear regression model (1.3) (McCullagh and Nelder 1989). A full discussion of the GLM is beyond the scope of this discussion since we focus on situations where the assumption of Gaussian residual errors is applicable.

2.1 Strategies for modeling non-linearity

If the relationship between response and (one of the) explanatory variables severely departs from linearity, then the following, more general, regression form might be considered

$$y_i = f(x_i) + e_i, \tag{2.1}$$

where $f(\bullet)$ may represent any mathematical function, for example, in linear regression, we have $f(x_i) = \beta_0 + x_i\beta_1$. A possible extension of linear regression can be achieved by adding higher degree (i.e. greater than 1) polynomial terms of the same variable x_i (Ruppert et al. 2003):

$$f(x_i) = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + x_i^3\beta_3 + \cdots. \tag{2.2}$$

Since $f(x_i)$ remains to be a linear combination of several covariates, the least squares method introduced above is applicable for estimation. Generally speaking, a quadratic or cubic polynomial function is sufficient to describe data with a simple non-linear relationship; for example, a quadratic polynomial function is often used to model the simple monotonic growth of a tree (e.g. Sillanpää et al. 2012). For more complicated situations, higher degree (i.e. >3) polyno-

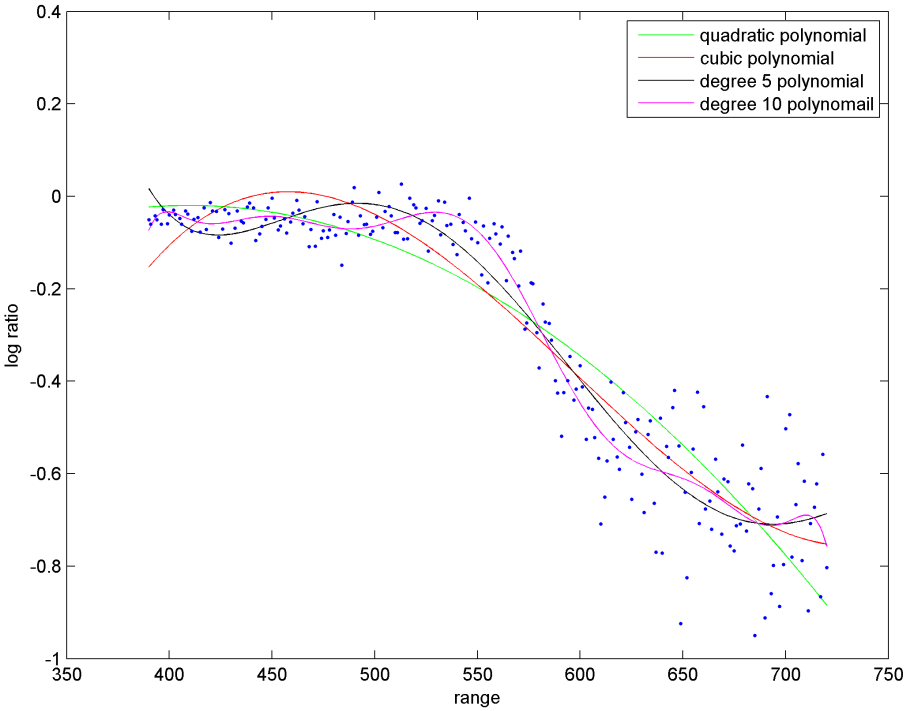

Figure 1: LIDAR data fitted by polynomial regression: estimated curves by polynomials with degree 2 (quadratic), 3 (cubic), 5 and 10 are shown in solid lines with green, red, black and magenta colors, respectively. Original data points are shown in blue dots.

mials might be applicable, but they may not provide substantial improvements in model fitting. Ruppert et al. (2003) provide a nice illustration of this problem in a LIDAR (light detection and ranging) data set that used the reflection of laser-emitted light to detect chemical bounds in atmosphere. The explanatory variable is the distance traveled by the light before it is reflected back to its source (represented as "range"), and the response variable is the logarithm of the received light from two laser sources (blue dots in Figure 1). Clearly, these data are associated in a non-linear pattern, and therefore a higher degree polynomial regression should be used to fit the data. Figure 1 also shows fitted curves by polynomials with degrees 2 (quadratic), 3 (cubic), 5 and 10, respectively in solid lines. Note that the lower degree polynomials (i.e. with orders 2, 3 and 5) do not adequately describe the sudden downturn shown in the middle part of the data, and furthermore they do not seem to fit the data particularly well in either upper or lower boundaries. While, the higher order polynomial function (i.e. degree=10) provides a generally good fits with the data, but the curve is complex and shows many un-necessary wiggles.

Briefly, a common disadvantage of polynomial regressions is that they often fail to properly capture the local trends of certain data that have quite sophisticated non-linear patterns. Next, we discuss a possible improved approach named spline basis extensions. A spline regression with

order s+1 or degree s (spline order=degree+1) is defined by

$$f(x_i) = \beta_0 + x_i\beta_1 + \cdots + x_i^s\beta_s + \sum_{k=1}^{K}(x-\zeta_k)_+^s\beta_{s+k}, \tag{2.3}$$

where $(x-\zeta_k)_+ = x-\zeta_k$ if $x > \zeta_k$ and is equal to 0 otherwise, $A < \zeta_1 < ... < \zeta_K < B$, and $x_i \in [A, B]$. Compared with a polynomial regression, a linear combination of spline bases or truncated power series $\sum_{k=1}^{K}(x-\zeta_k)_+^s\beta_{s+k}$ are further added in order to better describe the local behavior of the data. The values of ζ_k $(k = 1, ..., K)$, which specify the locations where those truncated spline bases are joined, are often refer to as (interior) knots. The number of knots and how they are placed over the range of the explanatory variables, as well as the order of the spline, need to be chosen by the user, and the combination of choices determines the quality of the curve fittings.

A popular variation of the standard spline approach is B-spline (De Boor 2001; Fahrmeir and Kneib 2011). A B-spline basis can be obtained by taking certain differences of the spline bases. B-spline bases are orthogonal, and therefore are numerically more stable especially for some large data sets. Fitting Ruppert et al.'s (2003) LIDAR data with B-splines provided an apparently improved fit compared with the polynomial regression (cf. Figure 1 and 2). For example, we examined four different settings of Knot numbers (K) and spline orders (s): (i) $K = 5$, $s = 2$, (ii) $K = 5$, $s = 4$ (iii) $K = 20$, $s = 2$ and (iv) $K = 20$, $s = 4$, where spline orders 2 and 4 correspond to linear spline and cubic spline, respectively. Since the data are quite evenly distributed over the x-axis, we specified the knot locations to be equal separated. The splines fit the data generally quite well even when the spline orders are low (Figure 2), but with the cubic spline providing a smoother fit than the linear spline at several of the curve peaks and valleys shown in the curve; however, when the number of knots increases, such differences between spline orders are not clear. By contrast, the choice of number of knots has a substantial impact on the smoothness of the curve, with the fitted curves with 5 knots smoother than curves with 20 knots (Figure 2).

Analogous to the linear multiple regression model (1.3), when multiple explanatory variables need to be considered, it is possible to extend (2.1) to an additive model (Ruppert et al. 2003; Hastie et al. 2009):

$$y_i = \sum_{j=1}^{p} f_j(x_{ij}) + e_i. \tag{2.4}$$

When many p curves need to be fitted simultaneously, choosing the most appropriate spline and knot parameters becomes a difficult task, which is rarely possible through data exploration and visualization. **Section 4** will include an introduction to some possible procedures for automatically determining the number of knots.

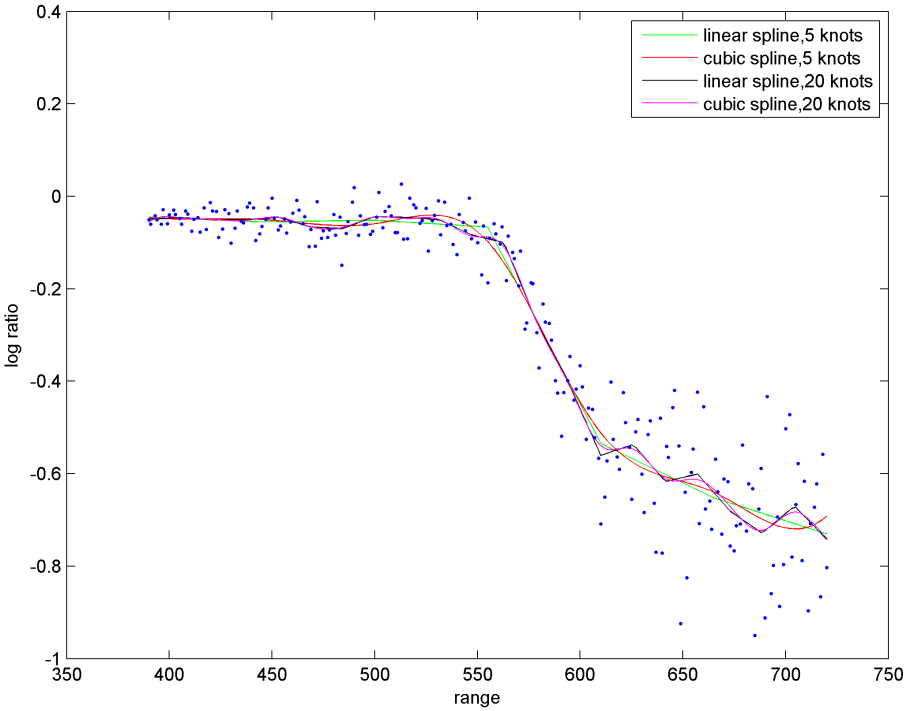

Figure 2: LIDAR data fitted by B-spline regression: estimated curves by B-splines with (i) orders=2 (linear), No. of knots=5, (ii) orders=4 (cubic), No. of knots=5, (iii) orders=20, No. of knots=2 and (iii) orders=20, No. of knots=4. are shown in solid lines with green, red, black and magenta colors, respectively. Original data points are shown in blue dots.

2.2 Linear mixed effects model

In many regression studies, the data might be collected from different sources, groups or clusters, with individuals within one cluster often more correlated with each other than the individuals collected from different clusters. Hence, it may not be appropriate to assume that all the individuals within a pooled sample are independently distributed as with model (1.1). In such circumstances, a linear mixed model (LMM) can be applied in order to account for the structure of the data by adding some cluster-specific random effects into the standard regression model (West et al. 2007). Assuming that among a total of N individuals (in the pooled data), there are m clusters (and that $m < n$), and in the ith cluster ($i = 1, ..., m$) there are n_i number of individuals ($N = \sum_{i=1}^{m} n_i$), then the data can be arranged as $(y_{ik}, \boldsymbol{x}_{ik})$, in which y_{ik} represents responses for clusters $i = 1, ..., m$, and within-cluster samples $k = 1, ..., n_i$, and $\boldsymbol{x}_{ik} = [1, x_{ik1}, ..., x_{ikp}]$ represents covariates for fixed effects. A linear mixed model is defined by

$$y_{ik} = \boldsymbol{x}_{ik}\boldsymbol{\beta} + \boldsymbol{z}_{ik}\boldsymbol{b}_i + e_{ik}, \tag{2.5}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_p]^T$ are the fixed effect coefficients, $\boldsymbol{z}_{ik} = [1, x_{ik1}, ..., x_{ikq}]$ are covariates for random effects ($q < p$) which may be chosen as a subset of the fixed effect covariates $\boldsymbol{x}_{ik}$ (Schelldorfer et al. 2011), $\boldsymbol{b}_i = [b_{i0}, b_{i1}, ..., b_{iq}]^T$ are the random effect coefficients, and e_{ik} are the residual error terms. For the random effects and residual error terms, we may assume $\boldsymbol{b}_i \overset{\text{i.i.d.}}{\sim} \text{MVN}(\boldsymbol{0}, \boldsymbol{\Lambda})$, and $\boldsymbol{e}_i = [e_{i1}, ..., e_{in_i}]^T \overset{\text{i.i.d.}}{\sim} \text{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$. Thus, individuals within one cluster are assumed to be correlated to some degree, and the within cluster correlations are introduced by the covariance structures $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}_i$. Furthermore, the random effects and error terms are assumed to be mutually independent.

In practice, a two-step algorithm can be used to estimate the parameters involved in LMM: (i) estimate the covariances $\boldsymbol{\Lambda}$ and $\boldsymbol{\Sigma}_i$ by the REML method, and (ii) assuming the variance components are known, then estimate fixed and random effect coefficients $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$ as solutions from mixed model equations. More details of this LMM approach can be found in references such as Ruppert et al. (2003).

The LLM methods have been used in various problems in quantitative genetics such as genome wide association studies (Kang et al. 2007), and phenotype prediction/genomic selection studies (De Los Campos et al. 2009). In those applications, the term "cluster" often refers to populations, pedigrees or families from where the data were collected. The focus of this thesis is on longitudinal studies for data sets with repeated measurements over time within each individual, and thus the clusters are the individuals.

2.3 Multivariate regression

Here we discuss another possible extension of the linear regression for multiple response data (i.e. data taken from the same individuals). Sometimes response variables are highly correlated, and yet it is often valuable to simultaneously consider the multiple responses within the same model. As an extension of the univariate regression model (1.3), a linear multivariate regression model can be applied:

$$y_{ik} = \sum_{j=1}^{p} x_{ij}\beta_{jk} + e_{ik}, \tag{2.6}$$

where y_{ik} is the ith observation of the kth response variable ($i = 1, ..., n$, $k = 1, ..., m$), x_{ij} is the ith observation of the jth explanatory variable ($i = 1, ..., n$, $j = 1, ..., p$), β_{jk} is regression coefficient of jth explanatory variable and kth response. The residual errors are assumed to be Gaussian distributed: $\boldsymbol{e}_i = [e_{i1}, ..., e_{ik}]^T \overset{\text{i.i.d.}}{\sim} \text{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Here, the terms "multiple" and "multivariate" may cause some confusion. In this thesis, multivariate regression refers to a regression model with multiple response variables, which is distinct from a univariate regression with a single response variable. The term (univariate) multiple regression refers to a model with

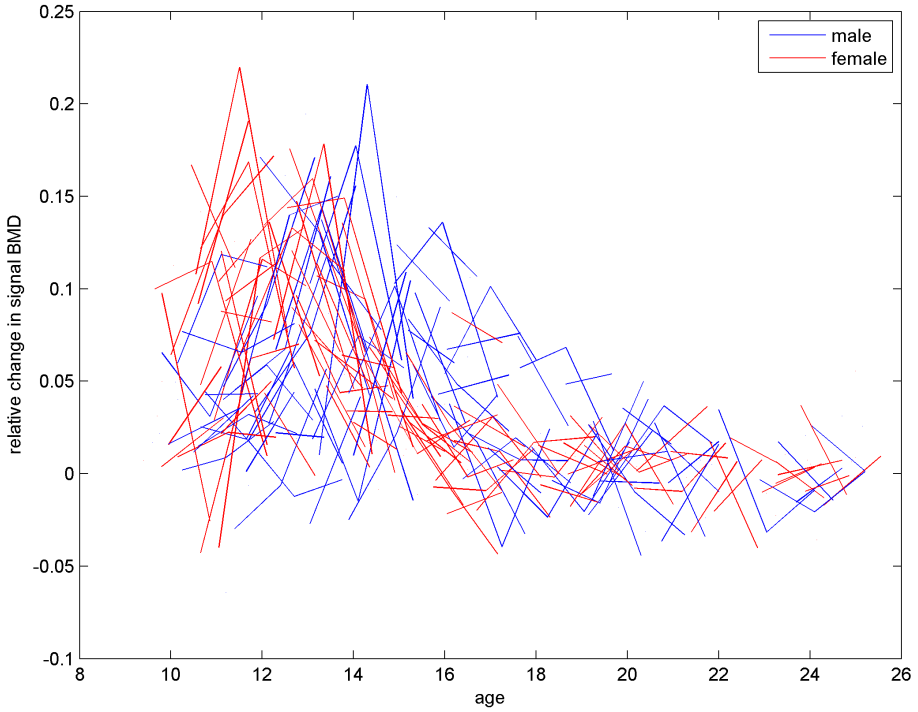

Figure 3: Profile plot of the BMD data: for each individual, his/her relative change in signal BMD is plotted against age. The individual trajectories of males and females are shown in blue and red solid lines, respective.

multiple explanatory variables, as opposed to a simple regression or marginal regression with a single explanatory variable.

It is possible to apply the ordinary least squares (OLS) method to estimate the regression coefficients in (2.6). Indeed, the OLS estimates of (2.6) are equivalent to the OLS estimates obtained by separately fitting k univariate regressions for one response at a time (Izenman 2008). Thus, the OLS approach could not take the covariance among multiple responses into account. In the next section, we discuss a possible modification of model (2.6) which is mainly applicable for the longitudinal data.

3 Regression techniques for longitudinal data

In biology, many quantitative traits such as height and weight, are dynamic in natural populations. A study that aims to quantify dynamic behavior of quantitative traits will use repeated measurements of the target trait taken from the same individual- longitudinal data (Diggle et al. 2002). For example, 1-3 repeated measurements on adolescent bone mineral density (BMD) were taken from 261 North American adolescents (age 9-25) (Bachrach et al. 1999; Hastie et al. 2009) (Figure 3). The impact that gender has on the relative change in signal BMD (see Figure 4) can

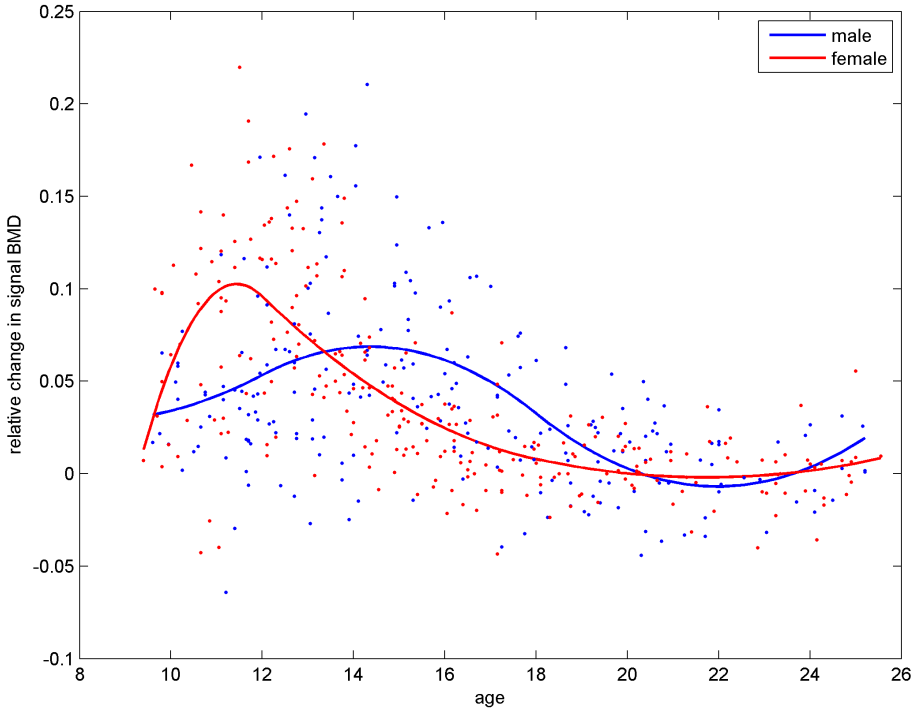

Figure 4: Profile plot of the BMD data by ignoring the labels for individuals: the original data points are shown in blue and red dots for males and female, respectively. The fitted curves by splines are shown in blue and red solid lines for males and females.

be quantified by separately fitting the following quadratic splines with knots specified at ages 12 and 18 to the male and female data, respectively: $f(t) = \alpha_0 + t\alpha_1 + t^2\alpha_2 + (t-12)_+^2\alpha_3 + (t-18)_+^2\alpha_4$. The fitted curves clearly indicate that a spurt in bone growth happens earlier in females than in males (Figure 4).

3.1 A Linear mixed model approach

While profile and/or scatter plots are useful tools for providing a rough description of longitudinal data, regression methods are needed to provide more precise estimation, testing and prediction of interesting quantities. In a longitudinal data set, the common assumption is that the within-individual variability is less than the between-individual variability and therefore it is natural to apply a linear mixed effects model (LMM) (2.5) where the individual can be treated as a cluster. A popular LMM approach for longitudinal data is random intercept and slope model:

$$y_{ik} = \alpha_0 + \alpha_1 t_{ik} + \alpha_{i0} + \alpha_{i1} t_{ik} + \boldsymbol{x}_{ik}\boldsymbol{\beta} + e_{ik}, \tag{3.1}$$

where y_{ik} is the response of kth repeated measurement of subject i ($i = 1, ..., n$, $k = 1, ..., m_i$, n is number of subjects, m_i is the number of repeated measures of the subject i), t_{ik} represents the

time points (e.g. recorded calender time, age...), α_0 and α_1 are (time relevant) fixed intercept and slope parameters, and α_{i0} and α_{i1} are the random intercept and slope parameters, $\boldsymbol{x}_{ik} = [\boldsymbol{x}_{ik1}, ..., \boldsymbol{x}_{ikp}]$ are the fixed effect covariates (other than time), $\boldsymbol{\beta}$ are the fixed effect coefficients of $\boldsymbol{x}_{ik}$, and e_{ik} are residual error terms. Note that in this work, we assume the covariates $\boldsymbol{x}_{ik}$ remain unchanged over time, so we may ignore the label k from $\boldsymbol{x}_{ik}$. Like in (2.5), the random effect and residual errors are assumed to be Gaussian distributed: $\boldsymbol{\alpha}_i = [\alpha_{i0}, \alpha_{i1}]^T \overset{\text{i.i.d.}}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Lambda}_{2\times 2})$, and $\boldsymbol{e}_i = [e_{i1}, ..., e_{im_i}]^T \overset{\text{i.i.d.}}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_i)$. The fixed effect part $\alpha_0 + \alpha_1 t_{ik}$ describes population level linear trend, and the random effect part $\alpha_{i0} + \alpha_{i1} t_{ik}$ describe the individual departure from the population trend. Some data such as the BMD data may have strong non-linear trends. In those cases, the model (3.1) can be extended by

$$y_{ik} = f(t_{ik}) + f_i(t_{ik}) + \boldsymbol{x}_{ik}\boldsymbol{\beta} + e_{ik}. \tag{3.2}$$

Here the fixed and random effects $f(t_{ik})$ and $f_i(t_{ik})$ can be specified by basis expansion approaches such as splines.

For example, analyzing Bachrach et al.'s (1999) BMD data using a mixed model allows us to include an age-gender interaction term in the model in addition to the additive effects. Specially, we fit the following mixed model:

$$y_{ik} = f(t_{ik}) + \alpha_{i0} + x_i\beta + x_i f^*(t_{ik}) + e_{ik}, \tag{3.3}$$

where, $f(t_{ik}) = \alpha_0 + t_{ik}\alpha_1 + t_{ik}^2\alpha_2 + (t_{ik} - 12)_+^2\alpha_3 + (t_{ik} - 18)_+^2\alpha_4$, and $f^*(t_{ik}) = t_{ik}\gamma_1 + t_{ik}^2\gamma_2 + (t_{ik} - 12)_+^2\gamma_3 + (t_{ik} - 18)_+^2\gamma_4$. The gender covariate x_i is coded as 1 for females, and 0 for males. The fixed effect term $f(t_{ik})$ describes the population trend of bone growth, and the interaction term $x_i f^*(t_{ik})$ describes the deviations from the overall trend by females. Since there are only 1-3 repeated measurements over 1 individual, only a random intercept term α_{i0} is used. We further assume $\alpha_{i0} \overset{\text{i.i.d.}}{\sim} N(0, \tau_0^2)$, and $e_{ik} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_0^2)$. Thus it is now clear that the gender variable has significant ($P < 0.05$) interactions with age (Table 1), further indicating gender influences bone growth.

Table 1: Analysis of the BMD data by an age × gender interaction LMM model (3.3): the estimated regression coefficients for the fixe effects and the corresponding statistical significance (p values) by a t test are shown.

variables	x_i (gender)	$x_i \times t_{ik}$	$x_i \times t_{ik}^2$	$x_i \times (t_{ik} - 12)_+^2$	$x_i \times (t_{ik} - 18)_+^2$
coefficients	2.79	-0.5104	0.02283	-0.02719	0.0061
p values	0.0003	0.0001	0.0001	10^{-5}	10^{-5}

3.2 A multilevel model approach

Next, we illustrate an alternative approach called multilevel model for analyzing longitudinal data, as this type of method has been used in several quantitative genetics studies (e.g. Heuven and Janss 2010; Sikorska et al. 2013). The multilevel model can be seen as a two step approach. In a first step, we fit the repeated measured responses y_{ik} of each subject with respect to time t_{ik} $(i = 1, ..., n,\ k = 1, ..., m_i)$:

$$y_{ik} = \psi_{i0} + \psi_{i1} t_{ik} + e_{ik}. \tag{3.4}$$

For simplicity, here we assume linearity between y_{ik} and t_{ik}. In a second step, we take ψ_{i0} and ψ_{i1} as latent response variables. Next, we fit

$$\psi_{i0} = \alpha_0 + \boldsymbol{x}_i \boldsymbol{\beta}_0 + \alpha_{i0}, \tag{3.5}$$

and

$$\psi_{i1} = \alpha_1 + \boldsymbol{x}_i \boldsymbol{\beta}_1 + \alpha_{i1}, \tag{3.6}$$

respectively. Here $\boldsymbol{x}_i$ represents covariates other than time, and α_{i0} and α_{i1} are residual terms. Alternatively, by considering ψ_{i0} and ψ_{i1} as correlated responses, (3.5) and (3.6) can also be simultaneously fitted by a multivariate regression model.

Next, we show how such a multilevel model is connected to the LMM. Substituting (3.5) and (3.6) back to (3.4), we have

$$\begin{aligned} y_{ik} &= \psi_{i0} + \psi_{i1} t_{ik} + e_{ik} \\ &= \alpha_0 + \boldsymbol{x}_i \boldsymbol{\beta}_0 + \alpha_{i0} + (\alpha_1 + \boldsymbol{x}_i \boldsymbol{\beta}_1 + \alpha_{i1}) t_{ik} + e_{ik} \\ &= \alpha_0 + \alpha_1 + \alpha_{i0} + \alpha_{i1} t_{ik} + \boldsymbol{x}_i \boldsymbol{\beta}_0 + \boldsymbol{x}_i \boldsymbol{\beta}_1 t_{ik} + e_{ik}. \end{aligned} \tag{3.7}$$

If we assume α_{i0} and α_{i1} as random effects, then the equation (3.7) becomes the mixed effects model (3.1) plus an extra interaction term $\boldsymbol{x}_i \boldsymbol{\beta}_1 t_{ik}$.

The difference between these two approaches is obvious: the multilevel model first fits the temporal trends of the original responses and the estimates the effects of covariates other than time on latent responses, which are summary statistics of the temporal trend; however the LMM estimates the temporal trend and the effects of the other covariates simultaneously.

3.3 A multivariate varying coefficient regression approach

Now let us focus on a special type of longitudinal data where the repeated measures of all the individuals $(i = 1, ..., n)$ are collected at the same time points $(t_1, ..., t_m)$. In such case, each

$\boldsymbol{y}(t_k) = [y_1(t_k), ..., y_n(t_k)]^T$ $(k = 1, ..., m)$ can be treated as a single response variable. Note that here we change the notation from y_{ik} to $y_i(t_k)$. Therefore, it is possible to apply the multivariate regression model (2.6) to the data:

$$y_i(t_k) = \beta_0(t_k) + \sum_{j=1}^{p} x_{ij}\beta_j(t_k) + e_i(t_k), \qquad \boldsymbol{e}_i = [e_i(t_1), ..., e_i(t_m)]^T \overset{\text{i.i.d.}}{\sim} \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{3.8}$$

In longitudinal contexts, this type of model is often referred to as a varying coefficient model (Ruppert et al. 2003; Fahrmeir and Kneib 2011). As we have mentioned earlier, the standard OLS estimates cannot take the dependency structure among the data into consideration, which is not optimal for parameter estimation in longitudinal data sets. In a longitudinal data, the usual assumption is that two repeated measurements (within the same subject) at nearby time points should be more similar than two measures taken from further apart. A popular strategy is to re-parameterize $\boldsymbol{\beta}_j = [\beta_j(t_1), ..., \beta_j(t_m)]$ $(j = 0, 1, ..., p)$ as a function over time (e.g. by spline basis expansions) instead of considering them as separate parameters. Additionally, when hypothesis testing, it is favorable to construct a test statistic that assess the whole function instead of the single parameters see e.g. Ma et al. (2002) and Xiong et al. (2011). We should also notice that the varying coefficient model (3.8) is closely connected to the mixed model (3.3) with interaction effects. In (3.8), all the covariates are modeled as interactions with time.

Furthermore, the ML estimate of the residual covariance matrix $\boldsymbol{\Sigma}$ may involve too many parameters when the number of time points is large. Under such circumstances it is easier to assume that the residual covariance matrix follows some parametric structure. The simplest setting, for example, is $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is an identity matrix. However, to better describe the temporal correlation among residuals, we may alternatively specify it to be the first order autoregressive structure $\boldsymbol{\Sigma}(s_1, s_2) = \frac{\sigma^2\rho^{|s_1 - s_2|}}{1-\rho^2}$ $(0 < \rho < 1,\ s_1, s_2 = 1, ..., m)$. The temporal correlation decays when the distance between two time points increases.

In the area of quantitative genetics, a series of methods based on the varying coefficient model has been developed for mapping dynamic traits with several different covariance structures (Ma et al. 2002; Liu and Wu 2009); reviewed by Wu and Lin (2006).

4 Model selection and regularization

We have discussed various regression techniques that are applicable to different situations. In general, most methods should be appropriate for data with a sufficiently large sample size (e.g. hundreds of individuals) and relatively few explanatory variables or bases (e.g. less than 10 variables). In many contemporary statistical studies, however, data sets comprise some hundreds or even thousands of explanatory variables- they are high dimensional data. Analyzing such high

dimensional data sets with traditional regression tools is questionable due to computational infeasibility. We therefore briefly discuss the main challenges posed when faced with attempting to analyze large data sets using either standard multiple linear regression (1.3) and/or least squares estimation (1.4), but it is worth noting that similar problems may occur when using other types of regression models (that we have discussed earlier). More detailed descriptions of such challenges in large scale regression analyses can be found in Hastie et al. (2009) and Izenman (2008).

First, from a computational point of view, standard OLS estimation involves calculation of the inverse of a $p \times p$ matrix $\boldsymbol{X}^T\boldsymbol{X}$. The matrix $\boldsymbol{X}^T\boldsymbol{X}$ can easily become ill-conditioned as p increases. In an extreme case, when $p > n$, $\boldsymbol{X}^T\boldsymbol{X}$ starts to become a singular matrix, and the inverse matrix does not exist. Second, from the model fitting standpoint, a linear model with a large number of explanatory variables may provide a good fit with the original data, but subsequently fit the new data poorly due to over-fitting (see e.g. **Section 1.2**).

Two strategies will likely improve model estimation. First, we may retain a few of the likely important explanatory variables into the model, and exclude the remaining, presumably unimportant, variables based on certain selection criteria- i.e. use model (variable) selection methods. Second, it may be possible to keep all explanatory variables in the model, but add a penalty to an OLS estimator in order to help inversion of ill-conditioned matrix, -the so called regularization methods (Izenman 2008). Regularization methods typically push the coefficients of unimportant variables towards zero (setting a regression coefficient to zero is equivalent to excluding the corresponding variable from the model), while keeping large coefficients of important variables. Hence regularization methods aim to reduce model complexity, similar to the aims of variable selection, and thus may be considered as a variable selection method.

Now let us extend the discussion of model/variable selection techniques to a broader class of regression techniques. In brief, the model selection can be used to achieve the following:

Selection of the fixed effects: Similar to the variable selection problem in a standard linear regression, in a linear mixed effects model (2.5) with high dimensional covariates for fixed effects, it is often beneficial to select a small number of important variables for fixed effects into the model. The variable selection can be used in multivariate regression (2.6) and additive models (2.4) as well, but there the selection of one variable corresponds to the selection of a group of regression coefficients instead of a single parameter.

Selection of knots or degree of smoothness: In **Section 2.1**, we demonstrated the importance of choosing the degree in a polynomial model (2.2), or choosing the number of knots in a spline model (2.3): a model where the degree polynomials are too low or where there are two few knots cannot describe the complex pattern of the data (i.e. is underfitted);

conversely, a model with too many bases will fit the data "too well" by showing a lot of unnecessary wiggles (i.e. is overfitted). Below in **Section 4.5**, we focus on discussing about the automatic strategies for selecting number of knots in B-splines.

Selection of random effects and covariance structure: In a mixed effects model, another important issue is to select covariates for the random effects and specifying correct covariance structures for both random effects and residual errors. Besides, specifying a good residual covariance structure is also important for multivariate regression (e.g. see Müller et al. (2013).

The publications (or submitted manuscripts) of this work focus on (i) variable selection in a standard linear regression model (Articles I and II), (ii) selection of fixed effects explanatory variables in a longitudinal mixed model (Article IV), and (iii) selection of both explanatory variables and knots in a multivariate varying coefficient model (Article III). Note, however, that the selection of random effects or covariance structure in the mixed model or varying coefficient model are not an essential topic of the thesis. In the remaining part of this section, we introduce some technical backgrounds of model selection from a frequentist statistics point of view in order to support the main articles in the thesis. Finally, we discuss Bayesian model selection.

4.1 Model selection criteria

Evaluating the success of a regression model requires useful tools for model assessment. For this purpose, a model may further refer to a subset of explanatory variables in a multiple linear regression or a number of knots in splines. Generally speaking, a model will be judged as good if it provides good predictability, is parsimious and is easy to interpret.

In order to assess a model based on its predictability, we may either evaluate its extra-sample or in-sample prediction error (Hastie et al. 2009). The extra-sample error is estimated under the assumption that there are two separate data sets: a training data set $(\boldsymbol{X}, \boldsymbol{y})$, and a test data set $(\boldsymbol{X}_{\text{new}}, \boldsymbol{y}_{\text{new}})$. The training data are used for estimating the regression coefficients (e.g. by ordinary least squares), and the test data are used to evaluate the prediction error. The in-sample prediction, by contrast, use a single data set under the assumption that there are new measurements of the responses $\boldsymbol{y}$ on the same input data $\boldsymbol{X}$.

A simple approach to estimate the extra-sample predictability is to divide a data set into two parts of roughly equivalent size: with one part of the data used for learning, and the remaining data then used for validation and evaluation of prediction error. Dividing a data set is only appropriate for large data sets. When the sample sizes are low, one can apply a K-fold cross validation (CV) method (Picard and Cook 1984). The CV divides the data into K equivalent

parts $(\boldsymbol{X}_k, \boldsymbol{y}_k)$ $(k = 1, ..., K)$, with each part $(\boldsymbol{X}_k, \boldsymbol{y}_k)$ used sequentially as a test set, and the remaining $K-1$ parts $(\boldsymbol{X}_{k-1}, \boldsymbol{y}_{k-1})$ used for training, and the prediction error is estimated as an average over the K runs: typically, a 5 or 10-fold CV is recommended in practice (Hastie et al. (2009).

Methods for estimating in-sample prediction error (IPE) often start from the point that the training error (or MSE) often gives an overly optimistic estimate of prediction error. Since, the IPE should be larger than MSE, we may represent their relationship by IPE=MSE+O, where O should be a positive value. Asymptotically, this relationship leads to the Akaike information criterion (AIC) (Akaike 1974):

$$\text{AIC} = \ln \text{MSE} + \frac{2}{n}\text{df}, \tag{4.1}$$

or from the maximum likelihood point of view, AIC can also be defined by

$$\text{AIC} = -\frac{2}{n} \ln \text{max-likelihood} + \frac{2}{n}\text{df}, \tag{4.2}$$

where df represents degree of freedom, the number of effective parameters in the model, and n is the sample size. In a multiple regression model (1.3), df is equivalent to the total number of regression coefficients, which is $p+1$. As the term $\frac{2}{N}\text{df}$ penalizes a model with more parameters, AIC is seen as an approach for achieving a compromise between obtaining the best model fit and keeping model complexity comparatively low.

An alternative choice for assessing model fit is Bayesian information criterion (BIC) (Schwarz 1978), defined by

$$\text{BIC} = \ln \text{MSE} + \frac{\ln n}{n}\text{df}, \tag{4.3}$$

or

$$\text{BIC} = -\frac{2}{n} \ln \text{max-likelihood} + \frac{\ln n}{n}\text{df}. \tag{4.4}$$

Note that as $\ln n > 2$, when $n > 7$, the BIC tends to favor more parsimonious model compared with AIC. As suggested by its name, the BIC has a Bayesian origin, and this aspect will be covered in the next section.

4.2 Stepwise selection methods

Now we start to discuss some automatic model selection procedures for a regression model. Here, we first focus on the standard multiple linear regression (1.3). Since the main aim of multiple regression is to select a subset of important explanatory variables to construct a good model, the problem is often referred to as "variable selection" instead of "model selection". Intuitively, one may apply "all best subset selection" (Kutner et al. 2004), an approach that simply enumerates

all possible combinations of explanatory variables. For each combination of variables, the above mentioned model selection criteria such as CV, AIC and BIC is used to measure the model goodness of fit, and the model with optimal selection score is selected as the best. Despite its simplicity, this approach is often not applicable in practice due to its heavy computation cost. For example, with p variables, the total number of possible models is 2^p; thus when $p = 100$, there are $2^{100} \approx 1.27 \times 10^{30}$ possible models to assess.

A computationally more effective alternative is a stepwise method, which includes forward selection and backward elimination of variables. Forward selection starts from a null model (i.e. with only the intercept term), and then adds explanatory variables one at a time into the model with the improvement of the model, judged, for example by CV, AIC, BIC or a t statistic. This process continues until explanatory variables cease to improve the model. In contrast, backward elimination begins with a full model (i.e. all the variables are involved), and then removes variables from the model, again one variable at a time, and continues with this process as long as the model improves, after which no more variables are deleted. Moreover, it is also possible to use a combination of forward selection and backward elimination in one algorithm (see Kutner et al. 2004).

4.3 Regularization methods

The stepwise methods and similar variants, are greedy algorithms in that, they can easily reach some local maxima. In addition, such discrete model search strategies are not stable, meaning that they may provide quite different results even if there is only a small change in the data sets (Hastie et al. 2009; Izenman 2008). Instead, some continuous procedures, mainly referring to regularization methods, are able to overcome these problems.

A classic regularization method is ridge regression (Hoerl and Kennard 1970), defined by

$$\hat{\boldsymbol{\beta}}_r = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2. \tag{4.5}$$

An l_2 penalty term $\lambda \sum_{j=1}^{p} \beta_j^2$ is added to the SSE function in order to shrink the regression coefficients towards zero. The tuning parameter λ determines the degree of shrinkage. In practice, an optimal value of λ ($\lambda > 0$) can be chosen by using any of the above mentioned model selection criteria. The ridge estimates can be analytically derived as

$$\hat{\boldsymbol{\beta}}_r = (\boldsymbol{X}^T\boldsymbol{X} + \lambda A)^{-1}\boldsymbol{X}^T\boldsymbol{y}, \tag{4.6}$$

where $\boldsymbol{X}$ and $\boldsymbol{y}$ are defined in the same way as in equation (1.4), and $A = \begin{pmatrix} 0 & \cdots & & 0 \\ \vdots & 1 & & \vdots \\ & & \ddots & \vdots \\ 0 & \cdots & & 1 \end{pmatrix}$.

The ridge regression has been criticized because it tends to over-shrink regression coefficients of some explanatory variables towards zero, even though those variables are considered to be important; in other words, ridge regression can fail to distinguish between important variables and un-important variables. In addition, due to the continuous nature of the l_2 penalty, ridge regression cannot shrink any regression coefficient exactly to zero, and therefore cannot provide a parsimonious model.

A somewhat newer approach is least absolute shrinkage and selection operator or LASSO (Tibshirani 1996), defined by

$$\hat{\boldsymbol{\beta}}_l = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{4.7}$$

In LASSO, an l_1 penalty $\lambda \sum_{j=1}^{p} |\beta_j|$ is used instead of the l_2 penalty. The l_1 norm, dis-continuing at zero, guarantees that some coefficients can be shrunk exactly to zero, with LASSO also benefitting by tending to shrink the coefficients of important variables than ridge regression. These properties can be better seen from a Bayesian point of view, which will be explained in **Section 5.2**. The LASSO solution is not analytically available, and requires sophisticated convex optimization algorithms such as least angle regression (LARs) (Efron et al. 2004) and the coordinate descent algorithm (Friedman et al. 2007; 2010).

The l_1 penalty can also be used for selection of fixed effects in a linear mixed effects model (see Schelldorfer et al. 2011).

4.4 LASSO and multiple hypothesis testing

A nice feature of LASSO is that it is able to shrink the coefficients of some explanatory variables exactly to zero, and therefore LASSO can be regarded as a variable selection tool. However, some theoretical and empirical studies have indicated that LASSO tends to choose too many variables into the model, so that some variables with quite small coefficients are included in addition to those variables with large effects (Bühlmann and Van De Geer 2011). This retention of variables with weak effect typically happens especially when the CV is used to choose the optimal value of the tuning parameter. Thus, if the choice of significant variables relies on LASSO variable selection, we have to tolerate some false positives (i.e. type 1 error rate). Hypothesis testing is needed to reduce the number of false positives. Since LASSO is often applied on high

dimensional data, a multiplicity adjustment might be needed when doing simultaneous testings on many parameters (Meinshausen et al. 2009).

In general, it is not a simple task to perform hypothesis testing based on the LASSO estimates. Constructing a test statistic for a LASSO estimate of a regression coefficient is not a straightforward task, because LASSO estimates do not asymptotically follow any standard parametric distribution. Some existing approaches such as those presented by Wasserman and Roeder (2009), Meinshausen et al. (2009), and Minnier et al. (2011) are based on sub-samplings or perturbations.

4.5 Model selection on B-splines

Now we move to the problem of selecting degree of smoothness or knots in spline regression. Here we focus on the B-spline bases, an orthogonalized version of the standard splines (2.3). We start from the case of single explanatory variable x_i. The regression model is specified as

$$y_i = \sum_{k=1}^{m} \psi_{ik}(x_i)\beta_k + e_i, \tag{4.8}$$

where $\psi_{ik}(x_i)$ $(k = 1, ..., m)$ are m B-spline bases. As shown in **Section 2.1**, the number of knots is the key factor determining the smoothness of the estimated curve, with too many or too few number of knots causing overfitting and underfitting respectively. One strategy to obtain a good model is to pre-specify a fairly large number of knots in the model to ensure that it does not under-fit the data. Because equation (4.8) has the same linear form as the standard multiple linear regression (1.3), the same type of variable selection or regularization methods can then be applied in order to avoid over-fitting. Here, we focus on the ridge regression with an l_2 norm penalty $\lambda \sum_{k=1}^{m} \beta_k^2$, because it has a simple analytical solution form. The ridge regression independently assign an individual squared penalty to the parameter of spline base with a common penalty factor λ. In some situations such as in longitudinal studies, the usual assumption is that the response data at nearby (time) points have more similar values compared to the data measured at further distances. In order to better describe such dependency structure, it is often preferable to alternatively use a fusion or difference penalty. The corresponding penalized B-spline regression is called p-spline (Eilers and Marx 1996). In some literature, these models often refer to non-parametric or semi-parametric regression models (Ruppert et al. 2003; Fahrmeir and Kneib 2011). The widely used first and second order difference penalties are $\lambda \sum_{k=2}^{m} (\beta_k - \beta_{k-1})^2$ and $\lambda \sum_{k=3}^{m} (\beta_k - 2\beta_{k-1} + \beta_{k-2})^2$, respectively. These differential penalties push the coefficients of the bases at adjacent knots to similar values, and thus smooth the estimated curve. We used the p-spline method to re-analyze the LIDAR data examples in

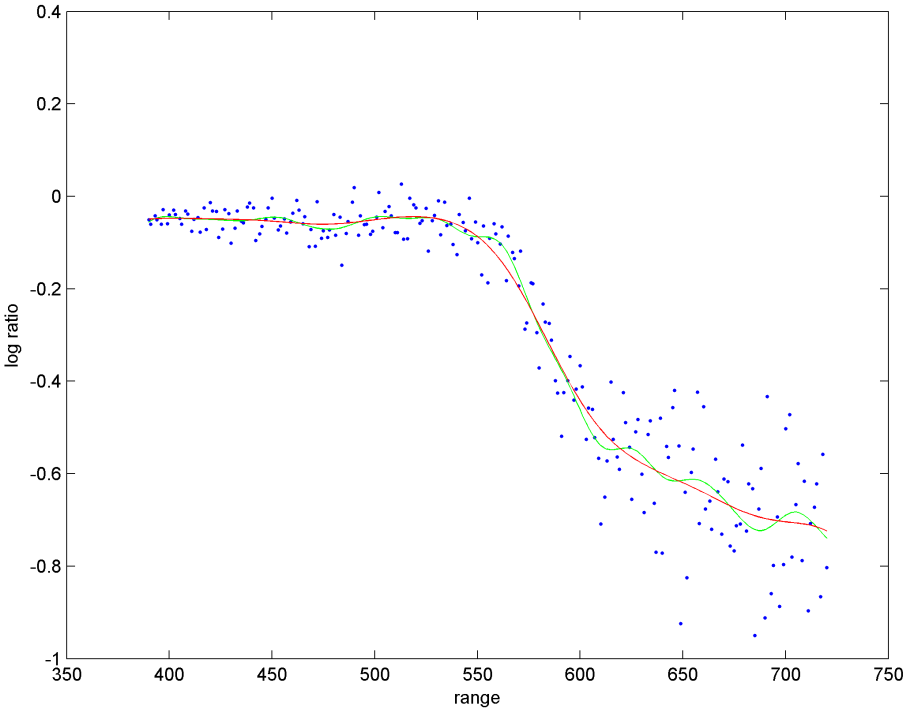

Figure 5: Comparison of fitted curves by p-spline with a second order difference penalty and B-spline without any penalty for the LIDAR data: estimated curves by p-spline and regular B-spline regressions are shown in red and green colors, respectively. Original data points are shown in blue dots.

Section 2.1. The cubic spline and number of knots were specified to be 20, and we choose the second order difference penalty. The 10-fold CV was used to select an optimal value of λ. Compared with the curve estimated by the B-spline method without any penalty, the penalized method provides much smoother fit without presenting any undue undulations in the curve (Figure 5). It is possible to extend the idea of p-spline to the additive model (2.4) and the varying coefficient model (3.8) by assigning the difference penalties to the coefficients of each explanatory variables. When the number of explanatory variables is large, a combination of the LASSO l_1 penalty and the difference penalty can be used to achieve variable selection and curve smoothing simultaneously (Daye et al. 2012).

5 Bayesian formulation and computation

So far, we have introduced the various regression models, and the relevant estimation, model selection and inference issues from a frequentist statistics point of view. Now, we introduce the Bayesian way of handling model selection and associated validation problems. In a Bayesian statistical model, there are three key factors: (i) a likelihood function $p(\boldsymbol{y}|\boldsymbol{\theta})$, the probabilistic description of data $\boldsymbol{y}$ conditional on the unknown parameters $\boldsymbol{\theta}$, (ii) a prior $p(\boldsymbol{\theta})$, the probabilistic

hypothesis of the parameters without knowing the data, and (iii) a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$, the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{y}$. Note that here a probability distribution is represented in the density form. The Bayes theorem tells that

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}. \tag{5.1}$$

Note that the denominator $p(\boldsymbol{y}) = \int p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is a normalizing constant. Thus, sometimes we might also express the Bayes theorem as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{5.2}$$

The posterior distribution combines the data likelihood with the prior information. This in principle differs from the frequentist statistics, where the inference is done merely based on the likelihood. Another difference is that in a Bayesian model, the parameters are often estimated as a whole posterior distribution, so that the uncertainties such as standard errors and credible intervals are directly estimable. However, frequentist approaches such as maximum likelihood, only provide point estimates of parameters and the level of uncertainty has to be estimated from the sample distribution.

In a Bayesian model, the choice of priors is an important issue. When good prior knowledge about the parameters is available, we may choose a relatively informative prior and this, may have a quite large impact on the posterior, especially when there are few data. In the absence of good knowledge about the parameters, a flat non-informative prior might be preferable. While, the posterior distribution shrinks to the likelihood when the prior is chosen to be flat, it can still be useful to take advantage of Bayesian computational tools. More details about Bayesian modeling and computation can be found in Gelman et al. (2004).

5.1 Marginal likelihood and BIC

Now we move to some Bayesian approaches for model selection. Following a Bayesian approach, we may treat a model M as a random variable as similar as the model parameters $\boldsymbol{\theta}$. Here we focus on discrete model space $M \in [M_0, M_1, ..., M_{N-1}]$ with the corresponding model parameters $\boldsymbol{\theta} \in [\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_{N-1}]$, where N is the total number of possible models. For instance, in a linear regression (1.3) with p explanatory variables, there are $N = 2^p$ possible models. We need to compute the posterior distribution of the model M:

$$p(M|\boldsymbol{y}) \propto p(M)p(\boldsymbol{y}|M), \tag{5.3}$$

where $p(M)$ is prior probability of the model, and $p(\boldsymbol{y}|M)$ is marginal likelihood, which is equivalent to the integral $\int p(\boldsymbol{y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}_m$, and coincidences with the denominator of (5.1). In practice, people often specify equal prior probabilities to the models, and in that case the posterior is equivalent to the marginal likelihood.

After computing the posterior distribution for all the models, we may seek the optimal model that gives the highest posterior probability. Alternatively, it is also possible to calculate the Bayes factor (BF) in order to compare two models M_1 and M_0.

$$\begin{aligned} \mathrm{BF}_{10}(\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_0)} \\ &= \frac{p(M_1|\boldsymbol{y})}{p(M_0|\boldsymbol{y})}\frac{p(M_0)}{p(M_1)}. \end{aligned} \tag{5.4}$$

As a 'rule of thumb' if BF_{10} is greater than 3 (or $2\ln BF_{10}$ is greater than 2), then model M_1 should be favorable over model M_0 (see Kass and Raftery 1995).

Clearly, the computation of the marginal likelihood is crucial for performing calculation of model posterior and BF. In many situations, the integral computation is often not tractable, which requires numerical solutions. Both stochastic sampling and determinist approximation methods are applicable. One possibility is using a Laplace approximation to the logarithm of the integral, which results in the following solution

$$\ln p(\boldsymbol{y}|M) \approx \ln p(\boldsymbol{y}|M, \hat{\boldsymbol{\theta}}) - \frac{\ln n}{2}\mathrm{df}, \tag{5.5}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate, n is the sample size, and df is the degree of freedom (Hastie et al. 2009). Note that $-2\ln p(\boldsymbol{y}|M)$ is equivalent to the above defined BIC form in (4.4). Below, we show an alternative solution for approximating the marginal likelihood by using a variational Bayes method.

5.2 Bayesian regularized linear model

We describe the Bayesian representation of the multiple linear regression model (1.3). As we have shown earlier, it is possible to write the equation (1.3) in the likelihood function form:

$$p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma_0^2) = (2\pi)^{-\frac{n}{2}}(\sigma_0^2)^{-\frac{n}{2}}\exp[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}x_{ij}\beta_j)^2]. \tag{5.6}$$

For the regression coefficients $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_j$ $(j = 1, ..., p)$, we may specify the following normal priors: $\boldsymbol{\beta}_0 \sim N(0, \tau_0^2)$ $(\tau_0^2 > 0)$, and $\boldsymbol{\beta}_j \sim N(0, \tau^2)$ $(\tau^2 > 0)$. For the residual variance, an Inverse-gamma prior is used: $\sigma_0^2 \sim \mathrm{IG}(a, b)$ $(a, b > 0)$. The normal and Inverse-gamma prior

are conjugate priors for $\boldsymbol{\beta}$ and σ_0^2, guaranteeing that the full conditional posterior distributions $p(\boldsymbol{\beta}|\sigma_0^2, \boldsymbol{y})$ and $p(\sigma_0^2|\boldsymbol{\beta}, \boldsymbol{y})$ are also in the normal and Inverse-gamma distribution form, respectively. Conjugate priors are important for constructing some efficient computational algorithms such as Gibbs sampling methods and variational approximation methods (which will be discussed in **Section 5.4**).

The next issue is to specify certain values for the hyper-parameters τ_0^2, τ^2, as well as a and b defined in the priors. If we want the priors on the regression coefficients to be flat and non-informative (i.e. with very limited impact on the posterior), then we may fix variance components τ_0^2 and τ^2 to be very large values, such as $\tau_0^2 = \tau^2 = 10^6$. An extreme choice is the improper uniform priors: $p(\beta_0) \propto 1_{(-\infty,+\infty)}$ and $p(\beta_j) \propto 1_{(-\infty,+\infty)}$, so that the priors densities are constant over the parameter space and have no influence on the posterior at all. In this case, the Bayesian mode point estimates of $\boldsymbol{\beta}$ are just equivalent to the OLS estimates. For the prior of the residual variance, a popular non-informative prior setting is to specify the hyper-parameters a and b to be quite small, such as $a = b = 0.0001$.

5.2.1 Bayesian ridge regression

Alternatively, when some regularization for the regression coefficients is needed, we should choose more informative prior or the regression coefficients of the explanatory variables β_j $(j = 1, ..., p)$. That means the value of τ^2 should be chosen not to be very large in order to make sure that the prior $p(\beta_j) = N(\beta_j|0, \tau^2)$ has non-negligible effect on the posterior evaluation, i.e. to shrink some coefficients toward zero. It is popular to treat the τ^2 as a random variable as well, and a hyper-prior $p(\tau^2)$ can be specified for it. By doing this, we obtain a following hierarchical prior setting for β_j as $p(\beta_j|\tau^2)p(\tau^2)$. For simplicity, here we specify $p(\beta_0) \propto 1_{(-\infty,+\infty)}$, $p(\sigma_0^2) \propto 1_{(0,+\infty)}$ and $p(\tau^2) \propto 1_{(0,+\infty)}$. The posterior becomes

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma_0^2, \tau^2|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma_0^2) \prod_{j=1}^{p} p(\beta_j|\tau^2)p(\tau^2)p(\beta_0)p(\sigma_0^2) \\ &\propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma_0^2) \prod_{j=1}^{p} p(\beta_j|\tau^2) \\ &\propto (\sigma_0^2)^{-\frac{n}{2}} \exp[-\frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2](\tau^2)^{-\frac{p}{2}} \exp[-\frac{1}{2\tau^2} \sum_{j=1}^{p} \beta_j^2]. \end{aligned} \quad (5.7)$$

Temporally, we assume σ_0^2 and τ^2 to be fixed. The maximization of the posterior with respect to $\boldsymbol{\beta}$ leads to:

$$\begin{aligned}
&\hat{\boldsymbol{\beta}} = \max_{\boldsymbol{\beta}} p(\boldsymbol{\beta}, \sigma_0^2, \tau^2 | \boldsymbol{y}) \\
\Leftrightarrow\ &\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} [-\ln p(\boldsymbol{\beta}, \sigma_0^2, \tau^2 | \boldsymbol{y})] \\
\Leftrightarrow\ &\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \frac{\sigma_0^2}{\tau^2} \sum_{j=1}^{p} \beta_j^2. \qquad (5.8)
\end{aligned}$$

Interestingly, $\hat{\boldsymbol{\beta}}$ is equivalent to the ridge regression estimate (4.5). We obtain a Bayesian interpretation of the ridge regression with l_2 penalty: a normal prior with zero mean and a common variance τ^2 on each coefficient β_j $(j = 1, ..., p)$ (Hsiang 1975). In the frequentist ridge regression, the tuning parameter $\hat{\lambda} = \frac{\sigma_0^2}{\tau^2}$ is explicitly selected by model selection criteria such as cross validation. In the Bayesian approach, σ_0^2 and τ^2 are considered as random variables similarly as the regression coefficients, and all the parameters are simultaneously estimated in a same procedure.

Furthermore, the Bayesian ridge regression model (5.7) can also be thought of as a linear mixed effects model (2.5). The intercept term β_0 can be seen as an fixed effect, and $\sum_{j=1}^{p} x_{ij}\beta_j$ can be seen as the random effects. This indicates that we can use any computational algorithm for LMM to solve the ridge regression problem. Conversely, we can treat any other mixed effects model as a Bayesian model, and apply the Bayesian computational tools on those models.

5.2.2 Bayesian LASSO

Similarly, the LASSO has a Bayesian interpretation (Tibshirani 1996). The l_1 penalty $\lambda|\beta_j|$ on each regression coefficient β_j corresponds to a Laplace (or double exponential) prior

$$p(\beta_j) = \frac{\lambda}{2} \exp(-\lambda|\beta_j|). \qquad (5.9)$$

The Laplace prior is a non-conjugate prior, and may cause some troubles for Bayesian computation. Inspired by the fact that the Laplace distribution is equivalent to a scale mixture of normals

$$\frac{\lambda}{2} \exp(-\lambda|\beta_j|) = \int_0^\infty \frac{1}{\sqrt{2\pi\tau_j^2}} \exp(-\frac{\beta_j^2}{2\tau_j^2}) \frac{\lambda^2}{2} \exp(-\frac{\lambda^2\tau_j^2}{2}) d\tau_j^2, \qquad (5.10)$$

a hierarchical conjugate prior $p(\beta_j|\tau_j^2)p(\tau_j^2) \propto N(\beta_j|0, \tau_j^2)Exp(\tau_j^2|\frac{\lambda^2}{2})$ are often used in practice to replace the Laplace prior (Figueiredo 2003; Park and Casella 2008; Yi and Xu 2008). In contrast with the ridge regression, here each β_j owns an individual level variance parameter τ_j^2. Thus, the LASSO is able to provide more flexible and adaptive estimates than provided

by a ridge regression, i.e. by shrinking less if the coefficient is large, and shrinking more if the coefficient is small.

5.2.3 Spike and slab priors

Another popular hierarchical prior setting for β_j is the spike and slab prior (Kuo and Mallick 1998; O'Hara and Sillanpää 2009):

$$p(\beta_j|r_j,\tau^2) = (1-r_j)1_{\{\beta_j=0\}} + r_j N(0,\tau^2), \tag{5.11}$$

where r_j ($r_j = 0, 1$) is a binary indicator variable. Furthermore, we may specify a conjugate Bernoulli prior to r_j, and an Inverse-gamma prior to τ^2. The spike and slab prior is a mixture of a point mass at zero and a normal distribution. When $r_j = 0$, we have $\beta_j = 0$, and variable is excluded from the model. When $r_j = 1$, we have $\beta_j \neq 0$, and the explanatory variable is believed to be important. The indicator plays a role on excluding the un-important variables, and enhances the effects of important variables. Another nice feature of the spike and slab prior is that the posterior mean estimate of r_j can be used as a posterior inclusion probability (PIP) in order to quantify how important the variable is.

The Bayesian representation of the linear mixed effect model and varying coefficient model for the longitudinal data can be found in Articles IV and III.

5.3 MCMC sampling

In Bayesian regularized regression models, the posteriors are intractable, and need numerical solutions. As with the frequentist maximum likelihood approach, it is possible to seek the (numerical) point estimates of the parameters which maximize the posterior; this is often called Maximum a posteriori (MAP) estimation. As we have pointed out earlier, there are some alternative methods in Bayesian statistics that are able to evaluate the whole posterior distribution instead of only producing the point estimates. Two such approaches involved in this thesis will be briefly introduced.

We first start with the Markov Chain Monte Carlo (MCMC) algorithms, which is a class of stochastic methods for simulating the posterior distribution. They target at generating a Markov Chain: a sequence of dependent samples, which converges to the target posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$.

5.3.1 Metropolis-Hastings sampling

Metropolis-Hasting (MH) sampler (Metropolis et al 1953; Hastings 1970) uses an acceptance/rejection rule to generate a Markov chain that converges to the target distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. Assuming an initial state $\boldsymbol{\theta}^{(0)}$ has been given, a proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(1)})$ is used to suggest a value $\boldsymbol{\theta}'$ for the next state. Next, we calculate the Metropolis-Hastings acceptance ratio

$$r = \frac{p(\boldsymbol{\theta}'|\boldsymbol{y})q(\boldsymbol{\theta}^{(1)}|\boldsymbol{\theta}')}{p(\boldsymbol{\theta}^{(1)}|\boldsymbol{y})q(\boldsymbol{\theta}'|\boldsymbol{\theta}^{(1)})}, \tag{5.12}$$

The proposal value is accepted by the new $\boldsymbol{\theta}^{(1)}$ with the probability $\min(r, 1)$, otherwise we set $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)}$. Following the same rule, we simulate a dependent sample with sufficient length, to guarantee that it converges to the target posterior. The convergence may be checked by either simple visual inspection or by more formal decision tools (Gelman et al. 2004). Another issue is to choose a good proposal density to make sure that the average acceptance ratio of the chain is neither too high nor too low. Robert and Casella (2004) provides details about some common choices of proposal densities.

5.3.2 Gibbs sampling

The Gibbs sampler (Geman and Geman 1984) is an alternative way of setting up an MCMC algorithm. In a Gibbs sampler we simulate each single component θ_j $(j = 1, ..., N)$ of $\boldsymbol{\theta} = [\theta_1, ..., \theta_N]$ successively from its full conditional distribution $p(\theta_j|\boldsymbol{y}, \boldsymbol{\theta}_{-j})$, where $\boldsymbol{\theta}_{-j} = [\theta_1, ..., \theta_{j-1}, \theta_{j+1}, ..., \theta_N]$. Though it was developed independently from the MH methods, the Gibbs sampler is in fact closely connected to the MH sampler, indeed presenting a special case of MH, where a new state generated from the proposal distribution is always accepted. Because of this, the Gibbs sampler should be faster and easier to be used in practice, and it should be a preferable choice for large data sets. A general requirement for using a Gibbs sampler is that the full conditional distributions for all the parameters are tractable and thus as many conjugate priors as possible should be used. For many problems, we could apply a combination of Gibbs and MH samplers, so that the Gibbs sampling is used for most of the parameters which are conjugate, and the MH sampling is only used for the non-conjugate parts (Gelman et al. 2004). Gibbs samplers for Bayesian LASSO and two other related models can be found in Article II, and the Gibbs sampler for a longitudinal linear mixed model with spike and slab priors for selecting fixed effects can be found in Article IV.

5.3.3 Posterior summarization

The MCMC approach generate posterior samples for each parameter θ_j $(j = 1, ..., n)$, which approximates its marginal posterior distribution. From the MCMC samples, it is easy to obtain both the point estimates such as posterior mean and posterior median, and uncertainty estimates such as standard error and credible interval for a parameter (Kyung et al. 2010).

5.4 Variational approximation

As a sampling based method, MCMC, is able to provide a quite accurate approximation to the exact posterior distribution of a high dimensional linear model but with huge time demands (Carbonetto and Stephens 2012). It is often preferable to use a faster method by sacrificing some estimation accuracy when dealing with some large data sets, with potentially good alternatives including some determinist approximation algorithms such as the Laplace approximation (Bishop 2006), variational Bayesian methods (Jaakkola and Jordan 2000; Beal 2003; Ormerod and Wand 2010) and expectation propagation (Minka 2001); however, Our focus is on the variational Bayes (VB) approach.

The VB method seeks a tractable free form of variational distribution $q(\boldsymbol{\theta}|\boldsymbol{y})$ which approximates the exact (intractable) posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$. One popular setting of $q(\boldsymbol{\theta}|\boldsymbol{y})$ is in a factorized form:

$$q(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{N} q(\theta_i|\boldsymbol{y}). \tag{5.13}$$

For simplicity, we assume that the approximate posterior is a product of the marginal posteriors of each single parameter θ_i. It is also possible to divide $\boldsymbol{\theta}$ into $[\boldsymbol{\theta_1}, ..., \boldsymbol{\theta_M}]$ $(M < N)$, where $\boldsymbol{\theta_j}$ $j = 1, ..., M$ might be a group of parameters.

We seek an estimate $\hat{q}(\boldsymbol{\theta}|\boldsymbol{y})$, that minimizes the Kullback-Leibler (KL) divergence (Kullback and Leibler 1951):

$$\mathrm{KL}(q||p) = \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}|\boldsymbol{y}) \ln \frac{q(\boldsymbol{\theta}|\boldsymbol{y})}{p(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta}. \tag{5.14}$$

This approach guarantees that the variational solution is at the closest distance to the true posterior in the probability sense. By minimizing the KL function with respect to the approximate marginal posterior of each single parameter: $q(\theta_i|\boldsymbol{y})$ $i = 1, ..., N$, we obtain

$$\hat{q}(\theta_i|\boldsymbol{y}) = \frac{\exp\{E_{\hat{q}(\boldsymbol{\theta}_{-i}|\boldsymbol{y})}[\ln\, p(\boldsymbol{\theta}, \boldsymbol{y})]\}}{\int_{\boldsymbol{\Theta}_i} \exp\{E_{\hat{q}(\boldsymbol{\theta}_{-i}|\boldsymbol{y})}[\ln\, p(\boldsymbol{\theta}, \boldsymbol{y})]\} d\theta_i} \tag{5.15}$$

$$\propto \exp\{E_{\hat{q}(\boldsymbol{\theta}_{-i}|\boldsymbol{y})}[\ln\, p(\boldsymbol{\theta}, \boldsymbol{y})]\}, \tag{5.16}$$

where $E_{\hat{q}(\boldsymbol{\theta}_{-i}|\boldsymbol{y})}[\ln p(\boldsymbol{\theta}, \boldsymbol{y})]$ is the expectation of the log-joint posterior distribution with respect to $\hat{q}(\boldsymbol{\theta}_{-i}|\boldsymbol{y}) = \prod_{j \neq i} \hat{q}(\theta_j|\boldsymbol{y})$, the product of distributions of all other partitions of $\boldsymbol{\theta}$ except θ_i. The optimization is done by updating (5.15) for $i = 1, ..., N$ iteratively, until convergence. The solution is guaranteed to at least converge to a local minima of the KL function. Similarly as the Gibbs sampler, in VB, it is beneficial to use conjugate priors to make sure that $\hat{q}(\theta_i|\boldsymbol{y})$ are recognized as known parametric distributions. In that case, the computation of the required moments/expectations becomes straightforward. Where such conjugate priors are not available, numerical integrations are needed, and this increases the computation demand. Alternatively, one may also consider the fixed form variational approximation to the non-conjugate part (Salimans and Knowles 2013; see also Article III). In Article II, we derived VB algorithms for several regularized regression models such as Bayesian LASSO.

In addition, the VB also provides an estimate to the marginal likelihood $p(\boldsymbol{y})$. We have

$$\begin{aligned}
\ln p(\boldsymbol{y}) &= \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}|\boldsymbol{y}) \ln \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta} - \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}|\boldsymbol{y}) \ln \frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{q(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}|\boldsymbol{y}) \ln \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta} + \mathrm{KL}(q(\boldsymbol{\theta}|\boldsymbol{y})||p(\boldsymbol{\theta}|\boldsymbol{y})) \\
&\geq \int_{\boldsymbol{\Theta}} q(\boldsymbol{\theta}|\boldsymbol{y}) \ln \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q(\boldsymbol{\theta}|\boldsymbol{y})} d\boldsymbol{\theta} \\
&\equiv L(q(\boldsymbol{\theta}|\boldsymbol{y})). \qquad (5.17)
\end{aligned}$$

We call $L(q(\boldsymbol{\theta}|\boldsymbol{y}))$ a lower bound of the logarithm of the marginal likelihood $\ln p(\boldsymbol{y})$. Since $\ln p(\boldsymbol{y})$ is a constant, the minimization of the KL function and the maximization of the lower bound happen at the same time. When $\hat{q}(\boldsymbol{\theta})$ approximates $p(\boldsymbol{\theta})$ well, we can also expect the lower bound $L(\hat{q}(\boldsymbol{\theta}|\boldsymbol{y}))$ to be a good approximation to the $\ln p(\boldsymbol{y})$. Besides, it is easy to calculate $L(\hat{q}(\boldsymbol{\theta}|\boldsymbol{y}))$ after obtaining $\hat{q}(\boldsymbol{\theta}|\boldsymbol{y})$ based on the formula:

$$\begin{aligned}
L(\hat{q}(\boldsymbol{\theta}|\boldsymbol{y})) &= \int_{\boldsymbol{\Theta}} \prod_{i=1}^{N} \hat{q}(\theta_i|\boldsymbol{y}) \ln p(\boldsymbol{\theta}, \boldsymbol{y}) d\boldsymbol{\theta} - \int_{\boldsymbol{\Theta}} \prod_{i=1}^{N} \hat{q}(\theta_i|\boldsymbol{y}) \sum_{i=1}^{N} \ln \hat{q}(\theta_i|\boldsymbol{y}) d\boldsymbol{\theta} \\
&= E_{\prod_{i=1}^{N} \hat{q}(\theta_i|\boldsymbol{y})}[\ln p(\boldsymbol{\theta}, \boldsymbol{y})] - \sum_{i=1}^{N} E_{\hat{q}(\theta_i|\boldsymbol{y})}[\hat{q}(\theta_i|\boldsymbol{y})]. \qquad (5.18)
\end{aligned}$$

Therefore, it is possible to use the variational lower bound as a model selection criterion just like BIC (Beal and Ghahramani 2003). Based on these ideas, Nott et al. (2012) developed a Bayesian variable selection method for the multiple linear regression model, which is connected to the stepwise regression. We further generalized their idea for performing variable selection in the varying coefficient model (see Article III).

A fast computation speed is the major advantage of the VB method, and for many of the regression models discussed above, it can usually converge within several hundreds of steps; by

contrast, MCMC methods often need simulations that proceed for more than 10000 steps. A drawback of the VB method, however, is that, it often underestimates the uncertainties, i.e. provides too narrow estimates of the standard errors. This might be problematic, if we want to construct some test statistics or credible intervals based on VB estimates.

5.5 Bayesian multiple hypothesis testing

The Bayesian regularized regression approaches directly provide both point estimates and their uncertainties, so that the construction of test statistics are convenient. However, it has been argued that the Bayesian approaches usually cannot automatically take the multiplicity adjustment into account (Berry and Hochberg 1999; Scott and Berger 2010) when dealing with high dimensional data, just like classic frequentist methods.

In this work, we consider two possible methods for multiplicity adjustment. The first method is permutation (Churchill and Doerge 1994), which constructs the test statistic by repeatedly reshuffling the response data. Since permutation usually requires much re-sampling of the data (i.e. for thousand times), it would be preferable to use this method together with the fast VB estimation procedure. More information is provided in Article II. Second, for the MCMC spike and slab regression, it is possible to apply a Bayesian false discovery rate (BFDR) based method by taking the advantage of the estimated posterior inclusion probabilities. A convenient thing is that the BFDR method relies on one MCMC chain, without any need for re-sampling. In our example analyses (see Article IV), its performance seems to be competitive to some LASSO related testing methods.

While these methods for hypothesis testing as well as some other possible approaches have been verified in some empirical studies, the theoretical results are still lacking in general. Thus, we may conclude that the Bayesian multiple hypothesis testing is an open research problem, and deserves more investigation.

6 Applications in quantitative genetics

We have introduced some background for understanding the statistical methodology side of the articles. In this section, we describe some problems in quantitative genetics that we want to solve by applying those regression models. Our study targets are quantitative traits containing continuous phenotype measurements in plants or animals, such as kernel weight in barley (Tinker et al. 1996), density and fiberwall thickness in wood of Scots pine (Article IV) and active probability of mouse behavior (Xiong et al. 2011).

6.1 QTL/association mapping

We want to identify a quantitative trait locus (QTL), a region of the genome that is associated with a quantitative trait. In practice, the measurable genotype data is a panel of genetic markers, such as amplified fragment length polymorphisms (AFLPs) or single nucleotide polymorphisms (SNPs) (Vignal et al. 2002). We typically consider biallelic markers, which have three possible genotypes AA, AB and BB. In some specific circumstances such as a back cross design, there might be only two genotypes AA and AB. Typically, a QTL is not exactly located at any marker, but when the marker density is high, the QTL should be closely correlated (or in linkage disequilibrium) with some markers. Thus, we may simply use marker positions as a proxy for QTLs (Xu 2003). We focus on QTL mapping problems with hundreds or thousands of genetic markers. A typical data set could be represented by (y_i,x_{ij}) $i = 1, ..., n$ and $j = 1, ..., p$, where y_i is phenotype measurement of individual i, and x_{ij} is the genotype data of the individual i and marker j, coded as 1 for AA, 0 for AB and -1 for BB. Intuitively, the multiple regression model (1.3) can be used to build the relation between phenotypes and genotypes. The regression coefficients, representing the effects of the markers on the trait, can be estimated by some regularization methods. The hypothesis testing can then be used to judge QTLs (Please read Articles I and II for more details).

Many traits such as wood density change during their developmental process of life. For dynamic traits where there have been repeated measurements over multiple time points, the above introduced three regression methods for the longitudinal data are applicable (see Articles III and IV).

6.2 Genomic selection

A side perspective of this thesis is genomic selection, which is a prediction problem (Meuwissen et al. 2001; De Los Campos et al. 2009). We have a training data set consisting of the individuals from old generations with phenotypes and genotypes measured, and a validation data set consisting of the younger individuals which only have genotypes. The training data is used to estimate the effects of the markers, and then the validation data is used for estimating the breeding values (see Article I for details).

7 Conclusions

We have proposed a series of Bayesian multiple regularization or variable selection methods for QTL analysis of complex traits. Our focus is on developing some variable selection methods for longitudinal data, which have not yet been widely applied in statistical genetics.

7.1 MCMC vs. Variational Bayes

We studied two algorithms for the Bayesian computation. The MCMC sampling algorithm provides accurate approximation to the full posterior distribution, but is relatively slow and typically, more explanatory variables (markers) require more MCMC steps to achieve convergence. On the other hand, the deterministic VB method converges much faster than MCMC, but often provides downward biased estimates to the uncertainties, which becomes problematic if we want to construct some test statistics or intervals. An alternative choice for approximate computation is expectation propagation (EP), but it has been argued that EP may provide upward biased estimates of uncertainties (Rue et al. 2009). Nevertheless, the VB method can serve as a good tool for fast exploration of the data. We suggest that the MCMC method is the preferable method for small data but that the VB method is a good, fast alternative for some large data sets when it is sufficient to obtain point estimates. It is also possible to use the VB estimates as the initial state of the MCMC sampling to ensure rapid convergence of the Markov chain.

7.2 Frequentist methods vs. Bayesian methods

We have considered both frequentist and Bayesian regularization methods here, and in this context, these methods do not appear particularly different. For example, the frequentist ridge regression, LASSO and mixed effects model all have Bayesian interpretations, and thus we may treat them as Bayesian models and use Bayesian computational tools to obtain the solutions. One notable difference between approaches is that by using Bayesian computational methods such as MCMC and VB, we approximate the full posterior distribution and often consider the posterior mean as the point estimate, while the solution of ML coincides with the (posterior) mode. This may partially explain why the results from some frequentist methods differ from the results of their corresponding Bayesian counterparts (see Article I).

7.3 Multiple loci methods vs. single locus methods

In some genetics literature, this type of regularized regression methods refers to methods that analyse multiple loci. In contrast, a single locus method refers to approaches using a marginal regression (defined in equation (1.1)) to analyze one marker at a time. The multiple loci methods are believed to be superior to the single locus methods, because they simultaneously estimate additive effects of multiple loci on one trait, which may better mimic the underlying biology. However, the single locus methods also have some advantages. First, quite mature multiple hypothesis testing methodologies have been developed from both frequentist (Dudoit and Van Der Laan 2008) and (empirical) Bayesian perspective (Efron 2010). Although some multiple

testing procedures have also been developed with the regularization methods, it seems that none of them has been accepted as a standard approach. Second, single locus methods benefits from being easily and quickly implemented. For example, contemporary genome-wide association mapping studies (GWAS) may comprise an excess of one million SNPs and are thus difficult to analyze using multiple loci methods due to the multimodality problem and the high computational demands. See, for example, Peltola et al. (2012) for an attempt to apply one multiple loci approach on huge GWAS data sets. By contrast, single locus methods are still applicable, and, indeed, become a standard choice for such purposes. It is also possible to combine the concepts of both methods by first using a single locus method to preselect a subset of several thousand loci with the smallest p-values, and then performing a multiple locus method on the subset of loci in order to obtain more accurate estimates; see Fan and Lv (2008) for an example of sure independence screening.

7.4 Comparison of three modeling strategies for longitudinal data

The statistical modeling of longitudinal QTL data is an important aspect of this thesis. As shown in earlier sections, we consider three closely connected longitudinal models: a linear mixed effects model, a (non-parametric) varying coefficient model, and a multilevel model. Overall, longitudinal models by jointly analyzing phenotype repeated measurements at multiple time points show greater statistical powers for QTL detection than those single trait methods by analyzing a single time point at a time. The LMM should be most general approach, and should be applicable in many different situations. The varying coefficient model is an time-gene interaction model, with a lot of time related parameters need to be estimated, so that it might be more suitable for the data with a sufficient number of time points. The multilevel model is a computationally easy approach of LMM, and can be more efficiently implemented in some big data sets. However, by separately estimating the effects of temporal trends and genetic markers, it might not provide as accurate estimates of uncertainties as the LMM approach (Sikorska et al. 2013).

7.5 Future work

There is still plenty of room in this thesis for improvement. For example, from the methodology point of view, it would be valuable to have more investigation into the multiple hypothesis testing issue. From the application point of view, it is possible to apply the current methods to other problems in genetics such as time series gene expression data.

Acknowledgments

I am grateful to Daniel Blande and Phillip Watts for giving constructive comments on the introduction part of the thesis. This work was supported by the research grants from the Finnish Population Genetics Doctoral Programme, the Doctoral Programme in Mathematics and Statistics in University of Helsinki, the Academy of Finland and the University of Helsinki's Research funds.

References

Akaike, H., 1974 A new look at the statistical model identification. IEEE T. Automat. Contr. 19: 716–723

Bachrach L. K., T. Hastie, M-C. Wang, B. Narasimhan, and R. Marcus, 1999 Bone mineral acquisition in healthy Asian, hispanic, black and caucasian youth. A longitudinal study. J. Clin. Endocrinol. Metab. 84: 4702–4712

Berry, D. A., and Y. Hochberg, 1999 Bayesian perspectives on multiple comparisons. J. Stat. Plan. Infer. 82: 215–227

Bishop, C. M., 2006 *Pattern Recognition and Machine Learning.* New York: Springer

Beal, M. J., and Z. Ghahramani, 2003 The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. Bayesian Stat. 7: 453–464

Beal, M. J., 2003 Variational algorithms for approximate Bayesian inference [PhD thesis]. University of London

Bühlmann, P., and S. Van De Geer, 2011 *Statistics for High-Dimensional Data: Methods, Theory and Applications.* New York: Springer

Carbonetto, P., and M. Stephens, 2012 Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. 7: 73–108

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics 138: 963–971

Daye, Z. J., J. Xie, and H. Li, 2012 A sparse structured shrinkage estimator for nonparametric varying-coefficient model with an application in genomics. J. Comput. Graph. Stat. 21: 110–133

De Boor, C. M., 2001 *A Practical Guide to Splines*. New York: Springer

De Los Campos, G.,H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182: 375–385

Diggle, P., P. Heagerty, K-Y. Liang, and S. Zeger, 2002 *Analysis of Longitudinal Data*. Oxford: Oxford University Press

Dudoit, S., and M. J. Van Der Laan, 2008 *Multiple Testing Procedures with Application to Genomics*. New York: Springer

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani, 2004 Least angle regression. Ann. Stat. 32: 407–451

Efron, B., 2010 *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. New York: Cambridge University Press

Eilers, P. H. C., and B. D. Marx, 1996 Flexible smoothing using B-splines and penalized likelihood. Stat. Sci. 11: 89–121

Fahrmeir, L., and T. Kneib, 2011 *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. New York: Oxford University Press

Fan, J., and J. Lv, 2008 Sure independence screening for ultrahigh dimensional feature space (with discussion). J. Roy. Stat. Soc. B. 70: 849–911

Figueiredo, M. A. T., 2003 Adaptive sparseness for supervised learning. IEEE Trans. Pattern. Anal. Mach. Intell. 25: 1150–1159

Friedman, J., T. Hastie, H. Höfling, and R. Tibshirani, 2007 Pathwise coordinate optimization. Ann. Appl. Stat. 1: 302–332

Friedman, J., T. Hastie, and R. Tibshirani, 2010 Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33: 1

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2004 *Bayesian Data Analysis (Second edition)*. London: Chapman and Hall

Geman, S., and D. Geman, 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE. Trans. Pattern. Anal. Mach. Intell. 6: 721–741

Goodman, S. N., 1999 Toward evidence-based medical statistics. 1: the P value fallacy. Ann. Intern. Med. 130: 995–1004

Hastie, T., R. Tibshirani, and J. H. Friedman, 2009 *The Elements of Statistical Learning (Second edition).* New York: Springer

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57: 97–109

Heuven, H. C. M., and L. L. G. Janss, 2010 Bayesian multi-QTL mapping for growth curve parameters. BMC Proc. 4: S12

Hoerl, A. E., and R. W. Kennard, 1970 Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12: 55–67

Hsiang, T. C., 1975 A Bayesian view on ridge regression. The Statistician 24: 267–268

Izenman, A. J., 2008 *Modern Multivariate Statistical Techniques.* New York: Springer

Jaakkola, T. S., and M. I. Jordan, 2000 Bayesian parameter estimation via variational methods. Stat. Comput. 10: 25–37

Kang, M. H., N. A. Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin, 2007 Efficient control of population structure in model organism association mapping. Genetics 178: 1709–1723

Kass, R. E., and A. E. Raftery, 1995 Bayes factors. J. Am. Stat. Assoc. 90: 773–795

Kullback, S., and R. A. Leibler, 1951 On information and sufficiency. Ann. Math. Stat. 22: 79–86

Kuo, L., and B. Mallick, 1998 Variable selection for regression models. Sankhya. B 60: 65–81

Kutner, M. H., C. J. Nachtsheim, and J. Neter, 2004 *Applied Linear Regression Models.* New York: McGraw-Hill

Kyung, M., J. Gill, M. Ghosh, and G. Casella, 2010 Penalized regression, standard errors, and Bayesian Lassos. Bayesian Anal. 2: 369–412

Liu, T., and R. Wu, 2009 A Bayesian algorithm for functional mapping of dynamic complex traits. Algorithms 2: 667–691

Ma, C., G. Casella, and R. Wu, 2002 Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics 161: 1751–1762

McCullagh, P., and J. Nelder, 1989 *Generalized Linear Models (Second edition).* Boca Raton: Chapman and Hall/CRC

Meinshausen, N., L. Meier, and P. Bühlmann, 2009 P-Values for high-dimensional regression. J. Am. Stat. Assoc. 104: 1671–1681

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953 Equations of state calculations by fast computing machines. J. Chem. Phys. 21: 1087–1092

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829

Minka, T. P., 2001 Expectation propagation for approximate Bayesian inference. Uncertnty. Artif. Intell. 17: 362–369

Minnier, J., L. Tian, and T. Cai, 2011 A perturbation method for inference on regularized regression estimates. J. Am. Stat. Assoc. 106: 1371–1382

Müller, S., J. L. Scealy, and A. H. Welsh, 2013 Model selection in linear mixed models. Stat. Sci. 28: 135–281

Nott, D. J., M. N. Tran, and C. Leng, 2012 Variational approximation for heteroscedastic linear models and matching pursuit algorithms. Stat. Comput. 22: 497-512

O'Hara R. B., and M. J. Sillanpää, 2009 A Review of Bayesian variable selection methods: what, how, and which? Bayesian Anal. 4: 85–118

Ormerod, J. T., and M. P. Wand, 2010 Explaining variational approximations. J. Am. Stat. Assoc. 64: 140–153

Park, T., and G. Casella, 2008 The Bayessian LASSO. J. Am. Stat. Assoc. 103: 681–686.

Patterson, H. D., and R. Thompson, 1971 Recovery of inter-block information with block sizes are unequal. Biometrika 58: 545–554.

Peltola, T., P. Marttinen, and A. Vehtari, 2012 Finite adaptation and multistep moves in the Metropolis-Hastings algorithm for variable selection in genome-wide association mapping studies. PLOS one 7: e49445

Picard, R. R., and R. D. Cook, 1984 Cross-validation of regression models. J. Am. Stat. Assoc. 79: 575–583

Robert, C. P., and G. Casella, 2004 *Monte Carlo Statistical Methods (Second Edition).* New York: Springer

Rue, H., S. Martino, and N. Chopin 2009 Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximation. J. R. Stat. Soc. B. 71: 319–392

Ruppert, D., M. P. Wand, and R. J. Carroll, 2003 *Semiparametric Regression*. New York: Cambridge University Press

Salimans, T, and D. A. Knowles 2013 Fixed-form variational posterior approximation through stochastic linear regression. Bayesian Anal. 8: 741–908

Schelldorfer, J., P. Bülhlmann, and S. Van De Geer 2011 Estimation for high-dimensional linear mixed-effects models using $\boldsymbol{l}_1$-penalization. Scand. J. Stat. 38: 197–214

Schwarz, G. E., 1978 Estimating the dimension of a model. Ann. Stat. 6: 461–464

Scott, J. G., and J. O. Berger, 2010 Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann. Stat. 38: 2587–2619.

Sikorska, K., F. Rivadeneira, P. J. F. Groenen, A. Hofman, A. G. Uitterlinden, P. H. C. Eilers, and E. Lesaffre, 2013 Fast linear mixed model computations for genome-wide association studies with longitudinal data. Stat. Med. 32: 165–180.

Sillanpää, M. J., P. Pikkuhookana, S. Abrahamsson, T. Knürr, A. Fries, E. Lerceteau, P. Waldmann, and M. R. Garcia-Gil, 2012 Simultaneous estimation of multiple quantitative trait loci and growth curve parameters through hierarchical Bayesian modeling. Heredity 108: 134–146

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B. 58: 267–288

Tinker, N. A., D. E. Mather, B. G. Rossnagel, K. J. Kasha, and A. Kleinhofs et al., 1996 Regions of the genome that affect agronomic performance in two-row barley. Crop. Sci. 36: 1053–1062

Vignal, A., D. Milan, M. SanCristobal, and A. Eggen, 2002 A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. 34: 275–305

Wasserman L., and K. Roeder, 2009 High dimensional variable selection. Ann. Stat. 37: 2178–2201

West, B. T., K. B. Welch, and A. T. Galecki, 2007 *Linear Mixed Models: A Practical Guide to Using Statistical Software*. New York: Chapman and Hall/CRC

Wu, R., and M. Lin, 2006 Functional mapping-how to map and study the genetic architecture of dynamical complex traits. Nat. Revs. Genet. 7: 229–237

Xiong, H., E. H. Goulding, E. J. Carlson, L. H. Tecott, C. E. McCulloch, and Ś. Sen, 2011 A flexible estimating equations approach for mapping function valued traits. Genetics 189: 305–316

Xu, S., 2003 Estimating polygenic effects using markers of the entire genome. Genetics 163: 789–801

Yi, N., and S. Xu, 2008 Bayesian LASSO for quantitative trait loci mapping. Genetics 179: 1045–1055