Empirical Bayes applications in biomedical high-dimensional prediction

Münch, M.M.

CHAPTER 1

Introduction

1.1 Biomedical high-dimensional prediction

High-dimensional data with tens or hundreds of thousands of variables are frequently part of biomedical (or other) studies nowadays. In this thesis the main focus is on clinical prediction from omics data, i.e., (near-)complete genetic or molecular profiles. A common aim in omics studies is prediction of binary or continuous quantities. Examples of predicted quantitites in clinical settings are disease status, drug efficacy, and therapy response. Often, a secondary goal is feature selection, without the loss of predictive power. Ultimately, the estimated, possibly sparse, predictors guide future research, or aid medical practitioners in their decision-making.

In addition to the primary data, researchers often have so-called complementary data on the features available (hereafter referred to as co-data, Neuenschwander et al., 2010). Co-data refer to any information on the features that does not involve the predicted outcome. Some examples of co-data in the omics field are: (i) p-values on the features in the primary data from a previous study. If the previous and current study outcomes are related, these p-values may contain relevant information on the features; (ii) genetic annotation on the genes involved. For example, for certain phenotypical outcomes, features on some chromosomes are more important than others; (iii) groupings of the primary data features based on expert knowledge, where groups known to be involved in the outcome may be more important. Traditionally, co-data are not included in statistical analyses, although relevant co-data have the potential to enhance model performance by guiding estimation and feature selection. However, care must be taken, as to not let co-data bias the estimation too much, or in the wrong direction. The methods proposed in this thesis, take the co-data into account through modelling of the prior and differential penalization of the model parameters. To avoid bias, the influence of co-data on estimation is

determined in a data-driven manner.

This thesis proposes three methods to construct predictors for continuous and binary outcomes based on a high-dimensional set of features. The three proposed methods are in agreement on estimation principles: all three are based on Bayesian models. Furthermore, they all rely on variational Bayes approximations, and, to some extent, empirical Bayes estimation of hyperparameters. They differ in their observational and prior model specifications.

1.2 Prior knowledge

An advantage of Bayesian modelling is that it allows for the inclusion of prior knowledge into the model. In Bayesian modelling, the prior knowledge is quantified as a prior distribution $\pi_\alpha(\theta)$ on the parameters θ, for some hyperparameter α. Any potentially beneficial co-data may then be modelled through the prior. The prior is multiplied with the data likelihood $p(y|\theta)$ and normalised to obtain the posterior distribution of interest:

$$p_\alpha(\theta|y) = \frac{p(y|\theta)\pi_\alpha(\theta)}{\int_{\boldsymbol{\Theta}} p(y|\theta)p_\alpha(\theta)d\theta} = \frac{p(y|\theta)p_\alpha(\theta)}{m_\alpha(y)} = \frac{\mathcal{L}_\alpha(\theta, y)}{m_\alpha(y)}. \tag{1.2.1}$$

Uncertainty of the prior knowledge is expressed as uncertainty in the prior distribution. A (nearly) flat prior then expresses a complete lack of prior knowledge. While attractive in theory, in high-dimensional prediction settings, a (nearly) flat prior leads to high (or even infinite) variance of the predictor. To counteract high variance, the prior distribution is, to some extent, concentrated around some prior expected value of the parameter. This deviation from flatness introduces bias. Generally the prior can be chosen such that the decrease in variance outweighs the increase in bias.

The choice of prior is important: the prior should balance bias and variance, such that the predictor is optimal. In practice, the choice of prior distribution family is based on convenience, or previous experience. The hyperparameter values that specify the distribution are either chosen on the basis of prior knowledge, or estimated. Choosing the hyperparameters by hand requires intricate prior knowledge of the modelled subject. Furthermore, a quantification from prior knowledge to hyperparmeter is necessary. If prior knowledge is both available and quantifiable, this approach is often preferable. If such prior knowledge is not available or difficult to quantify, estimation is a reasonable alternative. This is especially true in predictive modelling.

In this thesis, the hyperparameter is modelled as a function of the co-data c, i.e., $\alpha = f(c)$. In the simplest case, the co-data is categorical, so the features come in $g = 1, \ldots, G$ groups and c is a group index. Then, the functions cosidered are of the form $f(c) = \sum_{g=1}^{G} \phi_g \mathbb{1}_{\{c=g\}}$ and require to set or estimate the ϕ_g. In the case of continuous co-data, more elaborate and setting-specific modelling of the relation $\alpha = f(c)$ is required. Chapter 3 contains an example of such a continuous co-data model.

In a predictive setting, (frequentist) estimation of hyperparameters is often done through cross-validation. In (K-fold) cross-validation, the available data is divided into K subsets, or folds. For each fold k, the model is estimated on the remaining $K - 1$ folds, excluding fold k, for a grid of hyperparameters. Then, an empirical measure of predictive performance, or loss, $\ell_k(\alpha)$ is calculated on fold k for each value of α in the grid. Finally, the losses are averaged over the folds, and minimised

$$\hat{\alpha} = \underset{\alpha \in \text{grid}}{\operatorname{argmin}} \, K^{-1} \sum_{k=1}^{K} \ell_k(\alpha),$$

to obtain an estimate of the hyperparameter. This estimate balances bias and variance of the estimator to empirically maximise predictive performance.

1.3 Empirical Bayes

Although a versatile solution, cross validation requires l^K model evaluations, with l the size of the hyperparameter grid. In high dimensions, model evaluation is computationally expensive. Furthermore, the number of model evaluations increases exponentially with the dimension of α, thereby increasing the computational burden even further. An alternative to cross-validation of hyperparameters is empirical Bayes estimation. In empirical Bayes the prior is fit to the data, generally by means of marginal likelihood maximisation. The marginal likelihood, or model evidence, is the frequentist likelihood for the hyperparameter α. It is the denominator in the right-hand side of (1.2.1) and computed by integration of the product of data likelihood $p(y|\theta)$ and prior $\pi_\alpha(\theta)$ with respect to the random parameters:

$$m_\alpha(y) = \int_{\boldsymbol{\Theta}} p(y|\theta)\pi_\alpha(\theta)d\theta. \tag{1.3.1}$$

Maximisation of the (log) marginal likelihood with respect to the hyperparameter then gives the empirical Bayes estimate:

$$\hat{\alpha} = \underset{\alpha}{\operatorname{argmax}} \, \log m_\alpha(y). \tag{1.3.2}$$

Empirical Bayes is especially appealing in high dimensional models (van de Wiel et al., 2019). Empirical Bayes involves learning the hyperparameter of the prior distribution over the p-dimensional model parameter θ. Generally, the dimension of hyperparameter α does not grow with model parameter dimension p, so large p leads to more efficient estimation of α.

A common criticism of empirical Bayes, as compared to full Bayes, is the lack of error propagation. Once estimated, the hyperparameters are assumed fixed and known. A full Bayesian treatment expresses uncertainty in the hyperparameters through an extra layer of hyperpriors. The extra uncertainty in the hyperparameters then propagates through the model to increase the uncertainty in the model parameters of interest. In principal, the lack of hyperparameter uncertainty results in underestimated uncertainties for the model parameters and predictions. However, Carlin and Louis (2000) and van de Wiel et al. (2019) show that uncertainty quantification through credible intervals for empirical Bayes is competitive to full Bayes in terms of frequentist coverage probabilities.

Fong and Holmes (2020) show that the marginal likelihood is equivalent to a cross-validation score, averaged over all possible fold configurations. This correspondence between empirical Bayes estimation and cross validation implies that the estimated hyperparameter $\hat{\alpha}$ is optimised for prediction and the subsequent model parameters and predictions are competitive to the full Bayesian posterior point estimates in terms of predictive performance.

In contrast to the cross-validation approach, the empirical Bayes approach requires, in principal, just one model fit. Depending on the specific problem this may be much more computationally feasible, even for higher dimensionional α parameters. For most models, the bottleneck in this approach is computation of the integral in (1.3.1). Generally, it is not available in closed form, and the high dimension of θ complicates numerical or Monte Carlo approximations. Casella (2001) proposes and expectation-maximisation (EM) algorithm that computes (1.3.2) by iterating the steps

$$\alpha^{(k+1)} = \underset{\alpha}{\operatorname{argmax}}\, \mathbb{E}[\log \pi_\alpha(\theta)], \tag{1.3.3}$$

until convergence, where the expectation is with respect to the posterior $p_{\alpha^{(k)}}(\theta|y)$.

Computation of the expectation in (1.3.3) requires access to the posterior distribution $p_\alpha(\theta|y)$. In general (as well as in Casella, 2001) the posterior is approximated with Markov chain Monte Carlo (MCMC) samples. In high dimensions, MCMC becomes computationally demanding; even more so if repeated at every iteration (1.3.3). A fast alternative to MCMC approximation is variational Bayes.

1.4 Variational Bayes

This Section provides a concise introduction into variational Bayes. For a more complete review of the topic we refer the reader to Blei et al. (2017). Furthermore, Beal (2003) provides a more in-depth analysis and various examples for a wide range of applications. Variational Bayesian methods approximate the posterior with an alternative $q(\theta)$, that minimises the Kullback-Leibler divergence of the posterior from the approximation, where the Kullback-Leibler divergence is

$$\begin{aligned} D_{\mathrm{KL}}(q \,\|\, p_\alpha) &= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log p_\alpha(\theta|y)] && \text{(1.4.1a)} \\ &= \mathbb{E}_q[\log q(\theta)] - \mathbb{E}_q[\log \mathcal{L}_\alpha(\theta, y)] + \log m_\alpha(y) && \text{(1.4.1b)} \\ &= -\mathrm{ELBO}(q) + \log m_\alpha(y). && \text{(1.4.1c)} \end{aligned}$$

The quantity $\mathrm{ELBO}(q)$ is termed the evidence lower bound, as by the non-negativity of the Kullback-Leibler divergence, we have

$$\log m_\alpha(y) \geq \mathrm{ELBO}(q).$$

Inspection of (1.4.1) learns that minimisation of the Kullback-Leibler divergence with respect to q is equivalent to maximisation of the $\mathrm{ELBO}(q)$.

In this thesis, the mean-field variant of variational Bayes is used. In mean-field variational Bayes, the approximating posterior is assumed to factorise with respect to some partitioning of the parameters $\theta = \{\theta_1, \ldots, \theta_M\}$, i.e.,

$$p_\alpha(\theta|y) \approx q(\theta) = \prod_{m=1}^{M} q_m(\theta_m).$$

This factorisation results in

$$\begin{aligned} \mathrm{ELBO}(q) &= \mathbb{E}_q[\log \mathcal{L}_\alpha(\theta, y)] - \mathbb{E}_q[\log q(\theta)] \\ &= \int_{\Theta_1} \cdots \int_{\Theta_M} \prod_{m=1}^{M} q_m(\theta_m) \log \mathcal{L}_\alpha(\theta, y) d\theta_1 \cdots d\theta_M \\ &\quad - \int_{\Theta_1} \cdots \int_{\Theta_M} \prod_{m=1}^{M} q_m(\theta_m) \sum_{m=1}^{M} \log q_m(\theta_m) d\theta_1 \cdots d\theta_M. \end{aligned}$$

If we focus on θ_l, and denote $\mathbb{E}_{m \neq l}(\cdot)$ the expectation with respect to $\prod_{m \neq l} q_m(\theta_m)$,

we have

$$\begin{aligned}
\text{ELBO}(q) &= \int_{\Theta_1} \cdots \int_{\Theta_M} q_l(\theta_l) \prod_{m \neq l} q_m(\theta_m) \log \mathcal{L}_\alpha(\theta, y) d\theta_1 \cdots d\theta_M \\
&\quad - \int_{\Theta_1} \cdots \int_{\Theta_M} q_l(\theta_l) \prod_{m \neq l} q_m(\theta_m) \sum_{m=1}^{M} \log q_m(\theta_m) d\theta_1 \cdots d\theta_M \\
&= \int_{\Theta_l} q_l(\theta_l) \mathbb{E}_{m \neq l}[\log \mathcal{L}_\alpha(\theta, y)] d\theta_l - \int_{\Theta_l} q_l(\theta_l) \log q_l(\theta_l) d\theta_l \\
&= -D_{\text{KL}}(q \,\|\, \exp\{\mathbb{E}_{m \neq l}[\log \mathcal{L}_\alpha(\theta, y)]\}) + c,
\end{aligned}$$

where all terms not involving $q_l(\theta_l)$ have been combined in constant c. By the non-negativitiy of the Kullback-Leibler divergence, the ELBO(q) is maximised with respect to $q_l(\theta_l)$ at

$$q_l(\theta_l) \propto \exp\{\mathbb{E}_{m \neq l}[\log \mathcal{L}_\alpha(\theta, y)] \propto \exp\{\mathbb{E}_{m \neq l}[\log \tilde{p}_\alpha(\theta_l | \theta_{-l}, y)], \tag{1.4.2}$$

where θ_{-l} denotes all parameters, excluding θ_l.

If the full conditionals $\tilde{p}_\alpha(\theta_l | \theta_{-l}, y)$ are exponential family distributions, the computations simplify significantly. Exponential family full conditional densities may be written as:

$$\tilde{p}_\alpha(\theta_l | \theta_{-l}, y) \propto h(\theta_l) \exp\left[\eta_l(\theta_{-l}, y)^{\mathrm{T}} \theta_l\right], \tag{1.4.3}$$

with natural parameter $\eta_l(\theta_{-l}, y)$ and base function $h(\theta_l)$. Inserting (1.4.3) into (1.4.2) leaves us with

$$q_l(\theta_l) \propto h(\theta_l) \exp\left\{\mathbb{E}_{m \neq l}[\eta_l(\theta_{-l}, y)]^{\mathrm{T}} \theta_l\right\}. \tag{1.4.4}$$

In other words, the mean-field variational posterior is in the same exponential family as the full conditional distribution with natural parameter $\mathbb{E}_{m \neq l}[\eta_l(\theta_{-l}, y)]$. Many common models are full conditional exponential familiy models. In fact, the models introduced in Chapters 3 and 4 are full conditional exponential family models.

For exponential family full conditional models, the variational update (1.4.4) shows a connection between variational inference and Gibbs sampling, where samples are drawn iteratively from (1.4.3). It also highlights the difference with Gibbs sampling: the variational update (1.4.4) collapses all previous iterations into one expectated natural parameter $\mathbb{E}_{m \neq l}[\eta_l(\theta_{-l}, y)]$. As a result the variational posterior generally concentrates around the posterior expectation of the true model (and the converged Gibbs samples) and underestimates posterior variances.

The combination of empirical and variational Bayesian techniques is convenient. Many priors (and all priors considered in this thesis) are independent over the

parameters of interest, i.e., $\pi_\alpha(\theta) = \prod_{j=1}^{p} \pi_\alpha(\theta_j)$. Furthermore, many priors may be written as hierarchical mixtures of the form $\pi_\alpha(\phi|\psi)\pi(\psi)$, with $\theta = \{\phi, \psi\}$. The combination of these postulates leads to:

$$\mathbb{E}_q[\log \pi_\alpha(\theta)] = \sum_{j=1}^{p} \mathbb{E}_q[\log \pi_\alpha(\phi_j|\psi_j)], \tag{1.4.5}$$

where terms not involving α have been omitted. The expectation in the right-hand side of (1.4.5) is often easy to compute under the mean-field variational Bayes posterior. Consider, for example, the class of scale mixtures of Gaussians, with $\phi_j|\psi_j \sim \mathcal{N}(0, \alpha\psi_j)$. Then,

$$\mathbb{E}_q[\log \pi_\alpha(\theta)] = -\frac{p}{2}\log\alpha - \frac{\alpha^{-1}}{2}\sum_{j=1}^{p}\mathbb{E}_q(\psi_j^{-1}\phi_j^2) + c,$$

again combining terms not involving α in the constant c. Now hyperparameter update (1.3.3) becomes

$$\alpha^{(k+1)} = p^{-1}\sum_{j=1}^{p}\mathbb{E}(\psi_j^{-1}\phi_j^2), \tag{1.4.6}$$

where the expectation is with respect to $q(\theta)$. For many variational posteriors of the form $q(\theta) = q_\phi(\phi)q_\psi(\psi)$, the expectation in the right-hand side of (1.4.6) is straightforward to calculate.

1.5 Observational models & outline

In this thesis three different observational models are considered: logistic regression, multivariate linear regression, and factor regression.

Chapter 2 investigates logistic regression. Logistic regression models binary, or sums of m independent binary Bernoulli trials. The model then relates the outcomes to a p-dimensional fixed, covariate vector $\mathbf{x}$ through the logistic mean function:

$$y \sim \mathcal{B}\left(m, \text{expit}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})\right),$$

where $\mathcal{B}(m, \pi)$ denotes the Binomial distribution with number of trials m and probability π, and $\text{expit}(x) = \exp(x)/[1 + \exp(x)]$ is the logistic function. The model parameter $\boldsymbol{\beta}$ determines the strength and direction of the relation between $\mathbf{x}$ and y. Binary outcomes are common in biomedical high-dimensional prediction problems. They appear in, for example, diagnostics tests, prognostic modelling,

and therapy resonse studies. In this chapter, the co-data is categorical and included in the model through the prior variance of an elastic net prior. Simulations and applications to several cancer studies and one Alzheimer's study show that the inclusion of co-data does indeed benefit classification.

Chapter 3 considers multivariate continuous outcomes. The D outcomes are related to the fixed covariate vector $\mathbf{x}$ as independent Gaussians with identity link:

$$y_d \sim \mathcal{N}\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_d, \sigma_d^2\right),\ d = 1, \ldots, D.$$

Parameters $\boldsymbol{\beta}_d$ and σ_d determine the strength and direction of the relation between y_d and $\mathbf{x}$. Multivariate continuous outcomes occur in biomedical high-dimensional prediction in the form of, for example, expression quantitative trait loci (eQTL) studies, where gene expressions are explained with single-nucleotide polymorphisms (SNPs). Another application, presented in Chapter 3, is drug sensitivity screening, where molecular profiles are screened for sensitivity to multiple drugs simultaneously. Such screening programmes guide future drug research and are a first step in the direction of personalised medicine. The co-data that are included through a normal inverse Gaussian prior model are both categorical and continous. The multivariate drug response prediction application presented in the chapter benefits from the inclusion of these co-data.

The last observational models covered in Chapter 4, are two types of factor regressions. In linear factor regresssion, the continuous outcome y and p-dimensional feature vector $\mathbf{x}$ are both random and related to latent factors $\boldsymbol{\lambda}$:

$$\begin{aligned} y|\boldsymbol{\lambda} &\sim \mathcal{N}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\lambda}, \sigma^2), \\ \mathbf{x}|\boldsymbol{\lambda} &\sim \mathcal{N}_p(\mathbf{B}^{\mathrm{T}}\boldsymbol{\lambda}, \boldsymbol{\Psi}), \\ \boldsymbol{\lambda} &\sim \mathcal{N}_d(\mathbf{0}_{d\times 1}, \mathbf{I}_d), \end{aligned} \tag{1.5.1}$$

where $\boldsymbol{\Psi} = \mathrm{diag}(\psi_j)$, $j = 1, \ldots, p$. Factor loadings $\boldsymbol{\beta}$, $\mathbf{B}$, and variances σ^2 and ψ_j, $j = 1, \ldots, p$ determine the strength and direction of the marginal relationship between y and $\mathbf{x}$. The logistic factor regression extension for binary, or sums of m disjoint binary Bernoulli trials outcomes, exchanges (1.5.1) with:

$$y|\boldsymbol{\lambda} \sim \mathcal{B}\left(m, \mathrm{expit}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\lambda})\right).$$

These factor regression models are appropriate if the features and outcomes admit a lower dimensional latent representation. Examples are genes that are organised in functional networks, and a phenotype outcome that is driven by these functional networks. The Gaussian prior model includes categorical co-data through a prior

variance model. Two applications are considered: an influenze efficacy study and an oral cancer classification problem. In both applications, the inclusion of co-data improves performance of the model.

1.6 Contributions

This thesis makes several contributions: (i) it expands on the promising direction of data-driven inclusion of prior knowledge on the features in high-dimensional prediction problems, (ii) it illustrates several new and relevant application areas for variational Bayesian techniques, especially in combination with empirical Bayesian techniques, and (iii) it implements fast and easy-to-use software of the proposed methods in the form of `R` packages and an `R` shiny web app.

Contributions (i) and (ii) are more technical and promote the future development of methodology in the directions of this thesis, while contribution (iii) has a direct influence on applied research. The developed software allows applied researchers and consulting statisticians to apply the proposed methods in their research, thereby directly influencing biomedical research, or even clinical practice.