Convex Duality in Nonparametric Empirical Bayes Estimation and Prediction

by

Sile Tao

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences

University of Alberta

Abstract

The primary goal of this thesis is to implement the Kiefer-Wolfowitz nonparametric empirical Bayes method for models with multivariate response, using the idea of the dual algorithm outlined in a paragraph from Koenker and Mizera (2014). The approach of Kiefer-Wolfowitz was numerically elaborated by Koenker and Mizera (2014) and applied to the univariate normal means problem. For the problems with multivariate response, their method may be not numerically feasible. If the dual problem is considered instead, we are able to come up with an adaptive algorithm, which iteratively uses unequally spaced grids to approximate the prior. In this way, we can solve the dual problem without using overly many grid points. Another objective of the thesis is to facilitate the multivariate data-analytic application of the developed algorithm. To this end, we study Tweedie's formula, which can be used to compute the posterior mean, after the estimate of the prior is obtained. Finally, the formulation of the Koenker–Mizera dual has been justified in the discretized setting as the Lagrange dual of the original (discretized) formulation.

Acknowledgements

I would like to express my deepest and sincere gratitude to my thesis supervisor, Dr. Ivan Mizera for his constant encouragement and advice throughout my study. Without his warm support and immense knowledge, this thesis could not be completed.

My sincere thank also goes to Dr. Matus Maciak who explained ice hockey to me and found the data. Before that, I barely knew this game. I would also like to thank Dr. Linglong Kong, Xiaozhou Wang, Wenrui Ye and Qian Shi who gave me insightful comments and expressed interest on this study.

A special gratitude goes to my family, and especially my grandmother, Tianmei Huang and my parents, Zhenmin Tao and Yang Xia, for their continuous support throughout my life. This thesis would be impossible without them. I am also taking this opportunity to thank all my friends. Without them, life will become very dull.

Last, but not least, I would like to thank Dr. Tahir Choulli, Dr. Edit Gombay and Dr. Narasimha Prasad for being my thesis committee members.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The primary goal of this thesis is to implement the Kiefer-Wolfowitz nonparametric empirical Bayes method for models with multivariate response, using the idea of the dual algorithm outlined in a paragraph from Koenker and Mizera (2014):

> Note that although the primal formulation is infinite-dimensional in the objective (in F), the objective of the dual formulation is finite dimensional (in ν), and infinite-dimensionality appears only in the constraint. This offers a potential for certain refinements: instead of a uniformly spaced grid supporting an atomic measure meant to approximate F, we could instead work with an adaptive (and not necessarily uniformly spaced) collection of test points where the dual constraint is enforced. In fact, if we knew the locations of maxima for the function appearing in the constraint, we could simply select these test points at these locations. Such information is typically unavailable, but practical implementations may seek to refine the solution in an iterative manner by refining

the grid in regions identified by preliminary estimation.

The approach of Kiefer-Wolfowitz was numerically elaborated by Koenker and Mizera (2014) and applied to the univariate normal means problem. For the problems with multivariate response, whose applications are discussed below, their method may be not numerically feasible. The primal problem of maximizing the likelihood function can be reformulated as a convex problem and then efficiently solved by interior point methods. However, unlike in the univariate case, the number of grid points used for multivariate problem can be tremendously large and make the computational complexity of estimating the prior prohibitive. For example, if the dimension of the parameter space is three, the number of grid points needed will be then $1000 \times 1000 \times 1000$, which is about to make the problem intractable. If the dual problem is considered instead, we are able to come up with an adaptive algorithm, in Section 5.1, which iteratively uses unequally spaced grids to approximate the prior. In this way, we can solve the dual problem without using overly many grid points.

Another objective of the thesis is to facilitate the multivariate data-analytic application of the developed algorithm. To this end, we study Tweedie's formula in Section 2.3, which can be used to compute the posterior mean, after the estimate of the prior is obtained. Finally, as a side product, the formulation of the Koenker-Mizera dual has been justified in the discretized setting as the Lagrange dual of the original (discretized) formulation. This is the contents of Chapter 3.

The thesis is organized as follows. In Chapter 2, we review classical Bayes and empirical Bayes paradigms; we discuss both parametric and nonparametric empirical Bayes. We note that the setting of Kiefer and Wolfowitz, as an

example of nonparametric empirical Bayes paradigm, is a convex problem; then, duality results are applicable. We study the discretized and continuous dual problem in separate chapters, Chapter 3 and Chapter 4. Looking at the discretized dual problem, we find there is a potential to provide an algorithm which computes Kiefer-Wolfowitz's MLE in higher dimensional space. The adaptive algorithm and numerical experiments are discussed in Chapter 5.

Chapter 2

Empirical Bayes paradigm

This chapter introduces the main ideas in the empirical Bayes approach to statistical inference. The fundamental empirical Bayes ideas are provided in Section 2.1. Then we briefly review classical Bayes paradigm (Section 2.2). In Section 2.3, Tweedie's formula is introduced as a prediction method that assumes the existence of the prior but does not explicitly use the information within the prior. In Section 2.4, James-Stein estimator and hierarchical Poisson-Gamma model are studied as two examples for the parametric approach. Finally, in Section 2.5, we observe that Kiefer-Wolfowitz estimator can be justified as an example for nonparametric empirical Bayes paradigm.

2.1 The Problem of Multiple Prediction

We are concerned with the problem of estimating $\phi_i \in \mathbb{R}^p$, based on the observations $X_1, ..., X_n$ and

$$X_i \overset{ind}{\sim} f_i(\cdot|\phi_i) \tag{2.1}$$

for $i = 1, ..., n$ and $X_i \in \mathbb{R}^p$. The provision of different error distributions f_i is to allow for inclusion of covariates; without them all $f_i = f$, which will be assumed in what follows, unless the contrary is explicitly specified.

The ϕ_i's are viewed as drawn independently from a distribution Q. The performance of an estimator, or prediction $\hat{\phi}_i$ is evaluated based on the aggregate squared loss function

$$\sum_{i=1}^{n} \left\| \hat{\phi}_i - \phi_i \right\|^2 .$$

A classical application of this setting is the baseball predictions (Efron and Morris 1975,1977 and Brown 2008). The performance of a baseball player is measured by the batting average: the ratio of the number of hits H to the number of at-bats N. The number of hits naturally can be modelled as a binomial random variable $H \sim \text{Bin}(N, p)$ with an unknown parameter p that represents the player's latent ability. After taking the arcsin transformation $X = \arcsin(H/N)$, each X is approximately normally distributed with mean θ and variance $1/(4N)$, where $\theta = \arcsin(\sqrt{p})$. The parameter θ corresponds to ϕ in (2.1).

In the classical Bayesian setting, Q can be seen as a prior distribution. In such a case, the optimal prediction rule is the mean of the posterior density $E(\phi|x)$ derived via the Bayes rule. However, this can be worked out only if the prior distribution Q is known, which may not be true in applications. For example, to make the prior to reflect the past experience, we may need a large amount of data; in some fields, such a requirement is either impossible or too expensive. So a natural question arises: what if the prior is unknown or cannot be completely specified?

In such a situation, we can sometimes use the observed data to estimate

either the prior, or directly the prediction rule. When Q is not completely specified and depends on some unknown hyperparameter(s), say $\lambda \in \mathbb{R}^q$, we can handle the compound decision problem by first estimating λ from the marginal density of the observations and then computing the posterior mean. Another possible approach here is to use the hierarchical Bayes approaches, which treats λ as a random variable and forms prior distribution on the hyperparameters of the prior distributions, $\Lambda \sim \psi(\lambda)$, which is assumed to be known and not depending on any other unknown hyperparameters.

If the prior distribution Q is completely unknown, neither of the approaches above is applicable since both of them assume the parametric form of the prior is given. On the other hand, we can still work out the Bayes rule and it turns out to be an expression which mainly depends on the marginal distribution of X. If the error distribution $f(x|\phi)$ is taken from an exponential family, then the so-called Tweedie's formula shows we can either directly estimate the marginal density $f(x)$ or find a way to obtain the estimated prior $\hat{Q}$. In either case, the posterior mean can then be computed.

Brown (2008) applies both parametric and nonparametric empirical Bayes methodologies for prediction of the unknown θ_i, the transformed latent ability for player i. Each of them reduced the total squared error by about 50%, compared to the naïve estimator $\delta_0(X_i) = X_i$. Jiang and Zhang (2010) also suggest to add the covariates to the model and fit the partial linear regression

$$X_i = z_i^T \beta + \varepsilon_i + u_i,$$

where u_i is (approximately) normally distributed with variance $\sigma_i^2 = 1/(4N_i)$, the ε_i's are independently drawn from the unknown prior distribution and z_i

contains the information that whether the player is a so-called pitcher, the number of at-bats in the first half of the season and possible interactions. We then want to estimate $Z\beta + \varepsilon$. With this improvement, empirical Bayes estimators reduced about 80% of the total squared error.

Compared to the games like basketball or baseball, a game like hockey measures the performance of an individual player with the number of goals and assists. Given these numbers and the amount of time on ice for each player in a regular season, we are interested to know what are the predicted number of goals and assists in the playoffs. Each of the variables can be modelled with the Poisson process with an unknown parameter $\lambda = (\lambda_1, \lambda_2)$ that respectively represents the player's latent ability of scoring and assisting per unit time. To obtain the estimated λ_i, we can assume the two processes are independent and then the problem is equivalent to the univariate case. That is discussed in Section 5.3 where both the parametric and nonparametric Bayes approaches are used. However, the goals and the assists are considered likely to be positively correlated and the performance for an individual player is non-trivially measured by bivariate parameters. That is why we need a multivariate implementation.

2.2 Classical Bayes Paradigm

In the classical non-Bayesian approach, the parameter ϕ is unknown but fixed. Instead, in Bayesian statistics ϕ is considered to be a random variable described by a probability distribution Q, called prior distribution. The prior distribution, based on the experimenter's belief, is completely specified before data analysis. Once the prior is determined, we may take a random sample

$X_1, ..., X_n$ with distribution $f(\cdot|\phi)$ and then the Bayes rule says how to update our prior belief to posterior belief. If $\pi(\phi)$ is the density of the prior Q, then the posterior is

$$\pi(\phi|x) = \frac{f(x|\phi)\pi(\phi)}{f(x)}.$$

Suppose the loss is squared error. To obtain the Bayes estimator of ϕ, we choose a decision rule $d(x)$ to minimize the Bayes risk

$$\int_\Omega (\phi - d(x))^2 \pi(\phi|x) d\phi,$$

where Ω is the parametric space. Differentiating with respect to $d(x)$ and taking into account the posterior density integrates to 1, we obtain the Bayes estimator as the posterior mean

$$d(x) = E(\phi|x) = \int_\Omega \phi \frac{f(x|\phi)dQ(\phi)}{f(x)},$$

where $f(x) = \int_\Omega f(x|\phi)dQ(\phi)$.

The following theorem states that the Bayes rule always exists in most cases and is optimal with respect to the selected loss function.

Theorem 2.1: *Suppose the following assumptions hold for the problem of estimating $g(\Phi)$ with non-negative loss function $L(\phi, d)$.*

(a) There exists an estimator δ_0 with finite risk.

(b) For almost all x, there exists a value $\delta_Q(x)$ minimizing

$$E\left\{L\left[\Phi, \delta(x)\right] | X = x\right\}.$$

Then $\delta_Q(X)$ is a Bayes estimator.

Proof. See Lehmann and Casella (1998, page 228). □

The following example is due to Casella and Berger (2001, page 325).

Example. Let $X_1, ..., X_n$ be iid Bernoulli(p). Then $Y = \sum_{i=1}^{n} X_i$ is Binomial(n, p). The prior distribution on p is assumed to be Beta(α, β). The joint distribution of Y and p is

$$f(y,p) = \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}.$$

Notice that $f(p|y) \propto f(y,p)$ and then the posterior distribution is

$$f(p|y) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1},$$

which is Beta($y+\alpha, n-y+\beta$). If loss is squared error, the Bayes estimator of p is the posterior mean,

$$E(p|y) = \frac{y+\alpha}{\alpha+\beta+n}$$

or

$$E(p|y) = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right). \tag{2.2}$$

The (2.2) suggests the Bayes estimator in this case can be seen as a linear combination of the prior mean and the sample mean.

2.3 Tweedie's Formula

Suppose the random vector X and the parameter of interest ϕ are taken from $\mathbb{R}^p$ with $p \geqslant 1$. The parameter ϕ has a prior density q and the real-valued like-

lihood function of ϕ is taken from multivariate exponential family, as defined in (DasGupta, page 508),

$$f(x|\phi) = \exp\left\{\phi^T x - \psi(\phi)\right\} f_0(x),$$

where $\psi(\phi)$ is the cumulant generating function and $f_0(x) = f(x|\phi = 0)$.

Denote the marginal density as $f(x)$. A straightforward calculation gives the posterior density of ϕ given x,

$$f(\phi|x) = \exp\left\{x^T\phi - \lambda(x)\right\} [q(\phi)\exp\{-\psi(\phi)\}], \tag{2.3}$$

where $\lambda(x) = \ln\left(\frac{f(x)}{f_0(x)}\right)$. Notice that (2.3) is an exponential family with canonical parameter x and cumulant generating function $\lambda(x)$. For the multivariate exponential family, we have the result that the first derivative of the cumulant $\lambda(x)$ equals the expectation of ϕ given x. Then the Bayes rule has the form

$$E(\phi|x) = \frac{d\lambda(x)}{dx}, \tag{2.4}$$

where

$$\lambda(x) = \ln\left(\frac{f(x)}{f_0(x)}\right)$$

and

$$\frac{d\lambda(x)}{dx} = \left(\frac{\partial\lambda(x)}{\partial x_1}, ..., \frac{\partial\lambda(x)}{\partial x_p}\right).$$

The expression (2.4) is called Tweedie's formula and it was first provided by Robbins in 1956. Efron calls such an expression Tweedie's formula because that Robbins "credits personal correspondence with Maurice Kenneth Tweedie for an extraordinary Bayesian estimation formula". In some literature, this

formula is also referred to Robbins' formula but in this thesis we follow the terminology of Efron's paper. The formula (2.4) coincides for $p = 1$ with that derived in Efron (2011), who mentions a possibility of multivariate extension. As the latter is not readily available in the literature, we provide a multivariate version here.

Tweedie's formula says that the Bayes rule (2.4) depends directly on the marginal distribution of X; therefore, it is in principle not necessary to estimate the prior density q. Indeed, the marginal density $f(x)$ can be estimated by kernel density estimation or Lindsey's method (Efron, 2011).

Example. Suppose X is a p-dimensional random vector and μ is taken from $\mathbb{R}^p$. Given $X|\mu \sim N_p(\mu, \Sigma)$, where the covariance matrix Σ is known. It is clear that the canonical parameter $\phi = \Sigma^{-1}\mu$ and the Bayes rule is

$$E(\phi|x) = \frac{d \ln\left(\frac{f(x)}{f_0(x)}\right)}{dx}, \tag{2.5}$$

where

$$f_0(x) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{x^T \Sigma^{-1} x}{2}\right\}.$$

Then (2.5) becomes

$$\begin{aligned} E(\phi|x) &= \frac{d \ln(f(x))}{dx} - \frac{d \ln(f_0(x))}{dx} \\ &= \frac{d \ln(f(x))}{dx} + \Sigma^{-1} x. \end{aligned}$$

Now the Bayes rule for μ would be,

$$E(\mu|x) = x + \Sigma \frac{d \ln(f(x))}{dx}.$$

2.4 Parametric Empirical Bayes

Consider the setting defined in Section 2.1. If we can specify the parametric family of Q but leave certain hyperparameters unknown and the hyperparameters eventually are estimated from the data, this is called parametric empirical Bayes. The first major work in this area was made by Efron and Morris (1975, 1977). The procedure is first writing out the marginal distribution and then obtain the estimators for all hyperparameters. As soon as the prior is specified, the classical Bayesian follows and we can compute the posterior expectation without any trouble.

In this section, James-Stein estimator and hierarchical Poisson-Gamma model are studied as two examples of parametric empirical Bayes.

2.4.1 James-Stein Estimator

The approach leading to the James-Stein estimator assumes the observed density and the prior are both normal. To illustrate the outcome of this method, let us first consider a univariate case. Consider $\phi \sim N(0, a)$ with a unknown and the error distribution $f(x|\phi)$ we observed is $N(\phi, 1)$. Since we do not know the value of a in the prior $N(0, a)$, the classical Bayes approach cannot be used directly. However, we can follow empirical Bayes paradigm and extract the information about a from the marginal density of X.

It is not hard to see that the marginal distribution of X is again a normal distribution with mean 0 and variance $a + 1$. The Bayes estimator of ϕ_i is

$$\hat{\phi}_i = E(\phi_i|x_i) = \left(1 - \frac{1}{\hat{a} + 1}\right) x_i.$$

Using the method of moments, we may obtain a for an estimator

$$\hat{a} = \frac{\sum_{i=1}^{n} x_i^2}{n} - 1.$$

In the empirical Bayes, the unknown term $1/(a+1)$ is unbiasedly estimated by $(n-2)/\sum_{i=1}^{n} x_i^2$. This results in the James-Stein estimator

$$\hat{\phi}_i^{(JS)} = \left(1 - \frac{n-2}{\sum_{i=1}^{n} x_i^2}\right)_+ x_i,$$

where the notation $()_+$ is the positive part of function and it is defined as

$$f_+(x) = \max\{f(x), 0\}.$$

More generally, assume that $\phi_i \overset{ind}{\sim} N(M, A)$ and $X_i|\phi_i \overset{ind}{\sim} N(\phi_i, \sigma_0^2)$ with $i = 1, ..., n$ and $n \geqslant 4$, where the hyperparameters M and A are the mean and variance of the prior distribution. The marginal density of X_i is

$$X_i \sim N(M, A + \sigma_0^2)$$

and the posterior density

$$\phi_i|x_i \sim N(M + B(x_i - M), B\sigma_0^2),$$

where

$$B = \frac{A}{A + \sigma_0^2}.$$

Now the Bayes estimator of ϕ_i is

$$\hat{\phi}_i = M + B(z_i - M).$$

Although the values of A and B are unknown at the beginning, we can obtain the estimators from marginal density. Eventually, the James-Stein estimator acquires the form

$$\hat{\phi}_i^{(JS)} = \overline{x} + \left(1 - \frac{(n-3)\sigma_0^2}{S}\right)_+ (x_i - \overline{x}), \tag{2.6}$$

where $S = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

The estimator (2.6) shrinks each observed value x_i toward sample mean $\overline{x}$. The amount of shrinkage depends on other observations. This fact might counter our intuition because each observation $x_i|\phi_i$ is taken independently, but in most cases this type of shrinkage will reduce the total squared of error and improve the estimate of ϕ_i.

There is another very appealing theoretical property of this approach which is not so obvious at first glance.

Theorem 2.2: *For $n \geqslant 3$, the following is true that*

$$E_\phi\left\{||\hat{\phi}^{(JS)} - \phi||^2\right\} < E_\phi\left\{||\hat{\phi}^{(MLE)} - \phi||^2\right\}$$

for all ϕ.

Proof. See James and Stein (1961). □

The theorem says that for $n \geqslant 3$, the James-Stein estimator is always closer to ϕ than MLE; this fact severely shocked the statistical world.

We end up of this section with an example of Casella (1985) giving an intuitive justification of James-Stein estimator in the one-way analysis of variance (ANOVA).

Example. Suppose there are five treatments. Let $x_1, ..., x_5$ be observed means and $\theta_1, ..., \theta_5$ represent true means. Consider testing

$$\mathrm{H_0}\text{: all } \theta_i\text{'s equal}$$

versus

$$\mathrm{H_A}\text{: at least two of } \theta_i \text{ 's not equal.}$$

If $\mathrm{H_0}$ is true, then we should estimate θ with the sample mean $\overline{x}$, whereas if $\mathrm{H_A}$ is true, then we should estimate each θ_i with x_i. James-Stein estimator takes a compromise between these two extremes and can be seen as a linear combination of $\overline{x}$ and x_i (Figure 2.1). It becomes clear if we rewrite (2.6) as

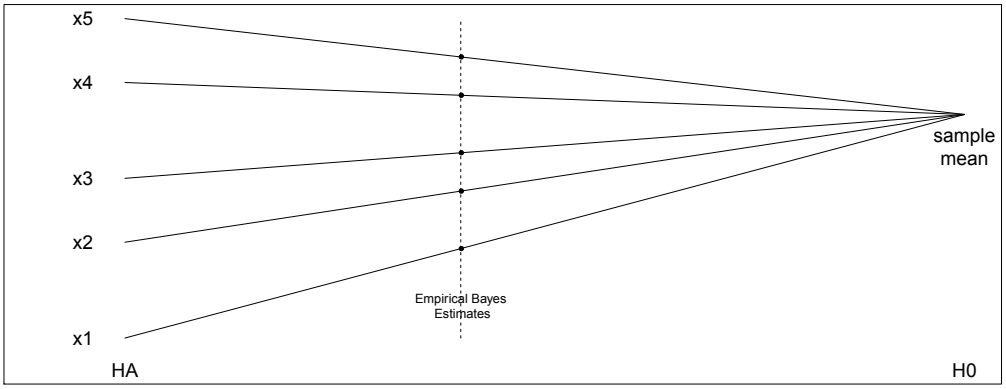
$$\hat{\phi}_i^{(JS)} = \left(\frac{(n-3)\sigma_0^2}{S}\right)\overline{x} + \left(1 - \frac{(n-3)\sigma_0^2}{S}\right)x_i. \tag{2.7}$$

From 2.1, we can see that James-Stein estimator affects extreme means x_1 and x_5 much more that if affects the ones that are close to $\overline{x}$. The amount of shrinkage depends the F statistic that test ANOVA null hypothesis. If there are n treatments, the F statistic is

$$F = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2}{(n-1)\hat{\sigma_0}^2},$$

where $\hat{\sigma_0}^2$ estimates σ_0^2. Here we assume σ_0^2 is known, hence the null hypothesis

Figure 2.1: The Empirical Bayes Estimator in the One-Way ANOVA



will be tested by

$$TS = \frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2}{(n-1)\sigma_0^2} \sim \frac{\chi^2_{n-1}}{(n-1)}$$

and (2.7) now can be written as

$$\hat{\phi}_i^{(JS)} = \left(\frac{n-3}{n-1}\right) TS^{-1}\overline{x} + \left[1 - \left(\frac{n-3}{n-1}\right) TS^{-1}\right] x_i.$$

As TS becomes large and the data favors H_A, $\hat{\phi}_i^{(JS)}$ puts more weights on x_i and less on $\overline{x}$, which consists of our intuition.

2.4.2 Hierarchical Poisson-Gamma Model

Poisson distribution is frequently used in the modelling of counts. In many cases, we will find that the observed variability is greater that the Poisson predicts. Such a phenomenon is called overdispersion. When the Poisson distribution is chosen, overdispersion is often a problem because in the model we assume mean and variance are identical, but usually they are not. One way to

deal with the overdispersion is to treat Poisson parameter as a random variable following Gamma distribution. Then it can be shown that the marginal distribution is negative binomial and it has larger variance than mean.

Suppose $X|\phi \sim \text{Poi}(\phi)$ and $\phi \sim \text{Gamma}(a, b)$, where a, b are positive and unknown. The marginal distribution of X is

$$f(x) = \frac{\Gamma(x+a)(\frac{b}{1+b})^x(\frac{1}{1+b})^a}{\Gamma(a)\Gamma(x+1)}$$

which is a negative binomial distribution with parameters a and $b/(1+b)$.

The method of moments estimators for a, b are $\widehat{a} = \overline{x}^2/(s^2 - \overline{x})$ and $\widehat{b} = (s^2 - \overline{x})/\overline{x}$, where $s^2 = \left(\sum_{i=1}^n x_i^2 - n\overline{x}^2\right)/(n-1)$. A straightforward calculation gives the posterior density

$$f(\phi|x) = \phi^{(x+a)-1}\exp\{-\frac{\phi}{(\frac{b}{1+b})}\}/[\Gamma(x+a)(\frac{b}{1+b})^{x+a}]$$

which is a gamma distribution with parameters $x+a$ and $b/(1+b)$. The Bayes estimator of ϕ_i is

$$\hat{\phi}_i = (1 - \frac{\overline{x}}{s^2})x_i + \frac{\overline{x}^2}{s^2}.$$

2.5 Nonparametric Empirical Bayes

The classical Bayes approach assumes the prior is completely known and set up before seeing any data. However, in practice this may not be feasible. The question is: if the prior Q is considered to exist but not known, is there a way to obtain an approximation to the unknown prior distribution function, or to the Bayes prediction rule, when the error distributions $f(\cdot|\phi_i)$ are given? Robbins (1956) was one of the first people who asked this question.

Suppose we observe $x_1, ..., x_n$ and the decision about ϕ_{n+1} is to be made. The key idea of the nonparametric empirical Bayes approach is to find a decision whose form depends on $x_1, ..., x_n$ such that the overall expected loss asymptotically equals to the minimum possible Bayes risk relative to Q. "[We] hope for large n...[we are] able to extract some information about [the prior] from the values...which have been observed, hopefully in such a way that [empirical Bayes] will be close to the optimal but unknown [Bayesian]." (Robbins, 1964) Given the squared loss function, to compute the posterior mean of ϕ_i, a possible approach is to find a way to estimate Q.

Let X be a random variable having probability density function or probability mass function depending on an hyperparameter ϕ,

$$L(\phi) = f(x|\phi)$$

and the parameter ϕ be a random variable with a prior distribution function Q. The marginal distribution of X is then the mixture

$$L(Q) = f(x|Q) = \int_{\Omega} L(\phi) dQ(\phi), \tag{2.8}$$

where Ω is the parameter space.

2.5.1 Kiefer-Wolfowitz Estimator

The method of Kiefer and Wolfowitz (1956) estimates the unknown prior via maximum likelihood. This amounts to

$$\max_{Q \in \mathcal{P}} L(Q) = \max_{Q \in \mathcal{P}} \int_{\Omega} L(\phi) dQ(\phi),$$

where $\mathcal{P}$ is the class of all probability measures on Ω. Once the estimator $\hat{Q}$ is obtained, the posterior mean can be computed as

$$\frac{\int_\Omega \phi L(\phi) dQ(\phi)}{\int_\Omega L(\phi) dQ(\phi)}.$$

Example. Assume $X_i \overset{ind}{\sim} \text{Poisson}(\phi_i)$, where $i = 1, ..., n$ and ϕ_i's are assumed to be taken from a distribution function Q. Kiefer-Wolfowitz MLE solves

$$\max_{Q\in\mathcal{P}} \left\{ \sum_{i=1}^{n} \ln\left(L_i(Q)\right) \right\} = \max_{Q\in\mathcal{P}} \left\{ \sum_{i=1}^{n} \ln\left(\int_\Omega \frac{\exp\{-\phi\}\phi^{x_i}}{x_i!} dQ(\phi) \right) \right\}.$$

Suppose the number of distinct data points is K. The Kiefer-Wolfowitz estimator can be rewritten as

$$\max_{Q\in\mathcal{P}} \left\{ \ln\left(\prod_{i=1}^{K} \left(L_i(Q)\right)^{n_i} \right) \right\} = \max_{Q\in\mathcal{P}} \left\{ \sum_{i=1}^{K} n_i \ln\left(L_i(Q)\right) \right\},$$

which is equivalent to

$$\min_{L(Q)\in\mathcal{M}} \left\{ -\sum_{i=1}^{K} n_i \ln\left(L_i(Q)\right) \right\}, \tag{2.9}$$

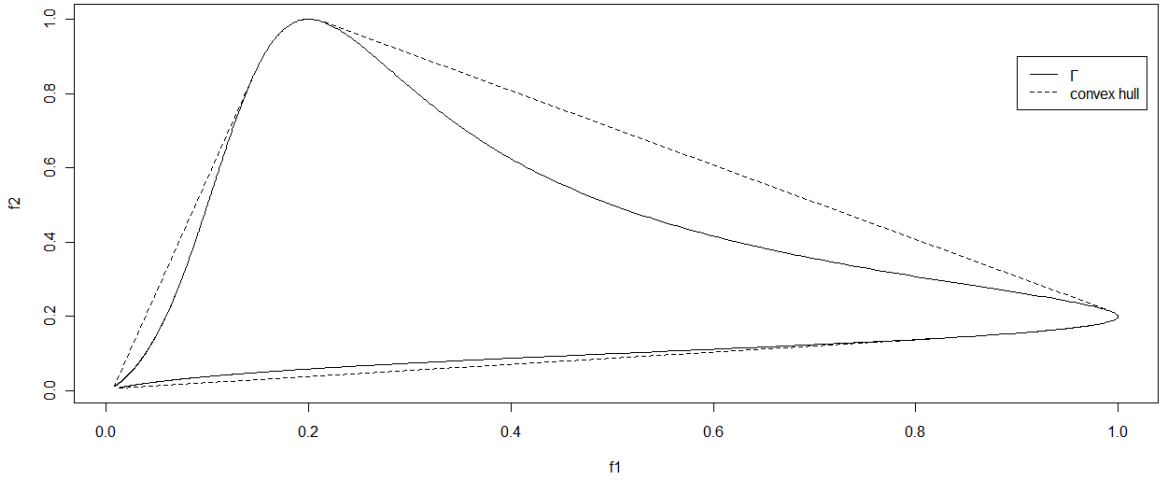
where $\mathcal{M}$ represents the set of mixture density vectors $\mathcal{M} = \{L(Q)|Q \in \mathcal{P}\}$ and $L(Q) = \{L_1(Q), ..., L_K(Q)\}$.

What does the set $\mathcal{M}$ look like? Let us denote the trace of the curve $L(\phi)$ as $\Gamma = \{L(\phi)|\phi \in \Omega\}$. Then we will have

$$L(Q) = \int_\Omega L(\phi) dQ(\phi).$$

Every element in $\mathcal{M}$ can be written as a convex combination of the elements in Γ. A fundamental result of convex geometry tell us that $\mathcal{M} = \text{conv}(\Gamma)$. Therefore, the set $\mathcal{M}$ is convex, because a convex hull itself is convex. The following example (Lindsay, 1995) can explain this.

Figure 2.2: The Likelihood Curve Γ for Two Cauchy Observations



Example. Let $L(\phi)$ be the Cauchy location density

$$\frac{1}{\pi}\left[1+(x-\phi)^2\right]^{-1}.$$

Given a pair of observations $(x_1, x_2) = (1, -1)$ and the location parameter $\phi \in [-10, 10]$, the curve has the form

$$\Gamma = \{(f_1, f_2)|\phi \in \Omega\} = \left\{\left(\left[1+(1-\phi)^2\right]^{-1}, \left[1+(-1+\phi)^2\right]^{-1}\right)|\phi \in \Omega\right\}.$$

The convex hull $\mathcal{M}$ of Γ is the region bounded by dashed lines (Figure 2.2).

Chapter 3

Duality: Discretized Version

In this chapter, we derive the duality theorem for the parameter spaces containing finitely many elements. This corresponds to the discretized version of the Kiefer and Wolfowitz method, which is used in implementations after all. From a theoretical point of view, the required mathematical formulation avoids sophisticated functional analysis. Later in Chapter 4, we use the result to obtain the continuous version (Duality Theorem 3) which has the analogous form as the discretized version studied in this chapter.

Section 3.1 introduces Lagrange duality as defined in Boyd and Vandenberghe (2004). Section 3.2 gives the derivation of the discrete duality theorem.

3.1 Lagrange Duality in Convex Optimization

We follow Boyd and Vandenberghe (2004). Consider an optimization problem in the standard form:

$$\min_x f_0(x)$$

$$\text{subject to } f_i(x) \leqslant 0, \quad i = 1, ..., m, \tag{3.1}$$

$$h_i(x) = 0, \quad i = 1, ..., p,$$

where f_0 is the objective function. We assume the domain is nonempty. The problem (3.1) is called convex when the objective f_0 is convex and subject to convex inequalities constraints $f_i(x) \leqslant 0,\ i = 1, ..., m$ and affine equalities constraints $h_i(x) = 0,\ i = 1, ..., p$.

In calculus, we learned how to use Lagrange's multipliers to solve the extremum problems with constraints. The similar idea, only concerning also inequality constraints, is applied here. Taking the constraints into account, we write the objective function with a weighted sum of the constraints: the Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ associated with the problem (3.1) is defined to be

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x).$$

The vectors λ and ν are called Lagrange's multiplier vectors. We define the Lagrange dual function as the minimum value of the Lagrangian over x:

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) = \inf_{x \in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right),$$

where the domain $D = \bigcap_{i=0}^{m} \operatorname{dom}(f_i) \cap \bigcap_{i=1}^{p} \operatorname{dom}(h_i)$. If the Lagrangian is unbounded below in x, the dual function takes the value $-\infty$.

We use the curled inequality symbol $\succeq$ (and its strict form $\succ$) to denote componentwise inequality (and strict inequality, respectively) between vectors.

It can be shown (Boyd and Vandenberghe, 2004, page 216) that for $\lambda \succeq 0$ and any ν, the dual function gives the lower bound on the optimal value p^* of

the problem (3.1)

$$g(\lambda, \nu) \leqslant p^*. \tag{3.2}$$

A natural question is: what is the greatest lower bound that can be obtained from the dual function (3.2)? This leads to the optimization problem

$$\max_{\lambda,\nu} g(\lambda, \nu) \tag{3.3}$$

$$\text{subject to } \lambda \succeq 0,$$

which is called the Lagrange dual problem associated with the problem (3.1); the original problem (3.1) is then called the primal problem.

If we denote the optimal value of the dual as d^*, from (3.2) it is straightforward to see that $d^* \leqslant p^*$. We would like to know when the equality will hold; it does not in general, but if the primal problem (3.1) is convex with the equality constraints $Ax = b$, we usually have strong duality $d^* = p^*$ under some simple conditions. One of those is called Slater's condition: there exists an $x \in \text{relint}(D)$, such that

$$f_i(x) < 0,\ i = 1, ..., m, \quad Ax = b$$

where

$$\text{relint}(D) = \left\{ x \in D | B(x, r) \bigcap \text{aff}(D) \subseteq D,\ \text{for some}\, r > 0 \right\}.$$

Theorem 3.1: *(Slater's theorem) If the problem is convex of the form*

$$\min f_0(x)$$

$$\text{subject to } f_i(x) \leqslant 0, \quad i = 1, ..., m,$$

$$Ax = b,$$

with $f_0, ..., f_m$ convex and satisfies Slater's condition, then strong duality holds.

For any optimization problem with differentiable objective and constraint functions satisfying strong duality, any pair of primal and dual optimal points, say $\tilde{x}$ and $(\tilde{\lambda}, \tilde{\nu})$, must satisfy the famous Karush–Kuhn–Tucker (KKT) conditions:

$$f_i(\tilde{x}) \leqslant 0, \ i = 1, ..., m,$$

$$h_i(\tilde{x}) = 0, \ i = 1, ..., p,$$

$$\tilde{\lambda}_i \geqslant 0, \ i = 1, ..., m,$$

$$\tilde{\lambda}_i f_i(\tilde{x}) = 0, \ i = 1, ..., m,$$

$$\nabla f_0(\tilde{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i \nabla f_i(\tilde{x}) + \sum_{i=1}^{p} \tilde{\nu}_i \nabla h_i(\tilde{x}) = 0.$$

Moreover, when the primal problem is convex, the KKT conditions are also sufficient for the points to be primal and dual optimal.

Theorem 3.2: *If a convex optimization problem with differentiable objective and constraint function satisfies Slater's condition, then the KKT conditions provide necessary and sufficient conditions for optimality: x is optimal if and only if there are (λ, ν) that, together with x, satisfy the KKT conditions.*

Theorem 3.2 can be used for checking primal and dual optimal and deriving the optimal solution from one to the other.

3.2 The Lagrange Dual

Consider now the setting of Kiefer-Wolfowitz estimation problem in which the parameter space contains finitely many elements. Denote the grid points as $\{u_1, ..., u_m\}$ and corresponding masses $\{\pi_1, ..., \pi_m\}$ with $\sum_{i=1}^{m} \pi_i = 1$. In such a case, the following duality theorem can be proved.

Notation. Let A be a n by m matrix consisting of the error density on the given grid points defined by $A_{ij} = f(x_i|u_j)$, where x_i's are observed values and u_j's are grid points. We write $\mathbf{1}_{m\times 1} = (1, ..., 1)^T$, $\pi = (\pi_1, ..., \pi_m)^T$ and the marginal density $g = A\pi$.

Theorem 3.3: *The primal problem*

$$\min_{g,\pi} \left\{ -\sum_{i=1}^{n} \ln(g_i) \,|A\pi = g,\ \mathbf{1}^T\pi = 1 \text{ and } \pi \succeq 0 \right\} \tag{3.4}$$

has the associated dual Lagrange problem

$$\max_{\nu} \left\{ \sum_{i=1}^{n} \ln(\nu_i) \,|A^T\nu \preceq n\mathbf{1},\ \nu \succ 0 \right\} \tag{3.5}$$

and Slater's condition is satisfied. If g^*, π^*, λ^* and ν^* are any points that satisfy the KKT conditions

$$\nu_i^* = 1/g_i^* \quad \lambda_i^*\pi_i^* = 0 \quad A\pi^* = g^* \quad \mathbf{1}^T\pi^* = 1 \quad \text{and} \quad \pi^* \succeq 0,$$

then (g^*, π^*) and (λ^*, ν^*) are primal and dual optimal.

Proof. First we derive the dual problem. For $\lambda \succeq 0$, the Lagrange dual function is given by

$$g(\lambda, \mu, \nu) = \inf_{g,\pi} L(g, \pi; \lambda, \mu, \nu)$$

$$= \inf_{g,\pi} \left\{ -\sum_{i=1}^{n} \ln(g_i)) + \nu^T(g - A\pi) + \mu(\mathbf{1}^T\pi - 1) + \lambda^T(-\pi) \right\}$$

$$= \inf_{g} \left\{ \sum_{i=1}^{n} (\nu_i g_i - \ln(g_i)) \right\} + \inf_{\pi} \left\{ \pi^T \left(-A^T\nu + \mu\mathbf{1} - \lambda \right) \right\} - \mu, \tag{3.6}$$

which can be determined analytically, since the infimum of the first term in (3.6) can be found by taking the derivative with respect to g and the minimum is attained when

$$g_i = 1/\nu_i$$

for $i = 1, ..., n$; the second term is a linear function which is bounded below only when it is identically zero. Therefore,

$$g(\mu, \nu) = \sum_{i=1}^{n} \{\ln(\nu_i) + 1\} - \mu = \sum_{i=1}^{n} \ln(\nu_i) + n - \mu,$$

if $A^T\nu + \lambda - \mu\mathbf{1} = 0$ and $-\infty$ otherwise. The Lagrange dual problem is to maximize this dual function g subject to $\lambda \succeq 0$, i.e.,

$$\max_{\mu,\nu} \sum_{i=1}^{n} \ln(\nu_i) + n - \mu \tag{3.7}$$

subject to $A^T\nu \preceq \mu\mathbf{1}$ and $\nu \succ 0$.

Now look at the inequality constraint in the dual problem and multiply the both sides by π^T,

$$\pi^T A^T \nu \leqslant \mu \pi^T \mathbf{1},$$

or

$$g^T \nu \leqslant \mu.$$

Also, we know that $\tilde{g}_i = 1/\nu_i$ and this yields $n \leqslant \mu$. On the other hand, the

matrix A consists of the values of error distribution and therefore it is bounded above. That is, there exists some real number $K > 0$, such that $A_{ij} < K$ for all i, j. Setting $v_i = 1/K$ for all i shows $\mu = n$ is feasible. Therefore, the dual problem (3.7) can be rewritten as (3.5).

Note that the primal (3.4) is a problem minimizing a convex function over a convex set which suggests the primal problem is convex. The Slater condition says that the optimal duality gap between (3.4) and (3.5) is zero if the inequality constraint is strictly feasible, i.e., there exist a π such that $\pi \succ 0$; we can see it is satisfied.

The last part of the proof follows the previous result: For a convex problem that satisfies Slater's condition, the KKT conditions provide necessary and sufficient conditions for optimality. Taking the derivative of L in g gives the 5th KKT condition:

$$\nu_i^* = 1/g_i^*.$$

Therefore, for any points that satisfy the KKT conditions

$$\nu_i^* = 1/g_i^*, \quad \lambda_i^* \pi_i^* = 0 \quad A\pi^* = g^* \quad \mathbf{1}^T \pi^* = 1 \quad \text{and} \quad \pi^* \succeq 0,$$

the pairs (g^*, π^*) and (λ^*, ν^*) are primal and dual optimal. □

Chapter 4

Duality: Continuous Version

The aim of the present chapter is to establish duality result when the parameter space Ω is continuous. The formulation of the dual problem can be deducted analogously from the discretized form derived in the previous chapter. Strong duality theorem, however, could be very hard to obtain by using general methods. Instead, it is better to use the results of Lindsay (1983), who studied maximum likelihood estimation for mixture distributions.

In Chapter 3, we proved that if the parameter space Ω is discrete, the primal problem

$$\min_{g,\pi} \left\{ -\sum_{i=1}^{n} \ln(g_i) \,|A\pi = g,\ \mathbf{1}^T\pi = 1 \text{ and } \pi \succeq 0 \right\}$$

and the associated dual problem

$$\max_{\nu} \left\{ \sum_{i=1}^{n} \ln(\nu_i) \,|A^T\nu \leqslant n\mathbf{1},\ \nu \succ 0 \right\}$$

are in the strong duality relationship, where A consists of the error density on the given grid points defined by $A_{ij} = f(x_i|u_j)$, where x_i's are observed

values and u_j's are grid points. If we look closely at the dual problem, we find that the constraints require that each component of $A^T\nu$ is less or equal to n. Denote the jth column of A as a_j and we have

$$a_j^T\nu = \sum_{i=1}^{n} f(x_i|u_j)\nu_i \leqslant n,$$

for all $j = 1, ..., m$. Each column of A corresponds to a grid point. When the parameter space contains an interval, as the number of grid points increases, the inequality constraints are likely to hold for all grid points and the associated dual problem is expected to have the form,

$$\max_{\nu}\left\{\sum_{i=1}^{n}\ln(\nu_i)\,|\sum_{i=1}^{n}\nu_i L_i(\phi) \leqslant n \text{ for all } \phi\right\}.$$

In Section 4.1, we sketch the proof of duality results for this formulation; the technical issues are postponed to Section 4.2.

4.1 The Continuous Duality Theorem

The result of Lindsay (1983) is built on the work of Böhning and Hoffmann (1981) who at that time are studying the problem of finding MLE for a certain class of discrete sampling models. The problem is to find a probability measure Q, such that the log likelihood function is maximized

$$\max_{Q\in\mathcal{P}}\left\{\sum_{i=1}^{K} n_i \ln\left(L_i(Q)\right)\right\} \tag{4.1}$$

or

$$\min_{L(Q)\in\mathcal{M}}\left\{-\sum_{i=1}^{K} n_i \ln\left(L_i(Q)\right)\right\}. \tag{4.2}$$

For simplicity, from now on we write the objective function as

$$l(p) = \sum_{i=1}^{K} n_i \ln(p_i),$$

where $p_i = L_i(Q)$ and $p \in \mathcal{M}$. Note that l is strictly concave on all positive values of its variables.

Since we are working on a convex set $\mathcal{M}$, it is natural to study the behaviour of the objective function l along paths between arbitrary points in the convex set. For this purpose, we give the following definition on directional derivative.

Definition: The directional derivative of l at y in direction z is defined as

$$d_1(y, z) = \lim_{\beta \to 0} \frac{l\left[(1-\beta)y + \beta z\right] - l(y)}{\beta},$$

where $\beta \in (0, 1)$, $y \succ 0$ and $z \succeq 0$. And $d_1(y, z) = +\infty$ for $y = 0$ in any arbitrary direction $z \succeq 0$.

Back to (4.1), we are interested to know the necessary and sufficient conditions for maxima of the concave function l and the following theorem proposed by Böhning and Hoffmann answered this question.

Theorem 4.1: *(Equivalence Theorem) For a differentiable concave function l, a is MLE if and only if $D_i l(a) = \nabla l(a) \cdot a$ for $i = 1, ..., n$.*

Proof. Suppose a is MLE, then $d_1(a, e_i) \leqslant 0$ for $i = 1, ..., n$ from the definition of directional derivative. Also notice that $d_1(a, a) = \sum_{i=1}^{n} a_i d_1(a, e_i) = 0$ and $a_i \geqslant 0$ for each i, then we conclude

$$d_1(a, e_i) = 0, \tag{4.3}$$

for each i.

Recall the fact that $d_1(y, z) = \nabla l(y) \cdot (z - y)$ (Appendix 1, Theorem 6.1), then

$$d_1(z, e_i) = \nabla l(a) \cdot (e_i - a). \tag{4.4}$$

Combing (4.3) and (4.4), we obtain $D_i l(a) = \nabla l(a) \cdot a$ for all i.

Conversely, suppose $D_i l(a) = \nabla l(a) \cdot a$ for $i = 1, ..., n$, then using Theorem 6.1 again, we find out that

$$d_1(a, e_i) = \nabla l(a) \cdot (e_i - a) = D_i l(a) - \nabla l(a) \cdot a \leqslant 0$$

for all i. For $p \in \mathcal{M}$, we have

$$d_1(a, p) = \sum_{i=1}^{n} p_i d_1(a, e_i) \leqslant 0,$$

since $p_i \geqslant 0$ for all i. For $y \succ 0$, we have the following inequality $\sup_{z \in \mathcal{M}} d_1(y, z) \geqslant l(z) - l(y)$ (Appendix 1, Theorem 6.3). Therefore,

$$0 = \sup_{p \in \mathcal{M}} d_1(z, p) \geqslant l(p) - l(a).$$

That is a is MLE.. □

Applying the equivalence theorem, Böhning and Hoffmann proved a duality theorem.

Theorem 4.2: *(Duality Theorem). For $y \succeq 0$, any $a \in \mathcal{M}$ is MLE if and only if a solves the primal problem:*

$$\min_{p \in \mathcal{M}} l(p)$$

subject to

$$d_1(y, e_i) \leqslant 0 \quad i = 1, ..., n.$$

Proof. If a is MLE, then by Equivalence Theorem, we will have

$$d_1(\mathrm{y}, \mathrm{a}) = \nabla l(y) \cdot (a - y) = \nabla l(y) \cdot (\sum_{i=1}^{n} a_i e_i - y)$$

$$= \sum_{i=1}^{n} a_i \nabla l(y) \cdot (e_i - y) = \sum_{i=1}^{n} a_i d_1(y, e_i) \leqslant 0.$$

Also, for any concave function l, we are able to show that $d_1(y, z) \geqslant l(z) - l(y)$ (Appendix 1, Theorem 6.2). It follows that

$$l(y) \geqslant l(y) + d_1(y, a) \geqslant l(y) + (l(a) - l(y)) = l(a).$$

That is, a is the solution to the primal problem. Conversely, suppose a is a solution of the primal problem, then by Theorem 6.3 in Appendix 1

$$0 = \sup_{1 \leqslant i \leqslant n} d_1(y, e_i) = \sup_{a \in \mathcal{M}} d_1(y, a) \geqslant l(a) - l(y),$$

for any $y \succ 0$. Hence, a is MLE. □

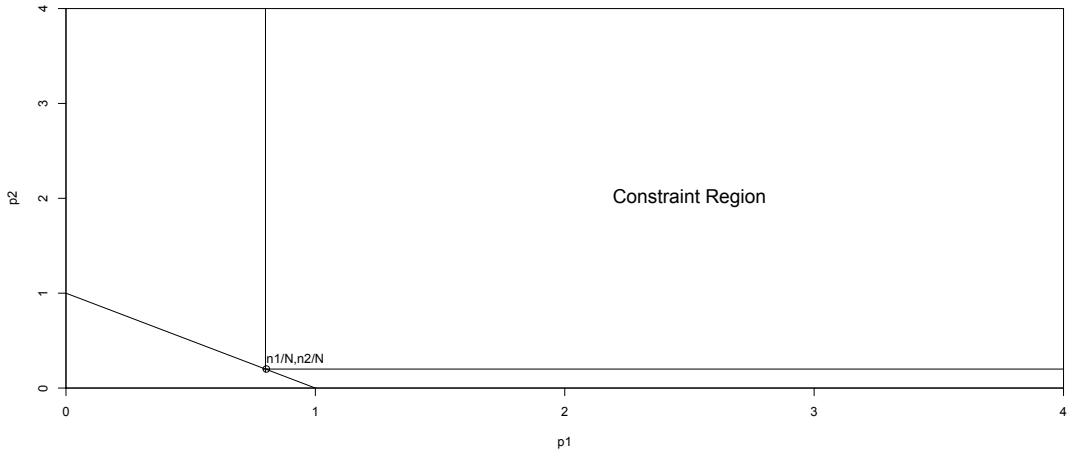
The following example is due to Böhning and Hoffmann (1982).

Example. Consider a n-cell multinomial, the kernel of the log likelihood function is

$$l(p) = \sum_{i=1}^{K} n_i \ln(p_i),$$

where n is the n-vector of observed frequency. The MLE can be worked out as $\hat{p} = n/N$, where $N = \sum_{i=1}^{K} n_i$. The primal problem is then to minimize l under

Figure 4.1: Geometric Demonstration of the Duality Theorem



the constraints $n_i/p_i \leqslant N$ (Figure 4.1). Geometrically, above the constraint region the objective function l is increasing under the direction (x, y) when x and y are nonnegative. Hence, it attains the minimum at the only vertex. According to the duality theorem, this point is the MLE.

Now we are ready to state Lindsay's duality result. Consider two equivalent problems:

Problem 1. Minimize $l(p)$ subject to $p \succeq 0$ and $d_1(p, L(\phi)) \leqslant 0$ for all $\phi \in \Omega$.

Problem 2. Maximize $l(\nu)$ subject to $\nu \succeq 0$ and $\sum_{i=1}^{n} \nu_i L_i(\phi) \leqslant n$ for all $\phi \in \Omega$.

Note that Problem 1 is equivalent to the one found by Böhning and Hoffmann.

Theorem 4.3: *(Duality Theorem 2) Suppose the trace of the curve* Γ *is compact and the mixture MLE is* $\hat{L} = L(\hat{Q})$ *, then* $p = \hat{L}$ *solves Problem 1 and* $\hat{\nu}_i = n_i/\hat{L}_i$ *solves Problem 2.*

Proof. Problems 1 and 2 are equivalent by the change of variable $(p_1, ..., p_K) =$

$(n_1/\nu_1, ..., n_K/\nu_K)$. Directly applying duality result of Böhning and Hoffmann gives the result that $p = \hat{L}$ solves Problem 1. □

Before we move on, I need to emphasize that the questions like the existence and uniqueness of the solution have not been discussed yet and will be discussed on Section 4.2.

Koenker and Mizera (2014), elaborating on Lindsay's result, derived the associated dual problem of (4.2) without assuming the likelihood curve Γ is closed.

Theorem 4.4: *(Duality Theorem 3) Let $\hat{Q}$ of (4.2) be an atomic probability measure, with at most K points of support. The locations $\hat{\phi}_j$ and the corresponding masses $\hat{p}_j$ can be found via the dual problem:*

$$\max_{\nu} \left\{ \sum_{i=1}^{n} \ln(\nu_i) \,|\, \sum_{i=1}^{n} \nu_i L_i(\phi) \leqslant n \text{ for all } \phi \right\}. \tag{4.5}$$

The solution $\hat{\nu}$ of (4.5) satisfies

$$\hat{\nu}_i = \frac{n_i}{\hat{L}_i} = \frac{n_i}{\sum_{j=1}^{m} L_i(\hat{\phi}_j)\hat{p}_j} \text{ for all i,}$$

where m is the number of grid points and $\hat{\phi}_j$ are exact if the equality in the dual constraint (4.5) holds.

Proof. In order to illustrate the idea of the proof, we assume $\Omega = (-\infty, \infty)$ and the number of distinct data points K is 1. The same arguments can be applied to the case $K > 1$.

To be able to apply Duality Theorem 2, we compactify $\mathbb{R}$ with ∞ and require the likelihood vector at ∞ is 0, i.e. $L(\infty) = 0$. Caratheodory's theorem

guarantees that the solution $\hat{Q}$ has no more than K points of support and we need to show that ∞ is not one of those atoms.

Suppose we have two solutions $\hat{Q}$ and $\hat{Q}^*$: $\hat{Q}$ has the support set $\{\hat{\varepsilon}_1, ..., \hat{\varepsilon}_J\}$, where $\hat{\varepsilon}_1 = \infty$ and the corresponding masses $\{\hat{\pi}_1, ..., \hat{\pi}_J\}$; $\hat{Q}^*$ has the support set $\{\hat{\varepsilon}_2, ..., \hat{\varepsilon}_J\}$ and the masses $\{\hat{\pi}_2^*, ..., \hat{\pi}_J^*\}$. If $\hat{\pi}_1 > 0$, then

$$L(\hat{Q}) = \sum_{j=1}^{J} L(\hat{\varepsilon}_j)\hat{\pi}_j = \sum_{j=2}^{J} L(\hat{\varepsilon}_j)\hat{\pi}_j < \sum_{j=2}^{J} L(\hat{\varepsilon}_j)\hat{\pi}_j^* = L(\hat{Q}^*).$$

It contrasts the fact that $\hat{Q}$ is MLE. □

4.2 Technical Issues

The following theorem is due to Lindsay (1983).

4.2.1 Existence and Support Size

Theorem 4.5: *(Existence and Support Size). Suppose that* Γ *is compact and* $\mathcal{M} = \text{conv}(\Gamma)$ *contains at least one point with positive likelihood, then there exists unique* $\hat{L} \in \partial\mathcal{M}$ *such that* $\hat{L}$ *maximizes the log likelihood function* l *over* $\mathcal{M}$*. The solution* $\hat{L}$ *can be expressed as* $L(\hat{Q})$*, where* $\hat{Q}$ *has no more than* K *points of support.*

Proof. For the continuous function l, it turns out that there is no stationary point in the compact set $\mathcal{M}$ so l approaches its maxima at the boundary. The uniqueness can be proved by contradiction. Suppose there are two distinct points $\hat{L}_1$ and $\hat{L}_2$ on the boundary of $\mathcal{M}$ which maximize the objective function

l. That is, $l(\hat{L}_1) = l(\hat{L}_2)$. The function l is strictly concave, for any $\alpha \in [0, 1]$,

$$l(\alpha\hat{L}_1 + (1-\alpha)\hat{L}_2) > \alpha l(\hat{L}_1) + (1-\alpha)l(\hat{L}_2) > l(\hat{L}_2).$$

This contradicts the fact that $\hat{L}_2$ is a maximum point. The second part is the direct consequence of a famous theorem of Caratheodory (Roberts and Varberg, 1973, page 76). □

The uniqueness of MLE $\hat{L}$ will be clear from the following example.

Example. Consider the upper set $U = \left\{p|l(p) \geqslant l(\hat{L})\right\}$. The strictly concavity of l implies the upper set U is convex and close (Lindsay, 1995). Let $L(\phi)$ be the Gaussian kernel

$$\exp\left\{-\frac{(x-\phi)^2}{2}\right\}.$$

Given a pair of observations $(x_1, x_2) = (1, -1)$ and $\phi \in [-10, 10]$. Then the likelihood curve has the form

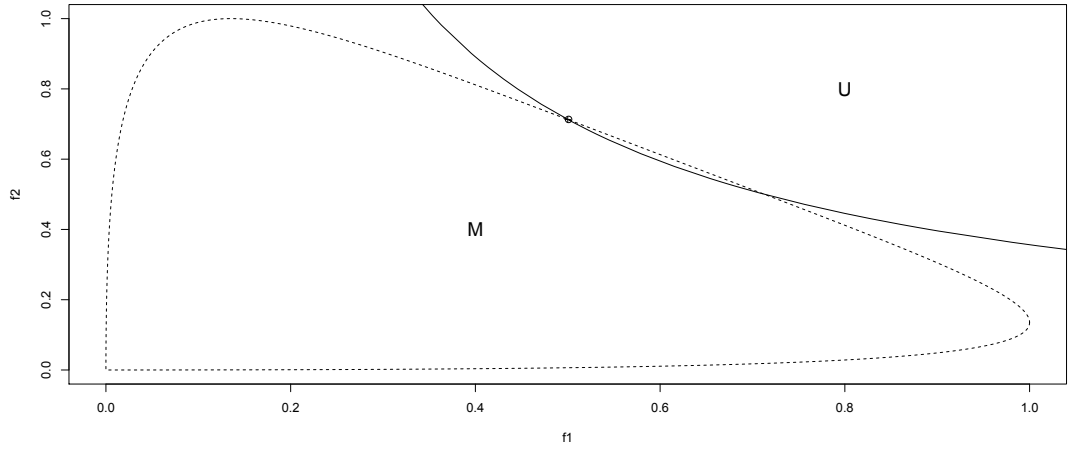
$$\Gamma = \{(f_1, f_2)|\phi \in \Omega\} = \left\{\left(\exp\left\{-\frac{(1-\phi)^2}{2}\right\}, \exp\left\{-\frac{(-1-\phi)^2}{2}\right\}\right)|\phi \in \Omega\right\},$$

the uniqueness can be seen as there is only one contact point between the upper set U and the convex hull $\mathcal{M}$ (Figure 4.2).

4.2.2 Identifiability

Given the unique estimated mixture likelihood vector $\hat{L}$, a natural question to ask is can we find the MLE $\hat{Q}$ via the equations $L(\hat{Q}) = \hat{L}$? If yes, is the solution $\hat{Q}$ unique? For univariate data points, Lindsay found out the answers

Figure 4.2: The Geometry of the Likelihood Maximization Problem



are positive for the both questions under the assumption of likelihood function $L(\phi)$ belonging to exponential family. However, the answer of identifiablity is unknown for multivariate case. Lindsay (1995) said that "the results [are] very difficult to obtain in higher dimensions, and hard to generalize outside the exponential family". Nevertheless, if the primary goal is to make prediction, this will not cause a problem. First of all, we note that the uniqueness theorem does not depend on the dimension of parameter space. In other words, the result that there is only one $\hat{L}$ that maximizes the log likelihood function l is also true when the dimension of the parameter space is greater than 1. If there exists distinct solutions, say $\hat{Q}_1$ and $\hat{Q}_2$, the uniqueness theorem guarantees their marginal densities are the same. That is, $L(\hat{Q}_1) = L(\hat{Q}_2)$. Tweedie's formula says the Bayes rule solely depends on the marginal density, hence the predicted value is unique.

In this section, I will follow the work of Lindsay (1983) and discuss univariate identifiability problem.

Recall Theorem 6.5 in Appendix 1, the mixture MLE $\hat{L}$ must lie in one of

the support hyperplanes of the mixture likelihood set $\mathcal{M} = \text{conv}(\Gamma)$, denoted as H. In symbols, we write

$$\hat{\text{L}} \in H = \left\{ z | \sum_{j=1}^{J} \nu_j z_j = n \right\},$$

where $\nu_j = n_j / \hat{L}_j$. The property of hyperplane gives the following theorem.

Theorem 4.6: *(Uniqueness). Assume the trace of the curve* Γ *is compact. Let* H *be a support hyperplane of* $\text{conv}(\Gamma)$ *containing* $\hat{L}$*, if* $H \cap \Gamma$ *consists of affinely independent vectors, then* $\hat{L}$ *on the boundary of* $\mathcal{M} = \text{conv}(\Gamma)$ *can be uniquely written as a convex combination of elements of* Γ *and the number of the elements is at most* K*.*

Proof. By the construction of L, the mixture MLE $\hat{L}$ lies in Γ. If $\hat{L} \in H$, then $\hat{L} \in H \cap \Gamma$. On the other hand, if $H \cap \Gamma$ consists of affinely independent vectors, then by Caratheodory's Theorem and affine independence, there exist unique $L(\varepsilon_1)$, $\text{L}(\varepsilon_2)$,... ,$L(\varepsilon_J)$, such that

$$\hat{L} = \sum_{j=1}^{J} \pi_j L(\varepsilon_j),$$

where $J \leqslant K$ and $\sum_{j=1}^{J} \pi_j = 1$. □

From Uniqueness Theorem, finding a likelihood function $L(\phi)$ to have an unique MLE $\hat{Q}$ is equivalent to finding $L(\phi)$ such that the both conditions hold:

(1) There are at most K solutions to

$$\sum_{i=1}^{K} \hat{\nu}_i L_i(\phi) = n$$

given the constraint

$$\sum_{i=1}^{K} \hat{\nu}_i L_i(\phi) \leqslant n,$$

where $\hat{\nu}_i = n_i/\hat{L}_i$.

(2) The set $\{L(\varepsilon_1), \mathrm{L}(\varepsilon_2), ..., L(\varepsilon_J)\}$ are affinely independent.

Let us consider the condition (1) at first. Recall the following result from mathematical analysis which can be proved by induction:

Lemma 4.1: *Let the polynomials $q_1(x), ..., q_K(x)$ be nonzero polynomials and with degree $a_1, ..., a_K$, and let the real constants $a_1, ..., a_K$ be distinct, then*

$$\sum_{i=1}^{K} q_i(x) \exp\{c_i x\}$$

has at most $K - 1 + \sum_{i=1}^{K} a_i$ real zeros.

Proof. See Pólya and Szegö, 1925, page 46. □

Suppose the function $f(\phi) = \sum_{i=1}^{K} \hat{\nu}_i L_i(\phi) - n$ is analytic in ϕ. In order to bound the number of roots to $f(\phi)$ at K, it is sufficient to show that $f'(\phi)$ has at most $2K - 1$ zeros, since maxima must alternate with minima. If $L(\phi)$ is the kernel of exponential family with the form $L(\phi) = \exp\{\phi x - \psi(\phi)\}$, then we have

$$f'(\phi) = \sum_{i=1}^{K} \hat{\nu}_i \left[x_i - \psi'(\phi)\right] \exp\left\{\phi x_i - \psi(\phi)\right\}.$$

Therefore, if $\psi'(\phi)$ is in the right form, the theorem above can be applied. For example, for normal density with mean ϕ and variance 1, we have $\psi'(\phi) = \phi$ and $f'(\phi)$ has at most $2K - 1$ zeros.

Now consider the condition (2). It is sufficient to show there exist K vectors

$L(\varepsilon_1), L(\varepsilon_2), ..., L(\varepsilon_K)$ which are linearly independent. Let

$$M = [L(\varepsilon_1), L(\varepsilon_2), ..., L(\varepsilon_K)]$$

be a $K \times K$ matrix with kth column $L(\varepsilon_k)$. It is equivalent to show $\det(M) \neq 0$. This is true if $L(\phi)$ is the kernel of exponential family (Karlin, 1968, page 18-20,117-120).

Chapter 5

Implementation

In practice, we have to work on the discretized parameter space Ω and estimate Q by a finite-dimensional approximation. When we look at the dual problem, we find there is an potential to give an algorithm which computes Kiefer-Wolfowitz's MLE in higher dimensional space. The primary goal of this chapter is to develop the adaptive algorithm which can be used in multivariate case.

In Section 5.1, I discuss the algorithm which can iteratively capture the support points with a grid of moderate size. A univariate numerical experiment is given in Section 5.2.

5.1 Adaptive Algorithm

It can be seen that Kiefer-Wolfowitz method also works for multivariate response and then it is natural to ask for an algorithm applied to this case. The univariate algorithm can be well done with a uniformly spaced grid as long as the grid is fine enough. However, for multivariate case, the number of grid

points grows quickly and make the computation prohibitive.

On the other hand, if we look at the primal and dual problems (4.2), (4.5), we find the objective function of the primal formulation is infinite-dimensional in Q, whereas the objective of the dual formulation is finite dimensional in ν and infinite-dimensionality is only in the constraint (Koenker and Mizera, 2014). This provides a possibility of iteratively using unequally spaced grids to approximate Q. Duality results tell us the maxima locate where the dual constraint is active. If such information is available, we could simply select the test points at these locations. Unfortunately, in practice such information is unavailable, we need to refine the solution in an iterative manner by looking at the grid identified by previous iteration.

Recall that the support points of $\hat{Q}$ locate at where the directional derivative is zero (Theorem 6.5). We start with an unequally spaced grid. Begin the loop. In each iteration, we look at the points where the critical condition (5.1) likely holds but throw out the rest of points. This can be done because the solution of the dual problem does not depend on the value of points out of the solution set. Next we solve the dual problem with the updated grid. Then we double the size of the gird in case of any solution is missing in the coming iterations. Go back to where we start the loop and continue. The algorithm stops when the index equals the desired number of iteration. The algorithm can be naturally generated to multivariate case.

The set $\{u_1, ..., u_m\}$ represents a grid, not necessarily equally spaced. Let A be a n by m matrix consisting with the error density on the given grid points defined by $A_{ij} = f(x_i|u_j)$, where x_i's are observed values and u_j's are grid points. To find the solution that the dual objective function is maximized, we

look at the set

$$\left\{u_j | a_j^T v = n\right\}, \tag{5.1}$$

where a_j is the jth column of A and $j = 1, ..., m$.

In practice, we look at where the equality in (5.1) nearly holds. That is, for a small $\varepsilon > 0$, the plausible solution lie within the set

$$\left\{u_j |\ |a_j^T v - n| < \varepsilon\right\}.$$

Denote the index by i and the number of iteration by I. The algorithm works as following:

(1) Set $m = 10$, $\varepsilon = 1$ and $i = 1$.

(2) **Repeat**

(3) Start with an unequally spaced grid $G = \{u_1, ..., u_m\}$ and generate matrix A.

(4) If $i > 1$: Calculate the difference $A^T v - n\mathbf{1}$ and update the grid G by look at the condition

$$|a_j^T v - n| \leqslant \left(1 - \frac{i}{I + 0.1}\right)\varepsilon \tag{5.2}$$

for each j.

(5) Maximize the dual problem with the grid G.

(6) Update m to $2m$. Increase i by one.

(7) **Until** i equals I.

In step 4, as the gap $|a_j^T v - n|$ getting close to zero, the number of points in the updated grid is much smaller than the ones in un-updated grid. That is, only the potential solutions are interested. The cost of computation in

optimization is therefore significantly reduced. From step 6, we can see the grid gets finer after each iteration so we are not likely miss any support point. As the number of iteration I goes to infinity, we will see the right-hand side of (5.2) goes to zero. All of these together suggest when the number of iteration is sufficiently large, we may capture the solution of the dual problem with a grid of moderate size.

5.2 Numerical Experiments

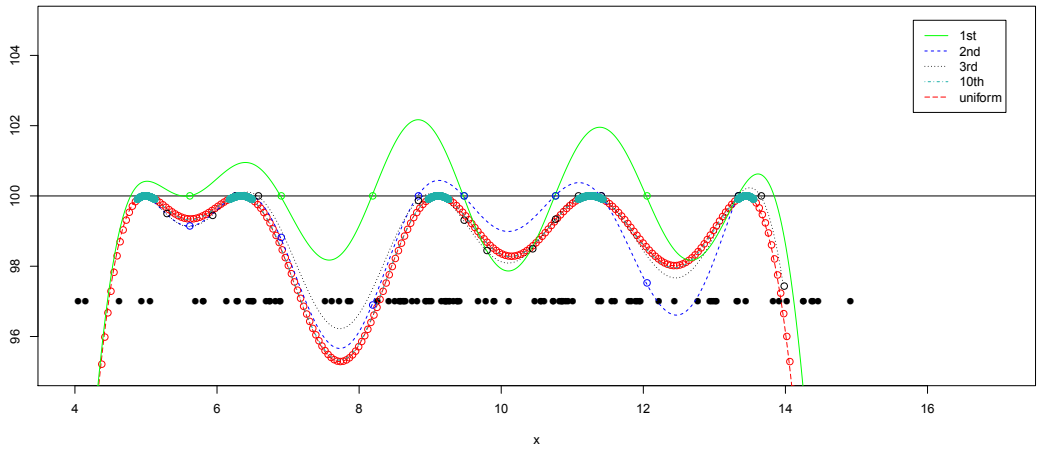
First we numerically verify that the solution of the dual problem locate at where the directional derivatives are zero and do not depend on the rest of the points in the grid. Then we demonstrate the performance of the adaptive algorithm in Bayesian estimation and compare it with the one using a uniformly spaced grid. The multivariate example is left to the future work.

The sample is randomly taken from Gaussian mixture model $X_i \overset{ind}{\sim} \mathrm{N}(\theta_i, 1)$ with prior distribution $\theta_i \overset{ind}{\sim} \mathrm{Unif}(5, 15)$, where $i = 1, ..., 100$.

Recall the adaptive algorithm is introduced based on the fact that the support points locate at where the directional derivatives are zero and do not depend on the rest of the points in the grid. To see this is numerically true, for each iteration we graph the vector $A^T v$ by only taking the points where the critical condition nearly holds and compare the curve to the one generated with a uniformly spaced grid. In Figure 5.1, we find the sequence of the curves eventually "converges" to the curve generated with a uniformly spaced grid. And the solution of the dual problem locate around where the curve touch the horizontal line. These are what we all expect to see.

The adaptive algorithm starts with 11 unequally spaced grids and the num-

Figure 5.1: Experiment



ber of iteration is set to be 10. In the end, the average number of grid used, after 30 repetition, is 937, compared to 1000 points used in the uniformly spaced algorithm.

When we look at the mixture density and Bayes rule, the two algorithms give very similar estimated results.

Figure 5.2: Adaptive Algorithm

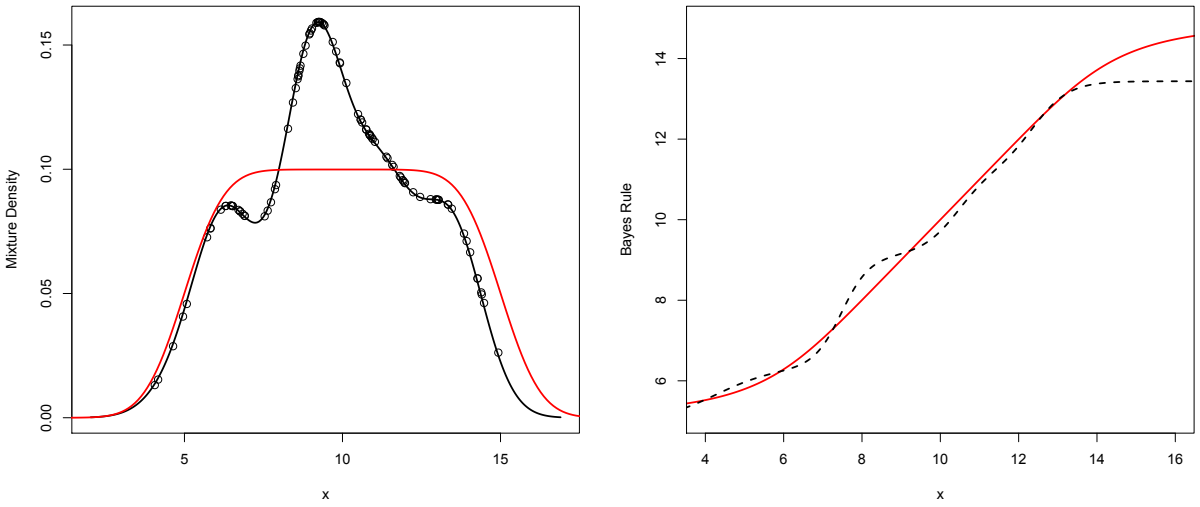
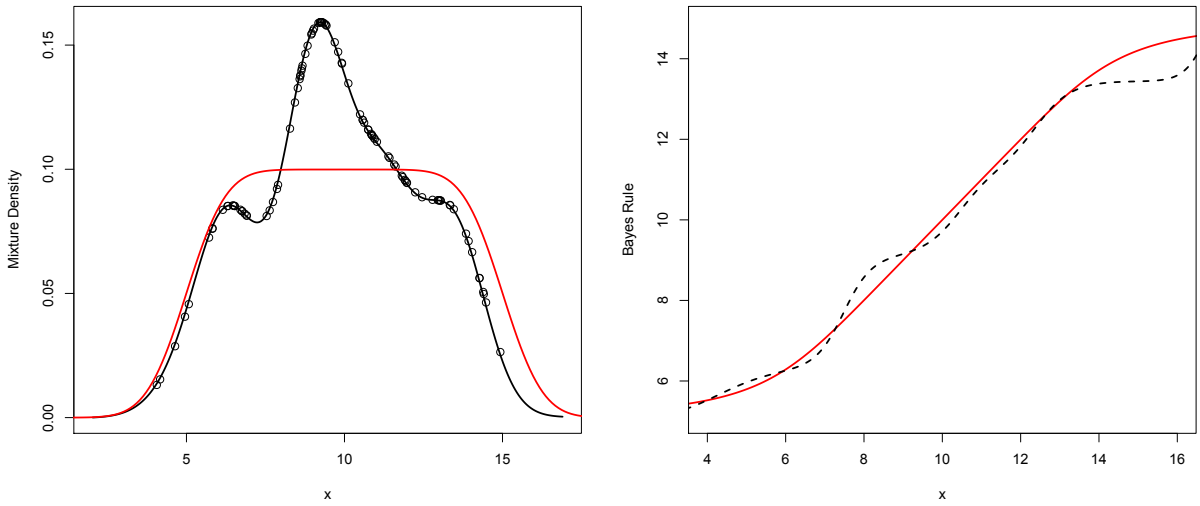


Figure 5.3: Algorithm with A Uniformly Spaced Grid



5.3 Empirical Ice Hockey

To test the performance of empirical Bayes paradigm, Brown (2008) delivered an example on prediction of the second-half season baseball batting average. As we know, baseball is a game that each player's performance can be measured by a single variable: batting average. In this sense, we may call baseball a univariate game. Nevertheless, for hockey, we need two variables, goals and assists, to describe players' performance. Hence, hockey is a bivariate game. In this chapter, I will only study the univariate case, that is, the two variables are assumed to be independent.

Consider n players involving in regular season and playoffs. For each player, given the total time on ice t and the number of goals and assists he earned by time t in regular season, denoted as G_t and A_t, we wish to estimate the expected number of goals and assists earned in playoffs.

Table 5.1 gives us a 2012-2013 regular season summary from the National

Hockey League. For each player, all the information we have now are the number of goals and assists he earns by time t.

Table 5.1: 2012-2013 Regular Season Summary

Player	Goal	Assist	Time On Ice (Seconds)
Adam Burish	1	2	10:34
Adam Cracknell	2	4	8:36
Adam McQuaid	1	3	14:17
...	...	...	...

5.3.1 Univariate Case

To make the story simple, we assume G_t and A_t are independent and only consider G_t from now on.

In practice, the family $\{G_t\}_{t\geqslant 0}$ can be modeled by Poisson process, then we will have

$$G_t \sim \text{Poi}(\lambda t),$$

where λ describes the expected number of goals per unit time in regular season.

To compare the performance of different methods, we calculate the sum of squared prediction error,

$$SSPE(\lambda) = \sum_{i=1}^{n} [\hat{\lambda}_i t_{2i} - g_{2i}]^2,$$

where $g_{2i} = g_{2i}(t)$ represents the number of goals the player i gained by time t_{2i} in the playoffs. In other words, the sum of squared prediction error is the Euclidean distance between the predicted goals up to time t_2 and the observed ones.

Description of estimators

Naïve estimator. We take the naïve estimator as the baseline, which simply uses the observed number of goals per unit time in regular season multiplied by time on ice in the playoffs as its prediction

$$\hat{\lambda}_0 = g_{1i}/t_{1i}.$$

We then are able to compute the sum of squared prediction error due to the naïve estimator

$$SSPE(\hat{\lambda}_0) = \sum_{i=1}^{n} [\hat{\lambda}_0 t_{2i} - g_{2i}]^2,$$

and then calculate the ratio

$$SSPE^*(\hat{\lambda}) = \frac{SSPE(\hat{\lambda})}{SSPE(\hat{\lambda}_0)}.$$

In this way, $SSPE^*(\hat{\lambda}_0) = 1$. If the ratio is far away from 1, it suggests the method is efficient and it significantly reduces the prediction errors.

EB (James-Stein (Subsection 2.4.1)). Suppose $G_t|\lambda \sim N(\theta t, \sigma^2)$ and $\lambda \sim N(M, A)$, where σ^2 is known. The posterior density is again a normal distribution with mean $M + B(tg_t - t^2 M)$ and variance $B\sigma^2$, where $B = A/(t^2 A + \sigma^2)$, then the James-Stein estimator (method of moments) of λ_i is

$$\hat{\lambda}_i = \frac{1}{t}\left(\overline{g_1} + \left(1 - \frac{\sigma^2}{S}\right)_+ (g_{1i} - \overline{g_1})\right),$$

where $S = \sum_{i=1}^{n} g_{1i}^2 - n\overline{g_1}^2$.

EB (Poisson-Gamma (Subsection 2.4.2)). Suppose $G_t|\lambda \sim \text{Poi}(\lambda t)$ and $\lambda \sim \text{Gamma}(a, b)$, then the Bayes estimator (method of moments) of λ_i under

the Poisson-Gamma model is

$$\hat{\lambda}_i = \frac{(1 - \frac{\overline{x}}{s^2})g_{1i} + \frac{\overline{x}^2}{s^2}}{t_{1i}},$$

where $s^2 = \left(\sum_{i=1}^{n} g_{1i}^2 - n\overline{g_1}^2\right)/(n-1)$.

NPEB (Tweedie (Section 2.3)). Suppose $G_t \sim \text{Poi}(\lambda t)$ and $\lambda \sim Q(\cdot)$, then the Bayes estimator of λ_i by using Tweedie's formula is

$$\hat{\lambda}_i \doteq \left(\frac{g_i + 1}{t_{1i}}\right) \frac{\hat{f}(g_i + 1)}{\hat{f}(g_i)},$$

where $\hat{f}(g_i)$ is the estimated marginal density of G_t using kernel density estimation.

NPEB (KW (Subsection 2.5.1)). Suppose $G_t \sim \text{Poi}(\lambda t)$ and λ follows an unknown prior Q, then the Kiefer-Wolfowitz estimator of λ_i is

$$\hat{\lambda}_i = \frac{\sum_{j=1}^{m} u_j L(g_{1i}|u_j)\hat{\pi}_j}{\sum_{j=1}^{m} L(g_{1i}|u_j)\hat{\pi}_j}.$$

The result of this is obtained by using interior point method implemented by Mosek.

Table 5.2 demonstrates the performance among different estimators. It can be seen that in this data set, Kiefer-Wolfowitz estimator outperforms other empirical Bayes estimators, both for goals and assists.

Table 5.2: Playoffs Prediction for All Players

Estimator	SSPE*(Goals)	SSPE*(Assists)
Naïve	1	1
EB(JS)	0.997	0.997
EB(PoiGam)	0.811	0.867
NPEB(Tweedie)	0.824	0.882
NPEB(KW)	0.803	0.861

Chapter 6

Conclusion

The primary goal of this thesis is to implement the Kiefer-Wolfowitz non-parametric empirical Bayes method for models with multivariate response. The current method may be not numerically feasible when the response variables are multivariate. Motivated by one paragraph from Koenker and Mizera (2014), we consider the dual problem instead and able to come up with an adaptive algorithm, which iteratively uses unequally spaced grids to approximate the prior. In the end, we can solve the dual problem without using overly many grid points. The algorithm has a potential to compute Kiefer-Wolfowitz's MLE in higher dimensional space. The numerical experiment and the field study on hockey both provide good performance.

The future work is to generate the adaptive algorithm to multivariate and be able to apply it to bivariate hockey game.

Appendix 1 Results in Vector Calculus

Theorem 6.1: *For $y \succ 0$ and $z \succeq 0$, $d_1(y,z) = \nabla l(y) \cdot (z-y)$.*

Proof. The proof is similar to univariate case. If a function l is differentiable at z then there exists a linear map $T_a : \mathbb{R}^n \to \mathbb{R}$ such that

$$l(a+v) = l(a) + T_a(v) + \|v\| E(a,v),$$

for $\|v\| < r$ for some $r > 0$, where $E(a,v) \to 0$ as $\|v\| \to 0$.

Then we follow the idea of Apostol (page 259) and write the difference

$$l\left[(1-\beta)y + \beta z\right] - l(y)$$

as

$$l\left[(1-\beta)y + \beta z\right] - l\left[(1-\beta)y\right] + l\left[(1-\beta)y\right] - l(y)$$

$$= T_{(1-\beta)y}\left(\beta z\right) + \|\beta z\| E((1-\beta)y, \beta z) + T_y\left(-\beta y\right) + \|-\beta y\| E(y, -\beta y). \quad (6.1)$$

Notice that the maps $T_{(1-\beta)y}$ and T_y are linear, then we can rewrite (3) as

$$\beta T_{(1-\beta)y}(z) + \|\beta z\| E((1-\beta)y, \beta z) - \beta T_y(y) + \|-\beta y\| E(y, -\beta y).$$

Now

$$\lim_{\beta \to 0} \frac{1}{\beta} \{l[(1-\beta)y + \beta z] - l(y)\} = T_y(z) - T_y(y)$$

$$= T_y\left(\sum_{i=1}^{n} z_i e_i\right) - T_y\left(\sum_{i=1}^{n} y_i e_i\right) = T_y\left(\sum_{i=1}^{n} (z_i - y_i) e_i\right) = \sum_{i=1}^{n} (z_i - y_i) T_y(e_i)$$

$$= \sum_{i=1}^{n} (z_i - y_i) D_i l(y) = \nabla l(y) \cdot (z - y).$$

Hence, $d_1(y, z) = \nabla l(y) \cdot (z - y)$. □

Theorem 6.2: *If l is concave, then $d_1(y, z) \geqslant l(z) - l(y)$ for $y, z \succ 0$ with $\sum_{i=1}^{n} y_i = 1$ and $\sum_{i=1}^{n} z_i = 1$.*

Proof. If f is concave, then by definition of concavity we immediately have

$$l\left[(1-\beta)y + \beta z\right] - l(y) \geqslant (1-\beta)l(y) + \beta l(z) - l(y) = \beta\left(l(z) - l(y)\right).$$

Divide β on the both sides and then take the limit as β goes to zero, then

$$d_1(y, z) = \lim_{\beta \to 0} \frac{1}{\beta} \left\{l\left[(1-\beta)y + \beta z\right] - l(y)\right\} \geqslant l(z) - l(y).$$

This completes the proof. □

Theorem 6.3: *If l is concave, then* $\sup_{1\leqslant i\leqslant n} d_1(y,e_i) = \sup_{z\in S} d_1(y,z) \geqslant l(z) - l(y)$ *for* $y \succ 0$ *with* $\sum_{i=1}^{n} y_i = 1$.

Proof. By Theorem 6.2, the inequality holds and we only need to show $\sup_{1\leqslant i\leqslant n} d_1(y,e_i) = \sup_{z\in S} d_1(y,z)$. Let $m = \sup_{1\leqslant i\leqslant n} D_i l(y)$, we will have

$$d_1(y,z) = \nabla l(y)\cdot(z-y) = \sum_{i=1}^{n} z_i D_i l(y) - \nabla l(y)\cdot y$$

$$\leqslant m\sum_{i=1}^{n} z_i - \nabla l(y)\cdot y = m - \nabla l(y)\cdot y.$$

The equality holds when $z = e_i$ and i is the index such that $D_i l(y) = m$. On the other hand, we have

$$d_1(y,e_i) = \nabla l(y)\cdot e_i - \nabla l(y)\cdot y \leqslant m - \nabla l(y)\cdot y.$$

We proved $\sup_{1\leqslant i\leqslant n} d_1(y,e_i) = \sup_{z\in S} d_1(y,z)$. □

For convenience, we write $D_Q(\phi) = d_1(L(Q), L(\phi))$.

Theorem 6.4: *The following three statements are equivalent:*

(1) $\hat{Q}$ *maximizes* $\ln(L(Q))$.

(2) $\hat{Q}$ *minimizes* $\sup_{\phi} D_Q(\phi)$.

(3) $\sup_{\phi}\{D_{\hat{Q}}(\phi)\} = 0$.

Proof. Suppose $\hat{Q}$ maximizes $l(Q)$, we have $d_1(\hat{L}, L(Q)) \leqslant 0$ for all Q, so (1) implies (3). The statements (2) and (3) are equivalent by noticing that

$$\sup_{\phi} D_Q(\phi) \geqslant d_1(L(\phi), L(\phi)) = 0.$$

To prove (3) implies (1), we will show $\sup_{\phi} D_Q(\phi) = 0$ only if $Q = \hat{Q}$. Theorem 6.2 says

$$d_1(L(Q), L(\phi)) \geqslant l(\phi) - l(Q),$$

for all Q, ϕ. Suppose $Q = Q^*$ and $Q^* \neq \hat{Q}$, then $l(\phi) > l(Q)$ for some ϕ and $D_Q(\phi) > 0$. □

Theorem 6.5: *The support of MLE $\hat{Q}$ lies in the set $\{\phi | D_{\hat{Q}}(\phi) = 0\}$.*

Proof. Suppose ε_j are support points of $\hat{Q}$, then

$$0 = D_{\hat{Q}}(\hat{Q}) = \int D_{\hat{Q}}(\phi) d\hat{Q}(\phi) = \sum_{j=1}^{J} D_{\hat{Q}}(\varepsilon_j)\pi_j,$$

where all π_j are positive. From Theorem 6.4 part 3, we have $D_{\hat{Q}}(\varepsilon_j) \leqslant 0$, therefore, $D_{\hat{Q}}(\varepsilon_j) = 0$ for all j. □

Bibliography

Böhning, D. and Hoffmann, K. M. (1982). Numerical techniques for estimating probabilities. *J. Statist. Comp. Sim.* **14** 283-293.

Boyd, S. and Vandenberghe, L. (2004). Convex Optimization. Cambridge University Press: Cambridge.

Brown, L. (2008). In-season prediction of batting averages: A field test of simple empirical Bayes and Bayes methodologies. *Ann. Appl. Statist.* **2** 113-152.

Casella, G. (1985). An introduction to empirical Bayes data analysis. *Amer. Statist.* **39** 83-87.

Casella, G. and Berger, R. (2001). Statistical Inference, 2nd ed. Pacific Grove, CA: Wadsworth.

DasGupta, A. (2011). Probability for Statistics and Machine Learning. New York: Springer.

Efron, B. (2010). Large-Scale Inference. Cambridge University Press: Cambridge.

Efron, B. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* **106** 1602-1614.

Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70** 311-319.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Sci. Amer.* **236** 119-127.

James, W. and Stein, C. (1961). Estimation with quadratic loss.

Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I. Univ. California Press: Berkeley, Calif. 361–379.

Jiang, W. and Zhang, C.-H. (2010). Empirical Bayes in-season prediction of baseball batting average. *Borrowing Strength: Theory Powering Application—Festschrift for L.D. Brown* (J.O. Berger, T.T. Cai, I.M. Johnstone, eds.) IMS Collections **6** 263-273.

Karlin, S. (1968). Total Positivity. Stanford University Press: Stanford, Calif.

Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.* **27** 887-906.

Koenker, R. and Mizera, I. (2014). Convex optimization, shape constraints, compound decisions, and empirical Bayes rules. *J. Amer. Statist. Assoc.* In press

Lehmann, E. and Casella, G. (1998). Theory of Point Estimation (2nd ed.) New York: Springer.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. *Ann. Statist.* **11** 86-94.

Lindsay, B. (1995). Mixture Models: Theory, Geometry and Applications. NSF-CBMS-IMS Conference Series in Statistics, Hayward, CA.

Pólya, G. and Szegö, G. (1925). *Aufgaben und Lehrsatze aus der Analysis, 2.* Springer, Berlin.

Robbins, H. (1956). An empirical Bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954-1955, vol. I.* University of California Press: Berkeley and Los Angeles. 157–163.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Ann. Math. Statist.*, vol 35, 1-20.

Roberts, A. and Varberg, D. (1973). *Convex Functions.* Academic, New York.