DISSERTATION

# MODEL SELECTION BASED ON EXPECTED SQUARED HELLINGER DISTANCE

Submitted by

Xiaofan Cao

Department of Statistics

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2007

COLORADO STATE UNIVERSITY

October 10, 2007

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY XIAOFAN CAO ENTITLED MODEL SELECTION BASED ON EXPECTED SQUARED HELLINGER DISTANCE BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Jan Hannig

William S. Duff

Hariharan K. Iyer (Adviser)

Haonan Wang (Co-Adviser)

F. Jay Breidt (Department Head)

ABSTRACT OF DISSERTATION

# MODEL SELECTION BASED ON EXPECTED SQUARED HELLINGER DISTANCE

This dissertation is motivated by a general model selection problem such that the true model is unknown and one or more approximating parametric families of models are given along with strategies for estimating the parameters using data. We develop model selection methods based on Hellinger distance that can be applied to a wide range of modeling problems without posing the typical assumptions for the true model to be within the approximating families or to come from a particular parametric family. We propose two estimators for the expected squared Hellinger distance as the model selection criteria.

In particular, the use of expected squared Hellinger distance is studied in ANOVA model selection problems where approximating models are typically submodels of the full factorial model. The properties of the expected squared Hellinger distance are explored under balanced model structure assuming independent and identically distributed normal error terms. A model selection strategy specific to ANOVA model selection problems based on one of the estimated expected squared Hellinger distance is proposed. This strategy is illustrated using a real data set and its performance is tested by simulation studies. An example of ANOVA model selection problem with non-normal error terms that follow two-parameter exponential distribution is discussed.

Model selection method based on estimated expected squared Hellinger distance is also applied to modeling the $p$-values from the microarray data analysis. The problem of estimating false discovery rate (FDR) from the distribution of $p$-values arising from statistical tests of differential gene expression in a microarray experiment is considered. A finite mixture model is studied in which one component is uniform on [0,1] corresponding to equally expressed genes and one or more additional components correspond to differentially expressed genes. Two different mixture families are explicitly investigated for estimating false discovery rate – a mixture of Beta densities and a mixture of Uniform densities. In both cases, the Minimum Hellinger distance is used to provide robust estimates of the mixture components. For the Beta mixture model we choose the number of Beta components by comparing the estimated expected squared Hellinger distance. The performance of the proposed methods is illustrated through a case study involving data from a published microarray experiment.

Xiaofan Cao
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Fall 2007

## ACKNOWLEDGEMENTS

I would like to first express my deepest gratitude to Dr. Hari Iyer for guiding me through my doctorate process with his knowledge, enthusiasm, and perspective. I consider myself very lucky to have had Dr. Iyer as my adviser, teacher and mentor. What I learned from him in and out of classroom has inspired me and will continue to be my lifelong treasure.

I would like to sincerely thank my co-adviser, Dr. Wang Haonan. Dr. Wang has offered his instruction and ideas whenever needed. It would have been very different without his generous help.

I would also like to thank Dr. Jan Hannig and Dr. William Duff for serving on my committee and providing valuable suggestions on this dissertation. Dr. Ann Hess provided me data and computation instructions in estimating false discovery rate in microarray data analysis. Dr. Yi-Ching Yao also offered important comments and insights on the dissertation.

Special thanks go to my parents. I would not have gone this far without their full support.

DEDICATION

To my family.

CONTENTS

# Chapter 1

## INTRODUCTION TO MODEL SELECTION

Model Selection is a process that a statistician would face routinely, such as deciding which of several candidate distributions fits the data well or which subset of the variables should be used as predictors in a linear regression problem. There is extensive literature on model selection. In their book titled "Model Selection" (Linhart and Zucchini, 1986), *Linhart and Zucchini* discussed general methods of model selection in different situations and provided detailed explanations and examples. *Burnham and Anderson* (2002) also gave a thorough review on latest developments in model selection in their book "Model Selection and Multi-Model Inference". *McQuarrie and Tsai* (1998) discusses model selection techniques for univariate and multivariate regression models, univariate and multivariate autoregressive models, nonparametric (including wavelets) and semi-parametric regression models, and quasi-likelihood and robust regression models in their book "Regression and time series model selection".

A model is an abstract mathematical representation of a process. We assume that there is a true model governing a process. If one is asked to select among competing candidate models the "best" one that represents a given process, one would naturally hope to choose the model that is the most similar to or the least different from the true model. This motivated researchers to consider approaches - among

which there's no ultimate "best" one - to measure the similarity or discrepancy between the true model and a candidate model.

The idea described above, model selection based on **Discrepancies**, is the one that *Linhart and Zucchini* focused on in their book (Linhart and Zucchini, 1986) and is also the one that we will explore in detail in the following chapters. This strategy, according to *Linhart and Zucchini*, selects the model which is estimated to be the "most appropriate" in the circumstances, namely, the background assumptions, the sample size, and the specific requirements of the user. To be more specific, the candidate model family which minimizes the expected discrepancy is selected. It is not necessary to assume that this family contains the true model. There are many discrepancy measures to choose from, among which are Kullback-Leibler discrepancy (Kullback, 1959), Hellinger distance (also known as Matusita's distance; Matusita, 1955, LeCam, 1970, and Beran, 1977), Kolmogorov-Smirnov discrepancy (Darling, 1957), Cramer-von Mises discrepancy, Pearson chi-squared and Neyman chi-squared discrepancies (for discrete data or grouped data), and Gauss discrepancy.

Kullback-Leibler discrepancy is a commonly-used and well-explored one, which often leads to simple criteria. For instance, it can be shown that the AIC criterion, originally proposed by Akaike (Akaike, 1973), is related to an estimate of the expected Kullback-Leibler discrepancy. AIC consists of two terms, a log-likelihood term and a penalty term penalizing on the number of parameters in the approximating model that need to be estimated. Many competitors and variants of AIC were introduced since the 70's. AICc (Hurvich and Tsai, 1989) and AICb (Cavanaugh and Shumway, 1997) were proposed to decrease the bias in small-sample applications. TIC (Takeuchi, 1976) is basically AIC but with a different penalty term, which does

not require the assumption that the true model is within the candidate model families. *Schwarz* (1978) proposed BIC and SIC based on a Bayesian approach. *Ishiguro and Sakamoto* (1991) proposed WIC and *Ishiguro et al* proposed EIC (1997), both of which are extensions to AIC utilizing bootstrap methods. Mallows $C_p$ method (Mallows, 1966) for model selection in regression problems may also be viewed as a discrepancy based model selection procedure.

Bootstrap methods (Efron, 1979), cross-validatory methods (Mosteller and Tukey, 1968), and other Monte Carlo methods are also receiving more attention in model selection lately. These resampling methods are often used in conjunction with model selection based on discrepancy and help circumvent the technical problems sometimes encountered in deriving the expected discrepancy and an estimator for it. The idea of using the bootstrap to improve the performance of a model selection rule had been suggested and investigated by *Efron* (1983). *Chung et al* (1996) explored the application of bootstrap methods on estimating expected Kullback-Leibler discrepancy. *Shao* (1996) applied bootstrap methods and cross-validation methods to model selection in linear regression scenario.

Another common approach to fitting a model is to "select the simplest approximating family which is not inconsistent with the data". This approach predates discrepancy based methods and is based on **Hypothesis Tests**. The application of this strategy in many situations has been thoroughly explored and experience has been accumulated. However, in practice, people tend to forget the assumption of this method, "Assuming that the selected family of models holds (and there is no evidence to suggest that it does not)" (Linhart and Zucchini, 1986). Model selection in linear regression is the area where practitioners are generally interested and much research has been done.

**Model Selection Based on Discrepancy**

We define a model selection problem as follows. Choose the one from the approximating (candidate) models that is the closest to the true model. The idea "the closest" is conveyed by the smallest discrepancy between the true model and the approximating model. Discrepancy is also referred to as "distance" in the literature. This term "distance" is somewhat misleading since many discrepancies are in fact not distances or metrics. We will follow *Linhart and Zucchini*'s definition of a discrepancy (Linhart and Zucchini, 1986).

Suppose that we have $n$ independent observations on $d$ variables and that each observation can be regarded as a realization of a $d$-dimensional random vector having distribution function $F$. Let $M$ be the set of all $d$-dimensional distribution functions. Each member of $M$ is a fully specified model. Let $G$ be an approximating model. Then a *discrepancy* between the true model $F$ and the approximating model $G$ is a functional, $\Delta$, on $M \times M$ which has the property

$$\Delta(F, G) \geq \Delta(F, F) \qquad \text{for} \quad G, F \in M,$$

with equality if and only if $G = F$ almost everywhere with respect to the Lebesgue measure. In many cases approximating models are given as a *family of models*, $G_\theta$ with $\theta \in \Theta$, which is a subset of $M$ whose individual members are identified by the vector of parameters $\theta = (\theta^{(1)}, \ldots, \theta^{(p)})^T$. We will use $\Delta(\theta)$ to represent $\Delta(F, G_\theta)$. A *fitted model*, $G_{\hat{\theta}}$, is a member of a family of models $G_\theta$, $\theta \in \Theta$, which is associated with a prescribed estimator of the parameters using the observations. The *overall discrepancy* is defined as $\Delta(F, G_{\hat{\theta}})$.

Since $\Delta(F, G_{\hat{\theta}})$ is a random variable due to the randomness of $\hat{\theta}$, it is the expected overall discrepancy, $E_F \Delta(F, G_{\hat{\theta}})$, that one wishes to compare among the

candidates. In practice, the true distribution $F$ is unknown and $E_F\Delta(F, G_{\hat{\theta}})$ is thus unknown. We are interested in finding an estimate for the true expected discrepancy. The estimation for the expected discrepancy is not necessarily straightforward. One may need to resort to bootstrap or cross-validation methods. Asymptotic methods is another option when the properties of $\hat{\theta}$ are known, which leads to *criteria* that are estimates of $E_F\Delta(F, G_{\hat{\theta}})$ and are usually easy to compute.

Several questions need to be answered before solving a model selection problem. They are:

a. What are the approximating models.

b. What is the estimating method that we should use when the parameters of the approximating models are not fully specified.

c. What discrepancy should we use.

In this dissertation, we will assume that the answers to the first two questions are given. That is, the approximating models are given with prespecified estimation methods. Some examples of the these estimation methods are maximum likelihood method, minimum discrepancy method, least squares method, and method of moments. Some of the discrepancy measures that we can consider in the last question are listed below:

***Kullback-Leibler Discrepancy***

$$\Delta_{K-L}(\theta) = E_F \log \frac{f(x)}{g_\theta(x)}$$

***Gauss Discrepancy***

$$\text{Discrete} : \Delta_G(\theta) = \Sigma_x (f(x) - g_\theta(x))^2$$

$$\text{Continuous} : \Delta_G(\theta) = \int_x (f(x) - g_\theta(x))^2 dx$$

***Kolmogorov-Smirnov Distance***

$$\Delta_{K-S}(\theta) = \sup_x |F(x) - G_\theta(x)|$$

Among the above mentioned discrepancies, Kolmogorov-Smirnov distance is a metric, in other words, it has the following properties:

**(M1)** $\Delta(F, G) \geq 0$ and $\Delta(F, G) = 0$ if and only if $G = F$ *a.e.* (*non − negativity*),

**(M2)** $\Delta(F, G) = \Delta(G, F)$ (*symmetry*),

**(M3)** $\Delta(F, G) \leq \Delta(F, H) + \Delta(H, G)$ (*triangle inequality*).

Different discrepancies define similarity/dissimilarity from different angles. Choosing different discrepancies may lead to different results and there is no right or wrong to that. The motivation behind our answer to the last question is based on a review of the limitation of the Kullback-Leibler (K-L) discrepancy, one of the most commonly used discrepancies.

The K-L discrepancy is related to log likelihood ratio and Fisher information, which makes it not only easy to be estimated but also favorable when it comes to derivation of asymptotic properties. The easy calculation of its estimate was quite appealing, especially before the recent development in computer sciences and Monte Carlo methods.

Note, however, that the K-L discrepancy (although often referred to as the K-L distance) is not a metric. Moreover, the K-L discrepancy is defined for two probability measures that dominate each other and $\infty$ otherwise. That is, the K-L

discrepancy between two distributions that has different supports will be infinity. Thus, the K-L discrepancy is good at discrimination, but not necessarily a good tool for approximation. For instance, the K-L discrepancy will discriminate two distributions instantly if the approximating distribution $G$ has a different support than that of the true distribution $F$, no matter how closely shaped they are. As shown in Example 1, this may not necessarily be a favorable property.

***Example 1*** Let the true distribution $f$ be an Exponential distribution with the scale parameter $\gamma$ being 1. Suppose we have two candidate models: $g^{(1)}$ is two parameter Exponential distribution with the location parameter $\eta$ being 0.1 and scale parameter $\gamma$ being 1; $g^{(2)}$ is Exponential distribution with scale parameter being 2. Now, the K-L discrepancy between $f$ and $g^{(2)}$ is found to be 0.1931 and the K-L discrepancy between $f$ and $g^{(1)}$ is $\infty$ according to the definition. Therefore, if we choose the "best" model based on minimum K-L discrepancy, we would choose $g^{(2)}$. However, Figure 1.1 featuring the three density functions shows that $g^{(2)}$ actually has more similarity in shape with $f$.

Unlike the K-L discrepancy, the Hellinger distance is indeed a distance and is defined for densities that have different supports. In this dissertation, we choose to use the **Hellinger Distance**. The Hellinger distance has been applied to density estimation by many researchers. Relatively fewer attention has been paid to the application of the Hellinger distance in model selection. Mandal (2006) considered distinguishing between competing models based on pairwise Hellinger distance in experimental design with applications of global optimization. Birge (2004) considers model selection for Gaussian regression with random design and uses the Hellinger distance between two Gaussian distributions as the loss function instead of the typical loss function of the squared $\mathcal{L}_2$ distance. In this dissertation, however, we

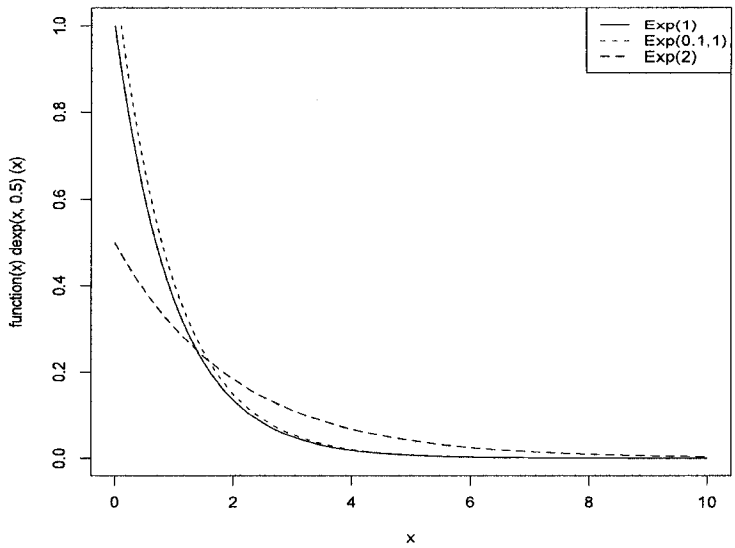


Figure 1.1: Three Density Curves: Exp(1), Exp(0.1,1), and Exp(1/2)

develop model selection methods based on expected squared Hellinger distance for general model selection problems. We do not assume any parametric distribution for the true model and the approximating models can come from any parametric families with any prespecified parameter estimation strategies.

This dissertation is organized as follows. In Chapter 2, we will introduce the Hellinger distance as the discrepancy that we choose to use in model selection and study the properties of the expected squared Hellinger distance. Also in this chapter, two estimators of the expected squared Hellinger distance are introduced and their large sample property is studied. In Chapter 3, simple illustrative examples of model selection using the estimated expected squared Hellinger distance are provided and the large sample property of the estimators is examined by simulation study. In Chapter 4, we study the model selection problem for balanced factorial ANOVA models and develop a model selection strategy based on theoretical considerations. The properties of the proposed strategy are examined using statistical

simulation. The method is also illustrated using real data. In Chapter 5, we consider the estimation of false discovery rate in Microarray data analysis, where the estimated expected squared Hellinger distance is used to select the best Beta mixture model to approximate the distribution of the $p$-values. Some concluding remarks and thoughts on future research direction are provided in Chapter 6.

## Chapter 2

# HELLINGER DISTANCE AND EXPECTED SQUARED HELLINGER DISTANCE

The Hellinger distance is also known as the Matusita's distance (Matusita, 1955, LeCam, 1970, and Beran, 1977). According to LeCam (1970), let $\mathcal{P}$ and $\mathcal{Q}$ be two probability measures on a $\sigma$-field $\alpha$. Let $\mu = \mathcal{P} + \mathcal{Q}$ and let $f$ and $g$ be the densities $f = d\mathcal{P}/d\mu$ and $g = d\mathcal{Q}/d\mu$. The Hellinger distance $H(\mathcal{P}, \mathcal{Q})$ is the square root of the squared Hellinger distance defined by

$$H^2(\mathcal{P}, \mathcal{Q}) = \int \left((dP)^{\frac{1}{2}} - (d\mathcal{Q})^{\frac{1}{2}}\right)^2 = \int (f^{\frac{1}{2}} - g^{\frac{1}{2}})^2 d\mu = 2[1 - \rho(\mathcal{P}, \mathcal{Q})], \qquad (2.1)$$

where

$$\rho(\mathcal{P}, \mathcal{Q}) = \int (d\mathcal{P} d\mathcal{Q})^{\frac{1}{2}} = \int \sqrt{fg} d\mu \qquad (2.2)$$

is also called the Hellinger affinity between $\mathcal{P}$ and $\mathcal{Q}$. Note that $0 \leq \int \sqrt{fg} d\mu \leq 1$ with equality on the left if and only if $f$ and $g$ are mutually singular and equality on the right if and only if $f$ and $g$ assign the same probability to each measurable set (Kraft 1955). That is, $0 \leq H^2(\mathcal{P}, \mathcal{Q}) \leq 2$, with the equality on the left if and only if $\mathcal{P} = \mathcal{Q}$ except for some set $A$ such that $\mu(A) = 0$ and the equality on the right if and only if $\mathcal{P}$ and $\mathcal{Q}$ are disjoint. The squared Hellinger distance between two probability distributions with density functions $f$ and $g$ can also be written as

$$H^2(f, g) = \|f^{\frac{1}{2}} - g^{\frac{1}{2}}\|^2 = \int (f^{\frac{1}{2}}(t) - g^{\frac{1}{2}}(t))^2 dt, \qquad (2.3)$$

where $\|\cdot\|$ denotes the $\mathcal{L}_2$ norm (Beran, 1977). We will use this form in this thesis. The Hellinger affinity can also be written as $< f^{\frac{1}{2}}, g^{\frac{1}{2}} >$, where $< \cdot >$ denotes the inner product.

The Hellinger distance is also defined in some literature (LeCam, 1973) as

$$H^2(P,Q) = \frac{1}{2}\int |(d\mathcal{P})^{\frac{1}{2}} - (d\mathcal{Q})^{\frac{1}{2}}|^2 = [1 - \rho(\mathcal{P},\mathcal{Q})],$$

so that the distance is now a value between 0 and 1. Some author also refer to the squared Hellinger distance as the Hellinger distance (Lu, Hui, and Lee, 2003). For the purpose of minimization, this makes no difference. The Hellinger distance is indeed a metric. The nonnegativity and symmetry of the Hellinger distance is straight forward from the nature of the $\mathcal{L}_2$ norm and the distance is 0 if and only if the two distributions are equal almost everywhere $\mu$. The triangle inequality of the Hellinger distance also follows from the property of the norm. Let $f$, $g$, and $h$ be 3 probability density functions,

$$\begin{aligned} H(f,g) &= \|f^{\frac{1}{2}} - g^{\frac{1}{2}}\| \\ &= \|(f^{\frac{1}{2}} - h^{\frac{1}{2}}) + (h^{\frac{1}{2}} - g^{\frac{1}{2}})\| \\ &\leq \|f^{\frac{1}{2}} - h^{\frac{1}{2}}\| + \|h^{\frac{1}{2}} - g^{\frac{1}{2}}\| \\ &= H(f,h) + H(h,g) \end{aligned}$$

The Hellinger distance is invariant under transformation. Let $X = \{X_{(1)}, \ldots, X_{(k)}\} \in A \subset \mathcal{R}^k$ be a $k$-dimensional random variable with probability density function $f$ and $Y\{Y_{(1)}, \ldots, Y_{(k)}\} \in B \subset \mathcal{R}^k$ with density $g$. Define a one-to-one continuous differentiable map function $T(\cdot) : \mathcal{R}^k \longmapsto \mathcal{R}^k$. Let

$$TX = \{t_1(X), \ldots, t_k(X)\} \quad \text{and} \quad TY = \{t_1(Y), \ldots, t_k(Y)\},$$

where $TX \in TA$ and $TY \in TB$. The distributions of the transformed variables are denoted as $f^\star$ and $g^\star$, respectively. Let $D_x = [t_{ij}(x)] = [\frac{\partial t_i}{\partial x_j}]$ and $D_y = [t_{ij}(y)] = [\frac{\partial t_i}{\partial y_j}]$ be the Jacobian matrices and let $J(x) = \det D_x$ and $J(y) = \det D_y$ be the Jacobian determinant. Suppose $J(x) \neq 0$ and $J(y) \neq 0$. Now, by Theorem 17.2 (Billingsley, 1995),

$$\begin{aligned} H(f,g) &= 2 - 2\int_{A\cap B} \sqrt{f(u)g(u)}du \\ &= 2 - 2\int_{A\cap B} \sqrt{f^*(Tu)\,|J(u)|\,g^*(Tu)\,|J(u)|}du \\ &= 2 - 2\int_{A\cap B} \sqrt{f^*(Tu)g^*(Tu)}\,|J(u)|\,du \\ &= 2 - 2\int_{TA\cap TB} \sqrt{f^*(t)g^*(t)}dt \\ &= H(f^*,g^*) \end{aligned}$$

Note that $\sqrt{f^*g^*}$ is non-negative.

A lot of work has been done in parametric estimation using Hellinger distance, namely, finding minimum Hellinger distribution estimators. It can be seen as a special case of model selection. After all, finding the minimum Hellinger distance estimator for a parametric model is equivalent to finding the model $g_{\theta_0}$ among a given parametric model family $\{g_\theta, \theta \in \Theta\}$ that is closest to the true model in Hellinger space. Beran defined MHDE, the minimum Hellinger distance estimator $\hat{\theta}_n$, as follows

$$\hat{\theta}_n = \arg\min\{H(\hat{f}, g_\theta)\}$$

where $\hat{f}$ is a suitable nonparametric estimator of the true density (Beran, 1977).

Beran mentioned that the estimator MHDE is related heuristically to the maximum likelihood estimator of $\theta$ if the true density is in fact some $g_{\theta_0}$, that is, if there's no misspecification. He proved that under certain regularity conditions

$$\hat{\theta}_n \xrightarrow{P} \theta_o$$

and found the limiting distribution of the MHDE. Beran made a comment on the efficiency of the MHDE: "The minimum Hellinger distance estimator may be regarded as a particular minimum distance estimator that is distinguished by being asymptotically efficient in regular models". Beran also studied the robustness property of MHDE. A lot of research has been done built on Baran's work. Some of these examples are: "Minimum Hellinger Distance Estimation for Finite Mixture Models" (Cutler and Cordero-Brana, 1996); "Minimum Hellinger Distance Estimation for Multivariate Location and Covariance" (Tamura and Boos, 1986); "Minimum Hellinger-Type Distance Estimation For Censored Data" (Ying, 1992); and etc.

## 2.1 Expected Squared Hellinger Distance

In a typical model selection problem, we are supposed to choose from approximating families rather than fully specified models with given parameters. In this case, model parameters need to be estimated from the data by some estimation method either prespecified or deemed appropriate for the problem. Thus, the Hellinger distance between the true distribution with density $f$ (operating model) and the approximating model with density $g$ and data-based estimator $\hat{\theta}$ becomes $H(f, g_{\hat{\theta}})$, where $\hat{\theta}$ is $p \times 1$ vector and $p \geq 1$. This distance is then a random variable due to $\hat{\theta}$. As pointed out by Linhart and Zucchini (1986), the distribution of the distance under the operating model determines the quality of a given procedure and "thus constitute the basis for comparing different fitting procedures". Instead of estimating the complete distribution of the distance, which is not always possible, one can estimate some characteristic of it such as the expectation. For the purpose of comparing among different approximating models, it satisfies to use the expected squared Hellinger distance in place of the expected Hellinger distance.

As the dimension of the approximating family increases, the approximation gets better while the estimation error increases. The distance between the true distribution and the approximating one based on any given data set does not penalize on the approximating dimension and thus is possible to be smaller for the approximating model that has more parameters than necessary. The expected distance, on the other hand, does penalize on the increased number of parameters that need to be estimated and thus balances between the approximation and estimation error.

Define the expected squared Hellinger distance as:

$$EH2 = E_{\hat{\theta}} H^2(f, g_{\hat{\theta}}) = E_{\hat{\theta}} \int \left( \sqrt{f(t)} - \sqrt{g_{\hat{\theta}}(t)} \right)^2 dt \tag{2.4}$$

where $\hat{\theta}$ is under the true distribution $F$. After the expectation, $EH2$ is a real number between 0 and 2 that depends on sample size $n$ only. It's easy to see that $EH2$ is also invariant under transformation. Note that $EH2$ can be written as:

$$EH2 = 2 - 2E\left[ \int \sqrt{f(t) g_{\hat{\theta}}(t)} dt \right] = 2 - 2 \int \sqrt{f(t)} E\left[ \sqrt{g_{\hat{\theta}}(t)} \right] dt, \tag{2.5}$$

given that the expectation and the integral exist. Note that in our discussion, the expectation $E[\cdot]$ is with respect to the true distribution $F$ unless otherwise specified. Define the model that has the smallest $EH2$ as the *true best model*, which is the model that is expected to be the "closest" in Hellinger metric to the true model.

### 2.1.1 Decomposition and Approximation of $EH2$

Now, we will study one way of decomposing and approximating $EH2$ as defined in equation (2.4). Assume that $\hat{\theta} - \theta_0 \xrightarrow{P} 0$. Then the squared Hellinger distance $H^2(f, g_{\hat{\theta}})$ between the true distribution $f$ and the approximating distribution $g_{\hat{\theta}}$ can be written as:

$$H^2(f, g_{\hat{\theta}}) = H^2(f, g_{\theta_0}) + H^2(g_{\theta_0}, g_{\hat{\theta}}) + 2 \int \left( \sqrt{f(t)} - \sqrt{g_{\theta_0}(t)} \right) \left( \sqrt{g_{\theta_0}(t)} - \sqrt{g_{\hat{\theta}}(t)} \right) dt \tag{2.6}$$

In this section we will show, under regularity conditions given in Assumptions A1-A2, that the second term $H^2(g_{\theta_0}, g_{\hat{\theta}})$ is $O_P(\frac{1}{n})$ and the last term

$$\int \left(\sqrt{f(t)} - \sqrt{g_{\theta_0}(t)}\right)\left(\sqrt{g_{\theta_0}(t)} - \sqrt{g_{\hat{\theta}}(t)}\right) dt$$

is $O_P(\frac{1}{\sqrt{n}})$. This is formally stated in Theorem 1.

Let $\underline{X} = \{X_1, X_2, ..., X_n\}$ be a data set of size $n$ where $X_i$'s are i.i.d. random variables with distribution function $F$. Let $g_{\hat{\theta}}$ be the approximating model where $\hat{\theta}$ is a data-based choice for the unknown parameter $\theta = (\theta^{(1)}, \ldots, \theta^{(p)}) \in \Theta$. Suppose that $g_\theta(t)$ has a second derivative at each point of an open set $S$ in $\mathcal{R}^p$ for every $t$ and denote, for $\theta^* \in \Theta$ and $t \in T$:

$$g'_{\theta^*}(t) = \left(\frac{\partial g_\theta(t)}{\partial \theta^{(1)}}, \ldots, \frac{\partial g_\theta(t)}{\partial \theta^{(p)}}\right)^T |_{\theta=\theta^*}$$

$$\dot{g}_{\theta^*}(t) = \left(\frac{\partial \sqrt{g_\theta(t)}}{\partial \theta^{(1)}}, \ldots, \frac{\partial \sqrt{g_\theta(t)}}{\partial \theta^{(p)}}\right)^T |_{\theta=\theta^*}$$

$$\ddot{g}_{\theta^*}(t) = \left(\frac{\partial^2 \sqrt{g_\theta(t)}}{\partial \theta^{(i)} \partial \theta^{(j)}}\right) |_{\theta=\theta^*}, i, j = 1, 2, \ldots, p.$$

Let $g'^{(i)}_{\theta^*}(t)$ denote the $i^{th}$ element of $g'_{\theta^*}(t)$, $\dot{g}^{(i)}_{\theta^*}(t)$ denote the $i^{th}$ element of $\dot{g}_{\theta^*}(t)$, and $\ddot{g}^{(i,j)}_{\theta^*}(t)$ denote the $(i,j)^{th}$ element of $\ddot{g}_{\theta^*}(t)$.

For any given $t \in T$, expanding $\sqrt{g_{\hat{\theta}}(t)}$ around $\theta_0$ gives:

$$\sqrt{g_{\hat{\theta}}(t)} = \sqrt{g_{\theta_0}(t)} + (\hat{\theta} - \theta_0)^T \dot{g}_{\theta_0}(t) + R = \sqrt{g_{\theta_0}(t)} + \sum_i^p (\hat{\theta}^{(i)} - \theta_0^{(i)}) \dot{g}^{(i)}_{\theta_0}(t) + R$$

The remainder term $R$ is

$$R = (1/2)(\hat{\theta} - \theta_0)^T \ddot{g}_{\bar{\theta}}(t)(\hat{\theta} - \theta_0) = (1/2) \sum_i^p \sum_j^p \ddot{g}^{(i,j)}_{\bar{\theta}}(t)(\hat{\theta}^{(i)} - \theta_0^{(i)})(\hat{\theta}^{(j)} - \theta_0^{(j)})$$

where $\bar{\theta}$ is a point on the line segment joining $\hat{\theta}$ and $\theta_0$ (Apostol, 1974) for all $t$.

Moreover, define the following quantities:

$$\mathrm{B}_1 = (\hat{\theta} - \theta_0)^T \left[ \int \dot{g}_{\theta_0}(t) \dot{g}_{\theta_0}^T(t) dt \right] (\hat{\theta} - \theta_0)$$

$$\mathrm{B}_2 = (\hat{\theta} - \theta_0)^T \left[ \int \dot{g}_{\theta_0}(t) (\hat{\theta} - \theta_0)^T \ddot{g}_{\bar{\theta}}(t) dt \right] (\hat{\theta} - \theta_0)$$

$$\mathrm{B}_3 = (\hat{\theta} - \theta_0)^T \left[ \int \ddot{g}_{\bar{\theta}}(t) (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T \ddot{g}_{\bar{\theta}}(t) dt \right] (\hat{\theta} - \theta_0)$$

$$\mathrm{C}_1 = (\hat{\theta} - \theta_0)^T \int \left( \sqrt{f(t)} - \sqrt{g_{\theta_0}(t)} \right) \dot{g}_{\theta_0}(t) dt$$

$$\mathrm{C}_2 = (\hat{\theta} - \theta_0)^T \left[ \int \left( \sqrt{f(t)} - \sqrt{g_{\theta_0}(t)} \right) \ddot{g}_{\bar{\theta}}(t) dt \right] (\hat{\theta} - \theta_0)$$

We use the convention that the integral of a matrix is the matrix of the integrals of the elements. The following Assumptions A1 through A2 will be used in our discussion:

- (A1) Assume $\sqrt{n}(\hat{\theta} - \theta_0)$ converges in distribution to a real random variable $Y$.

- (A2) Assume $|\ddot{g}_{\theta}^{(i,j)}(t)| \leq M_1(t)$ for all $t$ if $\theta \in \mathcal{O}_{\delta'}(\theta_0)$ or $\theta \in \{\theta : |\theta - \theta_0| < \delta'\}$, where $M_1(t)$ is a finitely integrable function, $i, j = 1, \ldots, p$. Assume further that $M_1$ belongs to $\mathcal{L}_2$-space and $\dot{g}_{\theta_0}^{(i)} \dot{g}_{\theta_0}^{(j)}$ is finitely integrable for $i, j = 1, \ldots, p$.

**Theorem 1.** *Assuming Assumptions A1 and A2,*

$$H^2(g_{\theta_0}, g_{\hat{\theta}}) = O_p(\frac{1}{n})$$

*and,*

$$\int \left( \sqrt{f(t)} - \sqrt{g_{\theta_0}(t)} \right) \left( \sqrt{g_{\theta_0}(t)} - \sqrt{g_{\hat{\theta}}(t)} \right) dt = O_p(\frac{1}{\sqrt{n}}).$$

The proof of Theorem 1 follows Lemmas 1 and 2.

**Lemma 1.** *Assuming that Assumptions A1 and A2 hold, then (i),* $\mathrm{B}_1 = O_p(n^{-1})$*; (ii),* $\mathrm{C}_1 = O_p(n^{-1/2})$.

*Proof.* (i). From Assumption A1,

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{D} Y,$$

where $Y$ is a random variable. Hence, $\sqrt{n}(\hat{\theta}-\theta_0)$ is bounded in probability, i.e.,

$$\hat{\theta}-\theta_0 = O_p(\frac{1}{\sqrt{n}}).$$

Moreover, since function $h(X) = X^TAX$ is a continuous function with $A$ being a square constant matrix,

$$\left(\sqrt{n}(\hat{\theta}-\theta_0)^T\right)\mathbf{A}\left(\sqrt{n}(\hat{\theta}-\theta_0)\right) \xrightarrow{D} Y^TAY$$

where $\mathbf{A} = \int \dot{g}_{\theta_0}(t)\dot{g}^T_{\theta_0}(t)dt$ is a finite square matrix that does not depend on data or $n$ by Assumption A2, thus

$$\mathrm{B}_1 = (\hat{\theta}-\theta_0)^T\mathbf{A}(\hat{\theta}-\theta_0) = O_p(\frac{1}{n}).$$

(ii). Note that by the Cauchy-Schwarz inequality,

$$\left(\int\left(\sqrt{f(t)}-\sqrt{g_{\theta_0}(t)}\right)\dot{g}_{\theta_0}(t)dt\right)^2 \leq \|\sqrt{f}-\sqrt{g_{\theta_0}}\|^2\|\dot{g}_{\theta_0}\|^2 \leq 2\|\dot{g}_{\theta_0}\|^2$$

where $\|\dot{g}_{\theta_0}\|^2$ is finite by Assumption A2. And thus $\int\left(\sqrt{f(t)}-\sqrt{g_{\theta_0}(t)}\right)\dot{g}_{\theta_0}(t)dt$ is finite. Since

$$\sqrt{n}(\hat{\theta}-\theta)^T\int\left(\sqrt{f(t)}-\sqrt{g_{\theta_0}(t)}\right)\dot{g}_{\theta_0}(t)dt \xrightarrow{D} Y^T\int\left(\sqrt{f(t)}-\sqrt{g_{\theta_0}(t)}\right)\dot{g}_{\theta_0}(t)dt,$$

$$\mathrm{C}_1 = \left(\hat{\theta}-\theta_0\right)^T\int\left(\sqrt{f(t)}-\sqrt{g_{\theta_0}(t)}\right)\dot{g}_{\theta_0}(t)dt = O_p(\frac{1}{\sqrt{n}}).$$

□

**Lemma 2.** *Suppose that Assumptions A1 and A2 hold, then (i),* $\mathrm{C}_2 = O_p(n^{-1})$*; (ii),* $\mathrm{B}_2 = O_p(n^{-3/2})$*; (iii),* $\mathrm{B}_3 = O_p(n^{-2})$ .

*Proof.* (i). Note from the proof of Lemma 1, we have by Assumption A1 that

$$\hat{\theta} - \theta_0 = \frac{1}{\sqrt{n}} O_p(1),$$

which implies that $(\hat{\theta} - \theta_0)$ is $o_p(1)$ and thus $\hat{\theta}$ converges to $\theta_0$ in probability. Fix a $\delta$ such that $0 < \delta \leq \delta'$ and for any given $\epsilon > 0$, there exists an $N_\epsilon$ such that for any $n > N_\epsilon$,

$$P\left[|\hat{\theta} - \theta_0| < \delta\right] > 1 - \epsilon.$$

Note further that since $\bar{\theta}$ is a point on the line segment joining $\hat{\theta}$ and $\theta_0$ for each $t$,

$$|\hat{\theta} - \theta_0| = |\hat{\theta} - \bar{\theta}| + |\bar{\theta} - \theta_0| \geq |\bar{\theta} - \theta_0|.$$

Thus for all $t$,

$$P\left[|\bar{\theta} - \theta_0| < \delta\right] \geq P\left[|\hat{\theta} - \theta_0| < \delta\right] > 1 - \epsilon.$$

for all $n > N_\epsilon$. Define

$$\mathcal{A} = \{\omega : |\bar{\theta} - \theta_0| < \delta\}; \qquad \mathcal{B} = \{\omega : |\ddot{g}_{\bar{\theta}}^{(i,j)}(t)| \leq M_1(t), \quad \forall t\}$$

where $i, j = 1, \ldots, p$. By Assumption A2, $\mathcal{A} \subseteq \mathcal{B}$. Therefore, for any $n > N_\epsilon$

$$P[\mathcal{B}] \geq P[\mathcal{A}] > 1 - \epsilon.$$

Let

$$\mathcal{C} = \{\|\ddot{g}_{\bar{\theta}}^{(i,j)}\|^2 \leq \|M_1\|^2\}$$

where $i, j = 1, \ldots, p$. Then $\mathcal{B} \subseteq \mathcal{C}$ since if $\left(\ddot{g}_{\bar{\theta}}^{(i,j)}(t)\right)^2 \leq M_1^2(t)$ for all $t$, $\int \left(\ddot{g}_{\bar{\theta}}^{(i,j)}(t)\right)^2 dt \leq \int M_1^2(t)dt$ where $\int M_1^2(t)dt$ is assumed to be finite by Assumption A2. This in turn implies that for any $\epsilon > 0$, choose an $M$ such that $M = \|M_1\|^2$ and

$$P\left[\|\ddot{g}_{\bar{\theta}}^{(i,j)}\|^2 \leq M\right] \geq P[\mathcal{B}] > 1 - \epsilon.$$

for all $n > N_\epsilon$. That is, $\|\ddot{g}_{\bar{\theta}}^{(i,j)}\|^2$ is bounded in probability for all $i, j = 1, \ldots, p$.

Moreover, by the Cauchy-Schwarz inequality,

$$\left(\int \left(\sqrt{f(t)} - \sqrt{g_{\theta_0}(t)}\right) \ddot{g}_{\bar{\theta}}^{(i,j)}(t)dt\right)^2 \leq \|\sqrt{f} - \sqrt{g_{\theta_0}}\|^2 \|\ddot{g}_{\bar{\theta}}^{(i,j)}\|^2 \leq 2\|\ddot{g}_{\bar{\theta}}^{(i,j)}\|^2$$

Therefore, $\int \left(\sqrt{f(t)} - \sqrt{g_{\theta_0}(t)}\right) \ddot{g}_{\bar{\theta}}^{(i,j)}(t)dt$ is also bounded in probability.

Secondly, $n(\hat{\theta}-\theta_0)^{(i)}(\hat{\theta}-\theta_0)^{(j)}$ converges to $Y^{(i)}Y^{(j)}$ in distribution by continuous mapping, and is thus bounded in probability. Therefore,

$$\begin{aligned}
\mathrm{C}_2 &= (\hat{\theta} - \theta_0)^T \left[\int \left(\sqrt{f(t)} - \sqrt{g_{\theta_0}(t)}\right) \ddot{g}_{\bar{\theta}}(t)dt\right] (\hat{\theta} - \theta_0) \\
&= \sum_{i,j} (\hat{\theta} - \theta_0)^{(i)}(\hat{\theta} - \theta_0)^{(j)} \left[\int \left(\sqrt{f(t)} - \sqrt{g_{\theta_0}(t)}\right) \ddot{g}_{\bar{\theta}}^{(i,j)}(t)dt\right] \\
&= \left(\frac{2p}{n}\right) O_p(1)O_p(1) \\
&= O_p(\frac{1}{n})
\end{aligned}$$

(ii). We show in the proof of part (i) that by Assumptions A1 and A2, $\|\ddot{g}_{\bar{\theta}}^{(j,k)}\|^2 = O_p(1)$ for $j, k = 1, \ldots, p$. Moreover,

$$\left(\int \dot{g}_{\theta_0}^{(i)}(t)\ddot{g}_{\bar{\theta}}^{(j,k)}(t)dt\right)^2 \leq \|\dot{g}_{\theta_0}^{(i)}\|^2 \|\ddot{g}_{\bar{\theta}}^{(i,j)}\|^2$$

where $\|\dot{g}_{\theta_0}^{(i)}\|^2$ is finite by Assumption A2. Therefore, $\int \dot{g}_{\theta_0}^{(i)}(t)\ddot{g}_{\bar{\theta}}^{(j,k)}(t)dt$ is also bounded in probability. We also have from Assumption A1 and continuous mapping that for $i, j, k = 1, \ldots, p$,

$$n^{\frac{3}{2}}(\hat{\theta} - \theta_0)^{(i)}(\hat{\theta} - \theta_0)^{(j)}(\hat{\theta} - \theta_0)^{(k)} \xrightarrow{D} Y^{(i)}Y^{(j)}Y^{(k)}.$$

Thus,

$$\begin{aligned}
\mathrm{B}_2 &= (\hat{\theta} - \theta_0)^T \left[\int \dot{g}_{\theta_0}(t)(\hat{\theta} - \theta_0)^T \ddot{g}_{\bar{\theta}}(t)dt\right] (\hat{\theta} - \theta_0) \\
&= \sum_{i,j,k} (\hat{\theta} - \theta_0)^{(i)}(\hat{\theta} - \theta_0)^{(j)}(\hat{\theta} - \theta_0)^{(k)} \int \dot{g}_{\theta_0}^{(i)}(t)\ddot{g}_{\bar{\theta}}^{(j,k)}(t)dt \\
&= (3p)(n^{-\frac{3}{2}})O_p(1)O_p(1) \\
&= O_p(n^{-\frac{3}{2}})
\end{aligned}$$

(iii). As shown in the proof of part (i), $\|\ddot{g}_{\tilde{\theta}}^{(i,j)}\|^2 = O_p(1)$ by Assumptions A1 and A2. It implies that

$$\int \ddot{g}_{\theta_0}^{(i,j)}(t)\ddot{g}_{\tilde{\theta}}^{(k,l)}(t)dt = O_p(1)$$

where $i, j, k, l = 1, \ldots, p$, since

$$\left(\int \ddot{g}_{\tilde{\theta}}^{(i,j)}(t)\ddot{g}_{\tilde{\theta}}^{(k,l)}(t)dt\right)^2 \leq \|\ddot{g}_{\tilde{\theta}}^{(i,j)}\|^2 \|\ddot{g}_{\tilde{\theta}}^{(k,l)}\|^2$$

By Assumption A1,

$$n^2(\hat{\theta}-\theta_0)^{(i)}(\hat{\theta}-\theta_0)^{(j)}(\hat{\theta}-\theta_0)^{(k)}(\hat{\theta}-\theta_0)^{(l)} \xrightarrow{D} Y^{(i)}Y^{(j)}Y^{(k)}Y^{(l)}$$

where $i, j, k, l = 1, \ldots, p$. Thus,

$$\begin{aligned}
\mathrm{B}_3 &= (\hat{\theta}-\theta_0)^T \left[\int \ddot{g}_{\tilde{\theta}}(t)(\hat{\theta}-\theta_0)(\hat{\theta}-\theta_0)^T \ddot{g}_{\tilde{\theta}}^T(t)dt\right](\hat{\theta}-\theta_0) \\
&= \sum_{i,j,k,l}(\hat{\theta}-\theta_0)^{(i)}(\hat{\theta}-\theta_0)^{(j)}(\hat{\theta}-\theta_0)^{(k)}(\hat{\theta}-\theta_0)^{(l)} \int \ddot{g}_{\tilde{\theta}}^{(i,j)}(t)\ddot{g}_{\tilde{\theta}}^{(k,l)}(t)dt \\
&= (4p)(n^{-2})O_p(1)O_p(1) \\
&= O_p(n^{-2})
\end{aligned}$$

□

We now give the proof of Theorem 1.

*Proof.* Assuming A1-A2, based on Lemmas 1 and 2,

$$\begin{aligned}
H^2(g_{\theta_0}, g_{\hat{\theta}}) &= \int \left(\sqrt{g_{\theta_0}(t)} - \sqrt{g_{\hat{\theta}}(t)}\right)^2 dt \\
&= \int \left((\hat{\theta}-\theta_0)^T \dot{g}_{\theta_0}(t) + \frac{1}{2}R\right)^2 dt \\
&= \mathrm{B}_1 + \mathrm{B}_2 + \frac{1}{4}\mathrm{B}_3 \\
&= O_p(\frac{1}{n}) + \frac{1}{\sqrt{n}}O_p(\frac{1}{n}) + \frac{1}{n}O_p(\frac{1}{n}) \\
&= O_p(\frac{1}{n})
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\int \left(\sqrt{f(t)} - \sqrt{g_{\theta_0}(t)}\right)\left(\sqrt{g_{\theta_0}(t)} - \sqrt{g_{\hat{\theta}}(t)}\right) dt \\
&= \int \left(\sqrt{f(t)} - \sqrt{g_{\theta_0}(t)}\right)\left((\hat{\theta} - \theta_0)^T \dot{g}_{\theta_0}(t) + R\right) dt \\
&= (\hat{\theta} - \theta_0)^T \int \left(\sqrt{f(t)} - \sqrt{g_{\theta_0}(t)}\right) \dot{g}_{\theta_0}(t) dt \\
&\quad + (\frac{1}{2})(\hat{\theta} - \theta_0)^T \left[\int \left(\sqrt{f(t)} - \sqrt{g_{\theta_0}(t)}\right) \ddot{g}_{\bar{\theta}}(t) dt\right] (\hat{\theta} - \theta_0) \\
&= \mathrm{C}_1 + \frac{1}{2}\mathrm{C}_2 \\
&= O_p(\frac{1}{\sqrt{n}}) + \frac{1}{\sqrt{n}} O_p(\frac{1}{\sqrt{n}}) \\
&= O_p(\frac{1}{\sqrt{n}})
\end{aligned}$$

This completes the proof. □

**Remark 1.** *In the case where $f = g_{\theta_0}$ except for the sets with Lebesgue measure 0, the first term $H^2(f, g_{\theta_0})$ and the cross product term on the right hand side of Equation (2.6) are both equal to zero. Otherwise, the second term $H^2(g_{\theta_0}, g_{\hat{\theta}})$ and the cross product term vanish as sample size increases while the first term becomes dominant. In either case, the cross product term is either zero or almost zero when $n$ is large enough.*

So $EH2$ can be approximated by:

$$EH2 \simeq H^2(f, g_{\theta_0}) + E[H^2(g_{\theta_0}, g_{\hat{\theta}})] \tag{2.7}$$

The first term on the right hand side of Equation (2.7) can be thought of as representing the "approximation error", denoted as the *model error term*, while the second term as representing the expected "estimation error", denoted as the *penalty term*.

**Proposition 1.** *Assuming Assumptions A1-A2, assuming further that* $E(YY^T) = \Sigma$, *then*

$$nE\left[H^2(g_{\theta_0}, g_{\hat{\theta}})\right] \longrightarrow \int \dot{g}_{\theta_0}^T(t)\Sigma\dot{g}_{\theta_0}(t)dt$$

*as* $n$ *goes to infinity.*

*Proof.* By Lemmas 1 and 2,

$$n\left(H(g_{\theta_0}, g_{\hat{\theta}}) - \mathrm{B}_1\right) = o_p(1)$$

Therefore,

$$nE\left[H(g_{\theta_0}, g_{\hat{\theta}})\right] - E\left[n\mathrm{B}_1\right] \longrightarrow 0.$$

Moreover, by Assumption A1 and the continuous mapping,

$$n\mathrm{B}_1 = n(\hat{\theta} - \theta_0)^T A(\hat{\theta} - \theta_0) \xrightarrow{D} Y^T\mathbf{A}Y$$

where $\mathbf{A} = \int \dot{g}_{\theta_0}(t)\dot{g}_{\theta_0}^T(t)dt$ is a finite square matrix by Assumption A2. Then

$$E\left[n\mathrm{B}_1\right] = E\left[n(\hat{\theta} - \theta_0)^T\mathbf{A}(\hat{\theta} - \theta_0)\right] \longrightarrow E(Y^TAY).$$

Thus,

$$\begin{aligned}
E(Y\mathbf{A}Y^T) &= E\left[Y^T\left(\int \dot{g}_{\theta_0}(t)\dot{g}_{\theta_0}^T(t)dt\right)Y\right] \\
&= trace\left\{E\left[Y^T\left(\int \dot{g}_{\theta_0}(t)\dot{g}_{\theta_0}^T(t)dt\right)Y\right]\right\} \\
&= \int trace\left[\dot{g}_{\theta_0}^T(t)E(YY^T)\dot{g}_{\theta_0}(t)\right]dt \\
&= \int \dot{g}_{\theta_0}^T(t)\Sigma\dot{g}_{\theta_0}(t)dt
\end{aligned}$$

This completes the proof. □

## 2.2 Estimation of EH2

One of the tasks we are faced with in estimating the expected squared Hellinger distance $EH2$ is to find an estimator of the unknown true distribution $f$. Unless otherwise stated, we choose to use kernel density estimator, a nonparametric estimator. A kernel density estimator is given as

$$\hat{f}(x) = (nh_n)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right) \tag{2.8}$$

Here, $K$ is a function satisfying $\int K(x)dx = 1$, which we call the *kernel*, and $h_n$ is a positive number depending on $n$, usually called the *bandwidth* or *window width*. A slightly more compact formula for the kernel estimator can be obtained by introducing the re-scaling notation $K_{h_n}(u) = h_n^{-1}K(u/h_n)$. Equation (2.8) can be also written as the following

$$\hat{f}(x) = n^{-1} \sum_{i=1}^{n} K_{h_n}(x - X_i) = (nh_n)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$$

Usually $K$ is chosen to be a unimodal probability density function that is symmetric about zero. This ensures that $\hat{f}$ is itself also a density.

The estimated squared Hellinger distance $H^2(\hat{f}, g_{\hat{\theta}})$ is then given by

$$H^2(\hat{f}, g_{\hat{\theta}}) = \int (\hat{f}^{\frac{1}{2}}(t) - g_{\hat{\theta}}^{\frac{1}{2}}(t))^2 dt$$

The task remains in finding the distribution of $g_{\hat{\theta}}$ and the expectation of the squared Hellinger distance. When a close form solution can not be derived, this can be approximated by bootstrap method. In the following subsections, we will introduce two estimators: $BEEH2$ as an estimator for $EH2$ in the form of equation 2.5 and $PEEH2$ as an estimator of approximated $EH2$ in the form of equation 2.7.

### 2.2.1 BEEH2

In this section, we will introduce an estimator, $BEEH2$, of the $EH2$ in the form of equation 2.5. Let $\underline{X} = \{X_1, X_2, ..., X_n\}$ be a data set of size $n$ where $X_i$'s are i.i.d. random variables that follow unknown distribution function $F$ and $X \in \mathcal{X} \subseteq R^k$. Let $m$ be a fraction of $n$ and $\underline{X}^* = \{X_1^*, X_2^*, ..., X_m^*\}$ where $X_i^*$'s are i.i.d. random variables that follow $F_n(\cdot|\underline{x})$, which is the empirical distribution based on $\underline{x}$, such that:

$$F_n(t|\underline{X} = \underline{x}) = Pr\{X \leq t|\underline{X} = \underline{x}\} = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,t]}(x_i). \tag{2.9}$$

Let $\hat{\theta}(\cdot)$ be an estimation function for the parameter $\theta$ corresponding to an approximating density $g_\theta$. Denote $\hat{\theta} = \hat{\theta}(\underline{X})$, and $\hat{\theta}^* = \hat{\theta}(\underline{X}^*)$. Then the underline distribution for $\hat{\theta}$ is $F$ while that for $\hat{\theta}^*$ is $F_n$ given $\underline{x}$. While the true density $f$ is estimated by kernel density estimator $\hat{f}$, a natural estimator of $E_{\hat{\theta}}(\sqrt{g_{\hat{\theta}}(t)})$ for all $t$ is the bootstrap mean

$$E_*\left(\sqrt{g_{\hat{\theta}^*}(t)}\right) = \int \cdots \int \sqrt{g_{\hat{\theta}(u_1,...,u_m)}(t)} dFn(u_1) \ldots dFn(u_m) = \frac{1}{n^m}\sum_{*} \sqrt{g_{\hat{\theta}^*}(t)},$$

where $E_*[\cdot]$ denotes the expectation with respect to the joint empirical distribution function and $\sum_*(\cdot)$ represents the summation over all possible bootstrap samples of size $m$, $\underline{x}^*$ given $\underline{x}$. Note that $E_*\left(\sqrt{g_{\hat{\theta}^*}(t)}\right)$ is a random variable due to the randomness of $\underline{X}$. Theoretically, $E_*(\sqrt{g_{\hat{\theta}}(t)})$ for all $t$ can be calculated exactly by enumerating the $n^m$ possible samples of size $m$ from $F_n$. We therefore propose an estimator $EEH2$, of $EH2$:

$$EEH2 = E_*\left[H^2(\hat{f}, g_{\hat{\theta}^*})\right] = 2 - 2\int \sqrt{\hat{f}(t)} E_* \sqrt{g_{\hat{\theta}}(t)} dt$$

where $\hat{f}$ is a kernel density estimator. $EEH2$ depends on data and thus is a random variable. In our following discussion, we will consider the case where $m = n$ unless

otherwise specified. In situations where $n$ is large, we can take $m < n$ for efficient computation.

However, enumerating all possible samples under $F_n$ is not realistic except when $n$ is considerably small. One will have to resort to generating a large number, say, $M$, of bootstrap samples of size $n$ under $F_n$. Then the bootstrap samples can be obtained efficiently by re-sampling with replacement from the original data. For each bootstrap sample, $\underline{x}^*$, one can find $\hat{\theta}^*$ and the corresponding value $\sqrt{g_{\hat{\theta}^*}(t)}$ for all $t$. Then $E_*(\sqrt{g_{\hat{\theta}}(t)})$ for all $t$ can be approximated by the average $\frac{1}{M}\sum_{l=1}^{M}\left(\sqrt{g_{\hat{\theta}^{*(l)}}(t)}\right)$ where $M$ is the number of bootstrap samples generated and $\hat{\theta}^{*(l)}$ is the estimated parameter from the $l^{\text{th}}$ bootstrap sample.

We thus propose the *Bootstrap Estimated Expected Squared Hellinger Distance*, $BEEH2$ based on Bootstrap methods as follows

$$BEEH2 = 2 - 2\left\{\int \sqrt{\hat{f}(t)}\left[\frac{1}{M}\sum_{l=1}^{M}\left(\sqrt{g_{\hat{\theta}^{*(l)}}(t)}\right)\right]dt\right\} \tag{2.10}$$

### 2.2.2 Penalty Term Estimated Expected Hellinger Distance, $PEEH2$

Analogous to $EEH2$, we expect the distribution of $(\sqrt{g_{\hat{\theta}}(t)} - \sqrt{g_{\hat{\theta}^*}(t)})$, where $\hat{\theta}^* = \hat{\theta}(\underline{X}^*)$ is the estimator of the parameter from bootstrap sample under empirical distribution $F_n$, approximate that of $(\sqrt{g_{\theta_0}(t)} - \sqrt{g_{\hat{\theta}}(t)})$ and thus propose a natural estimator of the approximated $EH2$ in the form of Equations (2.7), $EEH2B$, as:

$$EEH2B = H^2(\hat{f}, g_{\hat{\theta}}) + E_*\left[H^2(g_{\hat{\theta}}, g_{\hat{\theta}^*})\right] = H^2(\hat{f}, g_{\hat{\theta}}) + \frac{1}{n^m}\sum_{*} H^2(g_{\hat{\theta}}, g_{\hat{\theta}^*})$$

where $\hat{f}$ is the kernel density estimator, $E_*[\cdot]$ and $\sum_*(\cdot)$ are as defined in Section 2.2.1. Note that $m = n$ unless otherwise specified. Again, since it is not practical to exhaust all the possible bootstrap samples, we propose an approximation of

*EEH2B*, *Penalty term Estimated Expected Squared Hellinger distance* or *PEEH2*, as follows:

$$PEEH2 = H^2(\hat{f}, g_{\hat{\theta}}) + \frac{1}{M}\sum_{l=1}^{M} H^2(g_{\hat{\theta}}, g_{\hat{\theta}^{*(l)}}) \tag{2.11}$$

where $\hat{\theta}^{*(l)}$ is based on the $l^{th}$ bootstrap sample taken by re-sampling with replacement from the data, $M$ is the number of bootstrap samples. The first and second term of Equation (2.11) estimate the model error term and penalty term in Equation (2.7), respectively.

### 2.2.3 Consistency of the estimators *BEEH2* and *PEEH2*

In this section, we will discuss the consistency property of the estimators along with the conditions required. We will first introduce a set of assumptions D1-D4:

- (D1) $K$ is nonnegative Borel measurable function on $R^d$ with $\int K(x)dx = 1$.
- (D2) $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d = \infty$.
- (D3) Assume $\sqrt{g_{\theta_0}}$ and $\sqrt{f}$ are finitely integrable.
- (D4) Assume that for any $\delta > 0$, $P\left[\sup_t |E_*\sqrt{g}_{\hat{\theta}^*}(t) - \sqrt{g}_{\theta_0}(t)| \geq \delta\right] \to 0$ as $n \to \infty$.

**Lemma 3.** *Assuming Assumptions D1 and D2,*

$$H(f, \hat{f}) \xrightarrow{P} 0.$$

*Proof.*

$$\int \left|\hat{f}(t) - f(t)\right| dt \xrightarrow{P} 0$$

as $n \to \infty$ under Assumptions D1 and D2 (Devroye, 1983). Therefore,

$$H(f, \hat{f}) \xrightarrow{P} 0$$

due to Steerneman (1982), $H(f, \hat{f}) \leq \int |\hat{f}(t) - f(t)| dt$. □

To facilitate the following proof, define $\widehat{EH2} = E_* \left[H^2(f, g_{\hat{\theta}^*})\right]$.

**Lemma 4.** *Assuming Assumptions D3 and D4,*

$$E_* H^2(g_{\theta_0}, g_{\hat{\theta}^*}) \xrightarrow{P} 0,$$

*and*

$$\widehat{EH2} \xrightarrow{P} H^2(f, g_{\theta_0}),$$

*Proof.* Observe that

$$\begin{aligned} \left|E_* H^2(g_{\theta_0}, g_{\hat{\theta}^*}) - H^2(g_{\theta_0}, g_{\theta_0})\right| &= 2\left|\int \sqrt{g_{\theta_0}(t)} \left(E_* \sqrt{g_{\hat{\theta}^*}(t)} - \sqrt{g_{\theta_0}(t)}\right) dt\right| \\ &\leq 2 \int \sqrt{g_{\theta_0}(t)} \left|E_* \sqrt{g_{\hat{\theta}^*}(t)} - \sqrt{g_{\theta_0}(t)}\right| dt \end{aligned}$$

By Assumption D4, for any $\delta$ and $\epsilon$, $\exists\ N_{\delta,\epsilon}$ such that

$$P\left[\sup_t \left|E_* \sqrt{g_{\hat{\theta}^*}(t)} - \sqrt{g_{\theta_0}(t)}\right| < \delta\right] > 1 - \epsilon$$

for any $n > N_{\delta,\epsilon}$. Let

$$\mathcal{A} = \left\{\omega : \sup_t \left|E_* \sqrt{g_{\hat{\theta}^*}(t)} - \sqrt{g_{\theta_0}(t)}\right| < \delta\right\}$$

and

$$\mathcal{B} = \left\{\omega : \int \sqrt{g_{\theta_0}(t)} \left|E_* \sqrt{g_{\hat{\theta}^*}(t)} - \sqrt{g_{\theta_0}(t)}\right| dt < \delta \int \sqrt{g_{\theta_0}(t)} dt\right\}.$$

Note that $\int \sqrt{g_{\theta_0}(t)}dt$ is finite by assumption D3. Then $\mathcal{A} \subseteq \mathcal{B}$. Thus, let $\xi = \delta \int \sqrt{g_{\theta_0}(t)}dt$, for any $\delta > 0$ and $\epsilon > 0$,

$$P\left[\int \sqrt{g_{\theta_0}(t)}\left|E_*\sqrt{g_{\hat{\theta}^*}(t)} - \sqrt{g_{\theta_0}(t)}\right| dt < \xi\right] > 1 - \epsilon$$

for all $n > N_{\delta,\epsilon}$. Therefore,

$$E_*\left[H^2(g_{\hat{\theta}^*}, g_{\theta_0})\right] = E_*\left[H^2(g_{\hat{\theta}^*}, g_{\theta_0})\right] - H^2(g_{\theta_0}, g_{\theta_0}) \xrightarrow{P} 0.$$

Similarly, Observe that

$$\begin{aligned}\left|\widehat{EH2} - H^2(f, g_{\theta_0})\right| &= \left|E_*\left[H^2(f, g_{\hat{\theta}^*})\right] - H^2(f, g_{\theta_0})\right| \\ &= \left|2\int \sqrt{f(t)}\left(E_*\sqrt{g_{\hat{\theta}^*}(t)} - \sqrt{g_{\theta_0}(t)}\right) dt\right| \\ &\leq 2\int \sqrt{f(t)}\left|E_*\sqrt{g_{\hat{\theta}^*}(t)} - \sqrt{g_{\theta_0}(t)}\right| dt\end{aligned}$$

let

$$\mathcal{C} = \left\{\omega : \int \sqrt{f(t)}\left|E_*\sqrt{g_{\hat{\theta}^*}(t)} - \sqrt{g_{\theta_0}(t)}\right| dt < \delta \int \sqrt{f(t)}dt\right\}.$$

Note that $\int \sqrt{f(t)}dt$ is finite by assumption D3. Then $\mathcal{A} \subseteq \mathcal{C}$. Thus, let $\xi' = \delta \int \sqrt{f(t)}dt$, for any $\delta > 0$ and $\epsilon > 0$,

$$P\left[\int \sqrt{f(t)}\left|E_*\sqrt{g_{\hat{\theta}^*}(t)} - \sqrt{g_{\theta_0}(t)}\right| dt < \xi'\right] \geq P[\mathcal{A}] > 1 - \epsilon$$

for all $n > N_{\delta,\epsilon}$. Therefore,

$$\widehat{EH2} - H^2(f, g_{\theta_0}) \xrightarrow{P} 0.$$

□

**Lemma 5.** *Assuming Assumptions D1-D3, then*

$$\widehat{EH2} - EEH2 \xrightarrow{P} 0.$$

*Proof.* Observe that $|\widehat{EH2} - EEH2| \leq \int \left|\sqrt{\hat{f}(t)} - \sqrt{f(t)}\right| E_* \left[\sqrt{g_{\hat{\theta}^*}(t)}\right] dt$, and by Cauchy-Schwarz inequality,

$$\begin{aligned}\left[\int \left|\sqrt{\hat{f}(t)} - \sqrt{f(t)}\right| E_* \left[\sqrt{g_{\hat{\theta}^*}(t)}\right] dt\right]^2 &\leq H^2(f,\hat{f}) \int \left(E_* \left[\sqrt{g_{\hat{\theta}^*}(t)}\right]\right)^2 dt \\ &\leq H^2(f,\hat{f}) \int E_* \left[g_{\hat{\theta}^*}(t)\right] dt \\ &= H^2(f,\hat{f}) E_* \left[\int g_{\hat{\theta}^*}(t) dt\right] \\ &= H^2(f,\hat{f})\end{aligned}$$

It is shown in Lemma 3 that $H(f,\hat{f}) \xrightarrow{P} 0$ by Assumptions D1 and D2. By continuous mapping, $H^2(f,\hat{f}) \xrightarrow{P} 0$. Therefore, it follows that $\widehat{EH2} - EEH2 \xrightarrow{P} 0$.

□

**Theorem 2.** *Assuming Assumptions A1-A2 and D1-D4, then* $EEH2 - EH2 \xrightarrow{P} 0$ *as* $n$ *goes to* $\infty$.

*Proof.* Note that the Hellinger distance is a metric and thus by triangle inequality,

$$|H(f, g_{\hat{\theta}}) - H(f, g_{\theta_0})| \leq H(g_{\theta_0}, g_{\hat{\theta}})$$

By Theorem 1, $H(g_{\theta_0}, g_{\hat{\theta}}) = O_p(1/n)$ under Assumptions A1-A2. It is easy to show that $H(g_{\theta_0}, g_{\hat{\theta}}) = o_p(1)$ under the same set of conditions. Thus,

$$H(f, g_{\hat{\theta}}) \xrightarrow{P} H(f, g_{\theta_0}),$$

which means that the distribution of $H^2(f, g_{\hat{\theta}})$ degenerates to that of $H^2(f, g_{\theta_0})$ by continuous mapping theorem. By Theorem 8.8 (page 58, Lehmann, 1998),

$$EH2 = E\left[H^2(f, g_{\hat{\theta}})\right] \longrightarrow H^2(f, g_{\theta_0}).$$

By Lemma 4,

$$\widehat{EH2} = E_*\left[H^2(f, g_{\hat{\theta}^*})\right] \xrightarrow{P} H^2(f, g_{\theta_0})$$

under Assumptions D3 and D4. Therefore, $\widehat{EH2} - EH2 \xrightarrow{P} 0$. Moreover, $\widehat{EH2} - EEH2 \xrightarrow{P} 0$ by Lemma 5. Theorem 2 follows since $EEH2 - EH2 = (\widehat{EH2} - EH2) - (\widehat{EH2} - EEH2)$. □

**Lemma 6.** *Assuming Assumptions A1-A2 and D1-D3, then*

$$H(\hat{f}, g_{\hat{\theta}}) \xrightarrow{P} H(f, g_{\theta_0})$$

*as n goes to* $\infty$

*Proof.* By triangle inequality of the Hellinger distance, we have

$$H(\hat{f}, g_{\hat{\theta}}) \leq H(\hat{f}, g_{\theta_0}) + H(g_{\theta_0}, g_{\hat{\theta}}) \leq H(f, g_{\theta_0}) + H(f, \hat{f}) + H(g_{\theta_0}, g_{\hat{\theta}})$$

At the same time,

$$H(f, g_{\theta_0}) \leq H(f, g_{\hat{\theta}}) + H(g_{\theta_0}, g_{\hat{\theta}}) \leq H(\hat{f}, g_{\hat{\theta}}) + H(f, \hat{f}) + H(g_{\theta_0}, g_{\hat{\theta}})$$

Thus, by Theorem 1 and Lemma 3,

$$H(\hat{f}, g_{\hat{\theta}}) \xrightarrow{P} H(f, g_{\theta_0})$$

under Assumptions A1-A2 and D1-D4. □

**Theorem 3.** *Assuming Assumptions A1-A2 and D1-D4, then* $EEH2B - EH2 \xrightarrow{P} 0$ *as n goes to* $\infty$.

*Proof.* Recall that

$$EEH2B = H^2(\hat{f}, g_{\hat{\theta}}) + E_* H^2(g_{\hat{\theta}^*}, g_{\theta_0})$$

By Lemma 6,

$$H^2(\hat{f}, g_{\hat{\theta}}) \xrightarrow{P} H^2(f, g_{\theta_0}).$$

By Lemma 4,

$$E_* H^2(g_{\hat{\theta}^*}, g_{\theta_0}) \xrightarrow{P} 0.$$

Moreover, as shown in the proof of Theorem 2,

$$EH2 = E\left[H^2(f, g_{\hat{\theta}})\right] \longrightarrow H^2(f, g_{\theta_0}).$$

Therefore,

$$EEH2B - EH2 \xrightarrow{P} 0$$

□

# Chapter 3

## MODEL SELECTION BASED ON $EH2$ - ILLUSTRATIVE EXAMPLES

The problem that we are interested in is a general one: competing approximating models are given with specified estimation methods, and we want to choose the one that is the "closest" to the true model based on a prespecified distance between distributions. We are required to make our decision based on an i.i.d. sample of size n, $\underline{X} = \{X_1, X_2, ..., X_n\}$, that comes from the true distribution. As discussed in Chapter 2, the distance we are interested in is the Hellinger distance. The approximating model that is the "closest" in terms of the Hellinger distance is the one that has the smallest $EH2$, which is termed "true best model". In general, the true best model can be estimated by the approximating model for which the estimated expected squared Hellinger distance, $BEEH2$ or $PEEH2$ as proposed in Chapter 2, is the smallest among all approximating models. In this chapter, we will study a few simple examples using $BEEH2$ as the model selection criterion. In chapter 4, we will propose a model selection strategy based on $PEEH2$ specifically for factorial ANOVA model selection problems, where often times the approximating models are "nested" or sub-models of other approximating models. In Chapter 5, we will study model selection problems in modeling $p$-values from Microarray data, using mixture distributions with various number of Beta components and applying $BEEH2$ to estimate the true best model.

### 3.1 Example 1: Density Modeling

One example of model selection is density estimation. An interesting case is when the competing approximating distributions are different only in the estimation methods, for instance, when the approximating models are $N(0, \sigma^2)$ with the estimators of $\sigma$ being $MHDE$ and $MLE$ respectively. The $BEEH2$'s can be calculated for the approximating distributions using each estimator based on the data and the one with the smallest $BEEH2$ will be chosen. As mentioned in Chapter 2 , the estimator MHDE is related heuristically to the maximum likelihood estimator of the parameter (vector) $\theta$ if the true density is in fact some $g_{\theta_0}$, that is, if there's no model misspecification. In this section, we consider one example to see if the MLE is the one that minimizes the expected squared Hellinger distance.

Suppose we have i.i.d. data set $\underline{X} = \{X_1, X_2, ..., X_n\}$ from some distribution $F$. Suppose we know the true distribution: $F = N(0, 1)$. Let the candidate model be $G = N(0, \sigma^2)$ where the variance is estimated by $k^2S^2$. Let $S^2 = \frac{\sum_i X_i^2}{n}$, the MLE of $\sigma^2$. We constrain ourselves to the cases where $n > 1$. We would like to find the $k$ that minimizes the expected squared Hellinger distance. Now, the squared Hellinger distance is:

$$\begin{aligned} H^2(f, g) &= 2 - 2\int \sqrt{f(x)g(x)}dx \\ &= 2 - 2\int \frac{1}{\sqrt{2\pi kS}} \exp\left\{-\frac{x^2}{4} - \frac{x^2}{4k^2S^2}\right\} dx \\ &= 2 - 2\sqrt{\frac{2kS}{(kS)^2 + 1}} \end{aligned}$$

The $k$ that minimizes the expected squared Hellinger distance, $k^* = \arg\min_k\{E[H^2(\phi_{0,1}, \phi_{0,k^2S^2})]\}$ where $\phi$ denotes the Normal density, can be approximated by numerical methods. Note that, $nS^2$ follows a $\chi^2$ distribution with $n$

degrees of freedom. Generate $m$ values $u$ from a $\chi^2(n)$ distribution and let $s_j^2 = \frac{u_j}{n}$, $j = 1, ..., m$. Let $k = \{k_1, \ldots, k_{1000}\}$ be a real value sequence from 0 to 2. Approximate $EH2$ for each $k_i$ using

$$E[H^2(\phi_{0,1}, \phi_{0,k^2S^2})] \simeq \frac{1}{m}\sum_{j=1}^{m}\left(2 - 2\frac{\sqrt{2k_is_j}}{\sqrt{(k_is_j)^2+1}}\right), \quad \text{for} \quad i = 1, ..., 1000$$

Then the $EH2$ is approximately minimized at

$$k_{i^*} = \arg\min_k\{\frac{1}{m}\sum_{j=1}^{m}(2 - 2\frac{\sqrt{2k_is_j}}{\sqrt{(k_is_j)^2+1}})\}.$$

A simulation with $n = 10$ and $m = 50$ shows that the $k$ is approximately 1.05. Figure 2 shows the approximated $EH2$ against $k$ and the vertical line locates where the the numerical minimization lies. We can see that, although close, MLE is not exactly the one that minimizes the expected squared Hellinger distance.

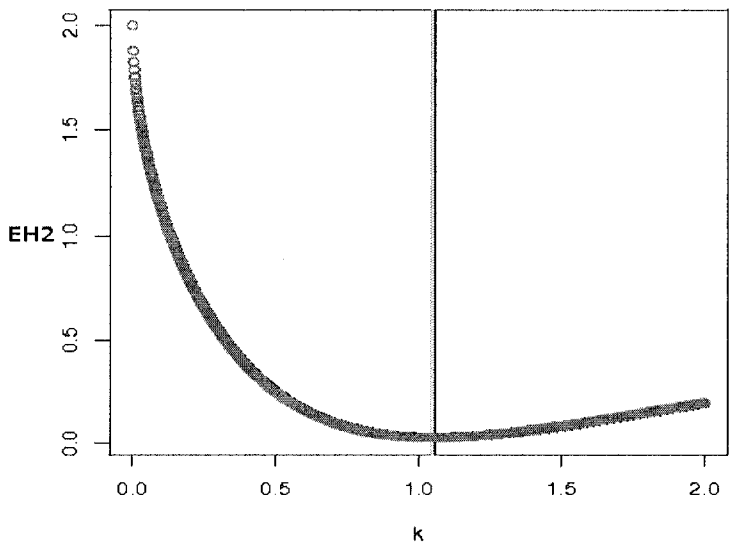


Figure 3.1: Example 1: Whether MLE Minimizes the $EH2$. *The true distribution is* $\Phi(0, 1)$ *and the approximating distribution is* $\Phi(0, k^2S^2)$. $EH2$ *is calculated based on* 50 *simulated values from* $\chi^2(10)$ *for every* $k$ *value from 0 to 2.*

**Example 2: Model Selection Among Two Approximating Families of Distributions**

A common model selection scenario is that for a given data set that comes from some unknown distribution, one needs to choose one from two (or more) competing approximating (candidate) distributions based on a prescribed criterion. Our selection strategy is to calculate $BEEH2$, for each distribution, from the data and choose the one with the smallest $BEEH2$. In this section, two simulated examples of such model selection problems will be discussed. In each of the two examples, $S = 1000$ data sets of size $n$=10, 30, 50, 80, and 150 are simulated from a true distribution (Lognormal or Exponential). Two approximating parametric families of models are given with specified estimators for the parameters. In both cases, kernel density estimators are fitted to the log-transformed data to avoid possible boundary problems near zero. $BEEH2$'s are calculated for the approximating models for every simulated data set. For each given sample size and each simulated data set, one of the models would be chosen based on the comparison of $BEEH2$'s, denoted as Choice.$BEEH2$. For a given sample size, the 1000 choices by $BEEH2$, Choice.$BEEH2$, are compared with the true best model given by $EH2$ ($EH2$ is calculated from 200 data sets simulated from the true distribution) and the success rate for a given $n$ (Suc.$BEEH2_n$) is the percentage of the times that the choices are the true best model:

$$\text{Suc.}BEEH2_n = \frac{\text{Number of Choice.}BEEH2 \text{ matching the true best model}}{S}$$

**Case 1: Data From Lognormal(0,1)**

In this example, data sets are generated from Lognormal(0,1) and the two approximating models are Gamma(2, $\beta$) and Weibull(1, $\lambda$) where the scale parameter

$\beta$ for Gamma is estimated by $\bar{X}/2$ and the scale parameter $\lambda$ for Weibull is estimated by sample mean $\bar{X}$. Note that when the shape is 1, Weibull distribution is simply the Exponential distribution with the same scale parameter. The number of Bootstrap samples $M$ for the calculation of $BEEH2$ is 1000. Figure 3.2 shows the density curves of lognormal(0, 1), Gamma(2,1.7/2), and Weibull(1,1.7), where 1.7 is the sample mean of a data set of size 150 generated from lognormal(0,1). In this example, $H^2(f, g_{\hat{\theta}})$ is also used as a criterion for comparison purposes. The success rates of choosing the true best model by both criteria are summarized in Table 3.2.

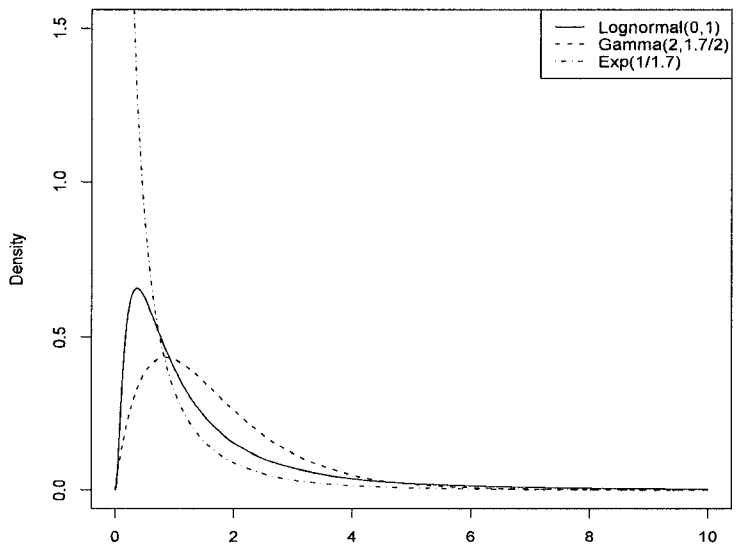


Figure 3.2: Three Density Curves: Lognormal(0,1), Gamma(2,1.7/2), and Weibull(1,1.7)

| *Distribution* | n=10 | n=30 | n=50 | n=80 | n=150 |
|---|---|---|---|---|---|
| Gamma(2, $\bar{X}/2$) | 0.1067 | 0.0754 | 0.0720 | 0.0682 | 0.0654 |
| Weibull(1, $\bar{X}$) | 0.0809 | 0.0554 | 0.0517 | 0.0493 | 0.0470 |

Table 3.1: $EH2$ Approximated from 200 Data Sets Generated from the true distribution Lognormal(0,1)

| *Estimator* | n=10 | n=30 | n=50 | n=80 | n=150 |
|---|---|---|---|---|---|
| $BEEH2$ | 0.563 | 0.686 | 0.738 | 0.801 | 0.859 |
| $H^2(\hat{f}, g_{\hat{\theta}})$ | 0.564 | 0.680 | 0.733 | 0.794 | 0.857 |

Table 3.2: Success Rates of Choosing The True Best Model Based on $BEEH2$ and $H^2(\hat{f}, g_{\hat{\theta}})$. *The true distribution is Lognormal(0,1), the two approximating families are Gamma(2, $\bar{X}/2$) and Weibull(1, $\bar{X}$).*

For all sample sizes, the true best model is the Weibull(1,$\bar{X}$) distribution (Table 3.1). We can see that the percentages of the time that the true best model is chosen increase as the sample size increases for both criteria. In this example, both $BEEH2$ and $H^2(\hat{f}, g_{\hat{\theta}})$ perform reasonably well. In more complicated problems, especially when many of the approximating models are sub-models of other approximating models, $H^2(\hat{f}, g_{\hat{\theta}})$ tends to favor larger models that contain the true model. More examples can be found in Chapter 4.

**Case 2: Data From Exponential(1)**

In this example, data sets are generated from Exp(1) and the two approximating families of models are Normal($\mu$, $\sigma^2$) and Lognormal($\mu_0$, $\sigma_0^2$). All the parameters are estimated by maximum likelihood method. The number of bootstrap samples $M$ for the calculation of $BEEH2$ is 1000. The density curves of Exp(1), Normal(1.03,0.99), and Lognormal(-0.50, 1.18) (the parameters of the approximating distributions are estimated from a random sample of size 150 from EXP(1)) is plotted in Figure 3.3. This is a case where one of the approximating distributions (Lognormal) is obviously closer to the true model than the other. Table 3.3 lists the $EH2$'s for both approximating distributions and all sample sizes. Not surprisingly, the percentage of choosing the true best model by $BEEH2$ is very high for even very small sample size and is 100% when sample size increases to 80. The success rates are summarized in Table 3.4.

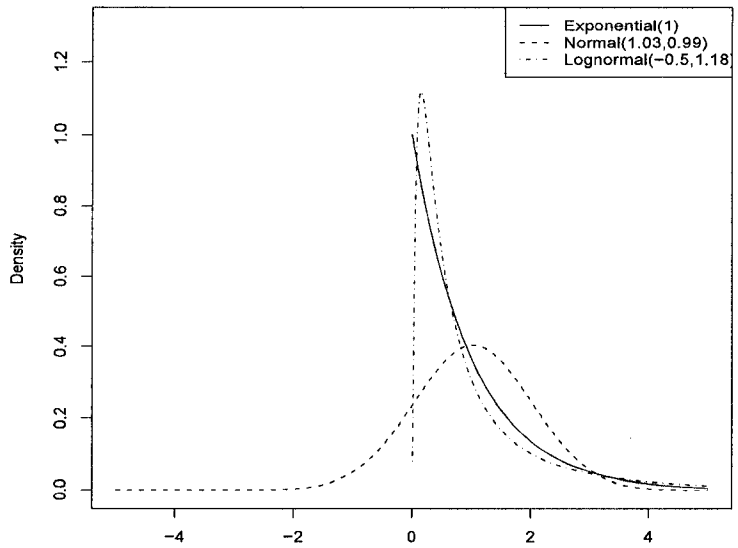


Figure 3.3: Three Density Curves: Exp(1), Normal(1.03,0.99), and Lognormal(-0.50, 1.18)

| *Distribution* | n=10 | n=30 | n=50 | n=80 | n=150 |
|---|---|---|---|---|---|
| Normal | 0.2797 | 0.2649 | 0.2603 | 0.2577 | 0.2585 |
| Lognormal | 0.0910 | 0.0694 | 0.0547 | 0.0520 | 0.0485 |

Table 3.3: $EH2$ Approximated from 200 Data Sets Generated from the true distribution Exp(1)

| *Estimator* | n=10 | n=30 | n=50 | n=80 | n=150 |
|---|---|---|---|---|---|
| $BEEH2$ | 0.943 | 0.997 | 0.999 | 1 | 1 |

Table 3.4: Success Rates of Choosing The True Best Model Based on $BEEH2$. *The true distribution is Exp(1), the two approximating families are Normal($\mu$, $\sigma^2$) and Lognormal($\mu_0$, $\sigma_0^2$), with the parameters being estimated by MLE.*

## 3.2 Simulation Example For Examining Convergence of $BEEH2$

In Chapter 2, we show that $EEH2 - EH2$ converges to zero in probability under some regularity conditions. In this section, we will check this result by a set of simulation examples. Since it not feasible to calculate $EEH2$ exactly, we will compute $BEEH2$ instead with the number of bootstrap samples being 200.

Let the true distribution be Lognormal(0,1) and the approximating distributions be Exponential, Gamma, Weibull, and Normal, respectively. We choose a series of sample sizes $n = 10, 20, ..., 500$. For each sample size, 100 data sets are generated from the true distribution. Thus, for each sample size, we have one $EH2$ and 100 $BEEH2$'s. These quantities are plotted against the sample sizes. Note that the bandwidths for fitting kernel density estimators are set to be depending on sample size $n$ only.

**Example 1: Lognormal(0,1) vs. Exp($\lambda$)**

Let the approximating distribution be Exp($\lambda$) with the rate parameter being estimated by $\hat{\lambda} = \frac{1}{\bar{X}}$.

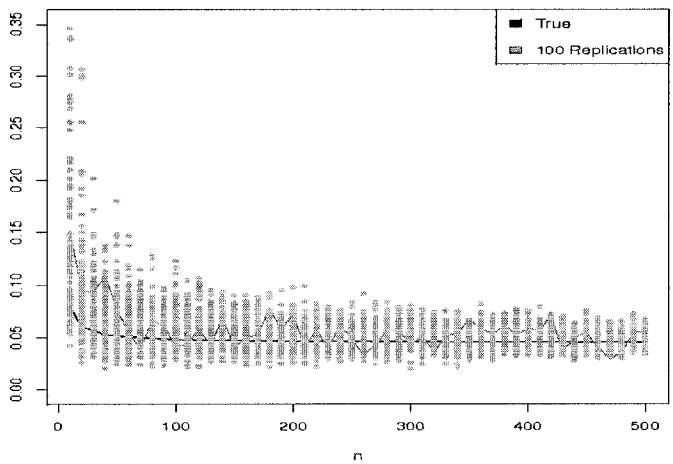


Figure 3.4: Convergence of $BEEH2$ to $EH2$. *The true distribution is Lognormal(0,1) and the approximating family is Exp($1/\bar{X}$); Sample size* $n = 10, 20, \ldots, 500$*; $BEEH2$ values are calculated for 100 data sets generated from the true distribution at each sample size.*

Figures 3.4 features $EH2$ and 100 realizations of $BEEH2$ and Figure 3.5 plots the mean and one standard deviation bounds of the $BEEH2$ values for sample sizes $n = 10, 20, ..., 500$. These figures show that $BEEH2$ values are oscillating around

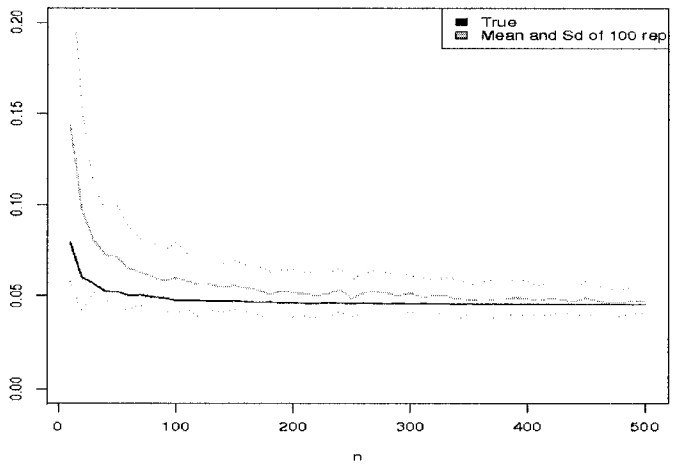


Figure 3.5: The Mean and 1 Standard Deviation Bounds of $BEEH2$ Values. *The true distribution is Lognormal(0,1) and the approximating family is Exp($1/\bar{X}$); Sample size* $n = 10, 20, \ldots, 500$*; $BEEH2$ values are calculated for 100 data sets generated from the true distribution at each sample size.*

the $EH2$ value and the spread of the $BEEH2$ values is getting smaller as sample size increases. In fact, the 1 standard deviation interval contains the true $EH2$ even at small samples sizes and the width of the interval within the bounds shrinks as the sample size increases. The mean of the $BEEH2$ values is also getting closer to $EH2$ as sample size increases.

**Example 2: Lognormal(0,1) vs. Gamma(2, $\beta$)**

Let the approximating distribution be Gamma(2, $\beta$) with $\hat{\beta} = \frac{\bar{X}}{2}$.

We can see from Figures 3.6 and 3.7 that the patterns are similar to those in the first example.

**Example 3: Lognormal(0,1) vs. Normal($\mu$,1)**

Let the approximating distribution be Normal($\mu$,1) with $\hat{\mu} = \bar{X}$.

Figures 3.8 and 3.9 show that the results are similar to those in the above examples, although the true $EH$ values are higher since a Normal distribution is

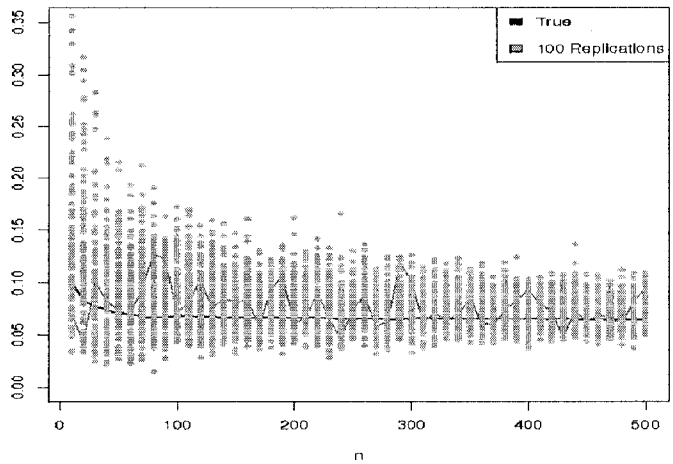


Figure 3.6: Convergence of $BEEH2$ to $EH2$. *The true distribution is Lognormal(0,1) and the approximating family is Gamma(2, $\bar{X}/2$); Sample size* $n = 10, 20, \ldots, 500$*; $BEEH2$ values are calculated for 100 data sets generated from the true distribution at each sample size.*

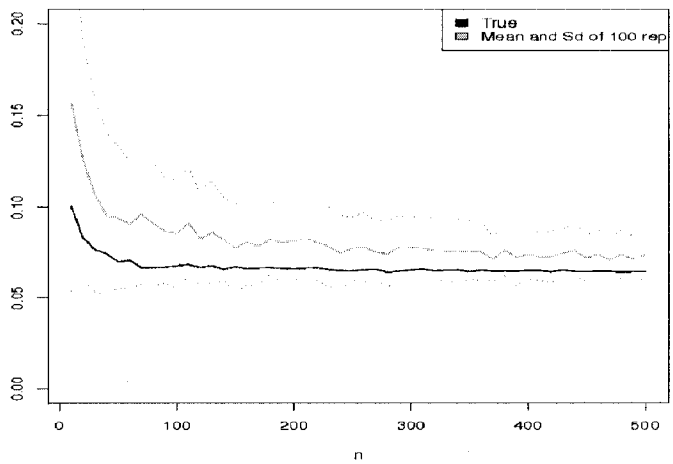


Figure 3.7: The Mean and 1 Standard Deviation Bounds of $BEEH2$ Values. *The true distribution is Lognormal(0,1) and the approximating family is Gamma(2, $\bar{X}/2$); Sample size* $n = 10, 20, \ldots, 500$*; $BEEH2$ values are calculated for 100 data sets generated from the true distribution at each sample size.*

obviously further away from the true distribution as compared to other approximating families.

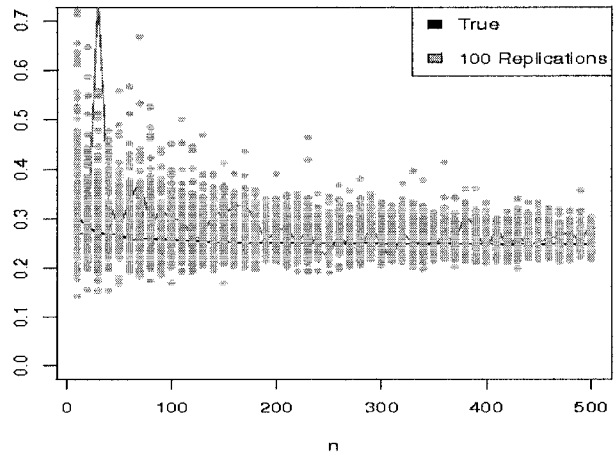


Figure 3.8: Convergence of $BEEH2$ to $EH2$. *The true distribution is Lognormal(0,1) and the approximating family is Normal($\bar{X}$, 1); Sample size* $n = 10, 20, \ldots, 500$*; $BEEH2$ values are calculated for 100 data sets generated from the true distribution at each sample size.*

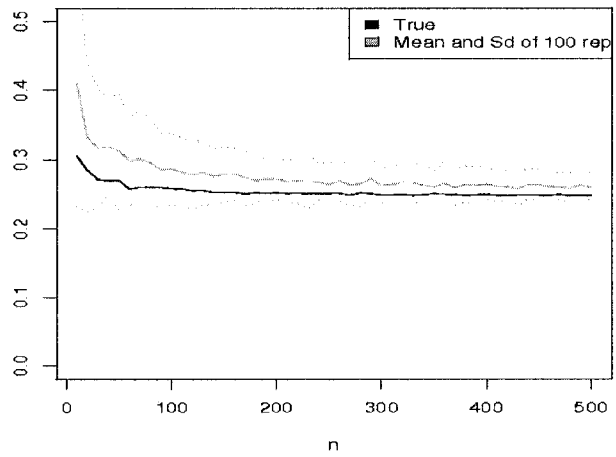


Figure 3.9: The Mean and 1 Standard Deviation Bounds of $BEEH2$ Values. *The true distribution is Lognormal(0,1) and the approximating family is Normal($\bar{X}$, 1); Sample size* $n = 10, 20, \ldots, 500$*; $BEEH2$ values are calculated for 100 data sets generated from the true distribution at each sample size.*

These simulation results confirm the convergence theorem in Chapter 2. Simulation studies on the convergence of $PEEH2 - EH2$ will be provided in Chapter 4.

# Chapter 4

# MODEL SELECTION BASED ON $EH2$ - APPLICATION IN ANOVA MODELS

## 4.1 Introduction

ANOVA model selection problems are particularly interesting in that the approximating models are often "nested" models, that is, some or all of the effects in one approximating model may be sub-models of other approximating model(s). The more effects a model has, typically, the better is the approximation. The decision as to whether a given effect should be included is often based on a test of the hypothesis that all levels of this effect are zero. More systematic procedures based on the $F$ test, such as Forward Selection, Backward Elimination, and Stepwise selection (Hocking, 1996) have been proposed to decide which effect(s) should be kept in the model. The forward selection method adds one variable at a time, stopping when it is determined that the remaining factors will not make a significant improvement in the model. The backward elimination method begins with the full model with all possible factors, and eliminates the factor that's considered to make the smallest contribution. The stepwise selection method is an improvement to the forward or backward method alone and a combination of these two. These methods were popular largely due to the fact that they involved little computation. These methods imply an order of importance on the variables that is generally meaningless. (Hocking, 1996).

Mallows (Mallows, 1966) proposed the statistic

$$C_p = \frac{RSS_k}{\hat{\sigma}^2} + 2p - N \tag{4.1}$$

for each subset of the combinations of the effects, where $RSS_k$ is the residual sum of squares associated with a subset model that has $k$ effects, $\hat{\sigma}^2$ is an estimator of $\sigma^2$ (usually the residual mean square for the full model), and $p = k + 1$, equal to the total number of parameters. The subset that minimizes $C_p$ is considered the best subset. He also pointed out that, those models with $C_p$-values that are approximately equal to the corresponding number of parameters usually have small prediction bias. Information based criteria such as AIC and AICc are also used.

The ANOVA models considered here are balanced fixed effects factorial models as defined in Hocking (1996). By **balanced** we mean that the numbers of observations in every cell are the same. In general, let $\mathcal{F} = \{F_1, F_2, \cdots F_k\}$ be the $k$ factors ($k$ denotes the number of factors) with $a_1, a_2, \ldots, a_k$ being the corresponding levels and $a_i > 1$ for $i = 1, \cdots, k$. Let $\Pi$ be the collection of $k$-tuples

$$\pi = (\pi_1, \pi_2, \ldots, \pi_k),$$

where $\pi_j \in \{1, 2, \ldots, a_j\}$, and $j = 1, \ldots, k$. Each element $\pi \in \Pi$ represents a particular combination of the levels of the $k$ factors which may be viewed as a cell in a $k$-way table corresponding to the $k$ factors. The response variable associated with $r^{th}$ replication in the cell $\pi$, denoted $Y_{\pi,r}$, can be expressed as

$$Y_{\pi,r} = \mu_\pi + \epsilon_{\pi,r}$$

where $\mu_\pi$ is the cell mean, and $r = 1, \cdots, n_{pi}$ (for balanced design, $n_1 = n_2 = \cdots = n_{pi} \equiv n$). The error term, $\epsilon_{\pi,r}$, is often assumed to be normally and independently

distributed with mean 0 and variance $\sigma^2$. Specially chosen contrasts of the cell means are referred to as **Factorial Effects**. Different ANOVA models considered here reflect different structures of the means, which can be expressed as combinations of some or all of the factorial effects. For simplicity, all the ANOVA models we consider in this chapter include the intercept.

Here, we propose model selection methods based on the *expected squared Hellinger distance* and apply these methods to ANOVA model selection problems. These methods are developed based on theoretical considerations under the assumptions that the true distribution is normal and the variance is known. This chapter is organized as follows. In Section 4.2, the properties of the expected squared Hellinger distance are studied and the model selection problem in ANOVA is discussed in detail. An estimator of the expected squared Hellinger distance, $PEEH2$, is proposed in Section 4.3, and the rationale behind the selection method based on $PEEH2$ is discussed. Section 4.4 displays the simulation results for checking the performance of this strategy in terms of choosing the "true best" model. In Section 5.4, this model selection method is applied to a real data problem for illustration. Section 4.7 deals with an example of a model selection problem in ANOVA models with error terms that follow two-parameter exponential distribution, in which case AIC and its variants can no longer be interpreted as the estimators of expected K-L discrepancy (since it does not exist).

## 4.2 ANOVA Model Selection with $EH2$

The selection among different ANOVA models can be interpreted as the selection among the approximating probability distributions. In a typical ANOVA

model selection problem, the approximating models are some or all possible factorial ANOVA models that include the full model and its sub-models. Each of the approximating model is associated with an approximating distribution for each cell $\pi$, namely, a normal distribution $N(\hat{\mu}_\pi, \hat{\sigma}^2)$ (density $\phi_{\hat{\mu}_\pi,\hat{\sigma}^2}$) with $\hat{\mu}_\pi$ being the corresponding mean structure estimated from data $\{Y_{\pi,1}, Y_{\pi,2}, \cdots, Y_{\pi,n}\}$ and $\hat{\sigma}^2$ being the MSE from the analysis of variance table. For the true distribution, we assume that the random variables $Y_{\pi,r}$'s in the cell $\pi$ are independently distributed with density function $f_{\theta_\pi}(x) = f(x - \theta_\pi)$, a family of densities with a location shift parameter. The Hellinger distance between the true model and the approximating model for cell $\pi$ is thus denoted as $H(f_{\theta_\pi}, \phi_{\hat{\mu}_\pi,\hat{\sigma}^2})$. The corresponding expected squared Hellinger distance, denoted as $EH2$, is given by

$$EH2(\pi) = E[H^2(f_{\theta_\pi}, \phi_{\hat{\mu}_\pi,\hat{\sigma}^2})]. \tag{4.2}$$

There are various ways to define the overall discrepancy between the true model and approximating model based on the cell-wise $EH2(\pi)$, e.g., the summation (equivalently, the average) and the product. Let $a = \prod_{i=1}^{k} a_i$ denote the number of all cells.

**Definition 1.** *We name* $\frac{1}{a}\sum_{\pi\in\Pi} EH2(\pi)$ *as the* overall expected squared Hellinger distance, *the average of the model error term in Equation (2.7) for all cells as the* overall model error, *and the average of the penalty term in Equation (2.7) for all cells as the* overall penalty.

### 4.2.1 An illustrative example of two-way ANOVA

In a two-way ANOVA model with factors $F_1$ and $F_2$, the response $Y_{(i,j),r}$, or simply denoted as $Y_{ijr}$, for the $r^{th}$ trial of the $i^{th}$ factor level in $F_1$ and the $j^{th}$ factor

level in $F_2$ can be stated as:

$$Y_{ijr} = \mu_{ij} + \epsilon_{ijr},$$

where $\mu_{ij}$ is referred to as the *cell mean*, while $\epsilon_{ijr}$ is the random error associated with $Y_{ijr}$. For this example, assume that $\epsilon_{ijr}$ is normally distributed for all approximating models. Let $\bar{\mu}_{..}$ denote the overall mean, $\bar{\mu}_{i.}$ denote the average of the cell means over all levels of factor $F_2$ for the $i^{th}$ level of factor $F_1$ and $\bar{\mu}_{.j}$ denote the average of the cell means over all levels of factor $F_1$ for the $j^{th}$ level of factor $F_2$.

The *Row effect*, *Column effect* and *Interaction* are defined as following

$$\begin{aligned} i-\text{th row effect}: &\quad \alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..} \\ j-\text{th column effect}: &\quad \beta_j = \bar{\mu}_{.j} - \bar{\mu}_{..} \\ (i,j)-\text{th interaction}: &\quad (\alpha\beta)_{ij} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..} \end{aligned}$$

Observing the convention that an interaction effect is in a model only when all corresponding main effects are also in the model, all possible two-way ANOVA models and their corresponding mean structures are shown in Table 4.1.

$$\begin{aligned} \text{Null model:} &\quad \mu_{ij} = \bar{\mu}_{..} \\ \text{Row Effect model:} &\quad \mu_{ij} = \bar{\mu}_{..} + \alpha_i \\ \text{Column Effect model:} &\quad \mu_{ij} = \bar{\mu}_{..} + \beta_j \\ \text{Main Effects model:} &\quad \mu_{ij} = \bar{\mu}_{..} + \alpha_i + \beta_j \\ \text{Full Model:} &\quad \mu_{ij} = \bar{\mu}_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} \end{aligned}$$

Table 4.1: Mean structures of all possible two-way ANOVA models

Assume that the data come from a column effect model with a known common variance $\sigma^2$ and that the levels of the two factors $a_1 = a_2 = 2$. And

$$Y_{i1r} \overset{i.i.d}{\sim} N(\mu_1, \sigma^2), \quad Y_{i2r} \overset{i.i.d}{\sim} N(\mu_2, \sigma^2), \qquad i = 1, 2 \quad r = 1, \ldots, n.$$

The limit (in probability sense) of the estimated cell means for cell $(1, 1)$ under five approximating models are illustrated in Table 4.2, assuming the true model is the column model. Recall that $\theta_0$ is what $\hat{\theta}$ converges to in probability. The estimated

cell means by the column effect model (true model), the main effects model, and the full model all converge to the true parameter, which implies that $f = g_{\theta_0}$ and therefore $H^2(f, g_{\theta_0}) = 0$ for all three models. It can be shown that this is true for all cells.

| | $\hat{\mu}_{ij}$ | $E[\hat{\mu}_{ij}]$ | $\theta_0$ |
|---|---|---|---|
| Cell (1,1) | | | |
| Null Model | $\bar{Y}_{...}$ | $\mu_1/2 + \mu_2/2$ | $\mu_1/2 + \mu_2/2$ |
| Column Effect Model | $\bar{Y}_{.1.}$ | $\mu_1$ | $\mu_1$ |
| Row Effect Model | $\bar{Y}_{1..}$ | $\mu_1/2 + \mu_2/2$ | $\mu_1/2 + \mu_2/2$ |
| Main Effects Model | $\bar{Y}_{1..} + \bar{Y}_{.1.} - \bar{Y}_{...}$ | $\mu_1$ | $\mu_1$ |
| Full Model | $\bar{Y}_{11.}$ | $\mu_1$ | $\mu_1$ |

Table 4.2: Expected Cell Means and $\theta_0$ for cell $(1, 1)$ When the True Model is the Column Effect Model

Now let us take a look at the penalty term in Equation (2.7) for those three models that has zero model error term. The squared Hellinger distance for cell $(1, 1)$ between $g_{\theta_0}$ and $g_{\hat{\theta}}$ is:

$$H^2(\phi_{\mu_1,\sigma^2}, \phi_{\hat{\mu}_{11},\sigma^2}) = 2 - 2\exp\{-\frac{(\hat{\mu}_{11} - \mu_1)^2}{8\sigma^2}\} \tag{4.3}$$

where $\hat{\mu}_{(1,1)}$ is estimated cell mean from the data based on the mean structure prescribed by the approximating model. Since all observations are independently distributed with identical distribution except for location shift, the sample means can be re-written as:

$$\bar{Y}_{ij.} = \mu_j + \frac{\sigma}{\sqrt{n}} Z_{ij};$$

where $Z_{ij} \overset{i.i.d}{\sim} N(0, 1)$. The estimated cell means for cell $(1, 1)$ according to the three models are therefore:

$$\begin{aligned}
\text{Column Effect Model}: &\quad \bar{Y}_{.1.} = \tfrac{\bar{Y}_{11.} + \bar{Y}_{21.}}{2} = \mu_1 + \tfrac{\sigma}{\sqrt{2n}} \tfrac{(Z_{11} + Z_{21})}{\sqrt{2}}; \\
\text{Main Effects Model}: &\quad \bar{Y}_{1..} + \bar{Y}_{.1.} - \bar{Y}_{...} = \mu_1 + \tfrac{\sqrt{3}\sigma}{2\sqrt{n}} \tfrac{(3Z_{11} + Z_{21} + Z_{12} - Z_{22})}{\sqrt{12}}; \\
\text{Full Factorial Model}: &\quad \bar{Y}_{11.} = \mu_1 + \tfrac{\sigma}{\sqrt{n}} Z_{11}
\end{aligned}$$

where $(Z_{11} + Z_{21})/\sqrt{2}$ and $(3Z_{11} + Z_{21} + Z_{12} - Z_{22})/\sqrt{12}$ both have standard normal distribution. The penalty term in Equation (2.7) corresponding to cell $(1, 1)$, $E[H^2(\phi_{\mu_1,\sigma^2}, \phi_{\hat{\mu}_{11},\sigma^2})]$, for the three models are listed below in increasing order:

$$\begin{aligned}
\text{Column Effect Model}: &\quad E[H^2(\phi_{\mu_1,\sigma^2}, \phi_{\bar{Y}_{1\cdot\cdot},\sigma^2})] = 2 - 2\sqrt{\tfrac{4n}{4n+1/2}} \\
\text{Main Effects Model}: &\quad E[H^2(\phi_{\mu_1,\sigma^2}, \phi_{\bar{Y}_{1\cdot\cdot}+\bar{Y}_{\cdot 1\cdot}-\bar{Y}_{\cdot\cdot\cdot},\sigma^2})] = 2 - 2\sqrt{\tfrac{4n}{4n+3/4}} \\
\text{Full Factorial Model}: &\quad E[H^2(\phi_{\mu_1,\sigma^2}, \phi_{\bar{Y}_{11\cdot},\sigma^2})] = 2 - 2\sqrt{\tfrac{4n}{4n+1}}
\end{aligned}$$

Note that this is also true for all other cells. In Theorem 5, we show this pattern holds in general.

**Remark 2.** *Observe that assuming that the true model is the column effect model and $\sigma$ known, the column effect model, the main effects model and the full model have zero overall model error term. Among these three models, the column effect model has the smallest overall penalty term.*

Further illustration of the relationship among the approximating models is provided in a lattice diagram shown in Figure 4.1. In a lattice diagram for all possible ANOVA factorial models, a model is contained by its ancestors and has the same number of effects as all the models at its level. The root of the lattice diagram is the simplest model, i.e., the null model. All other models contain the null model and thus are its ancestors. The full factorial model contains all other models and is at the top. In Figure 4.1, there are two paths that start from the top (the full factorial model) and end at the root (the null model).

In Section 4.2.2, we will show that the results seen in this example hold more generally.

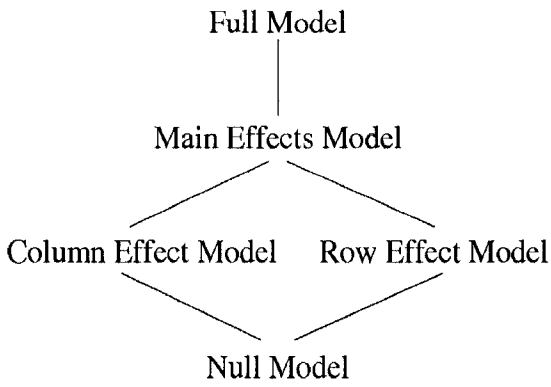


Figure 4.1: Lattice Diagram of 2-way ANOVA Models

### 4.2.2 Properties of the $EH2$ in Balanced ANOVA problems

In this section, we will follow the set of notation for factorial models used in Hocking (1996). Assume the design underlying the data vector is balanced with cell sample size $n$. Define $\mathcal{T}$ as the set of nonempty subsets of the set of factors $\mathcal{F} = \{F_1, F_2, \ldots, F_k\}$, with the $\{a_1, \ldots, a_k\}$ being the corresponding levels. In the example shown in Section 4.2.1, $\mathcal{F} = \{F_1, F_2\}$, $k = 2$, and $\mathcal{T} = \{\{F_1\}, \{F_2\}, \{F_1, F_2\}\}$. Any ANOVA model can be denoted as a subset of $\mathcal{T}$ plus $\emptyset$. Note that by default all models $\mathcal{D}$ include the empty subset $\emptyset$ and if $t \in \mathcal{D}$, so are all the subsets of $t$. Thus, $\mathcal{D}_1 = \{\emptyset, \{F_1\}, \{F_2\}, \{F_1, F_2\}\}$ represents all effects of a full 2-way factorial model while $\mathcal{D}_2 = \{\emptyset\}$ represents the null model. (Table 4.3).

| Models | $\mathcal{D}$ |
|---|---|
| Null Model | $\{\emptyset\}$ |
| Column Effect Model | $\{\emptyset, \{F_2\}\}$ |
| Row Effect Model | $\{\emptyset, \{F_1\}\}$ |
| Main Effects Model | $\{\emptyset, \{F_1\}, \{F_2\}\}$ |
| Full Model | $\{\emptyset, \{F_1\}, \{F_2\}, \{F_1, F_2\}\}$ |

Table 4.3: 2-way ANOVA models denoted by subsets of $\{\emptyset, \mathcal{T}\}$

For any $t \in \mathcal{T}$, $\underline{\mathcal{E}}_t$ denotes the vector of effect parameters. In order to define the factorial effects, we need the following notation. For any $t \in \mathcal{T}$, we first define matrix $\mathbf{H}_t$ as:

$$\mathbf{H}_t = \bigotimes_{i=1}^{k} B_i \tag{4.4}$$

where

$$B_i = I_t(F_i)\underline{\mathbf{S}}_{a_i} + (1 - I_t(F_i))\frac{1}{a_i}\underline{J}_{a_i}^T$$

where $\underline{\mathbf{S}}_{a_i}$ denotes the sum of squares matrix $\mathbf{S}_{a_i} = \mathbf{I}_{a_i} - (1/a_i)\mathbf{U}_{a_i}$ with the last row deleted, $\underline{J}_{a_i}$ is a vector of 1's with length $a_i$, $\mathbf{U}_{a_i}$ is a matrix of 1's with dimension $a_i \times a_i$, and $I_t(F_i) = 1$ if $F_i \in t$ and 0 otherwise.

Define the parameter vectors of the effects by

$$\underline{\mathcal{E}}_t = \mathbf{H}_t\underline{\mu}, \tag{4.5}$$

where $\underline{\mu}$ denotes the cell mean vector according to the full model. The estimated effect vector is

$$\hat{\underline{\mathcal{E}}}_t = \mathbf{H}_t\hat{\underline{\mu}}, \tag{4.6}$$

where $\hat{\underline{\mu}} = (\bigotimes_i^k \underline{J}_{a_i} \bigotimes \frac{1}{n}\underline{J}_n)^T\underline{y}$.

Denote

$$\mathbf{X}_t = \bigotimes_{i=1}^{k} \mathbf{Z}_i, \tag{4.7}$$

with

$$\mathbf{Z}_i = I_t(F_i)\Delta_{a_i}^T + (1 - I_t(F_i))\underline{J}_{a_i},$$

where $\Delta_{a_i} = (\mathbf{I}_{a_i-1}| - \underline{J}_{a_i-1})$. Note that $\mathbf{H}_t$ is a matrix of dimension $\prod_{\{i:F_i\in t\}}(a_i - 1) \times a$ and $\mathbf{X}_t$ is of dimensions $a \times \prod_{\{i:F_i\in t\}}(a_i - 1)$. More details of the notation can be found in Hocking (1996).

**Definition 2.** *Data is said to follow model $\mathcal{D}$ if $\underline{\mathcal{E}}_t = \underline{0}$, for all $t \in \mathcal{D}^C \subseteq \mathcal{T}$.*

**Remark 3.** *The cell mean vector according to model $\mathcal{D}$ is*

$$\underline{\mu}^{\mathcal{D}} = \mathcal{E}\underline{J} + \sum_{t \in \mathcal{D}} \boldsymbol{X}_t \underline{\mathcal{E}}_t, \tag{4.8}$$

*where $\underline{J}$ has length $a$ and $\mathcal{E}$ denotes the overall mean, the null effect or the effect parameter associated with the empty set. According to Definition 2, none of the effects $\{t \in \mathcal{D}^C\}$ contribute to the cell means specified by model $\mathcal{D}$.*

**Definition 3.** *$\mathcal{D}_0 \subseteq \mathcal{T}$ is said to be the* true model *if $\mathcal{D}_0$ is the set of $t$ such that $\underline{\mathcal{E}}_t \neq 0$.*

Observe that for any $t \in \mathcal{T}$, $\mathbf{X}_t$ is a full rank matrix since

$$r(\mathbf{X}_t) = \prod_{i}^{k} r(\mathbf{Z}_i) = \prod_{\{i:F_i \in t\}} (a_i - 1),$$

due to the fact that $r(\Delta_{a_i}) = a_i - 1$. Also note that if $t_1 \neq t_2$, then $\mathbf{X}_t^T \mathbf{X}_t = \mathbf{0}$, i.e., $\mathbf{X}_{t_1}$ and $\mathbf{X}_{t_2}$ are orthogonal to one another (Hocking, 1996). Suppose there are $s$ elements in $\mathcal{S}$ and let $\mathcal{S}^{(i)}$ denote the $i^{\text{th}}$ element in $\mathcal{S} \in \mathcal{T}$. Define

$$\mathbf{A}^{\mathcal{S}} = [\mathbf{X}_{\mathcal{S}^{(1)}} | \cdots | \mathbf{X}_{\mathcal{S}^{(s)}}], \quad \text{and} \quad \underline{\mathcal{E}}^{\mathcal{S}} = [\underline{\mathcal{E}}^T_{\mathcal{S}^{(1)}} | \cdots | \underline{\mathcal{E}}^T_{\mathcal{S}^{(s)}}]^T$$

$\mathbf{A}^{\mathcal{S}}$ is thus a full rank matrix with all the columns being linearly independent and $(\mathbf{A}^{\mathcal{S}})^T \mathbf{A}^{\mathcal{S}}$ is then non-singular.

**Lemma 7.** *Let $\mathcal{D}_0$ be the true model. If $\mathcal{S} \subseteq \mathcal{D}_0$, then*

$$\sum_{t \in \mathcal{S}} \boldsymbol{X}_t \underline{\mathcal{E}}_t \neq \underline{0}.$$

*Proof.* This lemma can be proved by showing that if $\sum_{t\in\mathcal{S}} \mathbf{X}_t\underline{\mathcal{E}}_t = \underline{0}$, then $\mathcal{S} \subseteq \mathcal{D}_0^C$. In fact, if

$$\sum_{t\in\mathcal{S}} \mathbf{X}_t\underline{\mathcal{E}}_t = \mathbf{A}^{\mathcal{S}}\underline{\mathcal{E}}^{\mathcal{S}} = \underline{0},$$

then

$$(\mathbf{A}^{\mathcal{S}})^T\mathbf{A}^{\mathcal{S}}\underline{\mathcal{E}}^{\mathcal{S}} = \underline{0} \quad \Rightarrow \quad \underline{\mathcal{E}}^{\mathcal{S}} = \underline{0} \quad \Rightarrow \quad \mathcal{S} \subseteq \mathcal{D}_0^C$$

The last implication is a consequence of definition 3. □

**Remark 4.** *The result of Lemma 7 can be extended to*

$$\sum_{t\in\mathcal{S}} \boldsymbol{X}_t d_t\underline{\mathcal{E}}_t \neq \underline{0}$$

*if* $S \in \mathcal{D}_0$, *where* $d_t$ *is a real number for* $t \in S$ *and* $d_t \neq 0$ *for at least one* $t \in S$.

The proof of the Remark 4 is similar to that of Lemma 7, only letting

$$\underline{\mathcal{E}}^{\mathcal{S}} = [d_{\mathcal{S}(1)}\underline{\mathcal{E}}^T_{\mathcal{S}(1)}|\cdots|d_{\mathcal{S}(s)}\underline{\mathcal{E}}^T_{\mathcal{S}(s)}]^T.$$

Note that in this case, it is also true that $\underline{\mathcal{E}}^{\mathcal{S}} = \underline{0}$ if and only if $\mathcal{S} \subseteq \mathcal{D}_0^C$.

**Lemma 8.** *Let* $\mathcal{S}_1$ *and* $\mathcal{S}_2$ *be subsets of the true model* $\mathcal{D}_0$. *Then*

$$\sum_{t\in\mathcal{S}_1} \boldsymbol{X}_t\underline{\mathcal{E}}_t = \sum_{t\in\mathcal{S}_2} \boldsymbol{X}_t\underline{\mathcal{E}}_t \tag{4.9}$$

*if and only if* $\mathcal{S}_1 = \mathcal{S}_2$

*Proof.* The sufficiency part is straightforward. The necessity part follows from Lemma 7 and Remark 4. In fact, if equation (4.9) holds, then

$$\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2} \mathbf{X}_t\underline{\mathcal{E}}_t + \sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C} \mathbf{X}_t\underline{\mathcal{E}}_t = \sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2} \mathbf{X}_t\underline{\mathcal{E}}_t + \sum_{t\in\mathcal{S}_1^C\cap\mathcal{S}_2} \mathbf{X}_t\underline{\mathcal{E}}_t,$$

which implies that one of the following must be true

- Conclusion A. Either the set $(\mathcal{S}_1 \cap \mathcal{S}_2^C) \cup (\mathcal{S}_1^C \cap \mathcal{S}_2)$ is empty, or
- Conclusion B. Otherwise,

$$\sum_{t \in (\mathcal{S}_1 \cap \mathcal{S}_2^C) \cup (\mathcal{S}_1^C \cap \mathcal{S}_2)} \mathbf{X}_t d_t \underline{\mathcal{E}}_t = \underline{0}$$

where $d_t = 1$ for $t \in \mathcal{S}_1 \cap \mathcal{S}_2^C$ and $d_t = -1$ for $t \in \mathcal{S}_1^C \cap \mathcal{S}_2$.

Conclusion B implies that

$$(\mathcal{S}_1 \cap \mathcal{S}_2^C) \cup (\mathcal{S}_1^C \cap \mathcal{S}_2) \subseteq \mathcal{D}_0^C$$

by Remark 4. But this result contradicts with the condition that $\mathcal{S}_1$ and $\mathcal{S}_2$ are subsets of the true model $\mathcal{D}_0$. Therefore, the set $(\mathcal{S}_1 \cap \mathcal{S}_2^C) \cup (\mathcal{S}_1^C \cap \mathcal{S}_2)$ must be empty, i.e.,

$$\sum_{t \in \mathcal{S}_1} \mathbf{X}_t \underline{\mathcal{E}}_t = \sum_{t \in \mathcal{S}_2} \mathbf{X}_t \underline{\mathcal{E}}_t \quad \Longrightarrow \quad \mathcal{S}_1 = \mathcal{S}_2$$

□

Now, the cell mean in cell $\pi \in \Pi$ according to any given model $\mathcal{D}$ can be written

$$\mu_\pi^{\mathcal{D}} = \mathcal{E} + \sum_{t \in \mathcal{D}} \mathbf{X}_t^\pi \underline{\mathcal{E}}_t = \mathcal{E} + \sum_{t \in \mathcal{D}} \mathbf{X}_t^\pi \mathbf{H}_t \underline{\mu}, \tag{4.10}$$

where $\mathbf{X}_t^\pi$ denotes the row corresponding to cell $\pi$ in matrix $\mathbf{X}_t$. Specifically, the row number in $\mathbf{X}_t$ corresponding to cell $\pi = \{\pi_1, \ldots, \pi_k\}$ is

$$\sum_{i=1}^{k-1} (\pi_i - 1) \prod_{j=i+1}^{k} a_j + \pi_k$$

Denote $\underline{\mu}^{\mathcal{D},\mathcal{D}_0}$ as the vector of parameters that $\underline{\hat{\mu}}^{\mathcal{D}}$ converges to in probability.

**Lemma 9.** *Let the true distribution of the $Y_{\pi,r}$, where $r = 1, 2, \ldots, n$ and $\pi \in \Pi$, be i.i.d. $N(\mu_\pi^{\mathcal{D}_0}, \sigma^2)$. For any given approximating model $\mathcal{D}$,*

$$\underline{\mu}^{\mathcal{D},\mathcal{D}_0} = \mathcal{E}\underline{J} + \sum_{t \in \mathcal{D} \cap \mathcal{D}_0} \boldsymbol{X}_t \underline{\mathcal{E}}_t.$$

*Proof.* It is well known that the least squares estimates of the cell means are unbiased:

$$E(\underline{\hat{\mu}}) = \underline{\mu}.$$

And thus by the weak law of large numbers,

$$\underline{\hat{\mathcal{E}}}_t = H_t\underline{\hat{\mu}} \xrightarrow{P} H_t\underline{\mu} = \underline{\mathcal{E}}_t,$$

By definition, $\underline{\mathcal{E}}_t$ is non-zero if $t \in \mathcal{D}_0$ and zero otherwise. Now, for any given model indexed by $\mathcal{D} \in \mathcal{T}$,

$$\underline{\hat{\mu}}^{\mathcal{D}} = \hat{\mathcal{E}}\underline{J} + \sum_{t\in\mathcal{D}} \mathbf{X}_t\underline{\hat{\mathcal{E}}}_t \xrightarrow{P} \mathcal{E}\underline{J} + \sum_{t\in\mathcal{D}\cap\mathcal{D}_0} \mathbf{X}_t\underline{\mathcal{E}}_t + \sum_{t\in\mathcal{D}\cap\mathcal{D}_0^C} \mathbf{X}_t\underline{\mathcal{E}}_t = \mathcal{E}\underline{J} + \sum_{t\in\mathcal{D}\cap\mathcal{D}_0} \mathbf{X}_t\underline{\mathcal{E}}_t.$$

□

**Proposition 2.** *Let the true distribution of the $Y_{\pi,r}$, where $r = 1, 2, \ldots, n$ and $\pi \in \Pi$, be i.i.d. $N(\mu_\pi^{\mathcal{D}_0}, \sigma^2)$. The corresponding distribution of an approximating model indexed by $\mathcal{D}$ is given by $N(\hat{\mu}_\pi^{\mathcal{D}}, \sigma^2)$, supposing $\sigma^2$ is known. Then the overall model error for model $\mathcal{D}$ is zero if and only if $\mathcal{D}_0 \subseteq \mathcal{D}$.*

*Proof.* The overall model error is zero if and only if the model error term in Equation (2.7) is zero for all cells $\pi \in \Pi$. The model error term for cell $\pi \in \Pi$ is

$$H^2(\phi_{\mu_\pi^{\mathcal{D}_0},\sigma^2}, \phi_{\mu_\pi^{\mathcal{D},\mathcal{D}_0},\sigma^2}) = 2 - 2\exp\{-\frac{(\mu_\pi^{\mathcal{D}_0} - \mu_\pi^{\mathcal{D},\mathcal{D}_0})^2}{8\sigma^2}\} \tag{4.11}$$

The value of Equation (4.11) is zero if and only if, for any $\pi \in \Pi$

$$\mu_\pi^{\mathcal{D}_0} = \mu_\pi^{\mathcal{D},\mathcal{D}_0}. \tag{4.12}$$

Due to Lemma 9, Equation (4.12) implies that

$$\mathcal{E}\underline{J} + \sum_{t\in\mathcal{D}_0} \mathbf{X}_t\underline{\mathcal{E}}_t = \mathcal{E}\underline{J} + \sum_{t\in\mathcal{D}\cap\mathcal{D}_0} \mathbf{X}_t\underline{\mathcal{E}}_t, \tag{4.13}$$

which in turn implies that $\mathcal{D} \cap \mathcal{D}_0 = \mathcal{D}_0$, or $\mathcal{D}_0 \subseteq \mathcal{D}$, due to Lemma 8. □

It follows naturally from the proof of Proposition 2 that the model error term will be greater than zero for an approximating model $\mathcal{D}$ which does not include all the effects in the true model $\mathcal{D}_0$.

**Proposition 3.** *Using the setting of Proposition 2, the model error terms of Equation (2.7) in any cell $\pi \in \Pi$ for two approximating models $\mathcal{D}_1$ and $\mathcal{D}_2$ are equal if and only if $\mathcal{D}_1 \cap \mathcal{D}_0 = \mathcal{D}_2 \cap \mathcal{D}_0$.*

*Proof.* The model error term of Equation (2.7) in any cell $\pi \in \Pi$ for the approximating model $\mathcal{D}_i$, $i = 1, 2$, is

$$H^2\left(\phi_{\mu_\pi^{\mathcal{D}_0},\sigma^2}, \phi_{\mu_\pi^{\mathcal{D}_i,\mathcal{D}_0},\sigma^2}\right) = 2 - 2\exp\{-\frac{(\mu_\pi^{\mathcal{D}_0} - \mu_\pi^{\mathcal{D}_i,\mathcal{D}_0})^2}{8\sigma^2}\}. \tag{4.14}$$

The difference between the two model error terms is

$$H^2\left(\phi_{\mu_\pi^{\mathcal{D}_0},\sigma^2}, \phi_{\mu_\pi^{\mathcal{D}_1,\mathcal{D}_0},\sigma^2}\right) - H^2(\phi_{\mu_\pi^{\mathcal{D}_0},\sigma^2}, \phi_{\mu_\pi^{\mathcal{D}_2,\mathcal{D}_0},\sigma^2}) \tag{4.15}$$

$$= 2\left[\exp\{-\frac{(\mu_\pi^{\mathcal{D}_0} - \mu_\pi^{\mathcal{D}_2,\mathcal{D}_0})^2}{8\sigma^2}\} - \exp\{-\frac{(\mu_\pi^{\mathcal{D}_0} - \mu_\pi^{\mathcal{D}_1,\mathcal{D}_0})^2}{8\sigma^2}\}\right] \tag{4.16}$$

$$= 2\left[\exp\{-\frac{(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{D}_2\cap\mathcal{D}_0}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t)^2}{8\sigma^2}\} - \exp\{-\frac{(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{D}_1\cap\mathcal{D}_0}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t)^2}{8\sigma^2}\}\right] \tag{4.17}$$

Equation (4.17) is zero if and only if

$$\sum_{t\in\mathcal{D}_1\cap\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t = \sum_{t\in\mathcal{D}_2\cap\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t.$$

By Lemma 8, Equation (4.2.2) is true if and only if

$$(\mathcal{D}_2 \cap \mathcal{D}_0) = (\mathcal{D}_1 \cap \mathcal{D}_0)$$

□

**Corollary 4.** *Using the setting of Proposition 2, the overall model errors for two approximating models $\mathcal{D}_1$ and $\mathcal{D}_2$ are equal if $\mathcal{D}_1 \cap \mathcal{D}_0 = \mathcal{D}_2 \cap \mathcal{D}_0$.*

Corollary 4 follows immediately from Proposition 3.

**Proposition 4.** *Using the setting of Proposition 2, the overall model error for* $\mathcal{D}_2$ *is greater than that for* $\mathcal{D}_1$ *if* $(\mathcal{D}_2 \cap \mathcal{D}_0) \subset (\mathcal{D}_1 \cap \mathcal{D}_0)$.

*Proof.* For simplicity of notation in this proof, denote $\mathcal{S}_1 = \mathcal{D}_1 \cap \mathcal{D}_0$ and $\mathcal{S}_2 = \mathcal{D}_2 \cap \mathcal{D}_0$. If $\mathcal{S}_2 \subset \mathcal{S}_1$, then

$$\mathcal{S}_1 = \mathcal{S}_2 \cup (\mathcal{S}_1 \cap \mathcal{S}_2^C)$$

where $\mathcal{S}_1 \cap \mathcal{S}_2^C = \mathcal{D}_1 \cap \mathcal{D}_2^C \cap \mathcal{D}_0$. Observe that

$$\left(\sum_{t\in\mathcal{S}_2} \mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C} \mathbf{X}_t\underline{\mathcal{E}}_t\right)^T = \left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C} \mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{S}_2} \mathbf{X}_t\underline{\mathcal{E}}_t\right)^T = \mathbf{0}$$

and

$$\begin{aligned}\left(\sum_{t\in\mathcal{D}_0} \mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C} \mathbf{X}_t\underline{\mathcal{E}}_t\right)^T &= \left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C} \mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{D}_0} \mathbf{X}_t\underline{\mathcal{E}}_t\right)^T \\ &= \left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C} \mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C} \mathbf{X}_t\underline{\mathcal{E}}_t\right)^T\end{aligned}$$

due to the fact that $\mathbf{X}_t$ is orthogonal to one another. Note that terms $\left(\sum_{t\in\mathcal{D}_0} \mathbf{X}_t^\pi\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_i} \mathbf{X}_t^\pi\underline{\mathcal{E}}_t\right)^2$ for all cells $\pi \in \Pi$ are the diagonal terms of the square matrix

$$\left(\sum_{t\in\mathcal{D}_0} \mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_i} \mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{D}_0} \mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_i} \mathbf{X}_t\underline{\mathcal{E}}_t\right)^T,$$

where $i = 1, 2$. Furthermore, based on the above observations,

$$\begin{aligned}
&\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_1}\mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_1}\mathbf{X}_t\underline{\mathcal{E}}_t\right)^T \\
&= \left(\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_2}\mathbf{X}_t\underline{\mathcal{E}}_t\right) - \sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C}\mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_2}\mathbf{X}_t\underline{\mathcal{E}}_t\right) - \sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C}\mathbf{X}_t\underline{\mathcal{E}}_t\right)^T \\
&= \left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_2}\mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_2}\mathbf{X}_t\underline{\mathcal{E}}_t\right)^T + \left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C}\mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C}\mathbf{X}_t\underline{\mathcal{E}}_t\right)^T \\
&\quad - \left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_2}\mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C}\mathbf{X}_t\underline{\mathcal{E}}_t\right)^T - \left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C}\mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_2}\mathbf{X}_t\underline{\mathcal{E}}_t\right)^T \\
&= \left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_2}\mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_2}\mathbf{X}_t\underline{\mathcal{E}}_t\right)^T - \left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C}\mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C}\mathbf{X}_t\underline{\mathcal{E}}_t\right)^T
\end{aligned}$$

Thus, the diagonal terms of $\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_1}\mathbf{X}_t\underline{\mathcal{E}}_t\right)\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_1}\mathbf{X}_t\underline{\mathcal{E}}_t\right)^T$ can be written:

$$\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_1}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t\right)^2 = \left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_2}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t\right)^2 - \left(\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t\right)^2$$

Since $\sum_{t\in\mathcal{S}_1\cap\mathcal{S}_2^C}\mathbf{X}_t\underline{\mathcal{E}}_t \neq \underline{0}$, the following inequality holds for $\pi \in \Pi$:

$$\left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_1}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t\right)^2 \leq \left(\sum_{t\in\mathcal{D}_0}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t - \sum_{t\in\mathcal{S}_2}\mathbf{X}_t^\pi\underline{\mathcal{E}}_t\right)^2,$$

with strict inequality holds for at least one cell. Thus, the value of the right hand side of Equation (4.17) is either zero or negative for every cell $\pi \in \Pi$. It follows that the overall model error for model $\mathcal{D}_2$ is greater than that for model $\mathcal{D}_1$. □

**Remark 5.** *If the overall model errors for two models at the same level in a lattice diagram are zero, the true model is a descendent of both models.*

Remark 5 follows from Proposition 2. If the overall model errors for two models, say $\mathcal{D}_1$ and $\mathcal{D}_2$, at the same level in a lattice diagram are zero, they must both contain the true model due to Proposition 2. Moreover, neither model can be contained by the other since they are on the same level. Therefore, the true model must be at a lower level and thus a descendent of both models $\mathcal{D}_1$ and $\mathcal{D}_2$.

**Remark 6.** *The overall model errors along any path from the full factorial model to the null model in a lattice diagram are non-decreasing.*

Remark 6 follows directly from Proposition 3. In fact, the overall model errors will be all zeros along any path if the true model is the null model but will be strictly increasing if the the true model is the full factorial model.

For the following discussion, we need the definition below.

**Definition 4.** *A set of models is said to be* intersection-closed *if the intersection of any two models in the set is also included in the set.*

**Definition 5.** *The set of all the models for which the overall model error is zero is referred to as the* unbiased group.

**Remark 7.** *The unbiased group is intersection-closed.*

Remark 7 follows from Proposition 2. In fact, suppose two models $\mathcal{D}_1$ and $\mathcal{D}_2$ are in the unbiased group, then both models contains the true model. Their intersection $\mathcal{D}_1 \cap \mathcal{D}_2$ is therefore also in the unbiased group since it contains the true model and thus has zero overall model error. The unbiased group can be thought as the group of all the models that contain the true model.

**Theorem 5.** *Using the setting of proposition 2, for any given sample size $n$ and approximating model $\mathcal{D}$, the penalty term in cell $\pi \in \Pi$ is*

$$E[H^2(\phi_{\mu_\pi^{\mathcal{D},\mathcal{D}_0},\sigma^2}, \phi_{\hat{\mu}_\pi^{\mathcal{D}},\sigma^2})] = 2 - 2\sqrt{\frac{4n}{4n + \frac{1}{a}\left(1 + \sum_{t\in\mathcal{D}} \prod_{\{i:F_i\in t\}}(a_i - 1)\right)}}. \quad (4.18)$$

*In particular, the overall penalty depends on the number of parameters in model $\mathcal{D}$ only.*

*Proof.* The squared Hellinger distance between $\phi_{\mu_\pi^{\mathcal{D},\mathcal{D}_0},\sigma^2}$ and $\phi_{\hat{\mu}_\pi^{\mathcal{D}},\sigma^2}$ in cell $\pi \in \Pi$ is

$$H^2(\phi_{\mu_\pi^{\mathcal{D},\mathcal{D}_0},\sigma^2}, \phi_{\hat{\mu}_\pi^{\mathcal{D}},\sigma^2}) = 2 - 2\exp\{-\frac{(\mu_\pi^{\mathcal{D},\mathcal{D}_0} - \hat{\mu}_\pi^{\mathcal{D}})^2}{8\sigma^2}\}.$$

$\hat{\mu}_\pi^{\mathcal{D}}$ can be written

$$\hat{\mu}_\pi^{\mathcal{D}} = \hat{\mathcal{E}} + \left(\sum_{t\in\mathcal{D}} \mathbf{X}_t^\pi \mathbf{H}_t\right)\hat{\underline{\mu}} = (\underline{c}_\pi^{\mathcal{D}})^T \hat{\underline{\mu}};$$

where

$$(\underline{c}_\pi^{\mathcal{D}})^T = \frac{1}{a}\underline{J}^T + \sum_{t\in\mathcal{D}} \mathbf{X}_t^\pi \mathbf{H}_t$$

Recall that $a = \prod_{i=1}^k a_i$. Now, the mean and variance of $\hat{\mu}_\pi^{\mathcal{D}}$ with respect to the true model are

$$E[\hat{\mu}_\pi^{\mathcal{D}}] = \mu_\pi^{\mathcal{D},\mathcal{D}_0}; \quad Var[\hat{\mu}_\pi^{\mathcal{D}}] = \frac{\sigma^2}{n}(\underline{c}_\pi^{\mathcal{D}})^T \underline{c}_\pi^{\mathcal{D}}.$$

The following random variable $Z$ has standard normal distribution:

$$Z = \frac{(\hat{\mu}_\pi^{\mathcal{D}} - \mu_\pi^{\mathcal{D},\mathcal{D}_0})}{\sigma\sqrt{(\underline{c}_\pi^{\mathcal{D}})^T \underline{c}_\pi^{\mathcal{D}}}/\sqrt{n}}.$$

The expected squared Hellinger distance thus becomes

$$\begin{aligned} E[H^2(\phi_{\mu_\pi^{\mathcal{D},\mathcal{D}_0},\sigma^2}, \phi_{\hat{\mu}_\pi^{\mathcal{D}},\sigma^2})] &= 2 - 2E[\exp\{-\frac{\sigma^2(\underline{c}_\pi^{\mathcal{D}})^T \underline{c}_\pi^{\mathcal{D}}}{8n\sigma^2} Z^2\}] \\ &= 2 - 2\left[1 - 2(-\frac{(\underline{c}_\pi^{\mathcal{D}})^T \underline{c}_\pi^{\mathcal{D}}}{8n})\right]^{-1/2} \\ &= 2 - 2\sqrt{\frac{4n}{4n + (\underline{c}_\pi^{\mathcal{D}})^T \underline{c}_\pi^{\mathcal{D}}}}. \end{aligned}$$

Furthermore,

$$(\underline{c}_\pi^{\mathcal{D}})^T \underline{c}_\pi^{\mathcal{D}} = \frac{1}{a^2}\underline{J}^T\underline{J} + \frac{1}{a}\sum_{t\in\mathcal{D}} \mathbf{X}_t^\pi \mathbf{H}_t \underline{J} + \frac{1}{a}\sum_{t\in\mathcal{D}} \underline{J}^T \mathbf{H}_t^T (\mathbf{X}_t^\pi)^T + (\sum_{t\in\mathcal{D}} \mathbf{X}_t^\pi \mathbf{H}_t)(\sum_{t\in\mathcal{D}} \mathbf{X}_t^\pi \mathbf{H}_t)^T \tag{4.19}$$

Note that the first term on the right hand side of the above Equation (4.19) becomes $\frac{1}{(\prod_i^k a_i)}$ while the second and third term become 0, since

$$\mathbf{H}_t \underline{J} = \underline{J}^T \mathbf{H}_t^T = 0$$

due to the fact that $\underline{\mathbf{S}}_{a_i}\underline{J}_{a_i} = \underline{J}_{a_i-1} - \underline{J}_{a_i-1} = 0$. Observe that for any $t \in \mathcal{T}$,

$$\mathbf{X}_t\mathbf{H}_t = \bigotimes_i^k \left( I_t(F_i)\Delta_{a_i}^T \underline{\mathbf{S}}_{a_i} + \frac{1}{a_i}(1 - I_t(F_i))\underline{J}_{a_i}\underline{J}_{a_i}^T \right).$$

Without loss of generality, if $t \neq t^*$, there exists at least one $i$ such that $F_i$ is in $t$ but not in $t^*$ and

$$\begin{aligned}
&\left( I_t(F_i)\Delta_{a_i}^T \underline{\mathbf{S}}_{a_i} + \frac{1}{a_i}(1 - I_t(F_i))\underline{J}_{a_i}\underline{J}_{a_i}^T \right) \left( I_{t^*}(F_i)\Delta_{a_i}^T \underline{\mathbf{S}}_{a_i} + \frac{1}{a_i}(1 - I_{t^*}(F_i))\underline{J}_{a_i}\underline{J}_{a_i}^T \right)^T \\
&= \Delta_{a_i}^T \underline{\mathbf{S}}_{a_i} \frac{1}{a_i} \underline{J}_{a_i}\underline{J}_{a_i}^T \\
&= \mathbf{0}
\end{aligned}$$

Therefore, if $t \neq t^*$,

$$\begin{aligned}
&(\mathbf{X}_t\mathbf{H}_t)(\mathbf{X}_{t^*}\mathbf{H}_{t^*})^T \\
&= \bigotimes_i^k \left( I_t(F_i)\Delta_{a_i}^T \underline{\mathbf{S}}_{a_i} + \frac{1}{a_i}(1 - I_t(F_i))\underline{J}_{a_i}\underline{J}_{a_i}^T \right) \left( I_{t^*}(F_i)\Delta_{a_i}^T \underline{\mathbf{S}}_{a_i} + \frac{1}{a_i}(1 - I_{t^*}(F_i))\underline{J}_{a_i}\underline{J}_{a_i}^T \right)^T \\
&= \mathbf{0}
\end{aligned}$$

The last term in Equation (4.19) for all $\pi \in \Pi$ are the diagonal terms of the following square matrix

$$
\begin{aligned}
(\sum_{t\in\mathcal{D}} \mathbf{X}_t\mathbf{H}_t)(\sum_{t\in\mathcal{D}} \mathbf{X}_t\mathbf{H}_t)^T &= \sum_{t\in\mathcal{D}}\sum_{t^*\in\mathcal{D}} \bigotimes_i^k \left( I_t(F_i)\Delta_{a_i}^T \underline{\mathbf{S}}_{a_i} + \frac{1}{a_i}(1 - I_t(F_i))\underline{J}_{a_i}\underline{J}_{a_i}^T \right) \\
&\qquad \left( I_{t^*}(F_i)\underline{\mathbf{S}}_{a_i}^T\Delta_{a_i} + \frac{1}{a_i}(1 - I_{t^*}(F_i))\underline{J}_{a_i}\underline{J}_{a_i}^T \right) \\
&= \sum_{t\in\mathcal{D}} \bigotimes_i^k [I_t(F_i)\Delta_{a_i}^T \underline{\mathbf{S}}_{a_i}\underline{\mathbf{S}}_{a_i}^T\Delta_{a_i} + (1 - I_t(F_i))\frac{1}{a_i^2}\underline{J}_{a_i}\underline{J}_{a_i}^T\underline{J}_{a_i}\underline{J}_{a_i}^T] \\
&= \sum_{t\in\mathcal{D}} \bigotimes_i^k [I_t(F_i)\frac{1}{a_i}(a_i\mathbf{I}_{a_i} - \mathbf{U}_{a_i}) + (1 - I_t(F_i))\frac{1}{a_i}\mathbf{U}_{a_i}]
\end{aligned}
$$

The diagonal terms of the above matrix are

$$
\frac{1}{a}\sum_{t\in\mathcal{D}} \prod_{\{i:F_i\in t\}} (a_i - 1), \quad \pi \in \Pi.
$$

Therefore, Equation (4.19) is simplified into

$$
(\underline{c}_\pi^{\mathcal{D}})^T\underline{c}_\pi^{\mathcal{D}} = \frac{1}{a}\left(1 + \sum_{t\in\mathcal{D}} \prod_{\{i:F_i\in t\}} (a_i - 1)\right),
$$

and the penalty term in cell $\pi \in \Pi$ becomes

$$
E[H^2(\phi_{\mu_\pi^{\mathcal{D}},\mathcal{D}_0,\sigma^2}, \phi_{\hat{\mu}_\pi^{\mathcal{D}},\sigma^2})] = 2 - 2\sqrt{\frac{4n}{4n + \frac{1}{a}\left(1 + \sum_{t\in\mathcal{D}}\prod_{\{i:F_i\in t\}}(a_i - 1)\right)}}
$$

Observe that $1 + \sum_{t\in\mathcal{D}}\prod_{\{i:F_i\in t\}}(a_i - 1)$ is the number of linearly independent parameters and the overall penalty increases as this number increases. □

An example with 2-way ANVOA models is given in Section 4.2.1, where penalties for the column effect model, the main effects model, and the full model are calculated.

**Remark 8.** *The overall penalty for any approximating model converges to 0 as* $n \to \infty$.

**Remark 9.** *The true model has the smallest penalty term and thus the smallest overall approximated EH2 on the right hand side of Equation (2.7) among the unbiased group.*

Observe that the true model is the smallest model among the unbiased group, in which all the models have zero overall model errors. It follows from Theorem 5 that it has the smallest overall penalty among the group.

**Theorem 6.** *Under the setting of Proposition 2, let $\mathcal{D}$ be any model including the intercept and $\mathcal{D}_0$ be the true model. Then the overall approximated EH2 on the right hand side of Equation (2.7) for $\mathcal{D} \cap \mathcal{D}_0$ is smaller than or equal to that for $\mathcal{D}$, with equality if and only if $\mathcal{D} \cap \mathcal{D}_0 = \mathcal{D}$.*

*Proof.* The result of the theorem is straightforward if $\mathcal{D} = \mathcal{D}_0$. If $\mathcal{D} \neq \mathcal{D}_0$, the intersection of model $\mathcal{D}$ and the true model $\mathcal{D}_0$, $\mathcal{D} \cap \mathcal{D}_0$, has less linearly independent parameters unless $\mathcal{D} \cap \mathcal{D}_0 = \mathcal{D}$. Then,

1. By Corollary 4, model $\mathcal{D} \cap \mathcal{D}_0$ has the same overall model error as model $\mathcal{D}$ does;

2. By Theorem 5, the overall penalty for model $\mathcal{D} \cap \mathcal{D}_0$ is either the same as (if and only if $\mathcal{D} \cap \mathcal{D}_0 = \mathcal{D}$) or otherwise smaller than that for model $\mathcal{D}$.

Therefore, the overall approximated expected squared Hellinger distance for model $\mathcal{D} \cap \mathcal{D}_0$ is either smaller than or equal to that for model $\mathcal{D}_0$. □

**Theorem 7.** *Using the setting of proposition 2, for any given sample size $n$ and approximating model $\mathcal{D}$, let*

$$\xi = \frac{1}{a}\left(1 + \sum_{t \in \mathcal{D}} \prod_{\{i:F_i \in t\}} (a_i - 1)\right)$$

*then*

$$EH2(\pi) = 2 - 2\sqrt{\frac{4n}{4n+\xi}}\exp\{-\frac{n(\mu_\pi^{\mathcal{D},\mathcal{D}_0} - \mu_\pi^{\mathcal{D}_0})^2}{2\sigma^2(4n+\xi)}\}.$$

*Proof.* The squared Hellinger distance between $\phi_{\mu_\pi^{\mathcal{D}_0},\sigma^2}$ and $\phi_{\hat{\mu}_\pi^{\mathcal{D}},\sigma^2}$ in cell $\pi \in \Pi$ is

$$H^2(\phi_{\mu_\pi^{\mathcal{D}_0},\sigma^2}, \phi_{\hat{\mu}_\pi^{\mathcal{D}},\sigma^2}) = 2 - 2\exp\{-\frac{(\mu_\pi^{\mathcal{D}_0} - \hat{\mu}_\pi^{\mathcal{D}})^2}{8\sigma^2}\}.$$

Note from the proof of Theorem 5 that the mean and variance of $\hat{\mu}_\pi^{\mathcal{D}}$ with respect to the true model are

$$E[\hat{\mu}_\pi^{\mathcal{D}}] = \mu_\pi^{\mathcal{D},\mathcal{D}_0}; \quad Var[\hat{\mu}_\pi^{\mathcal{D}}] = \frac{\sigma^2}{n}(\underline{c}_\pi^{\mathcal{D}})^T\underline{c}_\pi^{\mathcal{D}}.$$

The following random variable $X$ has normal distribution with mean $\mu_\pi^{\mathcal{D},\mathcal{D}_0} - \mu_\pi^{\mathcal{D}_0}$ and variance 1:

$$X = \frac{(\hat{\mu}_\pi^{\mathcal{D}} - \mu_\pi^{\mathcal{D}_0})}{\sigma\sqrt{(\underline{c}_\pi^{\mathcal{D}})^T\underline{c}_\pi^{\mathcal{D}}}/\sqrt{n}}.$$

And $X^2$ follows a non-central chi-square distribution with 1 degree of freedom and non-centrality parameter $\lambda$:

$$\lambda = \frac{(\mu_\pi^{\mathcal{D},\mathcal{D}_0} - \mu_\pi^{\mathcal{D}_0})^2}{\frac{\sigma^2}{n}(\underline{c}_\pi^{\mathcal{D}})^T\underline{c}_\pi^{\mathcal{D}}}$$

The expected squared Hellinger distance thus becomes

$$\begin{aligned}
E[H^2(\phi_{\mu_\pi^{\mathcal{D}_0},\sigma^2}, \phi_{\hat{\mu}_\pi^{\mathcal{D}},\sigma^2})] &= 2 - 2E[\exp\{-\frac{\sigma^2(\underline{c}_\pi^{\mathcal{D}})^T\underline{c}_\pi^{\mathcal{D}}}{8n\sigma^2}X^2\}] \\
&= 2 - 2\left[1 - 2(-\frac{(\underline{c}_\pi^{\mathcal{D}})^T\underline{c}_\pi^{\mathcal{D}}}{8n})\right]^{-1/2}\exp\{-\frac{\lambda\frac{(\underline{c}_\pi^{\mathcal{D}})^T\underline{c}_\pi^{\mathcal{D}}}{8n}}{(1+2\frac{(\underline{c}_\pi^{\mathcal{D}})^T\underline{c}_\pi^{\mathcal{D}}}{8n})}\} \\
&= 2 - 2\sqrt{\frac{4n}{4n+(\underline{c}_\pi^{\mathcal{D}})^T\underline{c}_\pi^{\mathcal{D}}}}\exp\{-\frac{n(\mu_\pi^{\mathcal{D},\mathcal{D}_0} - \mu_\pi^{\mathcal{D}_0})^2}{2\sigma^2(4n+(\underline{c}_\pi^{\mathcal{D}})^T\underline{c}_\pi^{\mathcal{D}})}\} \\
&= 2 - 2\sqrt{\frac{4n}{4n+\xi}}\exp\{-\frac{n(\mu_\pi^{\mathcal{D},\mathcal{D}_0} - \mu_\pi^{\mathcal{D}_0})^2}{2\sigma^2(4n+\xi)}\}.
\end{aligned}$$

where

$$\xi = (\underline{c}_{\pi}^{\mathcal{D}})^T \underline{c}_{\pi}^{\mathcal{D}} = \frac{1}{a}\left(1 + \sum_{t \in \mathcal{D}} \prod_{\{i:F_i \in t\}} (a_i - 1)\right),$$

from the proof of Theorem 5. □

Note that when the squared difference between $\mu_{\pi}^{\mathcal{D},\mathcal{D}_0}$ and $\mu_{\pi}^{\mathcal{D}_0}$ is less than or equal to $4\sigma^2$, $EH2$ is an increasing function in $\xi$. Otherwise, for fixed difference between $\mu_{\pi}^{\mathcal{D},\mathcal{D}_0}$ and $\mu_{\pi}^{\mathcal{D}_0}$ and any given $n$, $EH2$ will decrease in $\xi$ until $\xi = 2n(\frac{(\mu_{\pi}^{\mathcal{D},\mathcal{D}_0} - \mu_{\pi}^{\mathcal{D}_0})^2}{2\sigma^2} - 2)$ and then increase. In general, we observe that the expected squared Hellinger distance has a built-in penalty in the sense that there exists a number of parameters beyond which the expected squared Hellinger distance will increase (unless the number of parameters is bounded). Where the changing point occurs depends on the sample size and the amount of lack of fit.

**Theorem 8.** *Under the setting of Proposition 2, for any approximating model $\mathcal{D}$,*

$$EH2(\pi) \longrightarrow H^2\left(\phi_{\mu_{\pi}^{\mathcal{D}_0},\sigma^2}, \phi_{\mu_{\pi}^{\mathcal{D},\mathcal{D}_0},\sigma^2}\right)$$

*as $n \to \infty$ for $\pi \in \Pi$.*

*Proof.* Theorem 8 directly follow from Theorem 7. As $n \to \infty$

$$\begin{aligned} EH2(\pi) &= 2 - 2\sqrt{\frac{4n}{4n+\xi}}\exp\{-\frac{n(\mu_{\pi}^{\mathcal{D},\mathcal{D}_0} - \mu_{\pi}^{\mathcal{D}_0})^2}{2\sigma^2(4n+\xi)}\} \\ &\longrightarrow 2 - 2\exp\{-\frac{(\mu_{\pi}^{\mathcal{D},\mathcal{D}_0} - \mu_{\pi}^{\mathcal{D}_0})^2}{8\sigma^2}\} \\ &= H^2\left(\phi_{\mu_{\pi}^{\mathcal{D}_0},\sigma^2}, \phi_{\mu_{\pi}^{\mathcal{D},\mathcal{D}_0},\sigma^2}\right) \end{aligned}$$

□

**Remark 10.** *Under the setting of Proposition 2, the true best model will become the true model as $n \to \infty$.*

Remark 10 is a direct consequence of Theorem 8.

**Remark 11.** *For practical purposes, the true best model is unique.*

We can, however, construct examples where more than one model has the same smallest $EH2$.

**Definition 6.** *The union of the unbiased group and the set of models that has smaller $EH2$ than the true model, if any, is said to be the* target group.

Note that if the true best model is the true model itself, the target group is just the unbiased group and is thus intersection-closed. The true best model can be other than the true model only when the decrease in the overall penalty exceeds the increase in the overall model error for the true best model as compared to the true model. As sample size increases, the overall penalty will vanish and the overall model error becomes dominant. Therefore, when sample size increases, not only the target group will approach the unbiased group, the $EH2$'s for all the models in the target group will be close to zero and thus close to one another. Based on these properties, we propose a two-step procedure described in Section 4.3 to estimate the true best model by choosing the model that has the smallest overall penalty among the estimated target group.

All results in this section are based on the assumption that the variance of the true distribution is known. Although this assumption is not realistic in practice, these results motivate us to propose the model selection strategy described in the following section.

## 4.3 Model Selection Strategy Based On $PEEH2$

As noted earlier, all the models on any path from the full model to the true best model have $EH2$ that are close to one another when sample size is large enough. The error in estimating the $EH2$, i.e. the difference between $EH2$ and the estimated $EH2$, may mask the differences in $EH2$ between the true best model and other models in the target group. This motivates us to consider a group of models with estimated $EH2$ not significantly higher than the smallest estimated $EH2$. It is expected that when sample size is large enough, this group captures the feature of the unbiaed group. Within this group, the model with the smallest penalty term will be selected as our estimated true best model.

Based on the above comments, a two-step model selection procedure will be studied in this chapter:

1. *Find the estimated $EH2$ for all the approximating models;*

2. *Estimate the target group by finding the group of model(s) whose estimated $EH2$ are not significantly higher than the smallest among all approximating models. Include the models that are necessary to make the group intersection-closed. The one that has the smallest penalty term among the models in the estimated target group is the estimated true best model.*

### 4.3.1 Model Selection Based on $PEEH2$

In this chapter, we develop a set of model selection strategy using $PEEH2$, of which the first and second term estimate the model error term and penalty term in Equation (2.7), respectively. Note that $PEEH2$ is a random variable. To find the group of models for which the estimated $EH2$ are not significantly higher than

the smallest among the approximating models, we need to quantify the variation of $PEEH2$ for all the models. We thus propose to use the bootstrap variation of $PEEH2$ to approximate the variation of $PEEH2$. There are several possible ways to identify the target group. In this chapter, we consider both to compare the bootstrap confidence intervals of the models (Procedure *Group Identification I*) and to find the models of which $PEEH2$ is not more than some multiple of bootstrap standard deviations away from the smallest $PEEH2$ (Procedure *Group Identification II*). The model that has the smallest penalty term among the target group is then our best model. Our proposed strategy using $PEEH2$ to do model selection in ANOVA model problems is described in procedure *Selection* as follows:

**Procedure *Selection***

1. Calculate the sample cell means $\bar{y}_{\pi,\cdot}$ and compute the residuals $\{res_{\{\pi,r\}} = y_{\{\pi,r\}} - \bar{y}_{\{\pi,\cdot\}} : r = 1, 2, \cdots, n_\pi, \forall\ \pi \in \Pi\}$. Fit the kernel density estimator $\hat{f}$ on the residuals;

2. Calculate $PEEH2$

   - For each $\pi \in \Pi$, find $\hat{\theta}_\pi^{\mathcal{D}} = (\hat{\mu}_\pi^{\mathcal{D}}, (\hat{\sigma}^{\mathcal{D}})^2)$ according to the approximating model $\mathcal{D}$ and thus estimate the overall model error term by averaging the following quantity over all cells $\pi \in \Pi$:

   $$H^2\left(\hat{f}(t - \bar{y}_{\pi,\cdot}), \phi_{\hat{\mu}_\pi^{\mathcal{D}},(\hat{\sigma}^{\mathcal{D}})^2}\right);$$

   - Generate $M$ bootstrap samples by re-sampling with replacement from $\{res_{\{\pi,r\}}\}$ $M$ times and adding back the corresponding cell sample means $\bar{y}_{\{\pi,\cdot\}}, \pi \in \Pi$. Estimate the overall penalty by averaging the following

quantity over all cells $\pi \in \Pi$:

$$\frac{1}{M}\sum_{l=1}^{M} H^2\left(\phi_{\hat{\mu}_\pi^{\mathcal{D}},(\hat{\sigma}^{\mathcal{D}})^2}, \phi_{\hat{\mu}_\pi^{\mathcal{D}(l)},(\hat{\sigma}^{\mathcal{D}(l)})^2}\right);$$

3. Generate $M2$ bootstrap samples by re-sampling with replacement. For each of the bootstrap samples, repeat the previous step and get $M2$ $PEEH2^*$'s

4. Find the model that has the smallest quantity of $PEEH2$ among all the approximating models and denote it as $\hat{\mathcal{D}}_0$;

5. Estimate the target group according to Procedure *Group Identification I* or *Group Identification II.*

6. Among the models found in the previous step, the model that has the smallest penalty term will be chosen as the best model.

The approaches considered in this chapter to identify the target group are described below:

**Procedure *Group Identification I***

1. For each approximating model, find the $(1-\alpha)\%$ bootstrap confidence interval by locating the $(\alpha/2)^{\text{th}}$ and $(1-\alpha/2)^{\text{th}}$ percentiles of $PEEH2^*$'s;

2. Identify all the models of which the intervals overlap with that of model $\hat{\mathcal{D}}_0$;

3. Include models that are necessary for this group of model to form a intersection-closed set from the full model downwards;

**Procedure *Group Identification II***

1. Find the standard deviation, $SD^{\hat{\mathcal{D}}_0}$, of $PEEH2^*$'s;

2. Identify all the models of which $PEEH2$ is within $c(SD^{\hat{\mathcal{D}}_0})$ away from that of model $\hat{\mathcal{D}}_0$, where $c$ is a prespecified constant scaler;

3. Include models that are necessary for this group of models to form a intersection-closed set from the full model downwards;

The methods trying to identify the group of models that has the smallest $PEEH2$'s are based on heuristics and thus remain open for discussion.

## 4.4 Simulation Study: Model Selection Performance

In this section, performance of the model selection procedure using $PEEH2$ as described in Section 4.3.1 will be evaluated using statistical simulation. That is, we will see how often the true best model (the one that has the smallest $EH2$) is chosen by this strategy. We also calculate AIC for comparison.

### 4.4.1 Plan for the Simulation Study

$2 \times 2 \times 2$ ANOVA models will be considered in our simulation study. We will study three different situations:

1. The true model is within the set of approximating models and is indeed the true best model;

2. The true model is within the set of approximating models but is NOT the true best model;

3. The true model is not within the set of approximating models.

In each of the above situations, our simulation plan is as follows:

1. For every approximating model $\mathcal{D}$,

(a) Calculate $EH2(\pi) = 2 - 2\int \sqrt{f_{\theta_\pi}(x)} E\sqrt{\phi_{\bar{\mu}_\pi^{\mathcal{D}},(\hat{\sigma}^{\mathcal{D}})^2}(x)}dt$ for all cells $\pi \in \Pi$ by Monte Carlo approximation, respectively;

(b) Find the average $EH2$ over all cells $EH2(\pi)$;

2. Find the true best model, denoted as $TBM$, by locating minimum $EH2$;

3. Simulate $S$ samples of a given sample size $n$ from the true distribution. For every sample $\underline{Y}_i$, $i = 1, \cdots, S$,

(a) Choose the best model $BM_i$ according to Procedure *Selection* which incorporates either Procedure *Group Identification I* or *Group Identification II* described in Section 4.3.1;

(b) Let $Match_i$ assume value 1 if $BM_i = TBM$ and 0 otherwise;

4. Table the frequency of $Match$ for each model.

The simulation setting is $2 \times 2 \times 2$ balanced design with cell sample size $n$ while the true distribution is a normal distribution with common variance for all cells. The cell means and variance are $\mu_{ijk}$ and $\sigma^2$, while the density for each cell is $\phi_{\mu_{ijk},\sigma}$, where $i, j, k = 1, 2$. The effects of the 3-way full factorial model are:

$$
\begin{array}{ll}
\alpha_i = \bar{\mu}_{i..} - \bar{\mu}_{...} & i = 1,2 \\
\beta_j = \bar{\mu}_{.j.} - \bar{\mu}_{...} & j = 1,2 \\
\gamma_k = \bar{\mu}_{..k} - \bar{\mu}_{...} & k = 1,2 \\
(\alpha\beta)_{ij} = \bar{\mu}_{ij.} - \bar{\mu}_{i..} - \bar{\mu}_{.j.} + \bar{\mu}_{...} & i,j = 1,2 \\
(\alpha\gamma)_{ik} = \bar{\mu}_{i.k} - \bar{\mu}_{i..} - \bar{\mu}_{..k} + \bar{\mu}_{...} & i,k = 1,2 \\
(\beta\gamma)_{jk} = \bar{\mu}_{.jk} - \bar{\mu}_{.j.} - \bar{\mu}_{..k} + \bar{\mu}_{...} & j,k = 1,2 \\
(\alpha\beta\gamma)_{ijk} = \mu_{ijk} - \bar{\mu}_{ij.} - \bar{\mu}_{i.k} - \bar{\mu}_{.jk} + \bar{\mu}_{i..} + \bar{\mu}_{.j.} + \bar{\mu}_{..k} - \bar{\mu}_{...} & i,j,k = 1,2
\end{array}
$$

We consider simulation from mean structures that reflect the 3 situations mentioned at the beginning of this section. All 19 approximating models we are considering here are listed in Table ??, where the effects included in each of the models are

checked accordingly. We will refer to the model index used in this Table. Figure 4.2 gives the relationship among all the approximating models.

| | $\emptyset$ | $\{F_1\}$ | $\{F_2\}$ | $\{F_3\}$ | $\{F_1, F_2\}$ | $\{F_1, F_3\}$ | $\{F_2, F_3\}$ | $\{F_1, F_2, F_3\}$ |
|---|---|---|---|---|---|---|---|---|
| Model 1 | X | | | | | | | |
| Model 2 | X | X | | | | | | |
| Model 3 | X | | X | | | | | |
| Model 4 | X | | | X | | | | |
| Model 5 | X | X | X | | | | | |
| Model 6 | X | X | | X | | | | |
| Model 7 | X | | X | X | | | | |
| Model 8 | X | X | X | | X | | | |
| Model 9 | X | X | | X | | X | | |
| Model 10 | X | | X | X | | | X | |
| Model 11 | X | X | X | X | | | | |
| Model 12 | X | X | X | X | X | | | |
| Model 13 | X | X | X | X | | X | | |
| Model 14 | X | X | X | X | | | X | |
| Model 15 | X | X | X | X | X | X | | |
| Model 16 | X | X | X | X | X | | X | |
| Model 17 | X | X | X | X | | X | X | |
| Model 18 | X | X | X | X | X | X | X | |
| Model 19 | X | X | X | X | X | X | X | X |

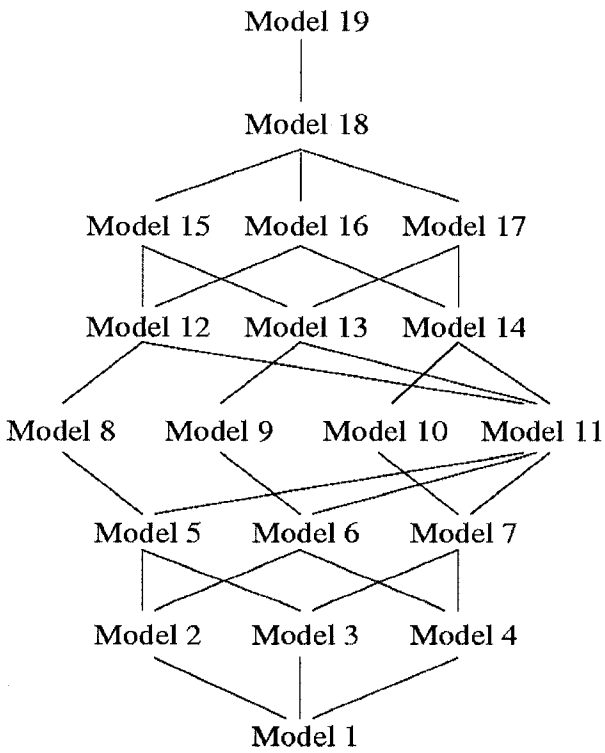


Figure 4.2: Lattice Diagram of 3-way ANOVA Models in Table ??

#### 4.4.1.1 Simulation Results

**Situation 1: Simulation from Model 8**

Data for each cell was independently simulated from normal distribution with cell mean structure described in Table 4.4:

| | $k=1$ | | $k=2$ | |
|---|---|---|---|---|
| | $j=1$ | $j=2$ | $j=1$ | $j=2$ |
| $i=1$ | $\mu$ | $\mu+B$ | $\mu$ | $\mu+B$ |
| $i=2$ | $\mu+A$ | $\mu+A+B+(AB)$ | $\mu+A$ | $\mu+A+B+(AB)$ |

Table 4.4: Mean Structure for Model 8

This mean structure corresponds to that of model 8 described in Section 5.4. Now, let $\mu = 0$, $\sigma = 1$, $A = 0.9$, $B = 0.4$, and $(AB) = 2.2$. Let the numbers of bootstrapped sample $M$ be 300 and $M2$ be 100, and the number of simulated data set $S$ be 100. The true best model is also Model 8.

From Table 4.5 we can see that the true best model is indeed the true model. Among the descendants of the true best model, the $EH2$ for model 5, which is contained in model 8, is the closest to that of the true best model. The true best model is chosen 75 times out of 100 by AIC, while the best performance of our methods, using procedure *Group Identification II* with $c = 1.5$, chooses the true best model 90 out of 100. In general, procedure *Group Identification II* outperforms procedure procedure *Group Identification I*, which is due to fact that the $PEEH2^*$'s tend to be screwed to the right and not symmetric around the corresponding $PEEH2$ when sample size is small. It is not surprising that when the confidence level or $c$ increases, model 5 is chosen more often. This is especially the case with procedure *Group Identification I.*

In order to assess the performance of $H(\hat{f}, g_{\hat{\theta}})$ in this situation, as mentioned in Chapter 3, the frequency of the time that any of the approximating models is

| | | Frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | PEEH2 | | | | | |
| | | | *Group Identification I* | | | *Group Identification II* | | |
| Model | $EH2$ | AIC | 50% CI | 80% CI | 95% CI | $c = 0.5$ | $c = 1.0$ | $c = 1.5$ |
| 1 | 0.32890 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.18986 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 3 | 0.26056 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.33039 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.08368 | 6 | 1 | 20 | 59 | 0 | 1 | 2 |
| 6 | 0.19161 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0.26212 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0.01684 | 75 | 79 | 78 | 40 | 73 | 89 | 90 |
| 9 | 0.19368 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0.26369 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0.08575 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | 0.01932 | 9 | 8 | 1 | 0 | 6 | 2 | 2 |
| 13 | 0.08815 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 0.08786 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 0.02221 | 9 | 3 | 0 | 0 | 6 | 2 | 1 |
| 16 | 0.02188 | 5 | 4 | 1 | 0 | 4 | 1 | 1 |
| 17 | 0.09026 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0.02477 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| 19 | 0.02826 | 4 | 4 | 0 | 0 | 11 | 5 | 4 |

Table 4.5: Frequency of Choosing Any Approximating Model by Different Methods

***Note:*** Data is simulated from model 8, $A = 0.9$, $B = 0.4$, and $(AB) = 2.2$. Cell sample size $n = 10$, numbers of bootstrapped samples $M = 300$ and $M2 = 100$, total number of simulated data $S = 100$. $EH2$ is approximated by Monte Carlo method.

chosen is also calculated. It turns out that the true best model, model 8, is never chosen out of the 100 times. The models that are chosen are model 16 (1 time), model 18 (5 times), and model 19 (94 times). The trend is clear that using $H(\hat{f}, g_{\hat{\theta}})$ as the model selection criterion will much more likely prefer larger models.

**Situation 2: Simulation from Model 19**

In this case, data comes from model 19 but for some arrangement of the parameters, the true best model, however, is not the true model itself. Cell mean structure of the model from which the data were simulated is described in Table 4.6:

| | $k=1$ | | $k=2$ | |
|---|---|---|---|---|
| | $j=1$ | $j=2$ | $j=1$ | $j=2$ |
| $i=1$ | $\mu$ | $\mu+B$ | $\mu$ | $\mu+B+C+(BC)$ |
| $i=2$ | $\mu+A$ | $\mu+A+B+(AB)$ | $\mu+A+C+(AC)$ | $\mu+A+B+C+$ $(AB)+(AC)+(BC)+(ABC)$ |

Table 4.6: Mean Structure for Model 19

| Model | $EH2$ | Frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | PEEH2 | | | | | |
| | | | *Group Identification I* | | | *Group Identification II* | | |
| | | AIC | 50% CI | 80% CI | 95% CI | $c = 0.5$ | $c = 1.0$ | $c = 1.5$ |
| $n = 10$ | | | | | | | | |
| 1 | 0.33298 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.23526 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.32509 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.20379 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 5 | 0.22483 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0.05185 | 0 | 12 | 47 | 86 | 1 | 7 | 17 |
| 7 | 0.19093 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0.22335 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0.04270 | 6 | 6 | 10 | 1 | 7 | 7 | 9 |
| 10 | 0.19299 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0.03253 | 9 | 32 | 34 | 12 | 16 | 30 | 29 |
| 12 | 0.03024 | 6 | 8 | 3 | 0 | 9 | 10 | 8 |
| 13 | 0.02318 | 16 | 24 | 5 | 1 | 20 | 21 | 20 |
| 14 | 0.03585 | 2 | 1 | 0 | 0 | 0 | 0 | 1 |
| 15 | 0.02103 | 39 | 7 | 1 | 0 | 28 | 11 | 10 |
| 16 | 0.03368 | 2 | 0 | 0 | 0 | 1 | 0 | 1 |
| 17 | 0.02664 | 4 | 0 | 0 | 0 | 3 | 3 | 0 |
| 18 | 0.02459 | 5 | 4 | 0 | 0 | 4 | 2 | 1 |
| 19 | 0.02884 | 11 | 6 | 0 | 0 | 11 | 9 | 4 |
| $n = 100$ | | | | | | | | |
| 1 | 0.33044 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.23006 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.32095 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.19878 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.21795 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 0.04170 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0.18398 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0.21469 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0.02979 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0.18418 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0.01976 | 0 | 0 | 0 | 23 | 0 | 0 | 0 |
| 12 | 0.01443 | 0 | 0 | 0 | 4 | 0 | 0 | 0 |
| 13 | 0.00737 | 0 | 5 | 60 | 65 | 3 | 4 | 11 |
| 14 | 0.02005 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | 0.00193 | 80 | 94 | 40 | 8 | 86 | 94 | 88 |
| 16 | 0.01472 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | 0.00768 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | 0.00224 | 11 | 1 | 0 | 0 | 6 | 2 | 1 |
| 19 | 0.00257 | 9 | 0 | 0 | 0 | 5 | 0 | 0 |

Table 4.7: Frequency of Choosing Any Approximating Model by Different Methods

***Note:*** Data was simulated from model 19, $A = 1$, $B = 0.3$, $C = 1.5$, $(AB) = 0.6$, $(AC) = 0.9$, $(BC) = 0.01$, and $(ABC) = 0.015$. Cell sample size $n = 10, 100$, numbers of bootstrapped samples $M = 300$ and $M2 = 100$, total number of simulated data $S = 100$.

Now, let $\mu = 0$, $\sigma = 1$, $A = 1$, $B = 0.3$, $C = 1.5$, $(AB) = 0.6$, $(AC) = 0.9$, $(BC) = 0.01$, and $(ABC) = 0.015$. Let the numbers of bootstrapped sample $M$ be 300 and $M2$ be 100, cell sample size be 10 and 100 and the number of simulated data set $S$ be 100. For this particular set of parameters, the true best model that has the smallest $EH2$ is model 15 rather than the true model itself for both sample sizes. Results of the simulation are in Table 4.7. When cell sample size is 10, none of the methods performs well. When cell sample size is 100, the performance of all methods improves a lot. In particular, when cell sample size is 100, method *Group Identification II* with confidence level 50% and method *Group Identification II* with all three choices of $c$ outperform AIC.

**Situation 3: Simulation from Model 4**

In this situation, the data is simulated from model 4, which is not one of the approximating models. We choose the approximating models to be models 1, 3, 7, 11, 14, 17, 18 and 19. These approximating models form a path from the full model (model 19) to the null model (model 1) that does not include the true model (model 4). On this path, the nearest ancestor of the true model is model 7. That is, model 7 has the least number of effects among the approximating models that contains the true model. Table 4.8 describes the cell mean structure of the true model.

| | $k=1$ | | $k=2$ | |
|---|---|---|---|---|
| | $j=1$ | $j=2$ | $j=1$ | $j=2$ |
| $i=1$ | $\mu$ | $\mu$ | $\mu+C$ | $\mu+C$ |
| $i=2$ | $\mu$ | $\mu$ | $\mu+C$ | $\mu+C$ |

Table 4.8: Mean Structure for Model 4

| Model | $EH2$ | Frequency | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | PEEH2 | | | | | |
| | | | *Group Identification I* | | | *Group Identification II* | | |
| | | AIC | 50% CI | 80% CI | 95% CI | $c = 0.5$ | $c = 1.0$ | $c = 1.5$ |
| n=10 | | | | | | | | |
| 1 | 0.13385 | 0 | 0 | 0 | 5 | 0 | 0 | 0 |
| 3 | 0.13580 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0.01155 | 65 | 69 | 97 | 95 | 63 | 83 | 93 |
| 11 | 0.01421 | 15 | 9 | 2 | 0 | 11 | 8 | 3 |
| 14 | 0.01703 | 7 | 8 | 1 | 0 | 8 | 2 | 1 |
| 17 | 0.02019 | 6 | 4 | 0 | 0 | 7 | 2 | 0 |
| 18 | 0.02290 | 4 | 7 | 0 | 0 | 6 | 2 | 2 |
| 19 | 0.02613 | 3 | 3 | 0 | 0 | 5 | 3 | 1 |
| n=100 | | | | | | | | |
| 1 | 0.13051 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0.13074 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0.00121 | 68 | 94 | 100 | 100 | 80 | 89 | 97 |
| 11 | 0.00148 | 10 | 4 | 0 | 0 | 2 | 0 | 2 |
| 14 | 0.00182 | 11 | 0 | 0 | 0 | 5 | 7 | 1 |
| 17 | 0.00211 | 2 | 2 | 0 | 0 | 4 | 0 | 0 |
| 18 | 0.00235 | 6 | 0 | 0 | 0 | 2 | 0 | 0 |
| 19 | 0.00274 | 3 | 0 | 0 | 0 | 7 | 4 | 0 |

Table 4.9: Frequency of Choosing Any Approximating Model by Different Methods

***Note:*** Data is simulated from model 4, $C = 1.5$. Cell sample size $n = 10, 100$, numbers of bootstrapped samples $M = 300$ and $M2 = 100$, total number of simulated data $S = 100$. $EH2$ is approximated by Monte Carlo method.

Now, let $\mu = 0$, $\sigma = 1$, $C = 1.5$. Let the numbers of bootstrapped samples $M$ be 300 and $M2$ be 100. Let the number of simulated data sets $S$ be 100. Two different cell sample sizes, 10 and 100, are considered. From Table 4.9 we can see that the true best model is model 7 for both sample sizes. Note that model 1 and 3, the only descendants of the true model among the approximating models, are quite far away from the target group in terms of $EH2$. That explains why higher confidence level and larger $c$ result in model 7 being more often chosen in

this case. The performance of both our procedures exceeds that of AIC, with that of procedure *Group Identification I* being the best. The performance of both AIC and our methods improves as the sample size increases.

#### 4.4.1.2 Discussion

In each situation considered in Section 4.4.1.1, the performance of our proposed model selection strategy with both grouping procedures improves as the cell sample size increases. Within each grouping procedure, as the confidence level or $c$ increases, our model selection strategy tends to more often choose the smallest model among the approximating models for which the $EH2$ are not far away from the smallest. Between the grouping procedures, *Group Identification I* with the three prespecified confidence levels tends to favor the smallest models among the approximating models for which the $EH2$ are not far away from the smallest. In two of the three simulation studies, procedure *Group Identification II* outperforms procedure *Group Identification I*, especially when sample size is small. It is not surprising that in the last situation, where the smaller approximating models outside of the target group have relatively very large $EH2$'s, procedure *Group Identification I* has better performance. Note that, procedure *Group Identification II* also gives satisfactory performance in the last situation.

## 4.5 Simulation Study: Convergence of $PEEH2$ to $EH2$

In this section, we check Theorem 3 in Chapter 2 using simulation under the framework of balanced factorial models. Since $EEH2B$ is not practical to be obtained exactly, we will calculate $PEEH2$ instead and let the cell sample size and bootstrap sample size both increase in $2 \times 2$ ANOVA model setting.

#### 4.5.0.3 General Plan of The Simulation

Continued from the setting described at the beginning of this chapter, we consider simulation from mean structures that represent possible $2 \times 2$ factorial ANOVA models with cell sample size $n$. Table 4.10 displays the mean structures for each cell associated with each of the 5 models. It can be shown that for any given cell, the Hellinger distances between the true model and the approximating models are independent of the parameters $\mu$ and $\sigma$ themselves. Therefore, without loss of generality, we let $\mu = 0$ and $\sigma = 1$. Table 4.11 shows the values of $A$, $B$, and $(AB)$ we choose for each model that we simulate data from. All together, we simulate data from 1 null model, 4 column effect models, 9 main effects models, and 6 full models.

| | Cell | | | |
|---|---|---|---|---|
| | $(1,1)$ | $(2,1)$ | $(1,2)$ | $(2,2)$ |
| Null Model | $\mu$ | $\mu$ | $\mu$ | $\mu$ |
| Row Effect Model | $\mu$ | $\mu + A$ | $\mu$ | $\mu + A$ |
| Column Effect Model | $\mu$ | $\mu$ | $\mu + B$ | $\mu + B$ |
| Main Effects Model | $\mu$ | $\mu + A$ | $\mu + B$ | $\mu + A + B$ |
| Full Model | $\mu$ | $\mu + A$ | $\mu + B$ | $\mu + A + B + (AB)$ |

Table 4.10: Models That Data Are Simulated From

#### 4.5.0.4 Simulation Results

The simulation setting is as follows:

- The approximating models are all 5 possible models described in Section 4.2.1;
- $S = 100$ data sets were generated from the true model;
- Let $h$, the bandwidth of the kernel density estimator be a function of $n$ only, namely, $h = (10n)^{-0.2}$;

| | | $A/\sigma$ | $B/\sigma$ | $(AB)/\sigma$ |
|---|---|---|---|---|
| Null Model | | | | |
| | Model 1 | 0 | 0 | 0 |
| Column Effect Model | | | | |
| | Model 1 | 0 | 0.1 | 0 |
| | Model 2 | 0 | 0.5 | 0 |
| | Model 3 | 0 | 1.0 | 0 |
| | Model 4 | 0 | 2.0 | 0 |
| | Model 5 | 0 | 5.0 | 0 |
| Main Effects Model | | | | |
| | Model 1 | 0.1 | 0.1 | 0 |
| | Model 2 | 0.1 | 1.0 | 0 |
| | Model 3 | 0.1 | 2.0 | 0 |
| | Model 4 | 1.0 | 0.1 | 0 |
| | Model 5 | 1.0 | 1.0 | 0 |
| | Model 6 | 1.0 | 2.0 | 0 |
| | Model 7 | 2.0 | 0.1 | 0 |
| | Model 8 | 2.0 | 1.0 | 0 |
| | Model 9 | 2.0 | 2.0 | 0 |
| Full Model | | | | |
| | Model 1 | -1 | -3 | 7 |
| | Model 2 | -3 | 1 | 1 |
| | Model 3 | 0 | 0 | -3 |
| | Model 4 | -3 | 3 | 1 |
| | Model 5 | 1 | 0 | -2 |
| | Model 6 | 3 | 1 | -4 |

Table 4.11: Choices of the Simulation Parameters

- Let the increasing sequences of cell sample size $n$ and number of Bootstrapped samples $M$ be $n = 100, 500, 1000$ and $M = 300, 1500, 3000$, respectively.

**True Model: Null** Each plot in Figure 4.3 depicts the trend of $PEEH2$ values relative to $EH2$ values as both $n$ and $M$ increase for one of the 5 approximating models, respectively. 100 $PEEH2$ values as well as their mean and standard deviations are plotted for $n = 100$ & $M = 300$, $n = 500$ & $M = 1500$, and $n = 1000$ & $M = 3000$. The $EH2$ values were also plotted in red for every cell sample size.

In the plots, the numbers 1 through 3 on the horizontal axis represents $n = 100$ $(M = 300)$, $n = 500$ $(M = 1500)$, $n = 1000$ $(M = 3000)$, respectively.

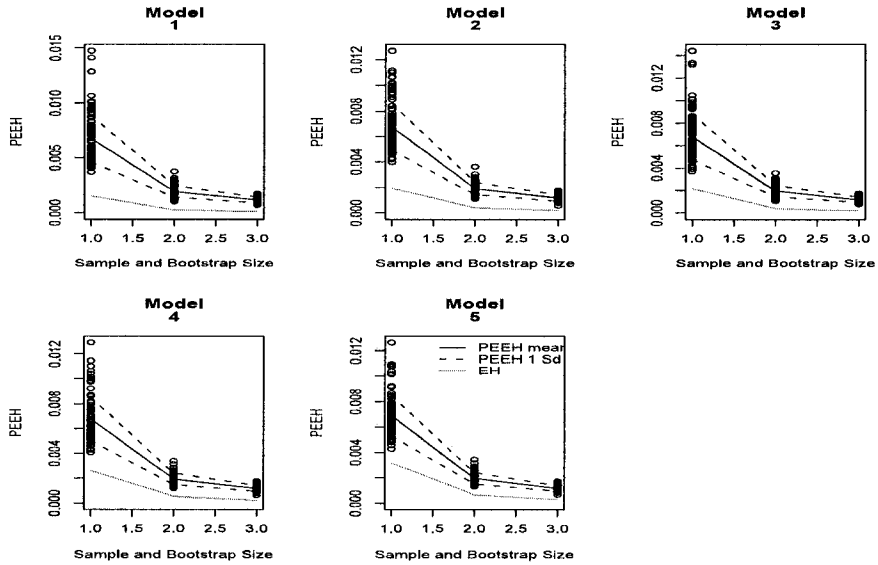


Figure 4.3: Convergence of $PEEH2$ to $EH2$ True Model Null

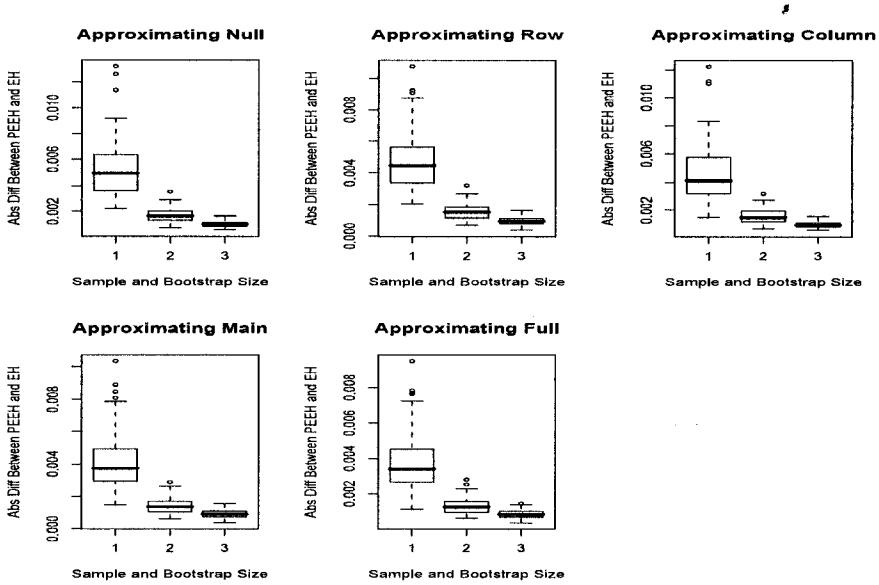


Figure 4.4: Boxplots of the Absolute Difference Between $PEEH2$ and $EH2$ True Model Null

Multiple boxplots of the absolute differences between $PEEH2$ and $EH2$ are plotted for each approximating model in Figure 4.4 where the horizontal axis indexes

the sample size/bootstrap size. We can see that the difference is getting close to zero while the variation is decreasing as the sample size and the number of bootstrap samples increase. Note that all the approximating models contain the true model and therefore their $EH2$'s are relatively small and close to one another. As shown in Table 4.12, the $EH2$ value for the null model is the smallest and increases as the the number of parameters increases for any given sample size. Meanwhile, the $EH2$ value between the true model and a given approximating model decreases as the sample size increases. For any given approximating model, the mean difference between $EH2$ and $PEEH2$ decreases as sample size and bootstrap size increases. For any given sample size, the $PEEH2$ values for different approximating models are close to one another, with the true best model - the Null model - not necessarily having the smallest mean $PEEH2$.

| | $EH2$ | | | | |
|---|---|---|---|---|---|
| | Null | Row | Column | Main | Full |
| $n = 100$ | 0.001498 | 0.001933 | 0.002165 | 0.002599 | 0.003164 |
| $n = 500$ | 0.000255 | 0.000396 | 0.000387 | 0.000527 | 0.000652 |
| $n = 1000$ | 0.000115 | 0.000180 | 0.000172 | 0.000237 | 0.000290 |
| | Mean of 100 $PEEH2$'s | | | | |
| | Null | Row | Column | Main | Full |
| $n = 100$ | 0.006736 | 0.006741 | 0.006771 | 0.006779 | 0.006901 |
| $n = 500$ | 0.001934 | 0.001957 | 0.001969 | 0.001939 | 0.001954 |
| $n = 1000$ | 0.001158 | 0.001160 | 0.001165 | 0.001167 | 0.001164 |

Table 4.12: $PEEH2$ and $EH2$ – True Model Null

**True Model: Column Effect** Each plot in Figures 4.5 through 4.8 depict the trend of $PEEH2$ values relative to $EH2$ values as both $n$ and $M$ increase for one of the 5 approximating models when the true models that the data are simulated from are models 1 through 5 described in Table 4.11 under column effect models, respectively. In the plots, the numbers 1 through 3 on the horizontal axis represent $n = 100$ ($M = 300$), $n = 500$ ($M = 1500$), and $n = 1000$ ($M = 3000$), respectively. It is shown from the figures that the sample mean of $PEEH2$ values does get close to $EH2$ while the variation of those decreases as the cell sample size and number of bootstrapped samples increase. Since the true model is the column effect model, we can see that the patterns of the models that do ont contain the true model (null model and row effect model) are alike while that of the other three models are similar.

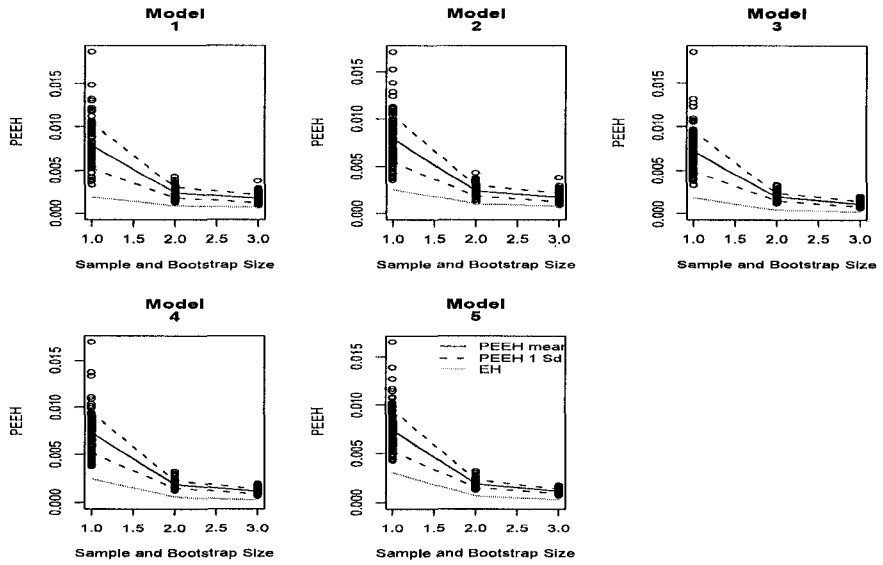


Figure 4.5: Convergence of $PEEH2$ to $EH2$ True Model Column Effect Model 1

Let us take a look at the first model that we simulated data from. The difference between the two cell means of the two columns is very small (0.1) relative to the

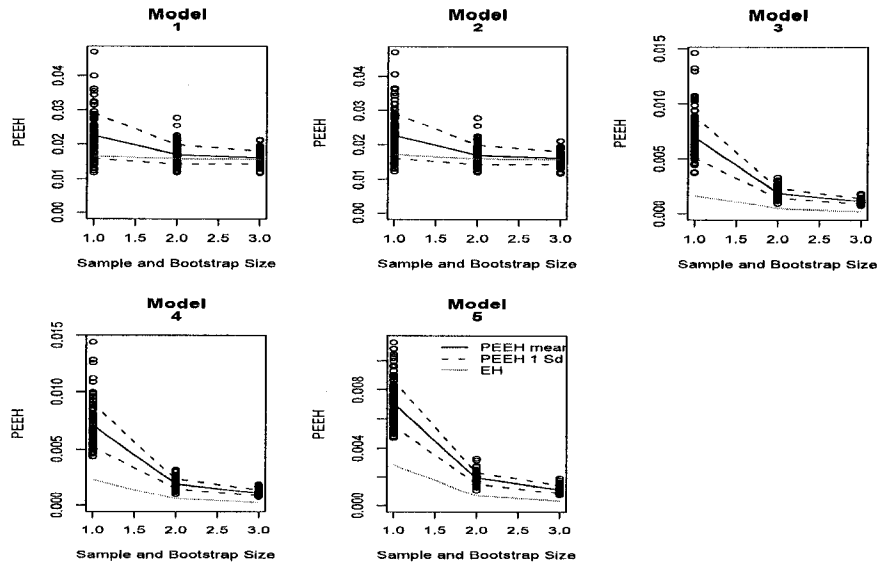


Figure 4.6: Convergence of $PEEH2$ to $EH2$ True Model Column Effect Model 2

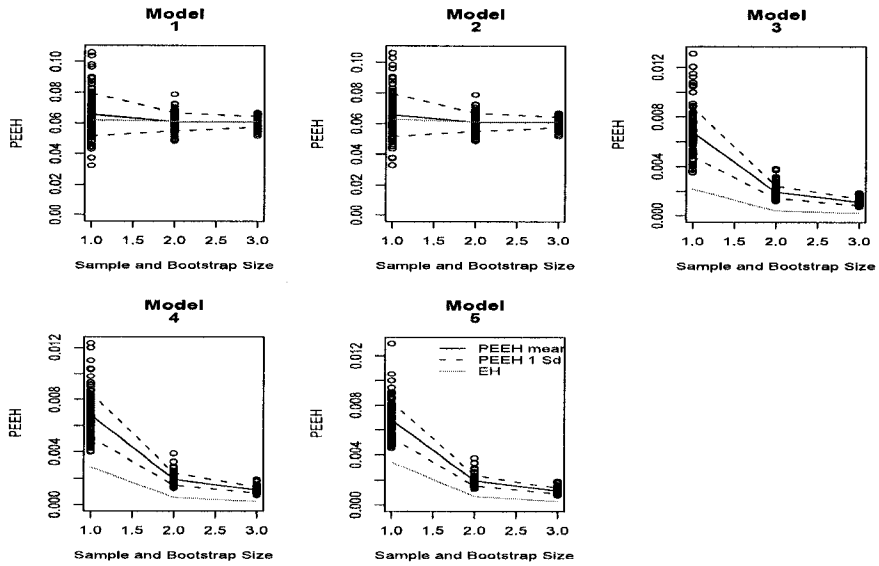


Figure 4.7: Convergence of $PEEH2$ to $EH2$ True Model Column Effect Model 3

variance of 1. This makes it hard to distinguish among the approximating models. Table 4.13 lists the $EH2$ values and the mean of 100 $PEEH2$ values for the 5 approximating models for the three sample sizes, respectively. We can see that as a result of this setting of parameters, $EH2$ values are relatively close to one another among the approximating models for any given sample size. The differences between

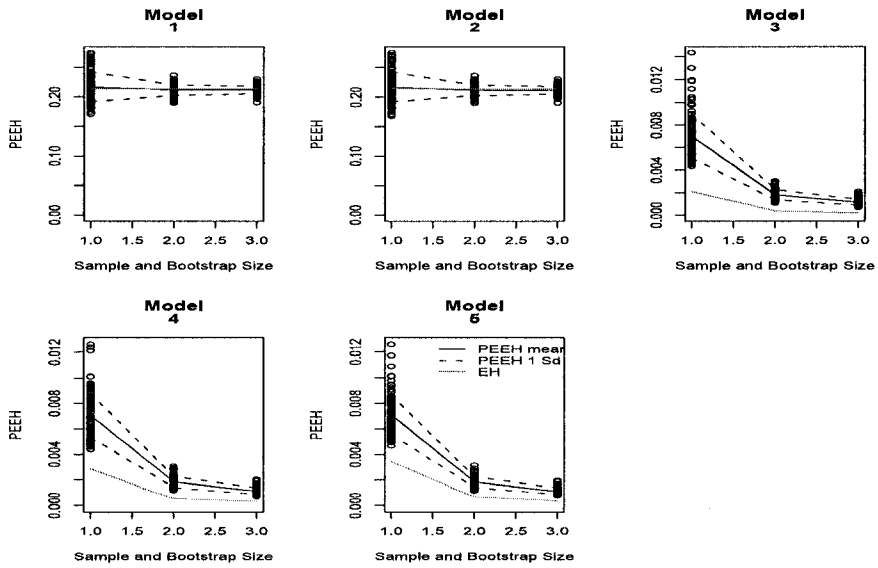


Figure 4.8: Convergence of $PEEH2$ to $EH2$ True Model Column Effect Model 4

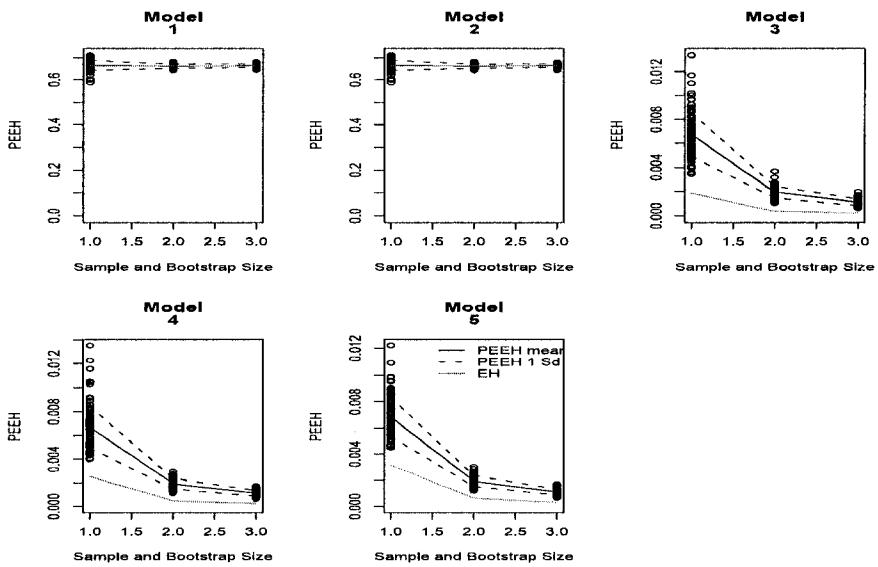


Figure 4.9: Convergence of $PEEH2$ to $EH2$ True Model Column Effect Model 5

$EH2$ and $PEEH2$ for any given approximating model decreases as the sample size and bootstrap size increase.

**True Model: Main Effects** Each plot in Figures 4.10 through 4.18 depicts the trend of $PEEH2$ values relative to $EH2$ values as both $n$ and $M$ increase for one of the 5 approximating models when the true models that the data are simulated

| | $EH2$ | | | | |
|---|---|---|---|---|---|
| | Null | Row | Column | Main | Full |
| $n = 100$ | 0.001862 | 0.002544 | 0.001762 | 0.002446 | 0.002995 |
| $n = 500$ | 0.000886 | 0.001022 | 0.0004 | 0.000537 | 0.000647 |
| $n = 1000$ | 0.000739 | 0.000810 | 0.000176 | 0.000246 | 0.000306 |
| | Mean of 100 $PEEH2$'s | | | | |
| | Null | Row | Column | Main | Full |
| $n = 100$ | 0.007877 | 0.007931 | 0.007205 | 0.007260 | 0.007347 |
| $n = 500$ | 0.002443 | 0.002416 | 0.001902 | 0.001875 | 0.001904 |
| $n = 1000$ | 0.001702 | 0.001713 | 0.001115 | 0.001126 | 0.001120 |

Table 4.13: $PEEH2$ and $EH2$ Values For True Model Column Effect and The Column Difference B=0.1

from are models 1 through 9 described in Table 4.11 under main effects models, respectively. In the plots, the numbers 1 through 3 on the horizontal axis represents $n = 100$ ($M = 300$), $n = 500$ ($M = 1500$), and $n = 1000$ ($M = 3000$), respectively.

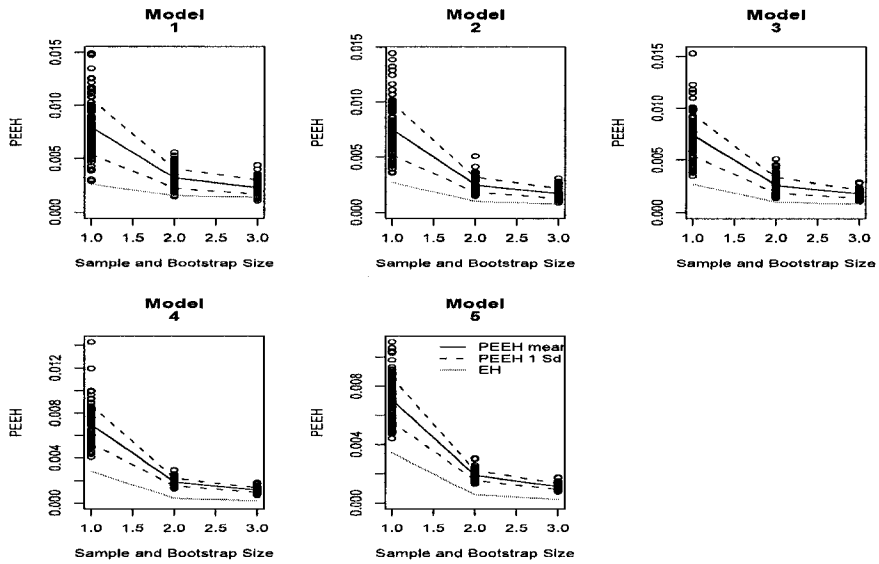


Figure 4.10: Convergence of $PEEH2$ to $EH2$ True Model Main Effects Model 1

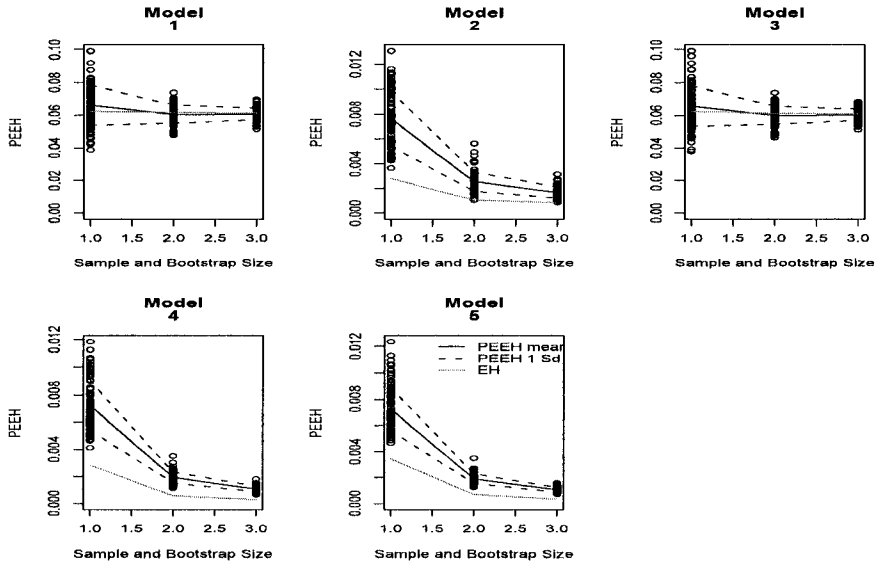


Figure 4.11: Convergence of $PEEH2$ to $EH2$ True Model Main Effects Model 2

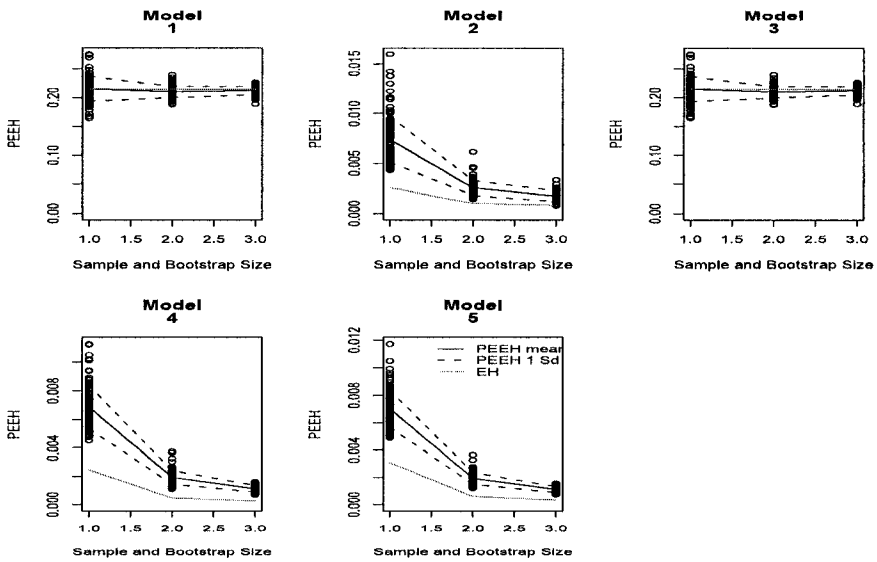


Figure 4.12: Convergence of $PEEH2$ to $EH2$ True Model Main Effects Model 3

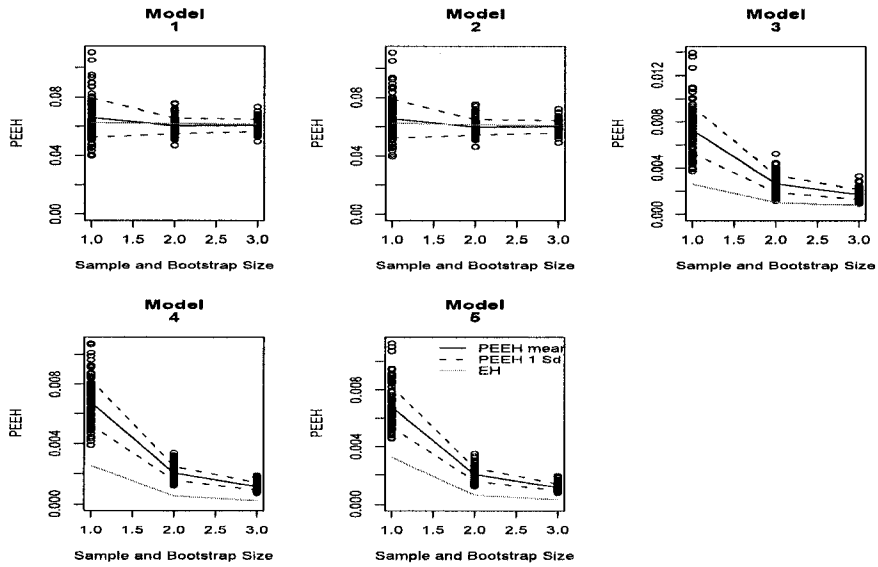


Figure 4.13: Convergence of $PEEH2$ to $EH2$ True Model Main Effects Model 4

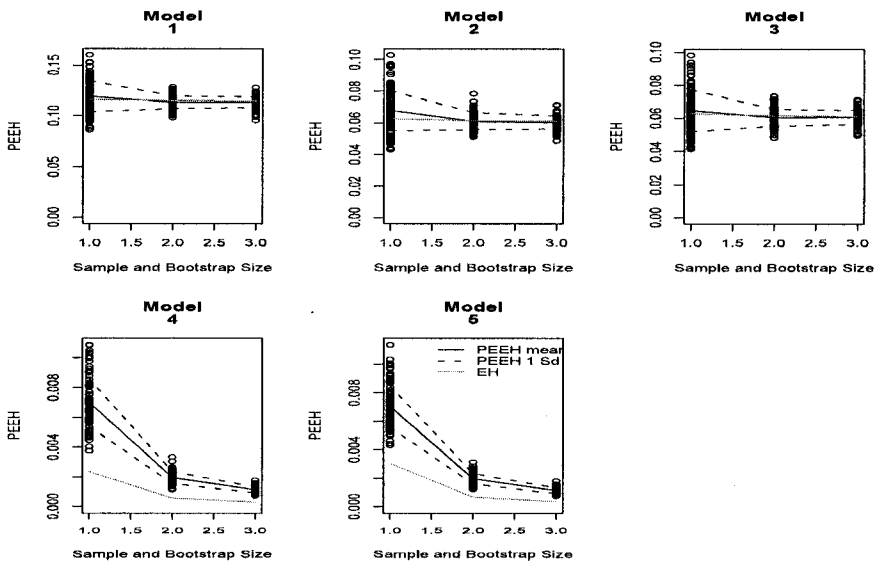


Figure 4.14: Convergence of $PEEH2$ to $EH2$ True Model Main Effects Model 5

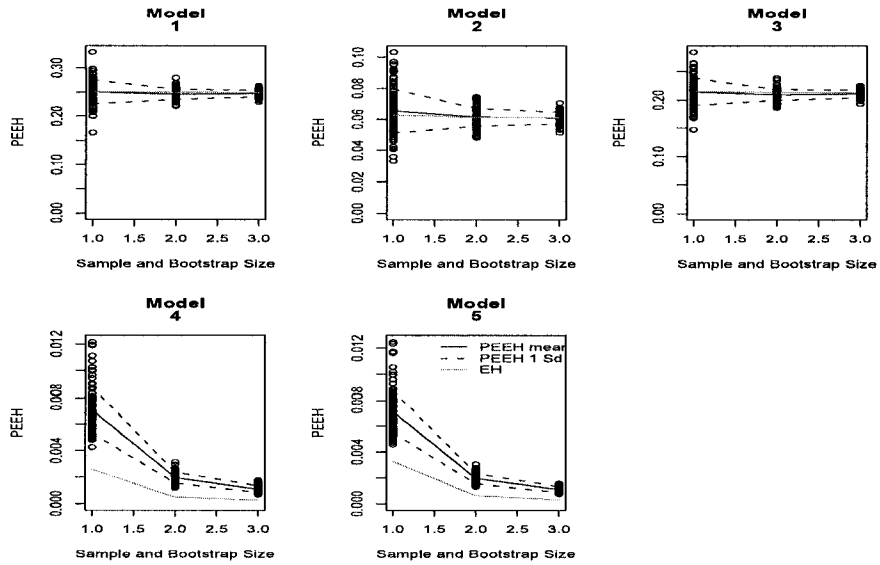


Figure 4.15: Convergence of $PEEH2$ to $EH2$ True Model Main Effects Model 6

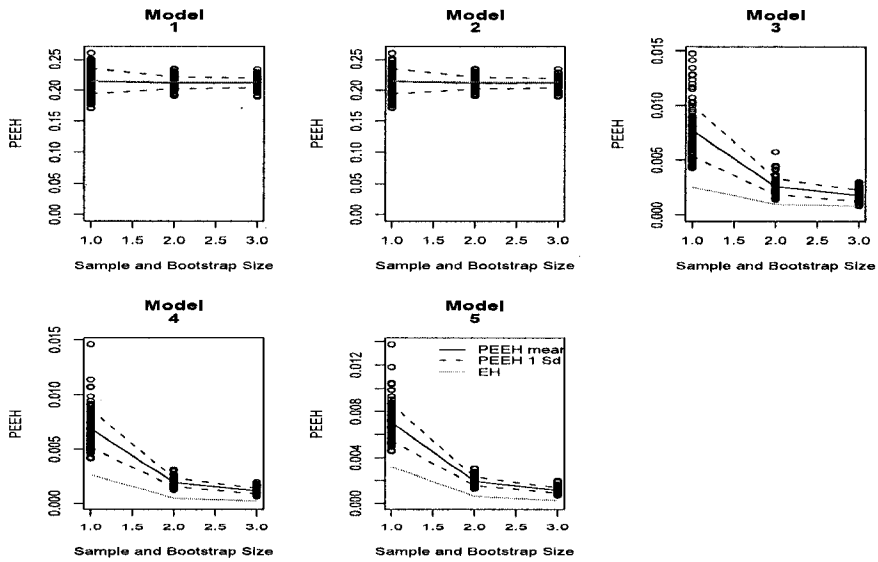


Figure 4.16: Convergence of $PEEH2$ to $EH2$ True Model Main Effects Model 7

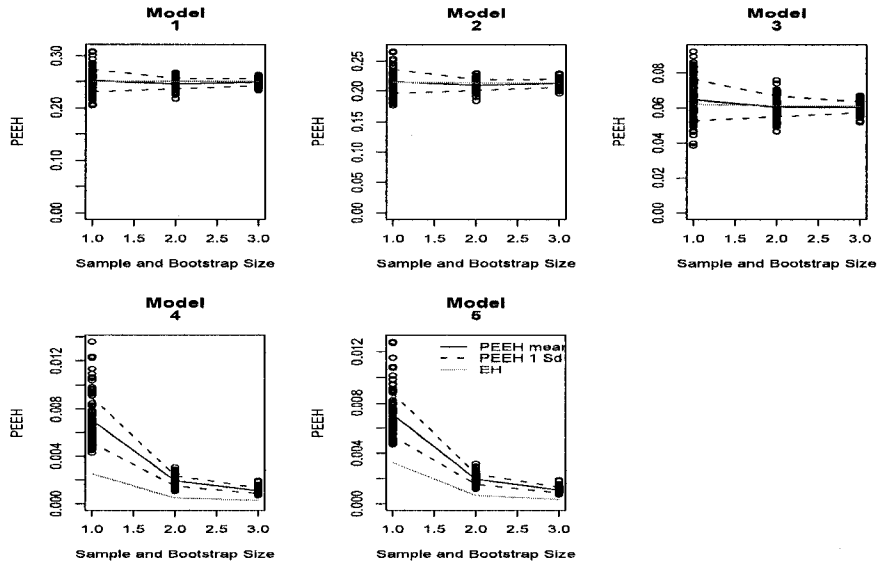


Figure 4.17: Convergence of $PEEH2$ to $EH2$ True Model Main Effects Model 8

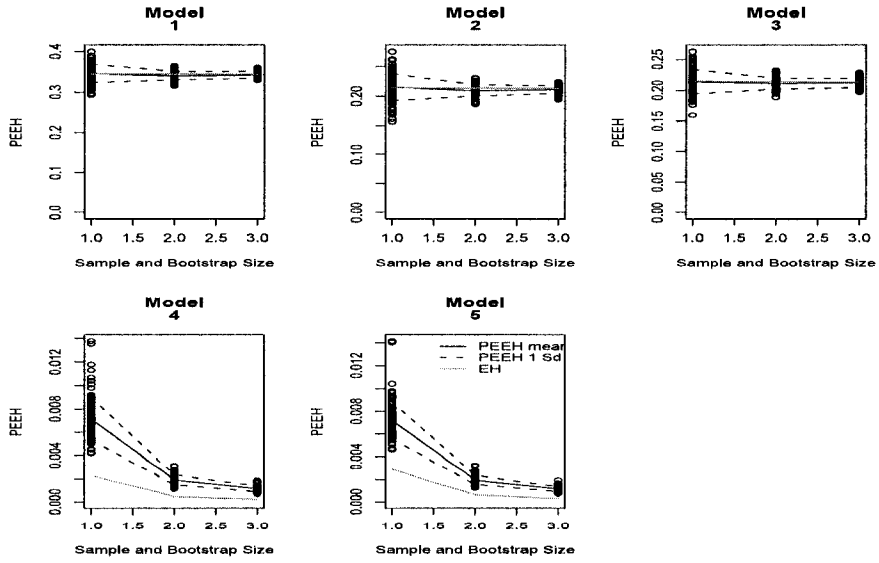


Figure 4.18: Convergence of $PEEH2$ to $EH2$ True Model Main Effects Model 9

Now let us take a look at two cases. One is that when the data is simulated from the third model listed in Table 4.11 under main effects models. The multiple boxplots of the absolute differences between $EH2$ and $PEEH2$ values for each approximating model are plotted in Figure 4.19. Again, the horizontal axis indexes the sample size/bootstrap size. Once again we can see the trend of the differences getting close to zero and the variation decreasing as the sample size and number of bootstrap sample increase.

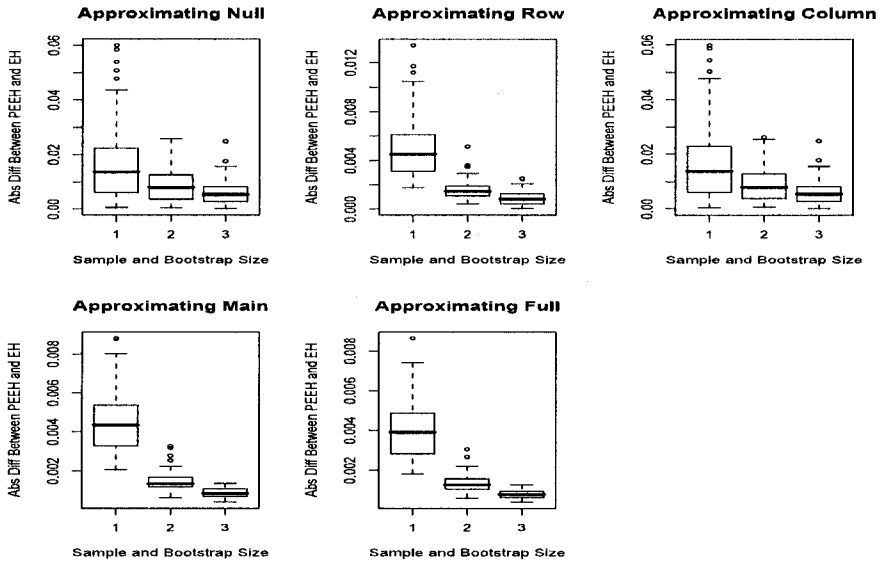


Figure 4.19: Boxplots of the Absolute Differences Between $EH2$ and $PEEH2$ – True Model Main Effects Model 3 and All Approximating Models

Another particular case is when the approximating model is the null model, Figure 4.20 plots the absolute differences between $EH2$ and $PEEH2$ values for each of the 9 true main effects models listed in Table 4.11. The same convergence trend is observed across all true models from which the data is simulated.

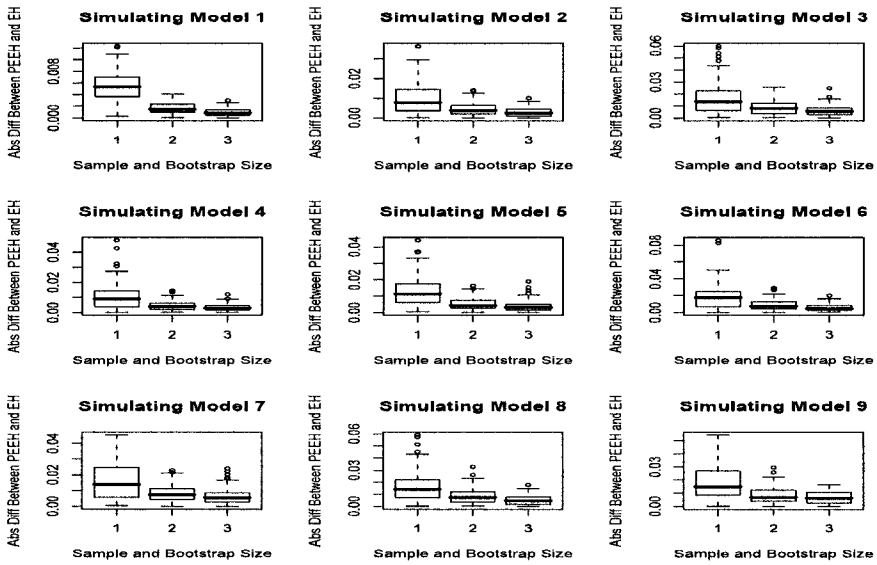


Figure 4.20: Boxplots of the Absolute Differences Between $EH2$ and $PEEH2$ – True Model All 9 Main Effects Models and Approximating Model Null

**True Model: Full** Each plot in Figures 4.21 through 4.26 depicts the trend of $PEEH2$ values relative to $EH2$ values as both $n$ and $M$ increase for one of the 5 approximating models when the true models that the data are simulated from are models 1 through 6 described in Table 4.11 under full models, respectively. In the plots, the numbers 1 through 3 on the horizontal axis represents $n = 100$ ($M = 300$), $n = 500$ ($M = 1500$), and $n = 1000$ ($M = 3000$), respectively.

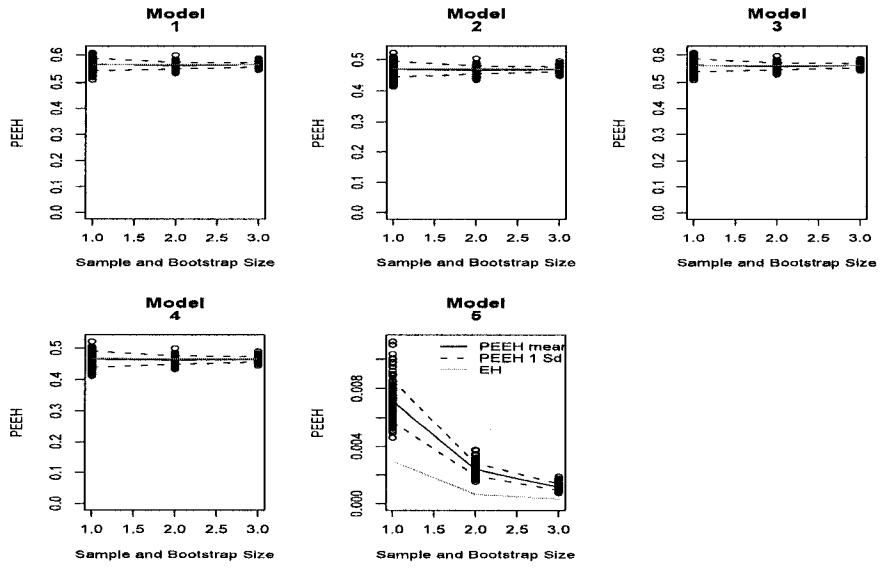


Figure 4.21: Convergence of $PEEH2$ to $EH2$ True Model Full Model 1

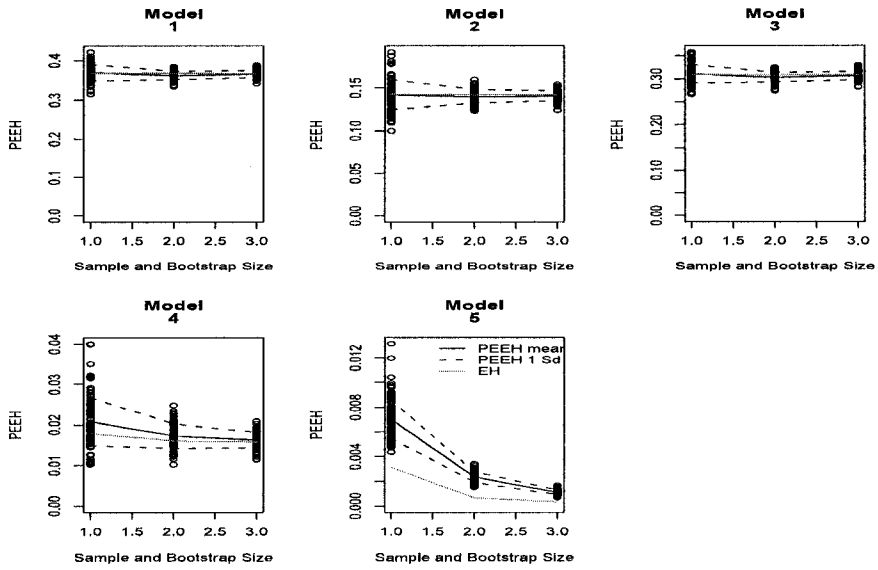


Figure 4.22: Convergence of $PEEH2$ to $EH2$ True Model Full Model 2

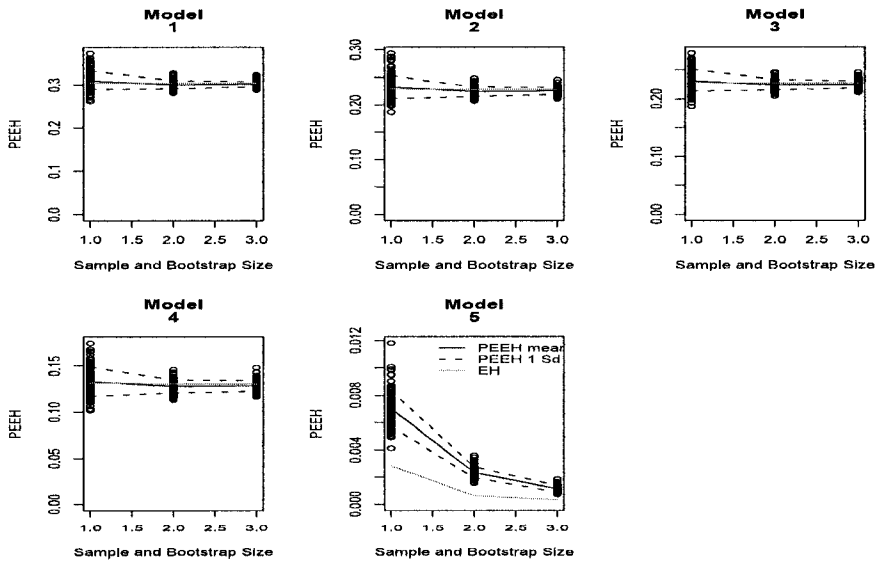


Figure 4.23: Convergence of $PEEH2$ to $EH2$ True Model Full Model 3

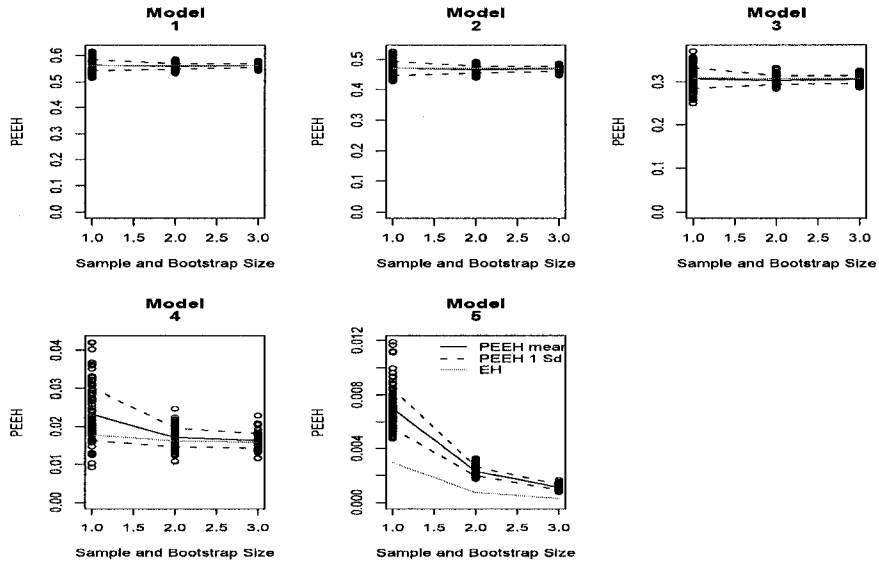


Figure 4.24: Convergence of $PEEH2$ to $EH2$ True Model Full Model 4

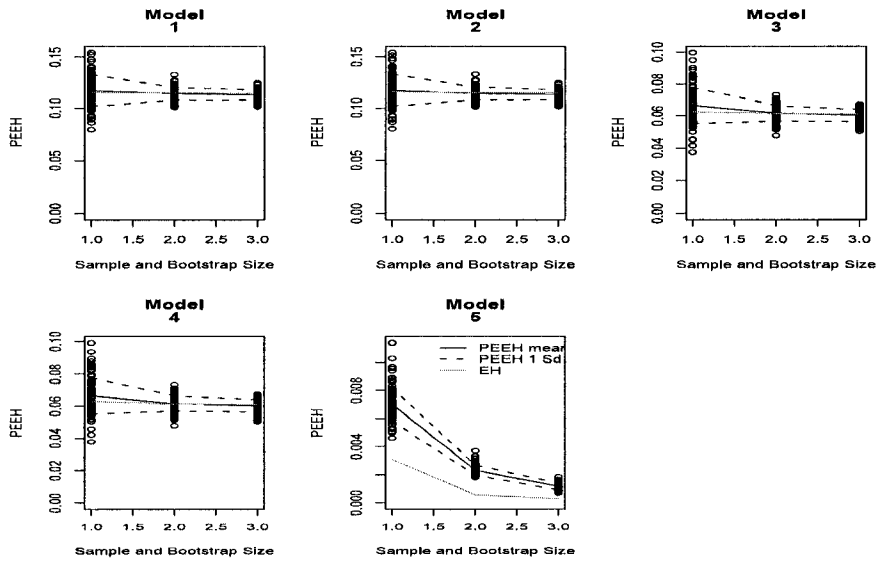


Figure 4.25: Convergence of $PEEH2$ to $EH2$ True Model Full Model 5

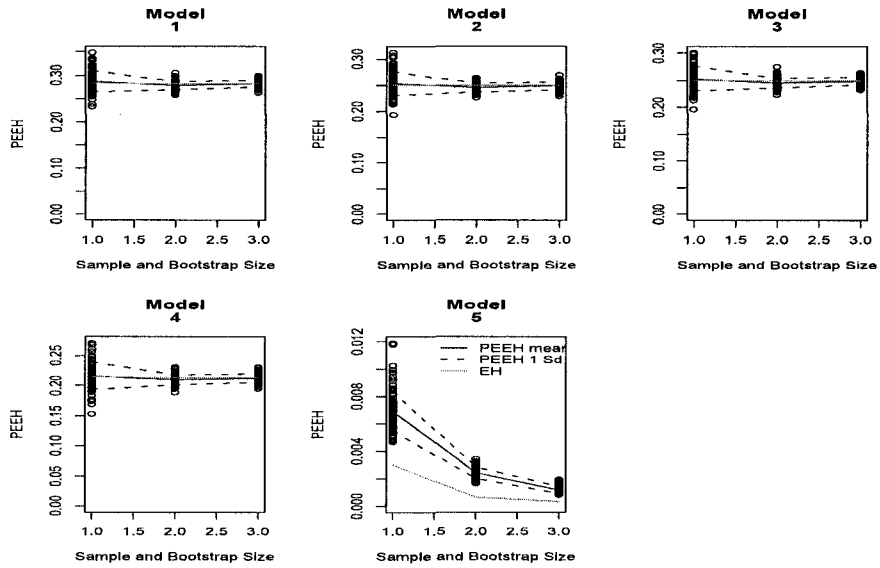


Figure 4.26: Convergence of $PEEH2$ to $EH2$ True Model Full Model 6

In summary, both mean and variance of the difference between $EH2$ and its estimator $PEEH2$ values decrease as both sample size and bootstrap size increase. The tendency of this difference converging to zero is observed.

## 4.6 Illustrative Example

Use a $2 \times 2 \times 2$ example from page 943 in "Applied Linear Statistical Model" (Neter et al, 1996). The effects of gender of subject (factor $F_1$), body fat of subject (measured in percent, factor $F_2$), and smoking history of subject (factor $F_3$) on exercise tolerance (Y) were studied in a small-scale investigation of persons 25 to 35 years old. Exercise tolerance was measured in minutes until fatigue occurs while the subject is performing on a bicycle apparatus. Three subjects for each gender-body fat-smoking history group were given the exercise tolerance stress test. Each factor has two levels and there are three replications (n=3) for each treatment. The data and the ANOVA table are displayed in Table 4.14 and 4.15, respectively.

| Gender | Body Fat | Smoking History | Data | | |
|---|---|---|---|---|---|
| Male | Low | Light | 24.1 | 29.2 | 24.6 |
| | | Heavy | 17.6 | 18.8 | 23.3 |
| | High | Light | 14.6 | 15.3 | 12.3 |
| | | Heavy | 14.9 | 20.4 | 12.8 |
| Female | Low | Light | 20.0 | 21.9 | 17.6 |
| | | Heavy | 14.8 | 10.3 | 11.3 |
| | High | Light | 16.1 | 9.3 | 10.8 |
| | | Heavy | 10.1 | 14.4 | 6.1 |

Table 4.14: 3-way ANOVA Example Data

We can see from Table 4.15 that all three main effects and one interaction term between Fat and Smoking are significant at a significance level of 0.05, which suggests model 14 listed in Table ??.

| Analysis of Variance | | | | | |
|---|---|---|---|---|---|
| Source | Sum Of Squares | DF | Mean Square | F Ratio | P Value |
| Gender | 176.584 | 1 | 176.584 | 18.915 | 0.000 |
| Fat | 242.570 | 1 | 242.570 | 25.984 | 0.000 |
| Smoking | 70.384 | 1 | 70.384 | 7.539 | 0.014 |
| Gender*Fat | 13.650 | 1 | 13.650 | 1.462 | 0.244 |
| Gender*Smoking | 11.070 | 1 | 11.070 | 1.186 | 0.292 |
| Fat*Smoking | 72.454 | 1 | 72.454 | 7.761 | 0.013 |
| Gender*Fat*Smoking | 1.870 | 1 | 1.870 | 0.200 | 0.660 |
| Error | 149.367 | 16 | 9.335 | | |

Table 4.15: 3-way ANOVA Example Analysis of Variance

All the approximating models are listed in Table ??. It is assumed that all random variables $Y_{ijkr}$ are independently and identically distributed with only different location parameters. Therefore, we can first find kernel density estimator $\hat{f}$ using all the residuals $res_{ijkr} = y_{ijkr} - \bar{y}_{ijk.}$ and then adding back $\bar{y}_{ijk.}$ to the grid to get $\hat{f}(t - \bar{y}_{ijk.})$. The approximating models for the $(ijk)^{th}$ cell are $\mathrm{N}(\hat{\mu}_{ijk}, \hat{\sigma}^2)$ with $\hat{\mu}_{ijk}$ being estimated sample mean reflecting the approximating model and $\hat{\sigma}^2$ being the MSE associated with the model.

Let $M$ be 500, the kernel be Epanechnikov and the bandwidth selected according to the the statistical software package R (R Development Core Team, 2006), default "to 0.9 times the minimum of the standard deviation and the interquartile range divided by 1.34 times the sample size to the negative one-fifth power (Silverman's "rule of thumb", Silverman (1986, page 48, eqn (3.31)) unless the quartiles coincide when a positive result will be guaranteed."

The $PEEH2$'s of all the approximating models are listed in the parenthesis in Figure 4.27. As discussed before, model 14 might be a naturally choice according to the analysis of variance table. Model 14 is chosen as the best model according to our model selection method based on procedure *Selection* incorporated with either

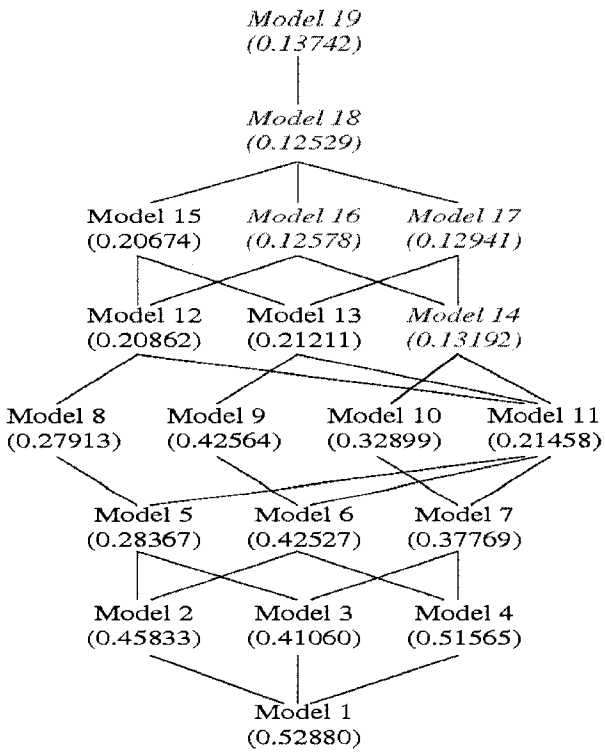


Figure 4.27: Lattice Diagram of 3-way ANOVA Models in Table ??

procedure *Group Identification I* (50%$CI$) or *Group Identification II* (c=0.5 and c=1). In fact, the target group is estimated to be $\{14, 16, 17, 18, 19\}$ in these cases. Model 14 is also the best model according to AIC and AICc. However, model 11 is chosen by our method with procedure *Group Identification I* with confidence level 80% and by procedure *Group Identification II* with $c = 1.5$; model 5 is chosen by procedure *Group Identification I* with confidence level 95%. Note that model 11 is a direct descendent to model 14, so is model 5 to model 11.

## 4.7 ANOVA Models With Two-parameter Exponential Distribution

In the previous sections, we discussed ANOVA models assuming normal distribution for the random error term and thus for the response. In many cases, however, this assumption does not hold true. For example, if we are modeling the

life-span or failure time of some electronic part, the distribution of the response maybe appropriately modeled by a *two-parameter exponential distribution* (Varde, 1969):

$$\beta(x|\gamma,\eta) = \frac{1}{\gamma}e^{-(x-\eta)/\gamma}I_{[\eta,\infty]}(x) \tag{4.20}$$

where $\gamma > 0$. It is clear that the mean is $\gamma+\eta$ while the variance is $\gamma^2$. When $\eta \geq 0$, the above distribution is also referred to as the *left truncated exponential* density function (Evans et al, 1980). This is especially meaningful when an individual has an unknown starting time. For instance, if a component is not used until a certain number of days after shipping out of the factory, then this number of days will be the location shift $\eta$ when modeling the failure time.

In this section, we will continue with the fixed effects factorial design, with the responses in each cell following the same exponential distribution with possibly different location parameters. The $r^{th}$ observation in cell $\pi$ can be expressed as:

$$Y_{\pi,r} = \eta_\pi + \epsilon_{\pi,r} \tag{4.21}$$

where $\epsilon_{\pi,r}$'s are assumed to be independently exponentially distributed with one parameter $\gamma$. The distribution of $Y_{\pi,r}$ is thus a two-parameter exponential with parameters $\gamma$ and $\eta_\pi$. Thus, the distributions in different cells are identical except for possibly different location parameters.

In this scenario, the Kullback-Leibler discrepancy is no longer well-defined since the approximating distribution may have different support than the unknown true distribution. Nonetheless, criteria such as AIC and AICC can still be calculated based on the maximum likelihood. But the interpretation of these criteria as the estimators of the expected K-L distance is no longer valid. On the other hand, model selection method based on expected squared Hellinger distance continues to be meaningful.

### 4.7.1 An Example of a 2-way Structure

In a simple balanced $2 \times 2$ ANOVA example, define $\eta_{..}$, $\eta_{i.}$, $\eta_{.j}$, and $\eta_{ij}$ such that:

$$\bar{\mu}_{..} = \eta_{..} + \gamma$$

$$\bar{\mu}_{i.} = \eta_{i.} + \gamma$$

$$\bar{\mu}_{.j} = \eta_{.j} + \gamma$$

$$\mu_{ij} = \eta_{ij} + \gamma$$

Then the main effects and the interaction effects are

$$\begin{aligned} i-\text{th row effect}: &\quad \alpha_i = \bar{\mu}_{i.} - \bar{\mu}_{..} = \eta_{i.} - \eta_{..}; \\ j-\text{th column effect}: &\quad \beta_j = \bar{\mu}_{.j} - \bar{\mu}_{..} = \bar{\eta}_{.j} - \bar{\eta}_{..}; \\ (i,j)-\text{th interaction}: &\quad (\alpha\beta)_{ij} = \mu_{ij} - \bar{\mu}_{i.} - \bar{\mu}_{.j} + \bar{\mu}_{..} = \eta_{ij} - \bar{\eta}_{i.} - \bar{\eta}_{.j} + \bar{\eta}_{..}; \end{aligned}$$

where $i, j = 1, 2$ with the constraints that

$$\sum_i \alpha_i = 0; \quad \sum_j \beta_j = 0;$$

$$\sum_j (\alpha\beta)_{ij} = 0, \quad i = 1, 2; \quad \sum_i (\alpha\beta)_{ij} = 0, \quad j = 1, 2; \quad \text{and} \quad \sum_{i,j} (\alpha\beta)_{ij} = 0.$$

Note that with the $2 \times 2$ structure,

$$-\alpha_1 = \alpha_2 = \alpha; \quad -\beta_1 = \beta_2 = \beta; \quad \text{and} \quad (\alpha\beta)_{11} = -(\alpha\beta)_{21} = -(\alpha\beta)_{12} = (\alpha\beta)_{22} = (\alpha\beta).$$

These effects are not involved with the parameter $\gamma$. Thus, the differences among the models are reflected by the different linear combinations of $\eta_{..}$, $\eta_{i.}$, $\eta_{.j}$, and $\eta_{ij}$.

In this simple model setting, we can find the maximum likelihood estimator (MLE) analytically.

#### 4.7.1.1 Finding MLE

Define the following quantities for $i, j = 1, 2$:

$$m_{ij} = \min(y_{ij1}, \ldots, y_{ijn}); \quad m_{..} = \min(y_{111}, \ldots, y_{11n}, y_{211} \ldots, y_{21n}, y_{121}, \ldots, y_{12n}, y_{221} \ldots, y_{22n})$$

$$m_{i.} = \min(y_{i11}, \ldots, y_{i1n}, y_{i21}, \ldots, y_{i2n}); \quad m_{.j} = \min(y_{1j1}, \ldots, y_{1jn}, y_{2j1}, \ldots, y_{2jn})$$

In general, the likelihood of data $\underline{Y} = \{Y_{ijr}; i, j = 1, 2; r = 1, \ldots, n\}$ is:

$$L(\{\eta_{ij}; i, j = 1, 2\}, \gamma | \underline{y}) = \gamma^{-4n} \exp\{\frac{4n}{\gamma}(\bar{y}_{...} - \bar{\eta}_{..})\} \prod_{i,j} I_{[\eta_{ij}, \infty]}(m_{ij}),$$

where $\bar{\eta}_{..} = \frac{1}{4}\sum_{ij} \eta_{ij}$. For fixed $\eta_{ij}$'s that satisfy $\prod_{i,j,r} I_{[\eta_{ij}, \infty]}(y_{ijr}) = 1$, the likelihood is maximized at

$$\hat{\gamma} = \bar{y}_{...} - \bar{\eta}_{..}$$

The solution of the maximization over parameters depends on the specific model structure. We will study the solution in the 5 possible approximating models one by one.

**Null Model** In case of the null model, the location parameter in all cell are the same:

$$\eta_{ij} = \eta_{..}, \quad i, j = 1, 2$$

Thus, the likelihood function becomes:

$$L(\eta_{..}, \gamma | \underline{y}) = \gamma^{-abn} \exp\{\frac{1}{\gamma}(4n)(\bar{y}_{...} - \eta_{..})\} I_{[\eta_{..}, \infty]}(m_{..})$$

and thus is maximized at $\hat{\eta}_{..} = m_{..}$ and $\hat{\gamma} = (\bar{y}_{...} - m_{..})$. The estimators of the location parameters thus become: $\hat{\eta}_{ij} = m_{..}$ for $i, j = 1, 2$.

**Row Effect Model** In case of the row effect model, the location parameter in the same row are the same:

$$\eta_{11} = \eta_{12} = \eta_{1.} = \eta_{..} - \alpha, \quad \eta_{21} = \eta_{22} = \eta_{2.} = \eta_{..} + \alpha$$

Thus, the likelihood function becomes

$$L(\eta_{1.}, \eta_{2.}, \gamma|\underline{y}) = \gamma^{-abn} \exp\{\frac{1}{\gamma}(4n)(\bar{y}_{...} - \eta_{..})\} \prod_i I_{[\eta_{i.},\infty]}(m_{i.})$$

and thus is maximized at $\hat{\eta}_{ij} = \hat{\eta}_{i.} = m_{i.}$ and $\hat{\gamma} = \bar{y}_{...} - \frac{1}{2}\sum_i m_{i.}$ for $i = 1, 2$.

**Column Effect Model** In similar argument as above, the MLE's are found to be:

$$\hat{\eta}_{ij} = \hat{\eta}_{.j} = m_{.j} \quad \text{and} \quad \hat{\gamma} = \bar{y}_{...} - \frac{1}{2}\sum_j m_{.j} \quad j = 1, 2.$$

**Main Effects Model** The case of main effects model needs more care. The location parameters are now $\eta_{ij} = \eta_{..} + \alpha_i + \beta_j$, which can be written in the forms as shown in Table 4.16. The likelihood now becomes

| | $j = 1$ | $j = 2$ |
|---|---|---|
| $i = 1$ | $\eta_{..} - (\alpha + \beta)$ | $\eta_{..} - (\alpha - \beta)$ |
| $i = 2$ | $\eta_{..} + (\alpha - \beta)$ | $\eta_{..} + (\alpha + \beta)$ |

Table 4.16: Location Parameters for Main Effects Model

$$L(\eta_{..}, \alpha, \beta, \gamma|\underline{y}) = \gamma^{-abn} \exp\{\frac{1}{\gamma}(4n)(\bar{y}_{...} - \eta_{..})\} \prod_{i,j} I_{[\eta_{..}+\alpha_i+\beta_j,\infty]}(m_{ij})$$

In order to maximize the likelihood, the parameters $\eta_{..}$, $\alpha$ and $\beta$ must satisfy:

$$\eta_{..} - (\alpha + \beta) \leq m_{11}; \quad \eta_{..} - (\alpha - \beta) \leq m_{12};$$

$$\eta_{..} + (\alpha - \beta) \leq m_{21}; \quad \eta_{..} + (\alpha + \beta) \leq m_{22}.$$

The above constrains can be written

$$\eta_{..} - m_{11} \leq (\alpha + \beta) \leq m_{22} - \eta_{..}, \quad \eta_{..} - m_{12} \leq (\alpha - \beta) \leq m_{21} - \eta_{..}.$$

To maximize the likelihood, we need:

$$\eta_{..} - m_{11} \leq m_{22} - \eta_{..}, \quad \eta_{..} - m_{12} \leq m_{21} - \eta_{..}.$$

The likelihood is thus maximized at $\eta_{..} = \min((m_{11}+m_{22})/2, (m_{12}+m_{21})/2)$. When $(m_{11}+m_{22})/2 = (m_{12}+m_{21})/2$, the solution of $\alpha$ and $\beta$ is unique. Otherwise, the solution is on a line segment of $\alpha+\beta = (m_{22}-m_{11})/2$ between $(m_{11}+m_{22})/2 - m_{12} \leq \alpha-\beta \leq m_{21}-(m_{11}+m_{22})/2$ if $(m_{11}+m_{22})/2 < (m_{12}+m_{21})/2$, or on $\alpha-\beta = (m_{21}-m_{12})/2$ between $(m_{12}+m_{21})/2 - m_{11} \leq \alpha+\beta \leq m_{22}-(m_{12}+m_{21})/2$ if $(m_{11}+m_{22})/2 > (m_{12}+m_{21})/2$. We will use the middle point on the line segment. The MLE for the scale parameter is then

$$\hat{\gamma} = \bar{y}_{...} - \frac{1}{4}\sum_{i,j}\hat{\eta}_{ij}$$

where $\hat{\eta}_{ij} = \hat{\eta}_{..} + \hat{\alpha}_i + \hat{\beta}_j$ for $i, j = 1, 2$ is determined as discussed above.

**Full Factorial Model** In this case, the MLE's are:

$$\hat{\eta}_{ij} = m_{ij} \quad i, j = 1, 2, \quad \hat{\gamma} = \bar{y}_{...} - \frac{1}{4}\sum_{i,j} m_{ij}$$

#### 4.7.1.2 Numerical Example

Let the true model be a $2 \times 2$ row effect model and data follows an Exponential distribution with the scale parameter being $\gamma$ and the location parameters as described in Table 4.17.

| | $j=1$ | $j=2$ |
|---|---|---|
| $i=1$ | 0 | 0 |
| $i=2$ | a | a |

Table 4.17: Location Parameter

Let $\gamma = 1$ and $a = 1$. Let the numbers of bootstrap samples be $M = 300$ and $M2 = 100$ and the number of simulated data sets be 100. For comparison purposes, we also calculated AIC, AICc, and BIC. The results for cell sample size 10, 30, 50, and 100 are listed in Table 4.18

| Model | Frequency | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | *Group Id I* | | | *Group Id II* | | |
| | *EH2* | AIC | AICc | BIC | 50% | 80% | 95% | 0.5 | 1.0 | 1.5 |
| $n = 10$ | | | | | | | | | | |
| 1 | 0.42677 | 0 | 0 | 0 | 100 | 100 | 100 | 42 | 90 | 99 |
| 2 | 0.16609 | 47 | 73 | 91 | 0 | 0 | 0 | 58 | 10 | 1 |
| 3 | 0.42224 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.20240 | 9 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.23469 | 44 | 22 | 8 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n = 30$ | | | | | | | | | | |
| 1 | 0.42091 | 0 | 0 | 0 | 6 | 36 | 95 | 0 | 0 | 1 |
| 2 | 0.10864 | 60 | 62 | 98 | 94 | 64 | 5 | 100 | 100 | 99 |
| 3 | 0.41943 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.12367 | 9 | 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.13884 | 31 | 29 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n = 50$ | | | | | | | | | | |
| 1 | 0.42249 | 0 | 0 | 0 | 1 | 1 | 21 | 0 | 0 | 0 |
| 2 | 0.07481 | 61 | 62 | 99 | 99 | 99 | 79 | 100 | 100 | 100 |
| 3 | 0.42149 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.08994 | 9 | 10 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.10174 | 30 | 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $n = 100$ | | | | | | | | | | |
| 1 | 0.41969 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 0.05575 | 53 | 55 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 3 | 0.41922 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0.06664 | 14 | 14 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0.07275 | 33 | 31 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 4.18: Frequency of Choosing Any Approximating Model by Different Methods

***Note:*** Data was simulated from row effect model, $a = 1$, $\sigma = 1$. Cell sample size $n = 10, 30, 50, 100$, numbers of bootstrapped samples $M = 300$ and $M2 = 100$, total number of simulated data $S = 100$.

For all 4 sample sizes, the true best model is the true model, the row effect model. The performance of our methods incorporated with both group identification procedures increases dramatically as the cell sample size increases from 10 to 30, except for *Group Identification I* with 95% confidence level. In general, our methods

with either procedure work reasonably well at larger sample sizes, for instance, $n \geq 30$. Smaller confidence interval as well as smaller $c$ leads to better performance in the corresponding group identification procedure. Between the group identification procedures, *Group Identification II* works better in smaller sample sizes, due to the fact that the bootstrap confidence intervals are not centered at the original $PEEH2$ for small sample sizes. In particular, at sample size 50, our method with *Group Identification II* with all three choices of $c$ chooses the true best model 100%. We also calculated AIC, AICc, and BIC. The overall best performance is given by BIC, which chooses the true best model 91% of the time at sample size of $n = 10$ when our methods largely prefer the null model. AIC and AICc, on the other hand, did not show a consistent improvement as sample size increases. In fact, the performance of AICc decreases, or stay the same, as sample size increases. Their performance is better than that of ours when sample size is small but falls far behind as it increases.

# Chapter 5

# MICROARRAY DATA ANALYSIS – FDR ESTIMATION AND MODEL SELECTION

## 5.1 Introduction

Microarrays allow scientists to monitor gene expression for thousands of genes simultaneously and help identify genes that are expressing differently under two or more conditions. The primary goal of many microarray experiments is to identify a group of differentially expressed genes. This is typically achieved by conducting appropriate statistical tests.

The dimension of the data resulting from microarray experiments provides some unique challenges and opportunities for data analysis. Typically, we test each gene for differential expression. Hence for even a simple experiment comparing two groups or treatments, thousands of tests will be performed and thousands of $p$-values will be generated. This can lead to a high number of false discoveries unless appropriate multiple testing adjustments are implemented. A well accepted method of multiple testing adjustment is through the control of the false discovery rate (FDR). The FDR is defined as the expected ratio of false positives to total positives (Benjamini and Hochberg, 1995). Let $V$ be the number of false positives and $R$ be the total number of rejected hypotheses. Then FDR is defined as $E(V/R)$ if $R > 0$ and is defined to be zero otherwise. Benjamini and Hochberg (1995) recommended a sequential

multiple testing adjustment to control FDR assuming independence of the individual tests. Later, they proposed an adaptive technique for controlling FDR based on the estimated number of true null hypotheses (Benjamini and Hochberg, 2000).

Various approaches to estimating $FDR$ from the observed (empirical) distribution of $p$-values have been reported in the literature. The large number of hypotheses tested in microarray experiments provides us an opportunity to model the distribution of $p$-values. Allison et al (2002) fitted a mixture of beta distributions plus a uniform (which itself belongs to the beta family) using maximum likelihood estimation. Pounds and Morris (2003) also fit a mixture of a beta plus a uniform (BUM) using maximum likelihood estimation. Dalmasso et al (2005) propose a location based estimator for estimating $FDR$. Broberg (2005) suggests a poisson regression approach to estimate $FDR$. Storey (2002) and Storey and Tibshirani (2003) introduce positive false discovery rate ($pFDR = E(V/R|R > 0)$) and the $q$ -value as an FDR based measure of significance. Tadesse et al (2005) propose controlling the pFDR using a Bayesian approach. Pounds and Cheng (2004) propose the spacings LOESS histogram method for estimating the conditional FDR (cFDR) where cFDR is defined as the expected proportion of false positives conditioned on having $k$ "significant" findings. Liao et al (2004) recommend using a mixture model for estimating local $FDR$. Local FDR was defined by Efron et al (2001) and Efron and Tibshirani (2002) as a measure of a specific gene's "significance" based on its specific $p$-value or test statistic. Tsai et al (2003) compare five different FDR measures and present a framework for modeling their distributions. Still more authors have concentrated on estimating the proportion of true null hypotheses, which is important for FDR estimation. See Nguyen (2004), Langaas and Lindqvist (2005), and Lai (2006). Broberg (2005) provides a good review and comparison of methods. Among

methods that fit mixture models to the $p$-value data, the BUM method, which uses a maximum likelihood approach, has been mentioned as performing competitively (at least for some of the scenarios considered) relative to the methods compared by Broberg (2005). The ultimate goal of these procedures is to take the observed $p$-values from a microarray experiment and select a group of genes identified as differentially expressed while maintaining a prespecified FDR.

Alternative approaches for fitting finite mixture models are also available in the general statistical literature. Several authors have shown that estimators obtained by minimizing the Hellinger distance between the theoretical and the empirical distributions possess certain robustness properties. However, Hellinger distance based methods have not been applied to estimation of FDR in microarray data analysis or other situations where a large number of hypotheses are being tested. In this chapter we propose two procedures for estimating the proportion of differentially expressed genes and FDR based on Hellinger distance as the measure of lack of fit between a theoretical mixture model and the data.

The first procedure considers the class of densities that are a mixture of a uniform $[0, 1]$ density plus one or more beta densities as candidate models (i.e. BUM models). The number of beta densities to include in addition to a uniform $[0, 1]$ density is determined using a model selection procedure that is also based on the Hellinger distance. The second procedure considers an approximating family of densities that are mixtures of a uniform $[0, 1]$ distribution and $M$ other uniform distributions on prespecified subintervals of $[0, 1]$. The parameters in this mixture model are also estimated using Hellinger distance. In each case, the fitted mixture model then provides an estimate of FDR for any given $p$-value threshold for declaring differential expression.

We apply both methods to the $p$-values resulting from analysis of data from Spira et al (2004) in which the effects of cigarette smoking on the human airway epithelial cell transcriptome have been studied. We focus on a subset of the data and compare current smokers versus healthy subjects who have never smoked. The histogram of $p$-values resulting from $t$-tests for comparing these two groups is shown in Figure 5.1. The shape of the distribution suggests that the mixture model of one uniform plus one or more beta distributions can provide a reasonable approximation.

This chapter is organized as follows. The next section discusses how one can obtain an estimate of FDR from a fitted mixture density model. Section 5.3 discusses the usage of $BEEH2$ as a criterion to decide on the number of beta components when working with beta mixture models. An algorithm for fitting a mixture of uniform densities involving one uniform $[0, 1]$ density and $M$ other uniform densities on disjoint subintervals of $[0, 1]$ is also provided in this section. The microarray data from Spira (2004) is considered in Section 5.4 to illustrate how the methods work in practice. The final section provides simulation study results, including the use of Efron's suggested empirical distribution for the $p$-values of the non-differentially expressed genes instead of the theoretical uniform distribution.

## 5.2 Estimating FDR from a Mixture Density Model

The use of mixture densities for modeling the distribution of $p$-values is based on the observation that the distribution of $p$-values from independent tests corresponding to true null hypotheses is uniform over the interval $[0, 1]$ whereas each $p$ value corresponding to a differentially expressed gene will follow a distribution that has more mass close to zero. The theoretical non-null distribution of a $p$-value from a differentially expressed gene will be determined by the magnitude of differential

expression as well as variability of this differential expression and is therefore unique to each differentially expressed gene. Nevertheless, a mixture model with a small number of components, one of which is the uniform density on $[0, 1]$, has been found to be generally adequate to model the empirical distribution of the $p$-values resulting from microarray data analysis (Allison et al, 2002). The uniform $[0, 1]$ component is viewed as the component corresponding to equally expressed genes and its associated mixing proportion is viewed as the proportion of equally expressed genes in the study.

Let $f$ and $F$ be the true probability density function and cumulative distribution function (cdf) of $p$-values respectively. The cdf $F$ can be written as

$$F(x) = q_0 G_0(x) + (1 - q_0) W(x), \tag{5.1}$$

where $G_0$ is the cdf of the uniform [0,1] component that corresponds to the genes that are equally expressed, and $W$ is the cdf corresponding to the genes that are differentially expressed. Moreover, it is assumed that $W$ has $[0, 1]$ as its support. The $FDR$ for a given test threshold level $\alpha$ is

$$FDR(\alpha) = \frac{q_0 \alpha}{F(\alpha)}. \tag{5.2}$$

The theoretical quantile $F(\alpha)$ is generally unknown. We propose to approximate it by $A(\alpha)$, where $A(\cdot)$ is a cdf with density $a(\cdot)$, an element from an approximating family of mixture models with one component being a uniform distribution. Without loss of generality, we assume that the density of $a(x)$ is a mixture of uniform distribution and $\nu$ nonuniform continuous densities $g_j(\cdot)$ , $j = 1, \ldots, \nu$, with $[0, 1]$ (or subintervals of $[0, 1]$) as their support; that is,

$$a(x) = p_0 + \sum_{j=1}^{\nu} p_j g_j(x), \quad \text{for} \quad x \in [0, 1] \tag{5.3}$$

and zero elsewhere. The $p_j$ are the mixing proportions satisfying $\sum_{j=0}^{\nu} p_j = 1$ and $0 \leq p_j \leq 1$ for all $j = 0, 1, \cdots, \nu$. Thus, the $FDR$ can be approximated by

$$p_0\alpha / A(\alpha).$$

The accuracy of this approximated FDR will depend on how close the approximating mixture density $a(\cdot)$ is to the true density $f(\cdot)$. A plug-in estimator for the FDR is given by

$$\widehat{FDR}(\alpha) = \frac{\hat{p}_0\alpha}{\hat{p}_0\alpha + \sum_{j=1}^{\nu} \hat{p}_j \hat{G}_j(\alpha)}, \tag{5.4}$$

where $\hat{p}_j, j = 0, \ldots, \nu$ are the estimated mixing proportions and $\hat{G}_j(\cdot)$ is the estimated cdf corresponding to the fitted density function $\hat{g}_j(\cdot)$.

In this chapter, we consider two approximating families – a beta mixture family in Section 5.2.1 and a uniform mixture family in Section 5.2.2.

### 5.2.1 Estimating FDR from a Mixture of Uniform and Betas

Here, each mixture component $g_j(x)$ is taken to be a beta probability density function with unknown parameters $(a_i, b_i)$, say $\beta(x; a_i, b_i)$. Thus, $a(x)$ can be expressed as

$$a(x) = p_0 + \sum_{j=1}^{\nu} p_j \beta(x; a_j, b_j), \; x \in [0, 1]. \tag{5.5}$$

It can be seen that the family of approximating densities is not identifiable. The reason is that different choices of the mixing probabilities and beta densities can result in the same density $a(\cdot)$. As a result, the value of the approximated $FDR$ depends on the particular choice for the mixing proportions. Note that a mixture representation of $A(\cdot)$ that has the largest value for $p_0$ among all mixture representations of $A(\cdot)$ will result in the largest value for the approximated FDR. We therefore

propose to use this representation since it will result in a conservative estimate of FDR.

Another important issue is the choice of the number of beta components (in addition to the uniform component) in the mixture model, i.e. the value of $\nu$. In Equation (5.4), the estimated values of the parameters of the mixture components are used to estimate $\hat{G}(\cdot)$. One needs to determine the number of beta components needed to model the $p$-value distribution adequately. Allison et al (2002) propose an approach that is based on hypothesis testing for the number of beta components. For this, they consider a *forward selection* approach. For $k = 0, 1, \ldots$, they test if a model with $k+1$ nonuniform beta components provides a better fit than a model with $k$ nonuniform beta components. If the test favors the larger model, then they proceed to compare that model with the model with an additional component, and so on, until the test accepts the smaller of the two models being compared. These authors used a parametric bootstrap method for conducting such tests. They report that, in their experience with several data sets, they have "yet to need more than one beta beyond the uniform".

### 5.2.2 Estimating FDR from a Uniform Mixture Model

As an alternative to the beta mixture approach, we can also approximate the unknown density function by a mixture of one uniform $[0, 1]$ density and $M$ other uniform densities $\frac{1}{d_i} I_{(b_{i-1}, b_i]}(x)$, $i = 1, \ldots, M$, where $0 = b_0 < b_1 < \cdots < b_M = 1$ and $d_i = b_i - b_{i-1}$, $M$ and $b_i$ $i = 1, \ldots, M$ are prespecified, and $I_{(r,s]}(x)$ is an indicator function that takes the value 1 if $x$ belongs to interval $(r, s]$ and 0 otherwise. That is, we approximate $f(x)$ by a mixture density $a(x)$ of the form

$$a(x) = p_0 + \sum_{i=1}^{M} p_i g_i(x), \text{ for } x \in [0, 1], \tag{5.6}$$

with $g_i(x) = \frac{1}{d_i} I_{(b_{i-1},b_i]}(x)$. When the subintervals are equally spaced on $[0,1]$, $d_i = 1/M$ for all $i = 1, 2, \ldots, M$. On the other hand, $d_i$'s may be different if we consider subintervals that are equally spaced on some transformed scale, such as square root of the $p$-values, to account for the fact that many distributions of the $p$-values have a large proportion of values close to zero.

The nonidentifiablility problem may also arise here. However, as in the case of beta mixtures, use of a mixture representation that has the largest value for $p_0$ will lead to a conservative estimate for $FDR$ and this is the representation we propose to use.

For mathematical convenience, the approximating family of densities in Equations (5.5) and (5.6) are denoted as $a_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}$ is the vector of parameters. In the beta mixture family of Equation (5.5), $\boldsymbol{\theta}$ is a vector of mixing proportions and the parameters associated with $\nu$ beta components; whereas, in the uniform mixture family of Equation (5.6), $\boldsymbol{\theta}$ is the vector $(p_0, p_1, \cdots, p_M)$.

In the following section, we will review the concept of Hellinger distance and discuss the procedures for estimating the parameter vector $\boldsymbol{\theta}$ of the approximating families discussed above.

## 5.3 Hellinger Distance and Mixture Model Estimation

The discrepancy between the empirical distribution of a set of data values and a theoretical distribution that is being used to model the data may be measured using any one of several distance measures between distributions. Such measures include the Kolmogorov-Smirnov distance the Kullback-Leibler (KL) distance , and the Hellinger distance . Donoho and Liu (1988) and Cao et al (1995) pointed out that minimum distance (discrepancy) estimators (MDEs) occupy an outstanding place

among the robust alternatives to the classical maximum likelihood (ML) method for point estimation. They explored MDEs for general finite mixture models. Beran (1977) studied the MHDE and its asymptotic properties and concluded that the MHDE was asymptotically efficient as well as robust. Tamura and Boos (1986) applied minimum Hellinger distance estimation for multivariate location and covariance, and stated that the robustness of the MHDE as measured by the breakdown point compares favorably against the previously studied M-estimators. Cutler and Cordero-Brana (1996) considered MHDEs for finite mixture models when the exact forms of the component densities are unknown in detail but are thought to be close to members of some parametric family. In particular, they studied examples where the component densities are approximated by normal densities. They addressed the issues of identifiability, existence, consistency and asymptotic normality of the MHDEs for finite mixture models and showed that the MHDEs are asymptotically efficient if the data come from a member of the parametric family and are robust to certain departures from the parametric family. Because of such robustness properties associated with MHDEs we propose to use them for modeling the distribution of $p$-values from microarray data analysis.

### 5.3.1 Minimum Hellinger Distance Estimation for Beta Mixture Model

A MHDE of $\boldsymbol{\theta}$ is defined as follows (Beran, 1977):

$$\hat{\boldsymbol{\theta}}_{MHD} = \arg\min_{\boldsymbol{\theta}} \{H(\hat{f}, a_{\boldsymbol{\theta}})\}. \tag{5.7}$$

where $\hat{f}$ is a suitable density estimator. In this chapter, a kernel density estimator is used in the beta mixture approach, and a histogram density estimator is used in the uniform mixture approach.

Considerable effort has been devoted by many authors for deciding the number of beta components in addition to the uniform component corresponding to equally expressed genes. Parker and Rothenberg (1988) determined the number of beta components by assessing the goodness-of-fit of each model using the Cramer-von Mises statistic, and stopping at the smallest value of $\nu$ that provided an "adequate" fit for the data. As discussed earlier, Allison and Gadbury et al (2002) proposed an approach based on hypothesis testing of a sequence of mixture models with increasing numbers of beta components.

In this chapter, we apply the model selection method based on the expected squared Hellinger distance. Specifically, we determine the value $\nu$ for which a model with $\nu$ nonuniform beta components has the smallest $BEEH2$. Let the fitted density be $a_{\widehat{\boldsymbol{\theta}}_\nu}$ with $\nu$ nonuniform beta components. $BEEH2$, as discussed in Chapter 2, is thus:

$$BEEH2 = \frac{1}{B}\sum_{l=1}^{B} H^2(\hat{f}, a_{\widehat{\boldsymbol{\theta}}^{*(l)}_\nu}).$$

where $B$ denotes the number of bootstrap sample here. The value of $\nu$ that yields the smallest $BEEH2$ may be chosen as the number of nonuniform beta components. A more practical approach for choosing the number of beta components is discussed below.

If two mixture models are close to one another, their Hellinger distances to density $f$ will also be close to each other. We use this fact in our proposed method for selecting the number of nonuniform beta components where we use the estimated expected squared Hellinger distance as the criterion of goodness of fit. To avoid choosing a mixture model that has $\nu$ beta components but is not in fact different from a mixture model with $\nu - 1$ or smaller number of beta components, we introduce a further step that favors the mixture model with the smallest number of beta

components. This is accomplished by choosing the mixture model whose $BEEH2$ is *not significantly different* from the minimum of all $BEEH2$ values corresponding to different numbers of beta components. More specifically, we approximate the distribution of $BEEH2$ by bootstrapping $B'$ samples from the original data and calculating $B'$ realizations of $BEEH2$ for each approximating mixture model. The mean and the standard deviation of the bootstrap distribution of $BEEH2$ are calculated for each candidate model. The model with the smallest number of beta components whose mean $BEEH2$ value is within one standard error of the smallest mean $BEEH2$ value will be chosen as the model to approximate the distribution of the $p$ values. A similar approach has been used by Hastie et al (2001, page 216) in the context of model selection and referred to as the *one-standard error rule.*

### 5.3.2 MHDE for the Uniform Mixture Model

For any given integer $M$, we divide the interval [0,1] into $M$ prespecified intervals as in (5.6). The corresponding histogram density estimator of $f(\cdot)$ is given by

$$\hat{f}_M(x) = \sum_{i=1}^{M} \frac{c_i}{d_i} I_{(b_{i-1},b_i]}(x), \tag{5.8}$$

where

$$c_i = \int_{(b_{i-1},b_i]} dF_n(x),$$

and $F_n$ is the empirical cdf of $f$. The squared Hellinger distance between $\hat{f}_M$ and the approximating uniform mixture density function $a_{\boldsymbol{\theta}}$ in Equation (5.6) can be written as:

$$H^2(\hat{f}_M, a_{\boldsymbol{\theta}}) = \sum_{i=1}^{M} \left[\sqrt{c_i} - \sqrt{p_0 d_i + p_i}\right]^2. \tag{5.9}$$

The mixing proportions $p_i$ are estimated by minimizing $H^2(\hat{f}_M, a_{\boldsymbol{\theta}})$.

For each fixed $p_0$ in the interval $[0,1]$, the minimization problem in (5.9) is accomplished using the Lagrange Multiplier method for which the objective function is

$$T(p_1,\ldots,p_M,\lambda) = \sum_{i=1}^{M}\left[\sqrt{c_i} - \sqrt{p_0 d_i + p_i}\right]^2 + \lambda(\sum_{i=0}^{M} p_i - 1). \tag{5.10}$$

For the time being we have not considered the nonnegativity constraints on the $p_i$ and we will return to this point later. For $i = 1,\ldots,M$, candidate values of $p_i$ that minimize the function in (5.10) may be obtained by equating to zero the partial derivatives of $T(p_1,\ldots,p_M,\lambda)$ with respect to the $p_i$'s and $\lambda$. In addition to the constraint $\sum_{i=0}^{M} p_i = 1$ this leads to the equations

$$\lambda + 1 = \frac{\sqrt{c_i}}{\sqrt{p_0 d_i + p_i}} \quad \text{for } i = 1,\ldots,M.$$

Multiplying the $i^{th}$ of the above equations by $p_i$ on both sides, we get

$$(\lambda+1)p_i = \frac{p_i\sqrt{c_i}}{\sqrt{p_0 d_i + p_i}},\ i = 1,\ldots,M, \tag{5.11}$$

from which it follows, by summing over $i = 1,\ldots,M$, that

$$\lambda + 1 = \frac{1}{1-p_0}\sum_{i=1}^{M}\frac{p_i\sqrt{c_i}}{\sqrt{p_0 d_i + p_i}}. \tag{5.12}$$

Using Equations (5.11) and (5.12) we get

$$p_i = \frac{p_i\sqrt{c_i}}{\lambda^\star\sqrt{p_0 d_i + p_i}} \tag{5.13}$$

where

$$\lambda^\star = \lambda + 1 = \frac{1}{1-p_0}\sum_{i=1}^{M}\frac{p_i\sqrt{c_i}}{\sqrt{p_0 d_i + p_i}}. \tag{5.14}$$

We thus propose the following algorithm for determining the values of $p_i$ , $i = 1,\ldots,M$, corresponding to a specified value for the mixing proportion $p_0$.

#### 5.3.2.1 Algorithm 1: Finding $\{\hat{p}_i\}$ for Any Given $p_0$

1. (*Initialization*) Select initial values for $p_i$, say $\{p_i^{(0)} : i = 1, \ldots, M\}$. These initial values must be nonnegative and must sum to $1 - p_0$. Define

$$a_{\boldsymbol{\theta},M}^{(0)}(x) = p_0 + \sum_{i=1}^{M} \frac{p_i^{(0)}}{d_i} I_{(b_{i-1},b_i]}(x).$$

2. ($\eta$-*Update*) For $t \geq 0$, given that $\{p_i^{(t)} : i = 1, \ldots, M\}$ are already available, calculate

$$\eta_i^{(t)}(p_0) = \frac{p_i^{(t)}\sqrt{c_i}}{\sqrt{p_0 d_i + p_i^{(t)}}}, \qquad i = 1, \ldots, M.$$

3. ($p_i$-*Update*) Calculate

$$p_i^{(t+1)} = \frac{(1 - p_0)\eta_i^{(t)}(p_0)}{\sum_{i=1}^{M} \eta_i^{(t)}(p_0)}, \qquad i = 1, \ldots, M.$$

   Define

$$a_{\boldsymbol{\theta},M}^{(t+1)}(x) = p_0 + \sum_{i=1}^{M} \frac{p_i^{(t+1)}}{d_i} I_{(b_{i-1},b_i]}(x)$$

   and compute $H(\hat{f}_M, a_{\boldsymbol{\theta},M}^{(t+1)})$.

4. Repeat steps (2) and (3) until convergence; that is, until the changes in the values of $p_i^{(t)}$ are smaller than a prespecified *tolerance*. Suppose this occurs when $t = t^\star$. At this point stop and report $p_i^{(t^\star+1)}$ as the MHDE for $p_i$ and $m_h = H(\hat{f}_M, a_{\boldsymbol{\theta},M}^{(t^\star+1)})$ as the minimum Hellinger distance achieved. Note that, given feasible starting values, the updated values for $\hat{p}_i^{(t+1)}$ in the iteration steps in Algorithm 1 are automatically nonnegative and sum to $1 - p_0$.

While Algorithm 1 gives us the MHDEs for $p_i$ corresponding to a specified value of $p_0$, there is no guarantee that the resulting estimated mixture distribution will fit

the data well. Values of $p_0$ that lead to mixture density solutions that do not fit the data well are deemed unacceptable. This leads to the question of how one might test the goodness of fit of a proposed mixture solution. The following algorithm suggests one possible method of testing the goodness of fit.

#### 5.3.2.2 Algorithm 2: Testing Consistency With the Data

To test the adequacy of the model obtained in the previous section, we use the estimated minimum Hellinger distance as a test static. The null distribution of this test statistic is approximated by conducting a parametric bootstrap as follows.

1. Generate $B$ random samples $\mathbb{Y} = \{y_1, \ldots, y_n\}$ from the following proposed density

$$p_0 + \sum_{i=1}^{M} \frac{\hat{p}_i}{d_i} I_{(b_{i-1}, b_i]}(x)$$

where $\hat{p}_1, \ldots, \hat{p}_M$ are the MHDEs of $p_1, \ldots, p_M$ obtained from Algorithm 1.

2. For the $b^{th}$ random sample $(1 \leq b \leq B)$, determine the empirical distribution

$$\hat{f}_{M,b}(x) = \frac{1}{nd_i} \cdot \#\{j : y_j \in (b_{i-1}, b_i]\}, \quad x \in (b_{i-1}, b_i], \quad i = 1, \ldots, M$$

and calculate the Hellinger distance, denoted by $MH_b^\star$, between $\hat{f}_{M,b}$ and the proposed density.

If

$$P^\star = \frac{\sum_{b=1}^{B} I_{(m_h, \infty)}(MH_b^*)}{B}$$

is smaller than a predetermined significance level $\alpha_0$ then conclude that the specified value of the mixing proportion $p_0$ does not lead to a solution that adequately fits the data.

As noted before, the mixing proportion $p_0$ is not uniquely determined. If $p_0^\star$ denotes the minimum among $c_i/d_i$, $i = 1, \ldots, M$, then it is easily verified that every value of $p_0$ in the interval $[0, p_0^\star]$ results in zero as the minimum value for $H^2(\hat{f}_M, a_{\boldsymbol{\theta}})$. However, a value of $p_0$ that is larger than $p_0^\star$ may lead to a value of $H^2(\hat{f}_M, a_{\boldsymbol{\theta}})$ that is not *significantly different* from zero. Our goal is to provide the largest such value for $p_0$. We implemented the *golden section* method, described below, to calculate such upper bound:

1. (*Initialization*) Denote $\phi = \frac{\sqrt{5}-1}{2}$, $a = p_0^\star$ and $b = 1$;

2. (*Update*) For the $i^{th}$ chosen candidate value of $p_0$, $p_{0,i} = a + (b - a)\phi$, the goodness-of-fit test as described in Algorithm 2 is carried out. If the value $p_0 = p_{0,i}$ is rejected, let $b = p_{0,i}$; otherwise, denote $a = p_{0,i}$.

3. Repeat Step 2 for $i = 1, 2, \ldots,$ until convergence occurs; that is, the value $b - a$ is sufficiently small. The largest candidate value such that the null hypothesis $p_0 = p_{0,i}$ is not rejected is denoted by $\hat{p}_U$ and is reported as a $1 - \alpha_0$ upper confidence bound for the mixing proportion $p_0$ and the corresponding $\hat{p}_i$ as the MHDEs for $\boldsymbol{p} = (p_1, \ldots, p_M)$. The corresponding value of $FDR$ obtained from Equation (5.4) is reported as an upper confidence bound for $FDR$ with confidence coefficient $1 - \alpha_0$.

In the next section, we will illustrate the methods proposed in this chapter by applying them to the microarry experiment of Spira et al (2004).

## 5.4 Real Data Example

Spira et al (2004) studied the effects of cigarette smoking on the human airway epithelial cell transcriptome and found a large number of genes whose expressions

are altered by cigarette smoking. There were 75 arrays (of type Affymetrix HG-U133A) corresponding to 75 different subjects. Out of the 75 subjects, 34 are current smokers and 23 are healthy persons who have never smoked. Each array consisted of 22283 probe sets. Each probe set can be roughly thought of as representing a single gene. The normal large-airway transcriptome was defined by the genes whose median probability of detection ($P_{(detection)}$) value was less than 0.05 across all 23 healthy never-smokers. More information above the way the microarray data was acquired can be found in Spira et al (2004). We obtained the data from *Gene Expression Omnibus* (GEO, internet site, 2007) and used the statistical software package `R` (`R` Development Core Team, 2006) for data analysis. Data were read and normalized using the functions *ReadAffy* and *mas5* (Irizarry et al, 2006) provided by the `R` package `Bioconductor` (Gentleman et al, 2004). Spira et al (2004) filtered the data according to the definition of normal large-airway transcriptome and found 7119 genes that are expressed across the majority of the healthy subjects. There is a difference between the set of genes selected by Spira et al (2004) and the set used here due to the use of different software packages and presumably different normalization and background correction algorithms. To closely match their filtering, we worked with the 6708 genes that are marked as "Present" for at least 11 of the 23 never smokers.

Independent two sample $t$-tests between current smokers and never smokers were conducted for each of the 6708 genes without assuming equality of variances. The histogram of the $p$-values is plotted in Figure 5.1. The shape of the distribution suggests that the mixture model of one uniform plus one or more beta distributions may provide a reasonable approximation. We fitted mixture models with uniform

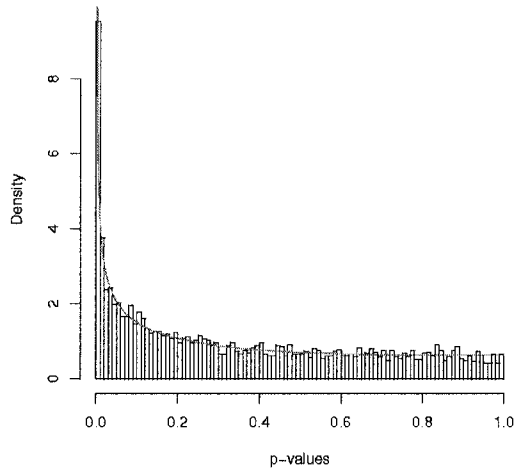


Figure 5.1: Histogram of the $p$-values and the fitted mixture density (solid line type) with Uniform plus two Beta components

plus 0 to 4 beta components respectively and the $BEEH2$ values were computed using $B = 500$ bootstrap samples. These $BEEH2$ values are listed in Table 5.1.

| Mixture | $\nu = 0$ | $\nu = 1$ | $\nu = 2$ | $\nu = 3$ | $\nu = 4$ |
|---|---|---|---|---|---|
| $BEEH2$ | 0.1001 | 0.00377 | 0.00058 | 0.00076 | 0.00063 |

Table 5.1: $BEEH2$ for the mixture models with uniform component plus $\nu$ nonuniform beta components where $\nu$ ranges from 0 to 4

The mixture model with two beta components has the smallest $BEEH2$, which is significantly smaller than that of the mixture model with one beta component based on our bootstrap results. The mixture model with one uniform and two beta components is therefore the chosen best model and the proportion of the uniform is estimated to be 0.6211. Table 5.2 lists all the estimated parameters.

Figure 5.1 shows a plot of the estimated mixture density against the density histogram of all $p$-values. The estimated $FDR$ corresponding to each possible significance threshold for $p$-values, i.e. $\alpha$ in Equation (5.2), is plotted as solid line type in Figure 5.2 where the horizontal line in the graph represents an FDR of 0.05.

| | Uniform | Beta Component 1 | | | Beta Component 2 | | |
|---|---|---|---|---|---|---|---|
| | $p_0$ | $p_1$ | $a_1$ | $b_1$ | $p_2$ | $a_2$ | $b_2$ |
| Estimator | 0.6211 | 0.0767 | 0.1586 | 1.00000004 | 0.3022 | 0.5521 | 4.0559 |

Table 5.2: Estimated parameters of the mixture density with 2 nonuniform Beta components

The FDR at significance level of 0.05 is estimated to be 0.1492. The significance level that ensures a FDR of 0.05 is found to be 0.00674 and 537 genes were found differentially expressed at this FDR level.

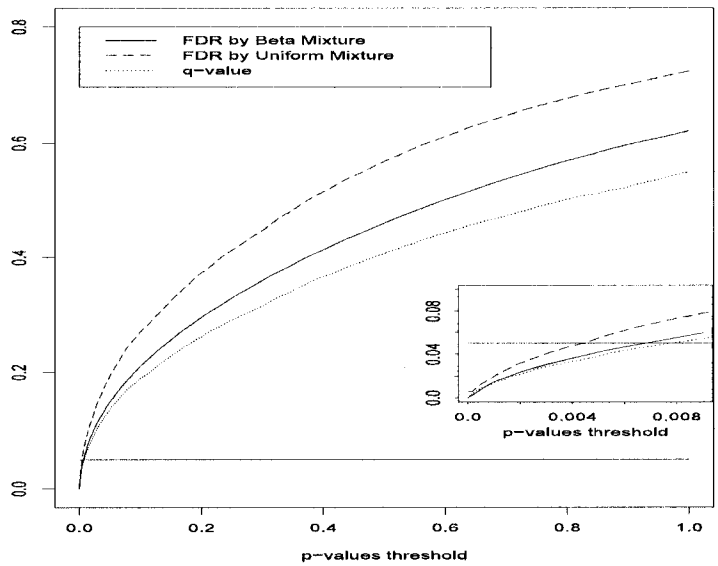


Figure 5.2: A graphical illustration of the estimated FDR using beta mixture (solid line), uniform mixture (dashed line) and $q$-values (dotted line) versus various thresholds for $p$-values ($x$-axis)

Next, we applied the uniform mixture method to this data. Due to the tendency of the $p$-value distribution to have higher density close to zero we used equally-spaced subintervals on the square root scale rather than on the original scale. The choice of the number of subintervals $M$ requires some care. Too small an $M$ obviously will lead to biased approximations while too large an $M$ will result in undersmoothed approximation of the true density. To choose a proper $M$, we followed the procedure

suggested by Linhart and Zucchini (1986, page 14) and applied it to the square root transformed $p$-values. The best choice of $M$ according to the Linhart and Zucchini procedure was found to be $M = 75$. The subinterval widths in the original scale become $d_i = (i/75)^2 - ((i-1)/75)^2$, for $i = 1, 2, \ldots, 75$ and are therefore unequal.

The estimated upper bound of the mixing proportion $p_0$ was found to be 0.723, the estimated upper bound for the FDR when a gene is claimed to be differentially expressed between current smokers and non-smokers at a $p$ -value less than 0.05 was found to be 0.1947, and the $\alpha$ was estimated to be about 0.00436 in order for the FDR to be no greater than 0.05. If we take the $\alpha$ level to be 0.00436, 461 genes will be identified as differentially expressed. The dashed line in Figure 5.2 provides a plot of the estimated FDR versus the significance threshold for $p$-values.

We also calculated $q$-values (Storey, 2002) using the R package *siggenes* (Schwender, 2006). The $q$-value associated with a $p$-value can be interpreted as the minimum FDR if we reject the null hypothesis at that particular $p$-value. The dotted line in Figure 5.2 is a plot of the $q$-values against the $p$-value threshold for significance. This is based on 0.5501 as the estimate, provided by *siggenes*, for the mixing proportion $p_0$. The maximum $p$-value corresponding to $q$-values that are less than or equal to 0.05 was found to be 0.00769 and 568 genes were found differentially expressed at this threshold.

Using the same $t$-test between current and never smokers, Spira et al (2004) found 97 differentially expressed genes at a $p$-value threshold of $1.06 \times 10^{-5}$. According to them, this threshold was selected "based on a permutation analysis performed to address the multiple comparison problem inherent in any microarray analysis". They also noted that they "chose a very stringent multiple-comparison correction"

and $p$-value threshold to "identify a subset of genes altered by cigarette smoking with only a small probability of having a false positive" (Spira et al, 2004).

Both of our proposed procedures result in a larger number of *significant* genes than what is reported in Spira et al (2004) and are closer to the results obtained using $q$ values. Not surprisingly, the uniform mixture approach yielded a smaller number of *significant* genes than the beta mixture approach since the former is essentially a nonparametric procedure. Both procedures yielded a smaller number of significant genes than the $q$-value approach since they used the largest value of $p_0$ that is consistent with the data. It is expected that, in general, the $q$-value approach would be the most liberal, the beta-mixture approach would be somewhat more conservative, and the uniform-mixture approach would be the most conservative among these three methods.

It has been reported in the literature that, due to various practical issues associated with the conduct of microarray experiments, the distribution of $p$-values of equally expressed genes may not be uniform. Efron (2004) and Nguyen (2004) suggest approaches for estimating the nonuniform $p$-value distribution associated with equally expressed genes. Both of the methods proposed here, the beta-mixture approach and the uniform-mixture approach, are easily generalized to this case. In place of the uniform [0, 1] component one simply needs to use the *estimated* nonuniform null $p$-value distribution and the proposed algorithms can be applied with very little change.

## 5.5 Simulation Study

In this section, our proposed methods in both beta mixture model and uniform mixture model are tested through simulation study. Uncorrelated array data

is simulated from mixture of uniform and beta distribution(s), the performance of the estimator $BEEH2$ of the $EH2$ and the MHDE for the beta mixture model parameters are studied, the uniform mixture model is also fitted. The empirical distribution of the $p$-values under null hypothesis suggested by Efron (2004) is applied in place of the theoretical uniform distribution.

### 5.5.1 Beta Mixture Model

**Model Selection Using $BEEH2$ – Single Date Set:**

We first generate a data set of size 1000, $\underline{U} = (u_1, u_2, \cdots u_n)$, of which 700 data points are generated from Uniform$[0, 1]$ and the other 300 are from Beta$(0.5, 2)$. Thus, this data set resembles a set of $p$-values from a group of genes of which 30% are indeed differentially expressed while the other 70% are not. The histogram of the data set is in Figure 5.3. To find the MHDE for the beta mixture model, we need to first find a kernel density estimator of the data. Fitting kernel density estimator directly is inappropriate due to the boundary issue especially near 0. We thus transform the data from $U \in [0, 1]$ to $Z \in \mathcal{R}$ by means of $Z = \tau(U) = \Phi^{-1}(u^{\frac{1}{3}})$ and fit the kernel density estimator after the transformation. Figure 5.4 depicts the the fitted kernel estimator on the transformed data. We can then fit the beta mixture model by minimum Hellinger distance and calculate $BEEH2$ from the transformed data. In fact, it can be shown that Hellinger distance is invariant under the transformation (Chapter 2).

Let $B = 500$, $BEEH2$ is calculated for each model. Table 5.3 shows that the true model and also the true best model, the mixture model with 1 beta component, has the smallest $BEEH2$ and thus was selected by our method. The $BEEH2$ for the distribution with only the uniform component is by far the largest. The rest

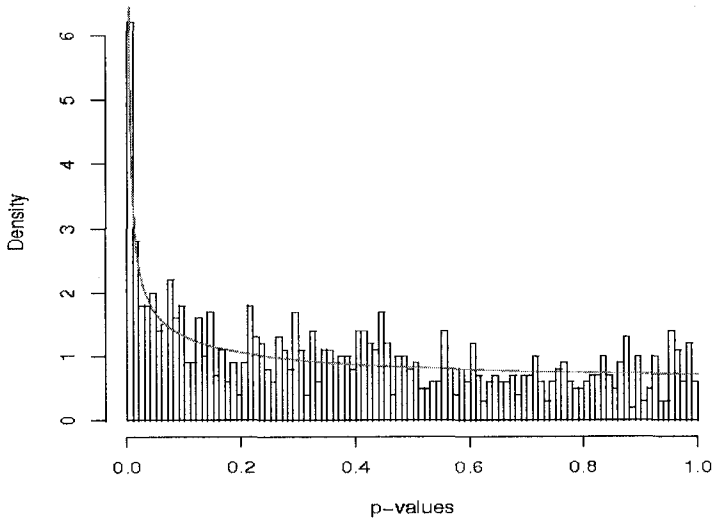


Figure 5.3: Histogram of the $p$-values with the fitted mixture density of one uniform plus one beta component. The true distribution is $f(u) = 0.7 + 0.3\beta(u; 0.5, 2)$.

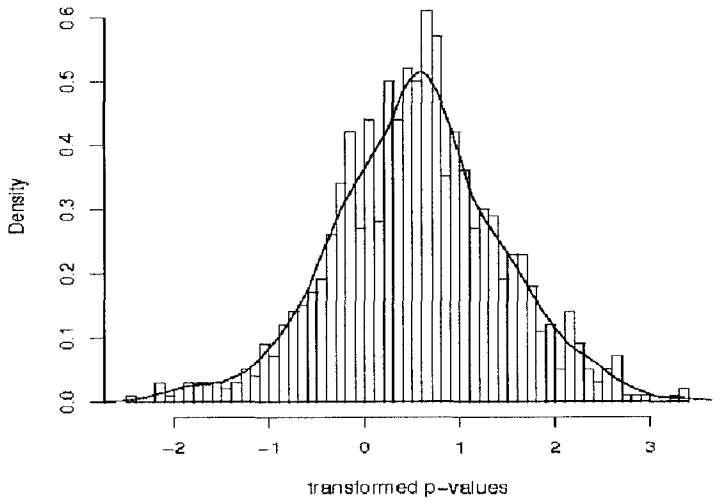


Figure 5.4: Histogram of the transformed $p$-values with the fitted kernel density estimator. The true distribution is $f(u) = 0.7 + 0.3\beta(u; 0.5, 2)$.

4 $BEEH2$'s are not that far from each other, which is not inconsistent with the fact that the fitted mixture densities with different numbers of beta components are pretty close.

| Mixture | 0 Beta | 1 Beta | 2 Beta | 3 Beta | 4 Beta |
|---|---|---|---|---|---|
| $BEEH2$ | 0.03486 | 0.00256 | 0.00259 | 0.00285 | 0.00290 |

Table 5.3: $BEEH2$ of the Mixture Models with 0-4 Beta Component(s). The true distribution is $f(u) = 0.7 + 0.3\beta(u; 0.5, 2)$.

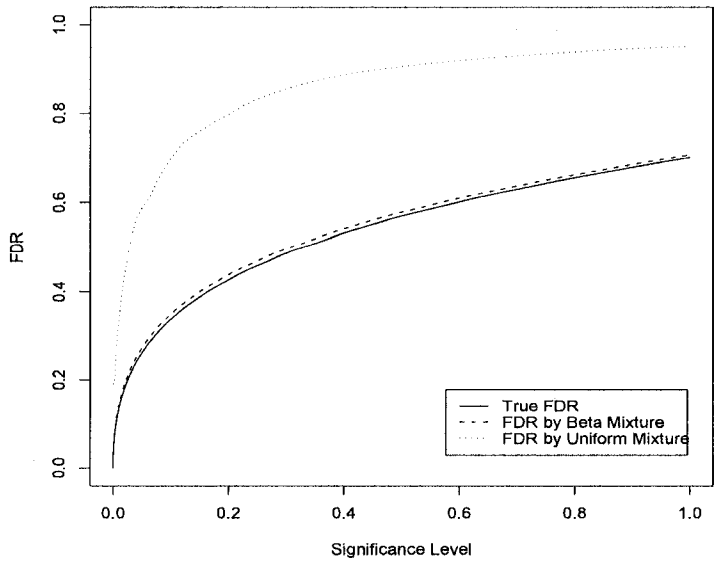


Figure 5.5: FDR respectively estimated from beta mixture model (1 uniform plus 1 beta), uniform mixture model, and from the true distribution. The true distribution is $f(u) = 0.7 + 0.3\beta(u; 0.5, 2)$.

The fitted model with one beta plus uniform from the initial data set is $g_{\hat{\theta}_1}(u) = 0.7059 + 0.2941\beta(u; 0.5056, 1.9463)$ for $u \in [0, 1]$. Figure 5.3 plots the above density on top of the histogram of the $p$-values. We can calculate the estimated FDR based on the above model for different $\alpha$, the test significance level or the "threshold" of the $p$-values for us to claim a significant difference. The estimated FDR is plotted in dashed line against the $p$-value threshold in both Figures 5.5 As shown in the plot, we can see that the estimated FDR curve is very close to the one according to the true distribution of the $p$-values (solid line). The corresponding $p$-value threshold for FDR of 0.05 is about 0.0012 by the true model and about 0.0011 according to the estimated one beta mixture model. If we set our test significance level at lower

than, say, 0.328, then about 51% or more of the significance results from the tests are estimated to be false positive, which is close to the percentage according to the true model (50%).

**Model Selection Using $BEEH2$ – Bootstrap Distribution of $BEEH2$:**

Now, consider a case in practice where the $BEEH2$ based on data for the mixture model with two beta components is slightly lower than that for the mixture model with one beta component. Do we have enough evidence to choose the former over the latter? To answer this question, we can implement the bootstrap interval discussed at the end of Section 5.3.1. For illustration purposes, we generate a data set from a 2 beta mixture model: $f(u) = 0.7 + 0.18\beta(u; 0.5, 2) + 0.12\beta(u; 0.98, 1.02)$, where the second beta component is close to the uniform distribution on $[0, 1]$. 50 samples are bootstrapped from the data set. For every bootstrapped data set, $BEEH2$ values are calculated for each of the mixture models with up to 4 beta components. The model with only the uniform component has much higher $BEEH2$ than others and thus is not of our interest. The means and one bootstrap standard deviation intervals of the 50 $BEEH2$ values for each of the other 4 mixture models are plotted in Figure 5.6. All the intervals include the minimum of mean $BEEH2$ values and therefore we chose the mixture model with one beta component since it has the smallest number of beta component(s).

**Model Estimation by Minimizing Hellinger Distance:**

To assess how well the estimators perform in terms of estimating the proportion of the equally expressed genes as well as the FDR for $\alpha = 0.05$, we simulated 500 data sets of $p$-values from the mixture distribution employed at the beginning of this session, 70% Uniform(0,1) and 30% Beta(0.5,2). The true FDR when the significance level is 0.05 is found to be 0.2613. The mixture model with one beta

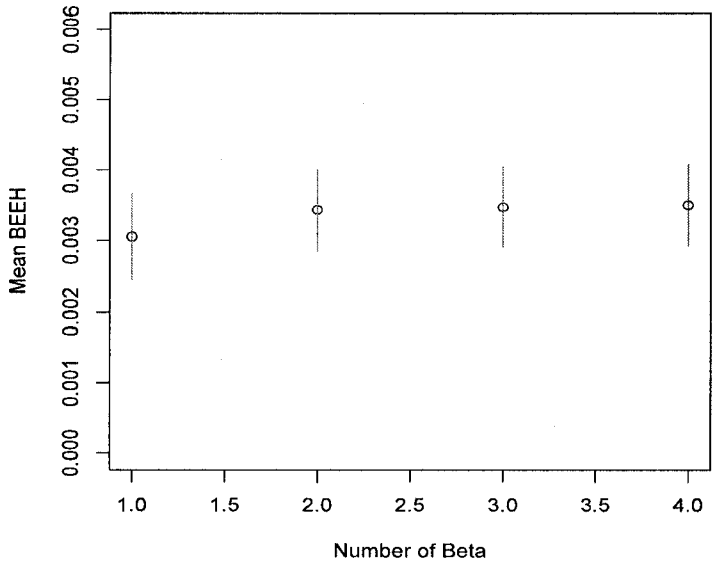


Figure 5.6: Bootstrap $BEEH2$ Intervals. $B' = 50$, the true distribution is $f(u) = 0.7 + 0.18\beta(u; 0.5, 2) + 0.12\beta(u; 0.98, 1.02)$.

component is fitted on the data sets. In addition to minimum Hellinger distance estimation, maximum likelihood method is also used. Table 5.4 gives the summary of the simulation. The averages of the estimators for $p_0$ and FDR for $\alpha = 0.05$ from both methods are quite reasonable and close to one another. It is noticed that the MHDE of the parameters and the corresponding estimated FDR have relatively less variation than the MLE and the corresponding estimated FDR.

| Estimator | $\hat{p}_0$ | | $\widehat{FDR}$ for $\alpha = 0.05$ | | |
|---|---|---|---|---|---|
| | Mean | Standard Dev. | 0.025 Quantile | Mean | 0.975 Quantile |
| MHDE | 0.7142 | 0.0399 | 0.2384 | 0.2586 | 0.2736 |
| MLE | 0.6902 | 0.097 | 0.1622 | 0.2570 | 0.3235 |

Table 5.4: Simulation Results From 500 Data Stes For Both MHDE and MLE. The true distribution is $f(u) = 0.7 + 0.3\beta(u; 0.5, 2)$ and the true FDR for $\alpha = 0.05$ is 0.2613

### 5.5.2 Uniform Mixture Model

We also apply the algorithms introduced in Section 5.3.2 to the same data set generated in the beginning of Section 5.5.1 (the true distribution is $f(u) = 0.7+0.3\beta(u; 0.5, 2)$). For the reasons we discussed in the previous section, we choose to divide the square root transformed $p$-values into $M$ equally spaced subintervals (bins). And thus the bin widths on the original scale is unequal. With $M = 20$, the 95% upper bound of the mixing proportion $p_0$ is estimated to be 0.9093. The estimated FDR is plotted as the dotted line in Figures 5.5.

### 5.5.3 Efron's Method on the Null Distribution of Equally Expressed Genes

Efron (2004) considers the choice of the distribution for the $p$-values under null distribution. Instead of the theoretical uniform distribution, an empirical distribution can be fitted in two steps on the inverse standard normal ($\Phi^{-1}$) transformed $p$-values:

$$z_i = \Phi^{-1}(p - \text{value}_i), \quad i = 1, 2, \ldots, n.$$

First fit $f(z)$ to the histogram count of the $z$'s by Poisson regression. And then the empirical distribution of the $Z$ is estimated to be a normal distribution with mean and standard deviation, say $\mu_0$ and $\sigma_0$, where $\mu_0$ and $\sigma_0$ are obtained by the center and half-width of the central peak:

$$\mu_0 = \arg\max f(z) \qquad \text{and} \qquad \sigma_0 = \left[-\frac{d^2}{dz^2}\log f(z)\right]_{\mu_0}^{-\frac{1}{2}}, \tag{5.15}$$

The corresponding empirical distribution of the $p$-values can thus be found. Since the $z$'s corresponding to the genes that are not differentially expressed are expected to be concentrated around 0 and the extreme $z$ values are more likely to

come from the differentially expressed genes, a truncation window is also considered when estimating the spread for empirical distribution. That is, only the set $\{z_i : |z_i| \leq w, i = 1, 2, \ldots, n\}$ will be used to determine $\sigma_0$, where $w$ is the truncation cut off point.

Now, applying the above mentioned method to the same data set we used in the beginning of Section 5.5.1. The empirical distribution of the $\Phi^{-1}$ transformed data is found to be $N(-0.325, 1.0934^2)$ with $w = 2$. Figure 5.7 plots the fitted distribution on the histogram of the $z$ values and Figure 5.8 plots both the empirical distribution (solid line) and the theoretical distribution (dashed line) of the $p$-values under null hypothesis against all the $p$-values.

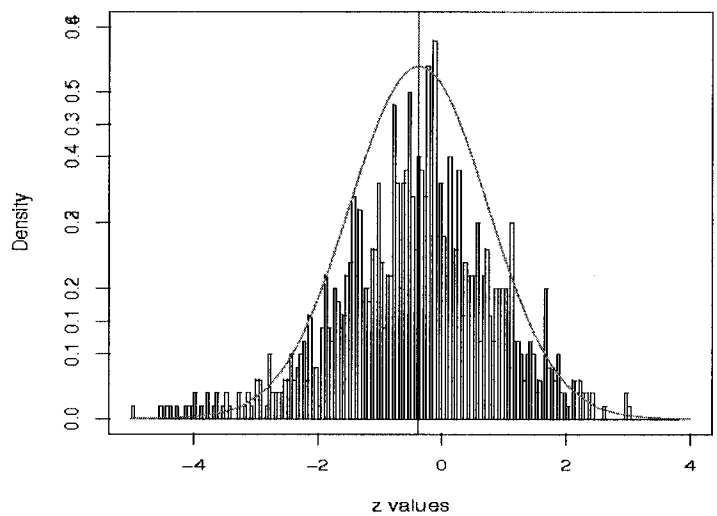


Figure 5.7: Histogram of the $z$ values with the fitted empirical null density. The true distribution is $f(u) = 0.7 + 0.3\beta(u; 0.5, 2)$.

We then fit a mixture model with the empirical distribution plus one beta distribution. The proportion of the $p$-values under null hypothesis is estimated to be 0.976. The estimated FDR is plotted as the dotted line in Figures 5.9 and 5.10 (in finer scale). The $p$-value threshold corresponding to an FDR of 0.05 is about

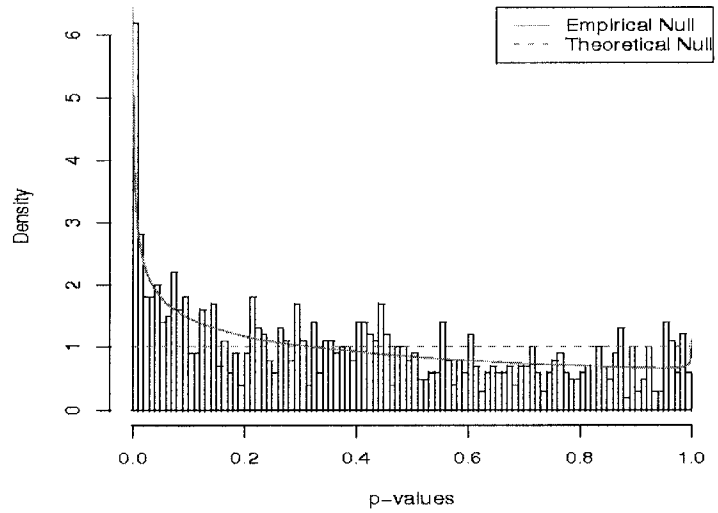


Figure 5.8: Histogram of the $p$-values with the fitted empirical null density and the theoretical uniform density for all $p$-values. The true distribution is $f(u) = 0.7 + 0.3\beta(u; 0.5, 2)$.

0.00035. If we use a threshold of 0.05, then about 67% of the time the null hypothesis is wrongly rejected, according to this model.

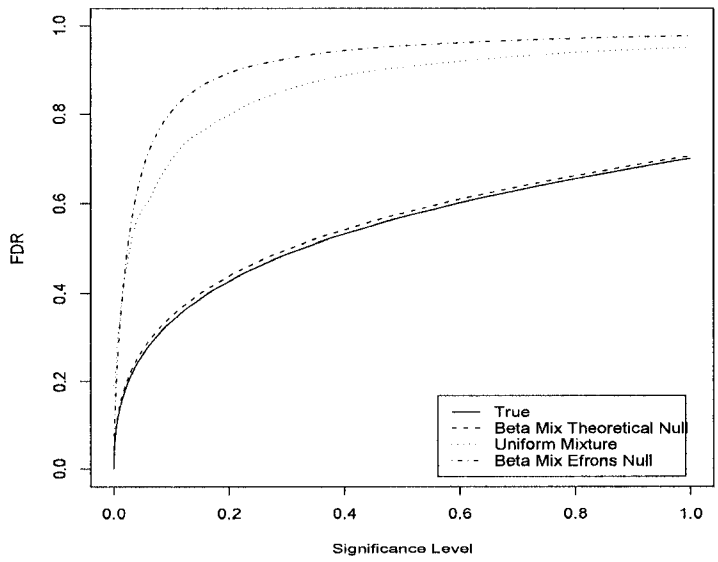


Figure 5.9: FDR respectively estimated from beta mixture model (1 uniform plus 1 beta), uniform mixture model, and mixture model with 1 empirical distribution for the null plus 1 beta. The true distribution is $f(u) = 0.7 + 0.3\beta(u; 0.5, 2)$

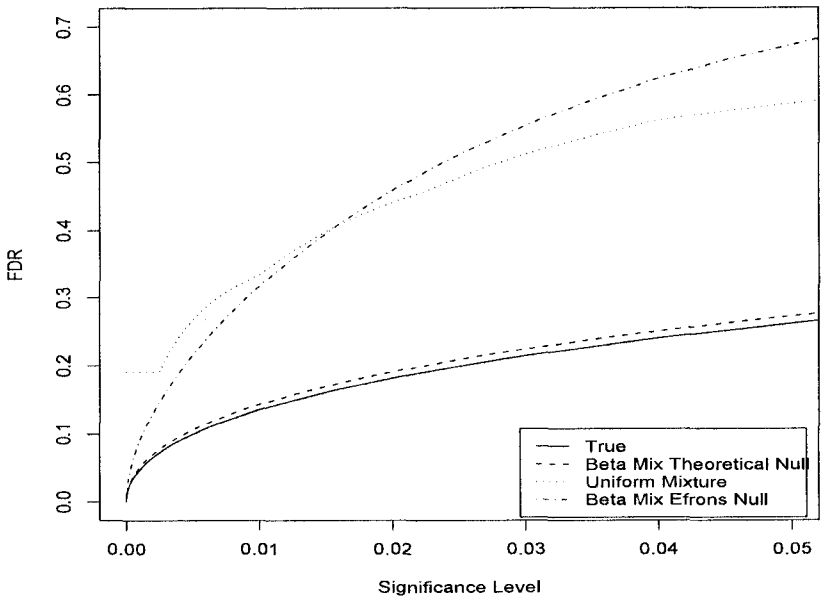


Figure 5.10: FDR respectively estimated from beta mixture model (1 uniform plus 1 beta), uniform mixture model, and mixture model with 1 empirical distribution for the null plus 1 beta – finer scale. The true distribution is $f(u) = 0.7 + 0.3\beta(u; 0.5, 2)$

**Correlated Data Sets**

We also consider the cases where the genes might be correlated. In stead of simulating $p$-values as done previously, we simulate 6708 standardized gene expressions for each of the 57 arrays, with 23 of them from control group and 34 from treatment group. Let $Y_{gi}$ represent the gene expression for the $g^{th}$ gene and the $i^{th}$ array, where $g = 1, 2, \cdots, 6708$ and $i = 1, 2, \cdots, 57$. We set up the simulation in a way so that the gene expressions for all 23 arrays in the control group independently come from a normal distribution with mean being $\mu + A_i$ and a gene-specific variances $\sigma_g^2$ while that for all 34 arrays in the treatment group independently come from a normal distribution with mean being $\mu + D_g + A_i$ and the same gene-specific variances $\sigma_g^2$. Of all $\{D_g : g = 1, 2, \cdots, 6708\}$, 80% of the randomly chosen $D_g$'s take the value of 0 while the other 20% being $D_g = T \cdot \sigma_g / \sqrt{57}$. This means that 20% of all the genes, or 1342 genes, are truly differentially expressed between the two groups. The array noise $A_i$, $i = 1, 2, \cdots, 57$ follows a normal distribution $N(0, \sigma^2/20)$ and $\sigma_g^2 = \sigma^2 \cdot U_g$ where $f \cdot U_g$ follows a chi-square distribution with degrees of freedom $f$.

Case I: $T = 5$

Let $\mu = 0$, $\sigma = 1$, $T = 5$, and $f = 30$. A data matrix with 6708 rows (genes) and 57 columns (arrays) are simulated and independent two-sample t-tests assuming unequal variances are conducted for each gene. Figure 5.11 plots the histogram of the $p$-values associated with the 80% equally expressed genes (EEG) on the top and that of the p-values for truly differentially expressed genes (DEG) on the bottom. We can see that the distribution on the upper panel resembles the uniform distribution on $[0, 1]$ while that on the lower panel can be approximated by a beta distribution with a mass near zero.

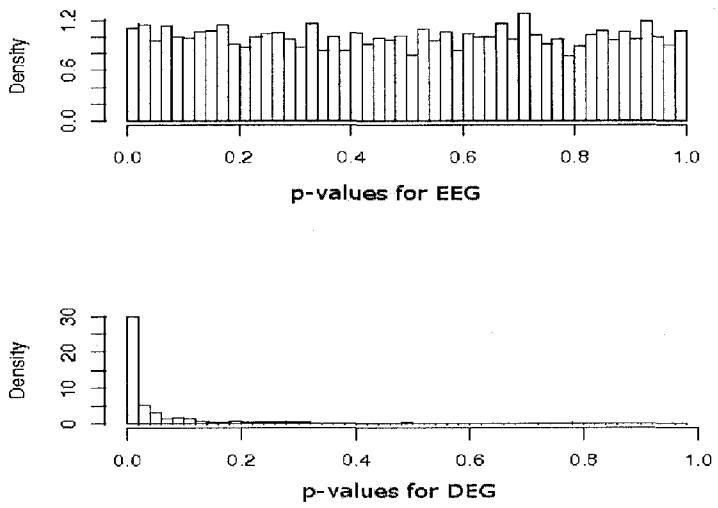


Figure 5.11: Histogram of the $p$-values of the equally expressed genes (EEG, upper panel) and for that of the differentially expressed genes (DEG, lower panel). Correlated data with $T = 5$.

We fit a mixture model with one uniform distribution and one beta distribution by minimizing Hellinger distance. The proportion of the the uniform distribution, $p_0$, which can also be interpreted as the proportion of the equally expressed genes is estimated to be 0.8131 and the beta distribution is estimated to be Beta(0.4092,16.6610). Figure 5.12 plots the estimated mixture density over all $p$-values and Figure 5.13 plots the estimated FDR over the significance levels. In order to achieve an estimated FDR of 0.05, we need to set the significance level at approximately 0.0047. 531 genes are identified as differentially expressed using this threshold.

We then apply Effron's method to get the distribution of the $p$-values of EEG's. The empirical distribution fitted on all $z$ values within the truncation window of $[-1.5, 1.5]$ is estimated to be $N(0.075, 1.1596^2)$. The truncation windows of $[-2, 2]$ and $[-3, 3]$ are also tried and the differences among the estimated variances are very small. Figure 5.14 plots the empirical distribution of the $z$ values under null hypoth-

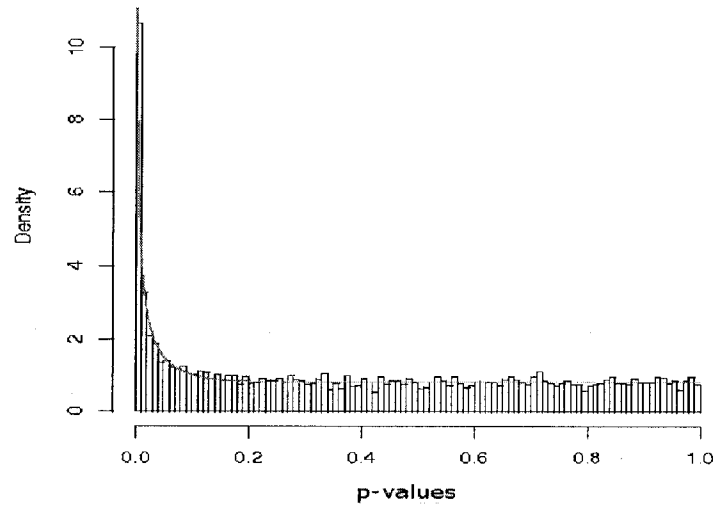


Figure 5.12: Histogram of all the $p$-values with the estimated mixture density of one theoretical null distribution and one beta distribution. Correlated data with $T = 5$.

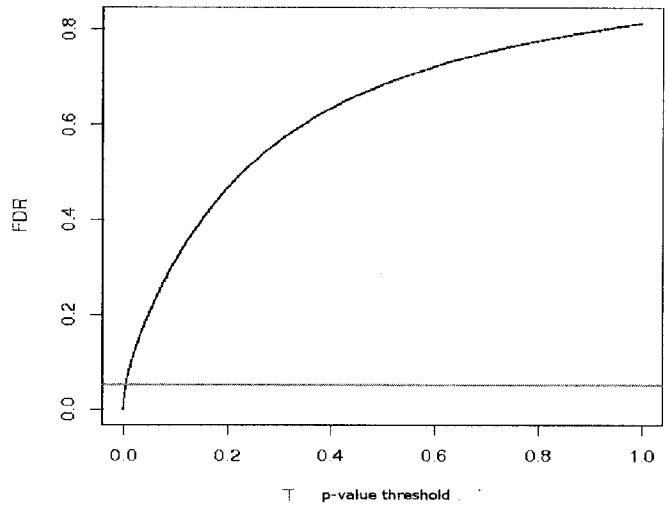


Figure 5.13: Estimated FDR based on the fitted beta mixture model with the theoretical null distribution and one beta distribution against the $p$-value threshold. Correlated data with $T = 5$.

esis on top of the histogram of all $z$ values. Figure 5.15 is Figure 5.14 transformed back into original $p$-values. We can see from the histogram that the distribution of the $z$ values is considerably asymmetric. After fitting the mixture model with the

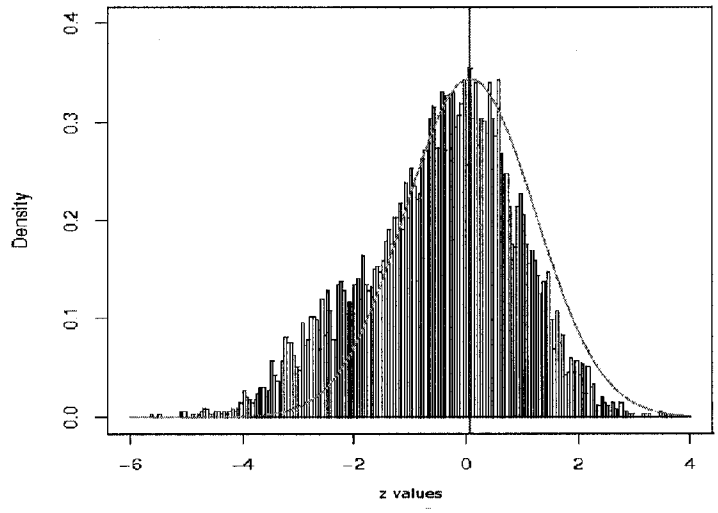


Figure 5.14: Histogram of all the $z$ values and the fitted empirical null density for the EEG. Correlated data with $T = 5$.

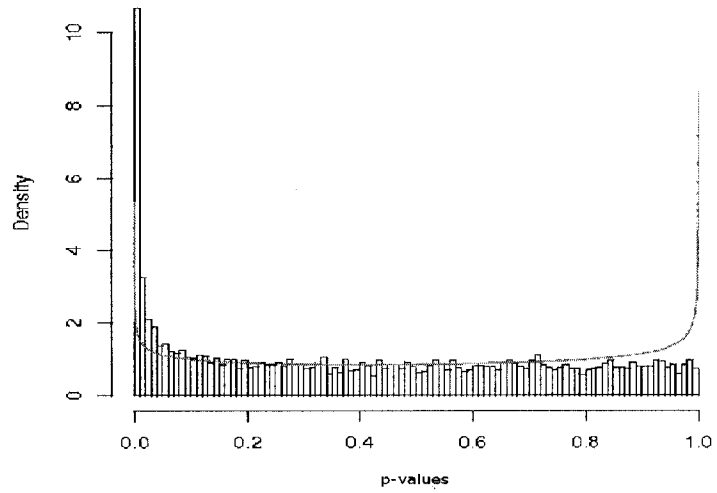


Figure 5.15: Histogram of all the $p$-values with the fitted empirical null density for the EEG. Correlated data with $T = 5$.

empirical null distribution, the proportion of the EEG is estimated to be 0.5473 and the beta density is estimated to be Beta(0.3631,1.0083). Figure 5.16 plots the estimated mixture density and Figure 5.17 plots the estimated FDR over corresponding significance levels. According to this model, we need to conclude significant differ-

ence for a particular gene between control and treatment groups only when the corresponding $p$-value is less than 0.0013, if we want to control the FDR at 0.05. And 303 genes are found differentially expressed using this threshold.

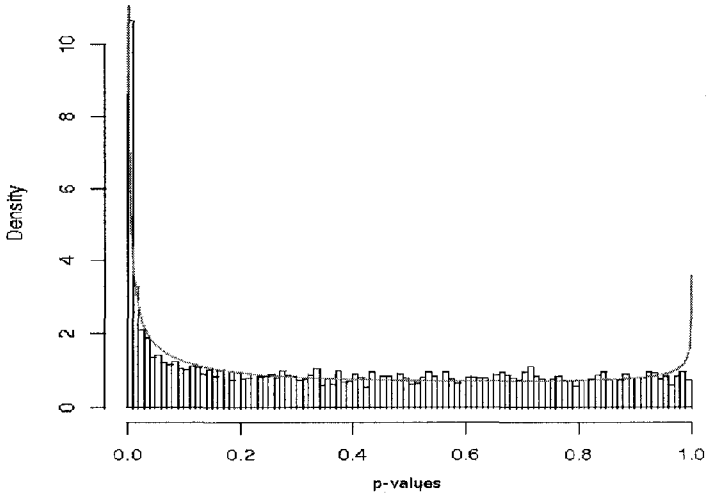


Figure 5.16: Histogram of all the $p$-values with the fitted mixture density of the empirical null distribution and one beta distribution. Correlated data with $T = 5$.

Case II: $T = 1.8$

Now change $T$ into 1.8 so that the true average fold change for those truly differentially expressed genes gets smaller and thus harder to detect. Figure 5.18 displays the histogram of the $p$-values under null hypothesis (upper penal) and that of the $p$-values under alternative hypothesis (lower penal). Again, it is plausible to approximate the former distribution with a uniform distribution on $[0, 1]$ and the latter with a beta distribution. The fitted mixture density is plotted in Figure 5.19. The proportion of the uniform is estimated to be 0.9643 and the estimated beta distribution is Beta(0.9999, 91.7611). The estimated FDR is plotted against significance level in Figure 5.20. With this data set, the estimated FDR is relatively high for any given significance level. The threshold associated with the FDR of 0.05

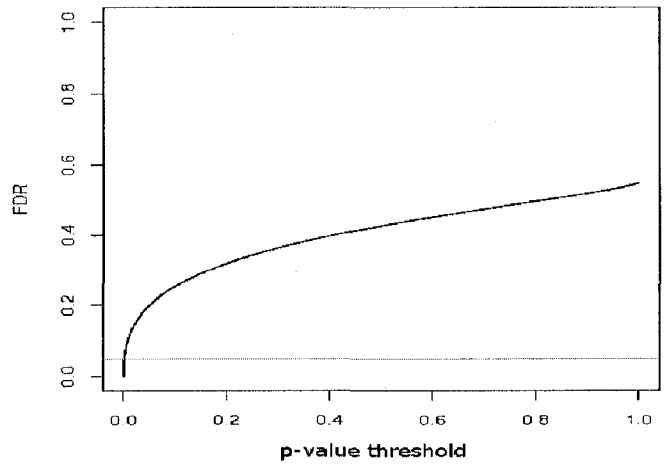


Figure 5.17: Estimated FDR based on the fitted mixture model with the empirical null distribution and one beta distribution against the $p$-value threshold. Correlated data with $T = 5$.

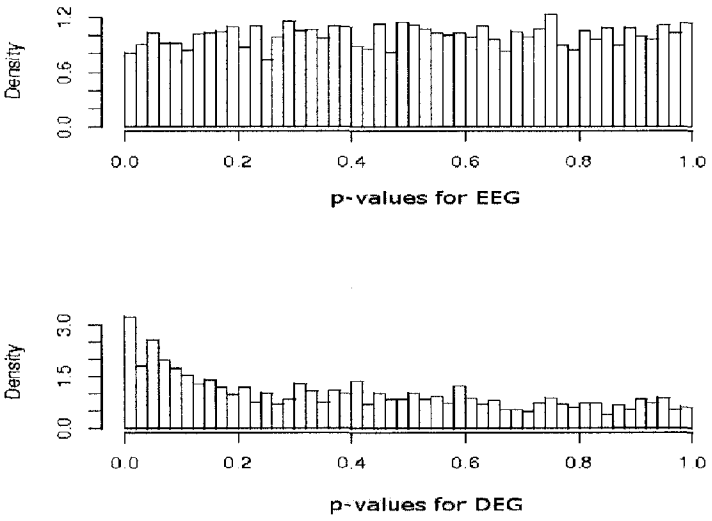


Figure 5.18: Histogram for the $p$-values of the equally expressed genes (EEG, upper panel) and for that of the differentially expressed genes (DEG, lower panel). Correlated data with $T = 1.8$

is estimated to be $1 \times e^{-12}$ and no genes are found to be differentially expressed using this significance level. If we choose the significance level to be 0.05, on the

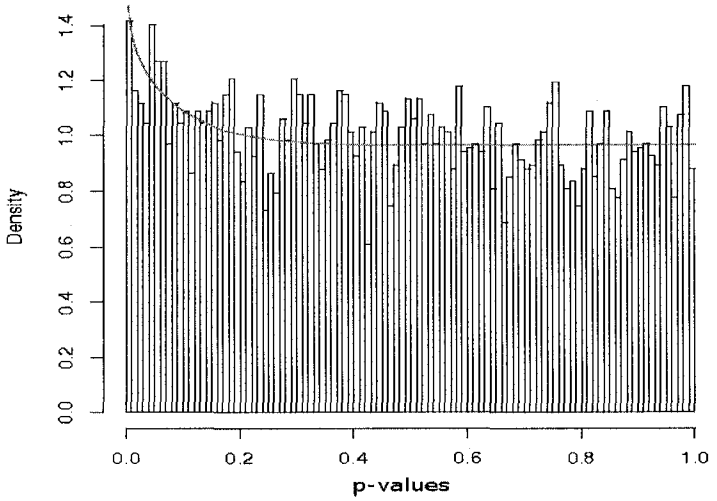


Figure 5.19: Histogram for all the $p$-values with the estimated mixture density of one theoretical null distribution and one beta distribution. Correlated data with $T = 1.8$

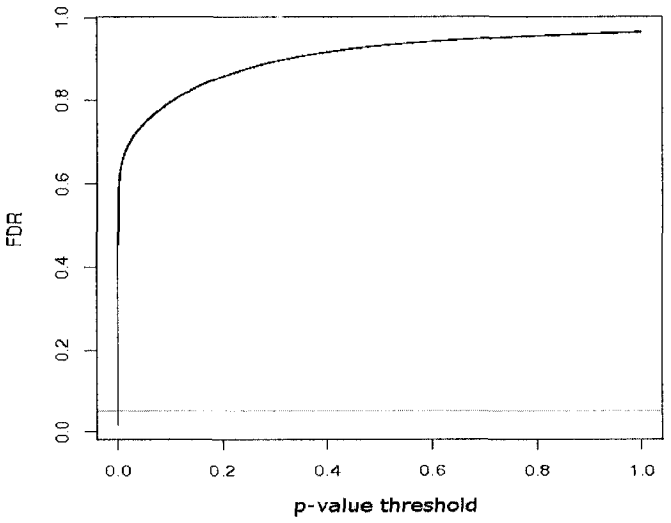


Figure 5.20: Estimated FDR based on the estimated mixture density with one theoretical null distribution and one beta distribution against the $p$-value threshold. Correlated data with $T = 1.8$

other hand, 412 genes are identified as differentially expressed among which about

74 percent are expected to be actually equally expressed between the control group and the treatment group according to this model.

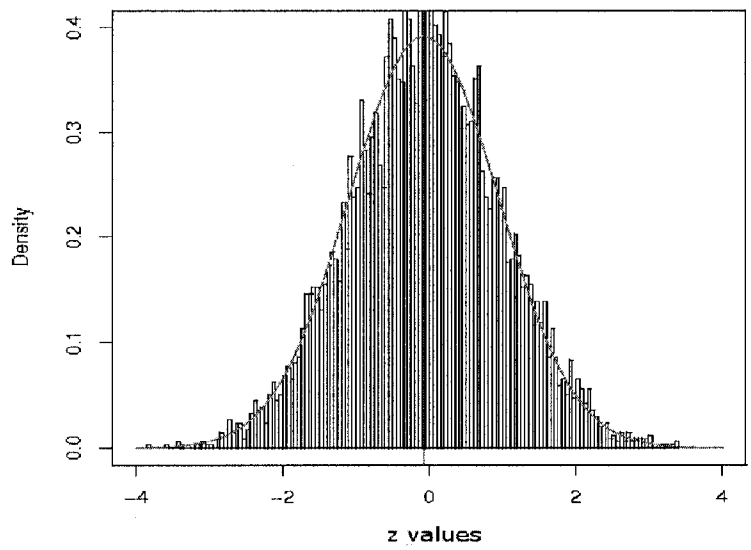


Figure 5.21: Histogram for all the $z$ values and the fitted empirical null density for the EEG. Correlated data with $T = 1.8$

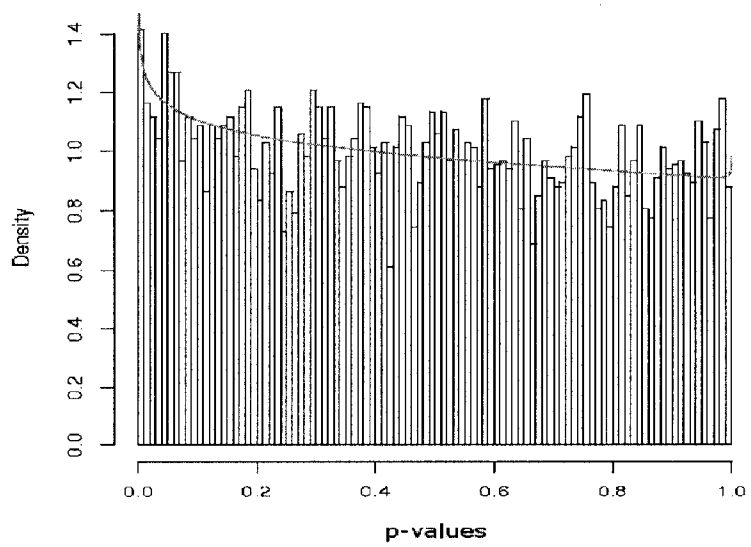


Figure 5.22: Histogram for all the $p$-values with the fitted empirical null density for the EEG. Correlated data with $T = 1.8$

Applying Efron's method, the empirical distribution of the $\Phi^{-1}$ transformed $p$-values for EEG is found to be $N(-0.075, 1.0187^2)$ with a truncation window of $[-1.5, 1.5]$. Figures 5.21 and 5.22 plot the empirical distribution over all $z$ values and all $p$-values respectively. It is interesting to notice that the empirical distribution that is supposed to be of the $p$-values for EEG only seems to fit the overall $p$-value quite well. After fitting the mixture model, $p_0$ is estimated to be 0.9980 and the fitted mixture density is plotted in Figure 5.23. Figure 5.24 plots the estimated FDR over the significance level. This curve of the estimated FDR first goes downwards for a short while at the left end and then climbs back upwards. Note from Figure 5.22 that the empirical distribution of the p-values under null hypothesis has a mass near zero. This contributes to the contra-intuitive part of the curve that goes downwards at the beginning. The estimated FDR is very high for all significance levels. The minimum estimated FDR is found to be around 0.8976.

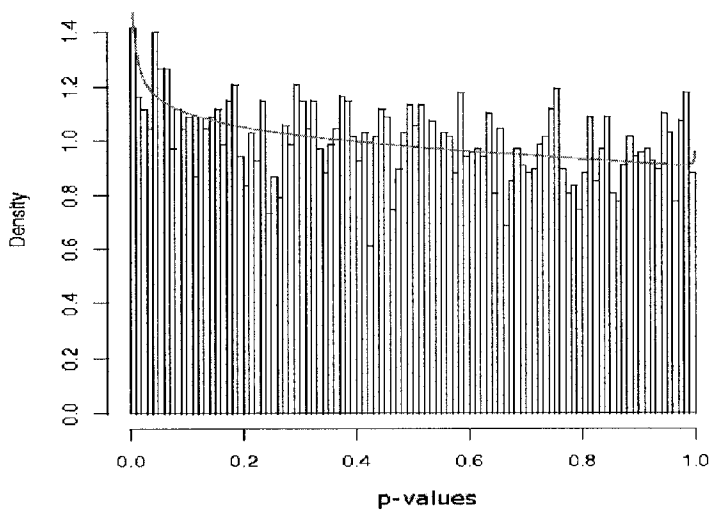


Figure 5.23: Histogram for the $p$-values with the fitted mixture density of the empirical null distribution and one beta distribution. Correlated data with $T = 1.8$.

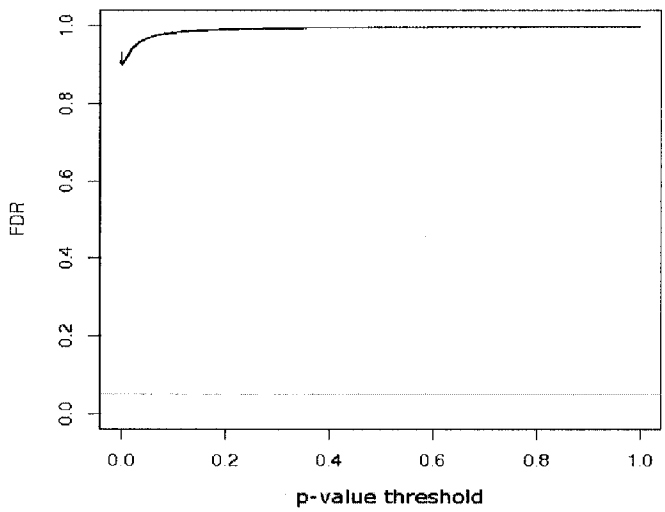


Figure 5.24: Estimated FDR based on the fitted mixture density with the empirical null distribution and one beta distribution against the $p$-value threshold. Correlated data with $T = 1.8$.

# Chapter 6

# CONCLUSIONS AND FUTURE WORK

## 6.1 Conclusions

In this dissertation, we considered general model selection problems. In these problems, the true model is unknown (and is not assumed to come from a parametric family) and one or more approximating parametric families of models are given along with strategies for estimating the parameters using data. We are required to select a parametric family and a corresponding estimating method (if more than one estimation method is considered) that results in an approximating model that is closest, in some sense, to the true model. Our decision is based on a set of observed data. The model selection methods we develop follow the principles of model selection based on distances or discrepancies. The Hellinger distance is the discrepancy we choose to use and "true best model" is the one among the approximating models that has the smallest expected squared Hellinger distance.

Two bootstrap-based estimators of the expected squared Hellinger distance, $BEEH2$ and $PEEH2$, are proposed in Chapter 2 and their large sample properties were investigated. Limited simulation studies were conducted to examine their small sample behavior and we concluded that the performance of the proposed methods are satisfactory in the situations examined.

Our model selection strategy was applied to problem of model selection among ANOVA models, where typically some of the approximating models are sub-models

of others. The properties of expected squared Hellinger distance under balanced ANOVA model settings where the error terms are independently and identically normally distributed with known variance were studied. This led to a specific model selection strategy using $PEEH2$. Limited simulation studies were carried out to evaluate the small sample performance which was deemed satisfactory in the settings considered.

In Chapter 5 we applied our model selection method to modeling the $p$-values from microarray data analysis. We considered two different mixture models for modeling the distribution of the $p$-values for differential expression. The first mixture model is one that uses a single Uniform $[0, 1]$ density and one or more nonuniform Beta densities. The second mixture model is a mixture model with a Uniform $[0, 1]$ distribution and $M$ other uniform distributions on disjoint subintervals of $[0, 1]$. Here we implement Hellinger distance estimation, instead of the often used method of maximum likelihood, in fitting such mixture models. For the Beta mixture model we compare $BEEH2$ to decide for the number of Beta components. Once the number of components is determined the parameters associated with these components are estimated and the false discovery rate (FDR) for any given significance level $\alpha$ can be computed from the estimated mixture Beta distribution. In the second approach with a mixture of uniform densities, we provide an iterative algorithm and a bootstrap testing procedure using which one can compute an upper confidence bound for the mixing proportion associated with the uniform component. This in turn leads to an upper bound on the FDR associated with any prespecified significance level $\alpha$ for declaring genes as differentially expressed. We have illustrated the application of the procedures by using a published microarray data set downloadable from GEO. Simulation studies are also carried out to test the performance

of the procedures. Finally, we implement the empirical distribution of $p$-values for equally expressed genes proposed by Efron (Efron, 2004) in fitting the mixture Beta distribution.

Our proposed model selection method based on estimated expected squared Hellinger distance is motivated by general model selection problems and can be applied to a wide range of specific modeling problems with minor adjustment. In fact, our method does not require the typical assumptions such as the true model being within the approximating family. Our simulation study shows that the performance of our proposed model selection method and procedures is satisfactory.

### 6.1.1 Future Work

Currently, the distribution of $H(f, g_{\hat{\theta}})$ is approximated by the bootstrap distribution of $H(f, g_{\hat{\theta}^*})$. The computation is challenging and time-consuming with large sample sizes. It is necessary to look at the application of other efficient approaches in approximating the distribution of $H(f, g_{\hat{\theta}})$. Two examples of such approaches that we can consider are the cross-validation method and the jackknife method.

In Chapter 4, we proposed our model selection strategy along with two grouping procedures. These grouping methods, which attempt to identify the group of models that are not far away from the one with the smallest $EH2$, are heuristic. It remains to explore more tools to identify such a group of models.

The application of our model selection method in survival analysis is also an interesting area yet to be explored. One attractive property of the Hellinger distance is that, unlike the K-L discrepancy, it is not subject to the constraint that the approximating distributions must have the same support as the true model. In survival analysis, it is not unusual to assume two-parameter Exponential distribution

models and thus the approximating distribution may have different support than the underlying true distribution. The Exponential distribution example in Chapter 4 is exploratory in this nature and more work on both the theoretical front and practical applications is needed. It should also be interesting to look at other non-Normal error terms besides the exponential case.

Finally, in this dissertation we considered balanced factorial ANOVA models with fixed factors. Expanding our investigations to more general mixed models is also a topic for future work.

## References


[1] Akaike, H. (1973), Information Theory And An Extension of The Maximum Likelihood Principle, *2nd International Symposium on Information Theory*, 267-281.

[2] Allison, D. B. and Gadbury, G. L. and Heo, M. and Fernandez, J.R. and Lee, CK. and Prolla, T.A. and Weindruch, R. (2002), A mixture model approach for the analysis of microarray gene expression data, *Computational Statistics and Data Analysis*, **39**, 1-20.

[3] Apostol, T. M. (1974), Mathematical Analysis, Addison-Wesley, Reading, MA.

[4] Benjamini, Y. and Hochberg, Y. (1995), Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society B*, **57**, 289-300.

[5] Benjamini, Y. and Hochberg, Y. (2000), On the adaptive control of the false discovery rate in multiple testing with independent statistics, *Journal of Educational and Behavioral Statistics*, **25**, 60-83.

[6] Beran, R. (1977), Minimum Hellinger Distance Estimates for Parametric Models, *The Annals of Statitics*, **5**, 445-463.

[7] Billingsley, P. (1995), Probability and Measure, Wiley, New York.

[8] Birge, L. (2004), Model Selection for Gaussian Regression with Random Design, *Bernoulli*, **10**, 1039-1051.

[9] Broberg, P. (2005), A comparative review of estimates of the proportion of unchanged genes and the false discovery rate, *BMC Bioinformatics*, **6**, 199.

[10] Burnham, K. and Anderson, D. (2002), Model Selection and Multi-Model Inference, Springer-Verlag, New York.

[11] Cao R. and Cuevas, A. and Fraiman, R. (1995), Minimum distance density-based estimation, *Computational Statistics and Data Analysis*, **20**, 611-631.

[12] Cavanaugh, J. E. and Shumway, R. II. (1997), A Bootstrap Variant of AIC for State-Space Model Selection, *Statistica Sinica*, **7**, 473-496.

[13] Chung, H. and Lee, K. and Koo, J. (1996), A Note on Bootstrap Model Selection Criterion, *Statistica and Probability Letter*, **26**, 35-41.

[14] Cutler A. and Cordero-Brana, O. I. (1996), Minimum Hellinger Distance Estimation for Finite Mixture Models, *Journal of the American Statistical Association*, **91**, 1716-1721.

[15] Darling, D. A. (1957), The Kolmogorov-Smirnov, Cramer-von Mises Tests, *The Annals of Mathematical Statistics*, **28**, 823-838.

[16] Dalmasso, C. and Broet, P. and Moreau, T. (2005), A simple procedure for estimating the false discovery rate, *Bioinformatics*, **21**, 660-668.

[17] Devroye L. and Penrod C. S. (1984), The Consistency of Automatic Kernel Density Estimates, *The Annals of Statistics*, **12**, 1231-1249.

[18] Donoho, D. L. and Liu, R. C. (1988), The 'Automatic' Robustness of Minimum Distance Functionals, *The Annals of Statistics* , **16**, 552-586.

[19] Efron, B. (1979), Bootstrap Methods: Another Look at the Jackknife, *The Annals of Statistics*, **7**, 1-26.

[20] Efron, B. (1983), Model Selection and The Bootstrap, *Mathematical Social Sciences*, **5**, 236-236.

[21] Efron, B. and Tibshirani, R. and Storey, J.D. and Tusher, V. (2001), Empirical Bayes Analysis of A Microarray Experiment, *Journal of the American Statistical Association*, **96**, 1151-1160.

[22] Efron, B. and Tibshirani, R. (2002), Empirical bayes methods and false discovery rates for microarrays, *Genetic Epidemiology*, **23**, 70-86.

[23] Efron, B. (2004), Large-Scale Simultaneous Hypothesis Testing: the choice of a null hypothesis, *Journal of the American Statistical Association*, **99**, 96-104.

[24] Everitt, B. S. and Hand, D. J. (1981), Finite Mixture Distributions, Chapman and Hall, London.

[25] Evans, I. G. and Nigm, A. H. M. (1980), Bayesian Prediction for the Left Truncated Exponential Distribution, *Technometrics*, **22**, 201-204.

[26] Gentleman, R. C. and Carey, V. J. and Bates, D. M. and Bolstad, B. and Dettling, M. and Dudoit, S. and Ellis, B. and Gautier, L. and Ge, Y. and Gentry, J. and Hornik, K. and Hothorn, T and Huber, W and Iacus, S. and Irizarry, R. and Li, F. L. C. and Maechler, M. and Rossini, A. J. and Sawitzki, G. and Smith, C. and Smyth, G. and Tierney, L. and Yang J. Y. H. and Zhang J. (2004), Bioconductor: Open software development for computational biology and bioinformatics, *Genome Biology*, **5**, R80.

[27] GEO, (2007). Internet site: http://www.ncbi.nlm.nih.gov/geo/.

[28] Hastie, T. and Tibshirani, R. and Friedman, J. (2001), The elements of statistical learning: data mining, inference, and prediction, Springer-Verlag, New York.

[29] Hocking, R. R., (1996), Methods and Application of Linear Models: Regression and the Analysis of Variance, Wiley, New York.

[30] Hurvich, C. M. and Tsai, C. L. (1989). Regression and Times Series Model Selection in Small Samples, *Biometrika*, **76**, 297-307.

[31] Ishiguro, M. and Sakamoto, Y. (1991), WIC: An Estimation-Free Information Criterion, *Research memorandum of the Institute of Statistical Mathematics*, Tokyo, 410.

[32] Ishiguro, M. and Sakamoto, Y. and Kitagawa, G. (1997), Bootstrapping Log Likelihood and EIC, an Extension of AIC, *Annals of Institute of Statistical Mathematics*, **49**, 411-434.

[33] Irizarry, R. A. and Gautier, L. and Bolstad, B. M. and Miller, C. with contributions from Astrand, M. and Cope, L. M. and Gentleman, R. and Gentry,J. and Halling, C. and Huber, W. and MacDonald J. and Rubinstein, B. I. P. and Workman, C. and Zhang, J. (2006), affy: Methods for Affymetrix Oligonucleotide Arrays, R package, version 1.12.2.

[34] Kraft, C. (1955), Some Conditions for Consistency and Uniform Consistency of Statistical Procedures, *University of California Publications in Statistics*, **2**, 125-142.

[35] Kullback, S. (1959), Information Theory and Statistics, John Wiley, New York.

[36] Lai, Y. (2006), A statistical method for estimating the proportion of differentially expressed genes, *Computational Biology and Chemistry*, **30**, 193-202.

[37] Langaas, M. and Lindqvist, B. H. (2005), Estimating the proportion of true null hypotheses, with application to DNA microarray data, *Journal of the Royal Statistical Society B*, **67**, 555-572.

[38] LeCam, L. (1970), On the Assumptions Used to Prove Asypmtotic Normality of Maximum Likeliood Esimates, *Annals of Mathematical Statistics*, **41**, 802-828.

[39] Lehmann, L. and Casella G. (1998), Theory of Point Estimation, Springer, New York.

[40] Liao, J. G. and Lin, Y. and Sevanayagam, Z. E. and Shih, W. J. (2004), A mixture model for estimating the local false discovery rate in DNA microarray analysis, *Bioinformatics*, **20**, 2694-2701.

[41] Linhart, H. and Zucchini, W. (1986), Model Selection, John Wiley, New York.

[42] Lu, Z. and Hui, Y. V. and Lee A. H. (2003), Minimum Hellinger Distance Estimation for Finte Mixtures of Poisson Regression Models and ITs Application, *Biometrics*, **59**, 1016-1026.

[43] Mallows, C. L. (1966), Choosing a Subset Regression, Joint Statistical Meetings, Los Angeles, CA

[44] Mandal, A. (2006), Some Contribution to Design Theory and Applications, Ph.D. Dissertation, Georgia Institute of Technology

[45] Matusita, K. (1955), Decision rules on the distance, for problems of fit, two-samples, and estimation, *Annals of Mathematical Statistics*, **26**, 631-640.

[46] McQuarrie, A. D. R. and Tsai, C. (1998), Regression and Time Series Model Selection, World Scientific Publishing Company, Singapore.

[47] Mosteller, F. and Tukey, J. W. (1968), Data analysis including statistics. In: Lindzey, G. and Aronson, E. (Eds.), *The Handbook of Social Psychology*, **2**, Addison-Wesley, Reading, MA.

[48] Neter, J. and Kutner, M. H. and Wasserman, W. and Nachtsheim, C. J. (1996), Applied Linear Regression Models, McGraw-Hill/Irwin, Homewood, IL.

[49] Nguyen, D. V. (2004), On estimating the proportion of true null hypotheses for false discovery rate controlling procedures in exploratory DNA microarray studies, *Computational Statistics and Data Analysis*, **47**, 611-637.

[50] Parker, R. A. and Rothenberg, R. B. (1988), Identifying Important Results From Multiple Statistical Tests, *Statistics in Medicine*, **7**, 1031-1043.

[51] Pounds, S. and Morris, S. W. (2003), Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values, *Bioinformatics*, **19**, 1236-1242.

[52] Pounds, S. and Cheng, C. (2004), Improving false discovery rate estimation, *Bioinformatics*, **20**, 1737-1745.

[53] R Development Core Team (2006), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

[54] Sackrowitz, H. and Samuel-Cahn, E. (1999), P values as random variables-expected p values, *Journal of the American Statistical Association*, **53**, 326-331.

[55] Schwarz, G. (1978), Estimating The Dimention of A Model, *Annals of Statistics*, **6**, 461-464.

[56] Shao, J. (1996), Bootstrap Model Selection, *Journal of the American Statistical Association*, **91**, 655-665.

[57] Schwender, H. (2006), siggenes: SAM and Efron's empirical Bayes approaches, R package, version 1.8.0.

[58] Silverman, B. W. (1986), Density estimation for statistics and data analysis, Chapman and Hall, London.

[59] Spira, A. and Beane, J. and Shah, V. and Liu, G. and Schembri, F. and Yang, X. and Palma, J. and Brody, J.S. (2004), Effects of Cigarette Smoke on the Human Airway Epithelial Cell Transcriptome, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 10143-10148.

[60] Storey, J. D. (2002), A direct approach to false discovery rates, *Jounal of the Royal Statistical Society B*, **64**, 479-498.

[61] Storey, J. D. and Tibshirani, R. (2003), Statistical significance for genomewide studies, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440-9445.

[62] Tadesse, M. G. and Ibrahim, J. G. and Vannucci, M. and Gentlemen, R. (2005), Wavelet Thresholding with Bayesian False Discovery Rate Control, *Biometrics*, **61**, 25-35.

[63] Takeuchi, K. (1976), Distribution of An Information Statistic and The Criterion for The Optimal Model, *Mathematical Science*, **153**, 12-18.

[64] Tamura, R. N. and Boos, D. D. (1986), Minimum Hellinger Distance Estimation for Multivariate Location and Covariance, *Journal of the American Statistical Association*, **81**, 223-229.

[65] Tsai, C.-A. and Hsueh, H.-m. and Chen, J. J (2003), Estimation of False Discovery Rates in Multiple Testing: Application to Gene Microarray Data, *Biometrics*, **59**, 1071-1081.

[66] van der Vaart A. W. (1998), Asymptotic Statistics, Cambridge University Press, Cambridge.

[67] Varde, S. D. (1969), Life Testing and Reliability Estimation for the Two Parameter Exponential Distribution, *Journal of the American Statistical Association*, **64**, 621-631.