A Modified Bayesian Power Prior Approach with Applications in Water Quality Evaluation

Yuyan Duan

Dissertation submitted to the Faculty of
Virginia Polytechnic Institute and State University
in fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Statistics

Keying Ye, Chair
Eric P. Smith, Co-Chair
Ilya A. Lipkovich
Samantha Bates Prins
Dan Spitzner

November, 2005
Blacksburg, Virginia

Keywords: Prior elicitation, Historical data, Power prior, Water quality standards.

A Modified Bayesian Power Prior Approach with Applications in Water Quality Evaluation

Yuyan Duan

(ABSTRACT)

This research is motivated by an issue frequently encountered in environmental water quality evaluation. Many times, the sample size of water monitoring data is too small to have adequate power. Here, we present a Bayesian power prior approach by incorporating the current data and historical data and/or the data collected at neighboring stations to make stronger statistical inferences on the parameters of interest.

The elicitation of power prior distributions is based on the availability of historical data, and is realized by raising the likelihood function of the historical data to a fractional power. The *power prior* Bayesian analysis has been proven to be a useful class of informative priors in Bayesian inference. In this dissertation, we propose a modified approach to constructing the joint power prior distribution for the parameter of interest, θ, and the power parameter, δ. The power parameter δ, in this modified approach, quantifies the heterogeneity between current and historical data automatically, and hence controls the influence of historical data on the current study in a sensible way. In addition, the modified power prior needs little to ensure its propriety. The properties of the modified power prior and its posterior distribution are examined for the Bernoulli and normal populations. The modified and the original power prior approaches are compared empirically in terms of the mean squared error (MSE) of estimated θ as well as the behavior of the power parameter. Furthermore, the extension of the modified power prior to multiple historical data sets is discussed, followed by its comparison with the random effects model.

Several sets of water quality data are studied in this dissertation to illustrate the implementation of the modified power prior approach with normal and Bernoulli models. Since the power prior method uses information from sources other than current data, it has advantages in terms of power and estimation precision for decisions with small sample sizes, relative to methods that ignore prior information.

Dedication

To my parents and my dear son, Larry.

Acknowledgements

I would like to express my deepest appreciation to my advisors, Dr. Keying Ye and Dr. Eric P. Smith for all their guidance and patience. They have been there for me every step of the way, encouraging and helping, as mentors and friends. I would also like to thank my committee members, Dr. Ilya A. Lipkovich, Dr. Samantha B. Prins, and Dr. Dan Spitzner, for their valuable comments and suggestions.

I feel so lucky to be surrounded by many loving friends, Zhengrong, Huizi, Younan, Mingjin, and Li Wang *etc.*, during the four and half years. All the happiness shared with them and Keying will stay in my memory forever.

My special thanks go to my husband, Bo, for his support over all these years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Water quality standards define conditions for acceptable water quality. For a particular constituent (e.g. pH, dissolved oxygen, biological oxygen demand), a numerical criterion is used to define the acceptable level. To assess water quality standards, measurements of water quality under the Clean Water Act are collected on a regular basis over a period of time. The data are analyzed to evaluate the percentage of samples in violation of the standard (i.e. below or above the numerical criterion). Approaches that have been discussed in the literature include the *US Environmental Protection Agency* (USEPA) raw score method, binomial test ([30]), acceptance sampling by variables ([31],[37], and [32]), and a Bayesian test on the percentile ([40]).

Unfortunately, because decisions are based on data from a limited time period (for water monitoring data, many current samples have only two years of data; and each year there might be only four observations available since in many sites, the measurements are taken quarterly), the sample size is often inadequate to provide necessary precision in parameter estimates and power for testing hypotheses. In such situations, *historical data*, a data set from similar studies or a data set from previous time periods on the same site, can be very helpful in interpreting the current status of water quality. Due to the nature of updating information sequentially, it is natural to use a Bayesian approach with an informative prior on the model parameters to incorporate the historical data into the current study. A traditional Bayesian approach in incorporating historical data is to construct an informative prior based on the historical data and then to combine this prior with the likelihood for the current data to yield the posterior distribution for statistical inference. This implies a simple pooling of current data and historical data, since the two data sets are equally weighted, and can be well

justified by assuming current and historical data are from the same population. However, the population parameters may change over time, or over different sites, although current and historical data are usually assumed to follow distributions in the same family. If the sample size of the historical data is much larger than that of current data and heterogeneity exists between the current and previous studies, historical data would dominate the analysis and the data pooling may result in misleading conclusions.

To address this issue, Ibrahim and Chen ([18]) proposed the concept of the *power prior*, based on the notion of the availability of historical data. The basic idea is to let a power parameter δ, with $(0 \leq \delta \leq 1)$, tell us how much historical data information are to be used in the current study. Ibrahim and Chen ([18]) and Chen *et al.* ([10]) demonstrated how to construct power priors and discussed the general conditions for the propriety of the posterior distribution. They also examined the power prior approach for generalized linear models, generalized linear mixed models, semiparametric proportional hazards models, and cure rate models with real data examples. The power prior methodology with a fixed δ has been well established by Ibrahim *et al.* ([19]). They gave a formal justification of the power prior as an optimal class of informative priors, for the case when δ is fixed. Furthermore, they pointed out that δ can be treated as a random variable, and defined a joint power prior on (θ, δ), where θ is the parameter of interest. We refer to this joint power prior proposed by Ibrahim and Chen ([18]) as the original power prior.

However, one problem with the original power prior approach occurs in the application of such priors to Bernoulli and normal mean models. The influence of the historical data is generally small, i.e., δ is close to 0, no matter how compatible the current and historical data are. In such a case, δ is underestimated, and the inference on θ is not much different from the inference ignoring the historical data. More importantly, in the original power prior approach, proportional likelihoods based on the historical data do not produce the same posteriors on model parameters.

Therefore, we propose a modified joint power prior for (θ, δ). In the modified approach, the power parameter quantifies the heterogeneity between current and historical data automatically, and hence controls the influence of historical data on the current study in a sensible way. In addition, the modified power prior needs little to ensure its propriety, and agrees with the likelihood principle.

The rest of this dissertation is organized as follows. In Chapter 2, we will review the literature on water quality evaluation and prior elicitation, especially the research done by

Ibrahim and Chen on the original power priors. In Chapter 3, the general development of the modified power prior approach is displayed in detail and certain properties of the approach under Bernoulli and normal populations are discussed. In addition, the modified and original power prior methods are compared with each other from several different aspects. Specifically, we investigate the behavior of the power parameter and how it is affected by the compatibility between historical and current samples as well as the availability of historical data. Furthermore, the trend of the posterior means of θ are compared under two power prior approaches. Various scenarios are covered for the comparison of two power prior methods in mean squared error (MSE). In Chapter 4, we investigate how to incorporate multiple historical data sets in the framework of modified power priors. Three power prior methods are proposed and compared. Then the one with the best performance among three power prior methods is further compared with the random effects model. In addition, some preliminary results are presented on the power priors applied to random effects models. Time-weighted power priors are introduced at the end of Chapter 4. In Chapters 5 and 6, we will illustrate the implementation of the modified power priors in water quality assessment and their advantage over alternative methods. Chapter 5 demonstrates the applications of power priors to normal models; Chapter 6 discusses the type I error rate and power when power priors are applied to a binomial model. Finally in Chapter 7, we summarize the properties of the modified power prior approach, and close the dissertation with proposal on future research.

Chapter 2

Literature Review

2.1 Water Quality Evaluation

One important problem in environmental statistics is the evaluation of air or water quality standards. Issues include the definition of standards ([1]), trend assessment ([16]) and the evaluation of data from locations to determine compliance. A standard for a chemical or pollutant is a qualitative or quantitative description of expectation for the chemical or pollutant. To implement such a standard often a numerical criterion is required. For example, the numerical criterion might be different for a lake used for drinking water than for a lake used for fishing. Also associated with a standard are expectations related to frequency, magnitude and duration. Air quality evaluation often involves the expected frequency of violation (for example the ozone standard, see Thompson *et al.* [35]). However, evaluation of water quality standards often involves a percentile view. For example, for dissolved oxygen, a site is expected to have 10% or fewer samples in violation.

Section 303(d) of the Clean Water Act requires states to review their water quality conditions using monitoring stations. Declaring a water segment impaired will initiate a complicated and potentially expensive process, the Total Maximum Daily Load (TMDL) process. To assess water quality standards, measurements of water quality (e.g. pH, dissolved oxygen, biological oxygen demand) are collected on a regular basis (e.g. quarterly) over a number of time periods and analyzed to evaluate the percentage of samples exceeding the standard. A common approach accepted by the *US Environmental Protection Agency* (USEPA) is the raw score approach that simply calculates the proportion of violations and

declares a violation if it exceeds 10%. Smith *et al.* ([30]) pointed out that the raw score approach does not control for error rates when using binary information to make the impairment determination. In this case, a Type I error is to declare a segment impaired when it is not, and a Type II error is to designate the segment as not impaired when in fact it is. Smith *et al.* ([30]) discussed the practical consequences and costs of both types of errors and also the tradeoff among them. Type I errors may result in unnecessary TMDL implementation costs, and a Type II error may pose a risk to human and ecological health. They suggested to use a binomial test, with which the error rates associated with impairment declarations (Type I and Type II error rates) may be evaluated and limited. They noted that the Section 303(d) water quality assessment process is essentially a statistical decision problem. Specifically, this can be set up as a hypothesis testing problem, with the null hypothesis being that the site is not impaired and the alternative hypothesis being that the site is impaired. We use the same binary information as in the raw score approach and assume a *binomial* population. If we use p to describe the true probability of impairment and let p_0 be the hypothesized probability of impairment under safe conditions, the impairment decision is based on the test $H_0 : p \leq p_0$ (not impaired) versus $H_1 : p > p_0$ (impaired), where p_0 is 0.1 based on EPA guidelines. Smith *et al.* ([30]) showed that the raw score method has a strong tendency to falsely list a site, and the binomial method with a Type I error rate of 0.05 has a tendency to not list sites that are actually impaired. However, the binomial method allows for control of both error rates, and error rates can be set at satisfactory levels with sufficient sample sizes.

Furthermore, Smith *et al.* ([30]) introduced a Bayesian approach to the binomial test, which is more flexible and also controls error rates (see also McBride and Ellis, [24]). The Bayesian approach uses information from other sources about the probability of violation, and a noninformative prior could be used on p when no such information is available. Based on the posterior distribution of p, a decision may be made using either a cutoff approach or an odds-ratio approach (Bayes factor). The Bayesian approach allows for control of the error rates through the choice of cutoff and prior distribution of p. The comparison between the frequentist and Bayesian approaches in Smith *et al.* ([30]) demonstrates the strong similarity in their results if an uniform prior on p is used.

However, all the approaches mentioned above do not fully use all the information provided by the data in the sense that only the binary data with "standard violation" or not is used in the analysis. An alternative approach is to make use of the actual measurements instead, and this could improve the accuracy of estimation. Smith *et al.* ([31]) suggested an

approach based on acceptance sampling by variables, pointing out that this would reflect the magnitude of violation. When the distribution of measurements is normal with unknown variance, a classical hypothesis test ($H_0 : p \leq p_0$ versus $H_1 : p > p_0$) is carried out to make the impairment determination. A reasonable test statistic t for a lower standard L is $(\overline{x} - L)/s$, and $(U - \overline{x})/s$ for an upper standard U, where $\overline{x}$ and s are sample mean and sample standard deviation. We reject H_0 if the test statistic is less than a cutoff value k. Since $t\sqrt{n}$ follows a non-central t distribution with $n-1$ degrees of freedom and non-centrality parameter $\lambda = -z_p\sqrt{n}$, it is easy to find that $k = t_{n-1,\lambda,\alpha}/\sqrt{n}$. An alternative approach suggested by Wallis ([37]) uses the same test statistic but avoids the use of the non-central t distribution by taking

$$k = \frac{z_\alpha\sqrt{2nz_{p0}^2 + 4n - 2z_\alpha^2} - 2nz_{p0}}{2n - z_\alpha^2}.$$

Smith *et al.* ([31]) also described two adjustment methods when positive autocorrelation in the data is present. They noted that a positive autocorrelation structure induces a smaller non-centrality parameter and hence a smaller cutoff point than the independence case. As a consequence, it increases the Type I error rate and reduces the sample size needed. Based on the properties of a non-central t distribution, a sampling plan could be set up to control the error rates. In Smith *et al.* ([31]), methods for estimating sample size were summarized for both independence and autocorrelation cases. Another approach may be based on tolerance intervals using a non-central t distribution, which is essentially equivalent to the variable acceptance approach described above. Tolerance intervals are intervals for a percentile of the distribution of measurements. See Smith ([32]) for a detailed description.

Ye and Smith ([40]) proposed to use a Bayesian test on the desired percentile of the distribution (e.g., 10th percentile) that is of interest to check the impairment of the monitoring site. By using Bayesian methodology, the quantity of interest (e.g., the percentile of the measurement distribution) can be naturally treated as a parameter and thus its posterior distribution can be used to make needed decisions.

Suppose a chemical concentration measurement follows a certain distribution. Then the raw data can be used to test a hypotheses about a chemical concentration

$$H_0 : L \geq L_0(\text{not impaired}) \quad \text{versus} \quad H_1 : L < L_0(\text{impaired}), \tag{2.1}$$

where L denotes a true lower percentile of the population distribution, and L_0 is the standard. To test the hypotheses in (2.1), one may consider rejecting the null hypothesis and hence declaring impairment when the posterior probability of the null hypothesis given the data is

small (e.g. < 0.05). Ye and Smith ([40]) discussed certain properties of this method when the distribution of measurements comes from a location-scale parameter model. Specifically, they gave the reference prior for the parameters, and derived the analytical expressions for Type I and Type II error probabilities. In addition, they investigated impact of transformation and the priori information for the hypotheses on our problem. When the underlying distribution is normal, the Type I error probability of the Bayesian approach using the raw measurements is shown to be quite close to that of the binomial method, but the Type II error probability is smaller with the Bayesian approach. The simulation results demonstrated the advantage of the raw measurement approach in terms of the error probabilities, compared to EPA's raw score method and the binomial method.

One challenge is that only a limited amount of sample data can be used to determine whether to list a segment as impaired or not. Because the assessment of site impairment is often required to be conducted on two year observations, the sample size may be inadequate to provide necessary precision in parameter estimates and thus decisions may be affected by variation. In such situations, "historical" data, data collected from previous time periods or at adjacent monitoring sites, can be very helpful in interpreting the current status of water quality. Due to the nature of updating information sequentially, it is straightforward to use a Bayesian approach with an informative prior on the model parameters to incorporate the historical data into the current study. In the next section, we will review how to elicit a prior distribution in a Bayesian analysis.

2.2 Prior Elicitation in Bayesian Analysis

2.2.1 Bayesian Analysis

Bayesian statistics is built on Bayes' rule, which defines the change in probability of an event A after another event B occurs, and the philosophy of viewing distribution parameters as random variables.

Suppose that given a parameter θ, the random variable X follows a distribution with density $f(x|\theta)$. Bayesian analysis is performed by combining the prior information of parameter θ and the sample information into the posterior distribution of θ given x. Bayesians assume that θ is also a random variable with density $\pi(\theta)$, called the prior of θ. This assumption is a key element of Bayesian statistics, and provides an innovative outlook towards

statistics. Furthermore, the conditional distribution of θ given the sample observations x is defined as the posterior distribution of θ given x, denoted $\pi(\theta|x)$. Then applying Bayes' rule on the random variable, we have

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\Theta f(x|\theta)\pi(\theta)d\theta},$$

provided that all the terms exist.

As the name "posterior distribution" indicates, $\pi(\theta|x)$ reflects the updated information about θ posterior to observing the sample x. Based on this posterior distribution, we can make inference on θ, e.g. point estimation, interval estimation, and hypothesis testing.

Prior elicitation is one of the most crucial issues in Bayesian data analysis. It is the most debated topic in theoretical research and is also a challenging issue to practitioners. Opponents of Bayesian approach criticize the arbitrariness in the choice of prior, whereas proponents praise it as a manageable way of introducing flexibility in Bayesian analysis ([22]). Berger ([2]) noted that whenever a practitioner can summarize historical or subjective information about the unknown parameter, an informative prior should be used. On the other hand, more often either historical or subjective information is unavailable, or incorporating such information into a prior distribution is difficult for a real problem, thus automatic or default prior distributions are needed. Noninformative priors are also called automated priors, default priors, vague priors, or priors of ignorance. Bayesian analysis with noninformative priors preserves the appearance of objectivity, and is being increasingly recognized by classical statisticians.

2.2.2 Noninformative Priors

In this section, we present a brief review of the most commonly used approaches to develop an noninformative prior. See, e.g., Kass and Wasserman ([22]), for a thorough discussion.

The *Uniform Prior*: A natural idea for choosing a noninformative prior is a uniform prior. A uniform prior assumes that θ is equally likely in a region, expressed as

$$\pi(\theta) \propto 1, \text{over the range of } \theta.$$

This choice was popularized by Laplace ([23]).

The *Jeffreys Prior*: This was proposed in Jeffreys ([21]), as a solution to the problem that the uniform prior does not yield an analysis invariant to parameter transformations.

The Jeffreys prior is simply defined as

$$\pi(\theta) \propto \sqrt{det(I(\theta))}, \tag{2.2}$$

given that the Fisher information matrix, $I(\theta)$, is defined and positive definite, where "det" stands for a determinant. It is easy to check that for any one-to-one transformation between two parameters θ and η, the Jeffreys priors on θ and η transform according to the change-of-variables formula. Thus different parameterizations do not cause any ambiguous results. Furthermore, Jeffreys ([21], pp.182-183) suggested a modification for general location-scale problems, which is to treat the location parameters separately from the rest. Suppose that there are location parameters $\mu_1, ..., \mu_k$, and an additional multidimensional parameter θ, then the modified Jeffreys prior is set as

$$\pi(\mu_1, ..., \mu_k, \theta) \propto \sqrt{det(I(\theta))},$$

where $I(\theta)$ is calculated holding $\mu_1, ..., \mu_k$ fixed ([21], pp.182-183). This is referred as the location Jeffreys prior in some literature, and (2.2) is referred as the nonlocation Jeffreys prior. However, in many multi-dimensional parameter problems, the Jeffreys prior yields poor performance because it frequently induces dependence among the parameters once the data are used, which also conflicts with the assumption of independence of the prior knowledge among parameters.

The *Reference Prior*: To overcome the difficulty of the Jeffreys prior in multiparameter problems, Bernardo ([7]) introduced the reference prior, which was further developed by Berger and Bernardo ([3], [4], [5], [6]). They defined a notion of "missing information" on θ in an experiment and developed a stepwise procedure for handling nuisance parameters.

Let $K(\pi(\theta|\underline{x}), \pi(\theta))$ be the Kullback-Leibler distance between the posterior and the prior densities, $K(\pi(\theta|\underline{x}), \pi(\theta)) = \int_\Theta \pi(\theta|\underline{x}) log(\pi(\theta|\underline{x})/\pi(\theta)) d\theta$. The reference prior method is motivated by the idea of maximizing the missing information in the experiment. The missing information is quantified by $E_{\underline{x}}(K(\pi(\theta|\underline{x}), \pi(\theta)))$, where the expectation is with respect to the marginal density $m(\underline{x}) = \int_\Theta f(\underline{x}|\theta)\pi(\theta)d\theta$. However, carrying out this maximization involves an asymptotic process using infinitely many independent and identical (i.i.d.) observations of the experiments, and a modification of $E_{\underline{x}}(K(\pi(\theta|\underline{x}), \pi(\theta)))$.

In many applied statistical problems, the parameter θ can be written in the form $\theta = (\theta_1, \theta_2)$, where θ_1 is a parameter of interest and θ_2 is a nuisance parameter. When there are no nuisance parameters and certain regularity conditions are satisfied, the reference prior turns out to be (2.2) for continuous parameter spaces and the uniform prior for finite

parameter spaces. When there exist nuisance parameters, the reference prior is considered satisfactory for making inference about θ_1 but may not be satisfactory for making inference about θ_2. In this case, the reference prior is often different from (2.2). When there is a partition $\theta = (\theta_1, \theta_2)$, Bernardo suggested a stepwise procedure. First, define $\pi(\theta_2|\theta_1)$ to be the reference prior for θ_2 with θ_1 fixed. Second, find the marginal density $f(\underline{x}|\theta_1) = \int f(\underline{x}|\theta_1, \theta_2)\pi(\theta_2|\theta_1)d\theta_2$. And then take $\pi(\theta_1)$ to be the reference prior based on the marginal density $f(\underline{x}|\theta_1)$. Finally a reference prior for (θ_1, θ_2) can be obtained by $\pi(\theta_1)\pi(\theta_2|\theta_1)$. Berger and Bernardo ([4], [5], [6]) and Ye and Berger ([39]) have extended this iterative algorithm to deal with parameters that are decomposed into any number of ordered groups. The ordering is decided by the importance of different groups. Generally, different groupings or orderings may yield different reference priors, and therefore the same model may have several reference priors.

The *Probability Matching Prior*: Since noninformative priors are characterized as "letting the data speak for themselves", it may be desirable to have posterior probabilities agree with sampling probabilities. Specifically, the probability matching prior, as its name implies, is obtained by matching the posterior probabilities of Bayesian credible intervals with repeated-sampling coverage probabilities of corresponding confidence intervals, at least asymptotically.

Suppose that θ is a parameter of interest and $l(\underline{x})$ and $u(\underline{x})$ have the posterior probability $1 - \alpha = Pr(l(\underline{x}) \leq \theta \leq u(\underline{x})|\underline{x})$. Thus $l(\underline{x})$ and $u(\underline{x})$ is a Bayesian credible interval with posterior probability $1 - \alpha$. On the other hand, we can treat θ as fixed and consider $l(\underline{x})$ and $u(\underline{x})$ in the sense of confidence intervals under the scheme of repeated sampling. The frequentist coverage probability of this interval can be calculated as $Pr(l(\underline{x}) \leq \theta \leq u(\underline{x})|\theta)$. We say a prior is a probability matching prior if it satisfies

$$Pr(l(\underline{x}) \leq \theta \leq u(\underline{x})|\theta) \approx Pr(l(\underline{x}) \leq \theta \leq u(\underline{x})|\underline{x}) = 1 - \alpha,$$

for all $\underline{x}$ and θ, asymptotically.

Welch and Peers ([38]) showed that in the one-dimensional case, the Jeffreys prior satisfies this equality in the order of $O(\frac{1}{n})$. Peers ([26]) and Stein ([33]) made some progress on examining multiparameter problems with parameters being partitioned into a parameter of interest and nuisance parameters. Based on Stein's paper, Tibshirani ([36]) suggested a prior that leads to accurate confidence intervals for the parameter of interest. Severini ([28]) showed that under certain circumstances, some priors will give Highest Posterior Density (HPD) regions that agree with their nominal frequentist coverage to order $n^{-3/2}$.

The *Maximal Data Information Prior* (MDIP): This method is motivated by emphasizing the information in the data density or likelihood function. Zellner ([41], [42]) and Zellner and Min ([45]) suggested choosing the prior $\pi(\theta)$ that maximizes the average information in the data density relative to that in the prior. Note that the negative entropy of the joint density $p(\underline{x}, \theta) = \pi(\theta) f(\underline{x}|\theta)$ breaks up into the average information in the data density and the information in the prior density, as shown below.

$$\begin{aligned} -H &= \int\int p(\underline{x}, \theta) \ln p(\underline{x}, \theta) d\underline{x} d\theta \\ &= \int I(\theta)\pi(\theta) d\theta + \int \pi(\theta) \ln \pi(\theta) d\theta, \end{aligned}$$

where $I(\theta) = \int f(\underline{x}|\theta) \ln f(\underline{x}|\theta) d\underline{x}$, negative entropy of $f(\underline{x}|\theta)$.

To emphasize the information in the data density, MDIP is derived by maximizing the difference

$$G = \int I(\theta)\pi(\theta) d\theta - \int \pi(\theta) \ln \pi(\theta) d\theta,$$

with respect to $\pi(\theta)$.

Note that $G = \int\int \pi(\theta|\underline{x}) \ln[f(\underline{x}|\theta)/\pi(\theta)] m(\underline{x}) d\theta d\underline{x}$. Thus MDIP can also be interpreted as the prior that maximizes the expected log ratio of the likelihood function to the prior density. Maximizing G subject to the condition $\int \pi(\theta) d\theta = 1$ gives the MDIP $\pi(\theta) \propto \exp(I(\theta))$. Zellner ([41], [42], [44]) and Zellner and Min ([45]) derived MDIPs for parameters of location-scale, normal mean, regression, autoregression, exponential, uniform, and other models. Although MDIPs are not parameterization invariant in general, Zellner ([44]) provided the side conditions that could be used to produce MDIPs that are invariant to specific classes of reparameterizations.

Two examples are given below using all five methods to elicit a noninformative prior.

Example 1. Suppose that X follows a binomial distribution with unknown parameter p and known n, where $0 \leq p \leq 1$. Then the uniform prior is $\pi(p) = 1$; both the Jeffreys' prior and the reference prior are of the form $\pi(p) \propto p^{-1/2}(1-p)^{-1/2}$; the MDIP is $\pi(p) \propto p^p(1-p)^{1-p}$; and the Jeffreys' prior achieves probability matching to $O_p(\frac{1}{n})$ ([38]).

Example 2. Suppose $X_1, ..., X_n$ is a random sample from a normal population $N(\mu, \sigma^2)$ with μ and σ^2 both unknown. Furthermore suppose $\phi = \mu/\sigma$ is the parameter of interest and σ is a nuisance parameter. Then both the uniform prior and MDIP are $\pi(\phi, \sigma) \propto 1$; the Jeffreys' prior is $\pi(\phi, \sigma) \propto 1/\sigma$; and the reference prior is $\pi(\phi, \sigma) \propto (2 + \phi^2)^{-1/2}\sigma^{-1}$. Sweeting ([34]) proposed a local probability matching prior parameters of interest based on

a data-dependent approximation to the order of $O_p(\frac{1}{n})$. In this case, a local probability matching prior is given by

$$\pi(\phi, \sigma) \propto (1 + \frac{1}{2}\phi^2)^{-1/2} \exp\left(\frac{1}{2}\phi s\overline{x}(s^2 + \frac{1}{2}\overline{x}^2)^{-1}\right),$$

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

2.2.3 Informative Priors

Although noninformative priors may be convenient and easier to specify, they can not be used in all applications. For example, proper priors are required to compute Bayes factors. Also, Kass and Wasserman ([22]) pointed out that many noninformative priors (e.g., Jeffreys prior, reference prior, coverage matching prior) are built from asymptotic arguments, and "it is dangerous to put faith in any default solutions" for small-sample problems. So an informative prior may be necessary for small-sample problems. In applied problems, subjective information about the parameter of interest is often available, or the investigator has access to previous studies measuring the same response and covariates as the current study. In such cases, an informative prior is recommended because it can make use of real prior information.

Before the recent breakthroughs in Bayesian computation techniques, a posterior distribution with a workable form was desired. Conjugate priors were the most commonly sought informative priors, because they produced posterior distributions that were from the same family as priors. The conjugate prior is still popular because elicitation can be done by choosing only one or two parameter values. However, many times it is not accurate. To overcome this difficulty, two modifications could be made: a finite mixture of conjugate priors and a conditional conjugate prior for a multi-dimensional parameter model.

The improvement in computational techniques allows us to use a non-conjugate prior with an approximate distribution form, which may summarize all known information. The known information includes subjective information and data from previous similar studies (called historical data). The difficulty of constructing an informative prior comes from the difficulty of incorporating subjective and historical information into a prior distribution. The first issue is how to convert subjective information into a prior distribution. If the parameter space Θ is discrete, the problem is simply to determine the subjective probability of each element of Θ. When Θ is continuous, several techniques can be used: the histogram

approach, the relative likelihood approach, matching a given functional form, and subjective construction of the CDF (cumulative distribution function). See Berger ([2]) for detailed description of these techniques. However, studies have shown that untrained people do quite poorly on eliciting probability distributions, because overconfidence concerning their prior knowledge often leads to elicited distributions being too concentrated.

The second issue is how to incorporate historical data into a prior distribution. A natural thought is to apply Bayesian analysis on historical data with a noninformative prior and then use the posterior distribution as an informative prior for the current data. This approach uses the historical data to update our prior information. The construction of the final posterior distribution implies a simple pooling of current data and historical data together, since the two data sets are equally weighted. This pooling can be well justified by assuming current and historical data from the same population. However, the population parameters may change over time, or over different settings, although current and historical data are usually assumed to follow distributions in the same family. If the sample size of the historical data is much larger than that of current data and heterogeneity exists between the current and previous studies, historical data would dominate the analysis and the data pooling may result in misleading conclusions. The question is then how much the historical data should be accounted for in the current study. To address this issue, Ibrahim and Chen ([18]) and Chen *et al.* ([10]) proposed the concept of *power priors*, which is based on the notion of the *availability* of historical data. We will give an extensive review on power prior distributions in the next section.

2.3 Power Prior Distributions

2.3.1 Introduction

The power prior approach provides a useful class of informative priors for Bayesian inference. The basic idea is to use the power parameter δ $(0 \leq \delta \leq 1)$ to control the influence of the historical data on the current study. The power prior is constructed by raising the likelihood function based on the historical data, denoted by D_0, to a suitable power to discount the historical data relative to the current data. The initial idea can be found in Diaconis and Ylvisaker ([12]) and Morris ([25]), but they only consider δ as a predetermined constant. Ibrahim and Chen developed this idea and extensively studied the theoretical properties of the power priors in Ibrahim and Chen ([18]), Chen *et al.* ([10]), Ibrahim *et al.* ([19]), and

Chen *et al.* ([11]).

Let θ denote the parameter of interest, and $L(\theta|D_0)$ be the likelihood function of θ based on the historical data. It is assumed that given θ, the historical data D_0 and current data, denoted by D, are independent random samples. $\pi(\theta)$ is taken as the initial prior before any historical information is gathered and usually it is a noninformative prior. Given δ, Ibrahim and Chen ([18]) define the power prior of θ for the current study as

$$\pi(\theta|D_0, \delta) \propto (L(\theta|D_0))^{\delta}\pi(\theta). \tag{2.3}$$

The parameter δ measures the portion of historical information needed in the current study and is described using the prior in (2.3). The case $\delta = 0$ means that no historical data should be used; while $\delta = 1$ gives equal weight to $L(\theta|D_0)$ and the likelihood of the current study $L(\theta|D)$, resulting in full incorporation of the historical data. Therefore, (2.3) can be viewed as a generalization of the usual Bayesian update of $\pi(\theta)$ (see discussion in Ibrahim and Chen, [18]). The power parameter δ can be interpreted as a precision parameter. For example, consider the case of a normal sample with known variance. Suppose that D_0 consists of n_0 observations, $X_1, X_2, \ldots, X_{n_0}$, from the normal population with unknown mean parameter θ and known variance σ^2. If the prior $\pi(\theta)$ is assumed to be uniform (non-informative), (2.3) implies a prior distribution of θ for the current data set D, $\theta|D_0, \delta \sim N(\bar{x}_0, \frac{\sigma^2}{\delta n_0})$, where $\bar{x}_0$ is the sample mean of the historical data. Hence, δ can be viewed as part of a precision parameter, because smaller δ implies larger power prior variance while larger δ means the smaller power prior variance.

The power prior $\pi(\theta|D_0, \delta)$ in (2.3) was initially elicited for fixed δ. Chen *et al.* ([10]) noted that, since δ is not necessarily pre-determined, we may extend it further to the case that δ is random. A random δ gives the investigator more flexibility in weighting the historical data. The power prior specification on (θ, δ) is then completed by specifying a prior distribution for δ. Ibrahim and Chen ([18]) proposed a joint power prior distribution for (θ, δ) of the form

$$\pi(\theta, \delta|D_0) \propto (L(\theta|D_0))^{\delta}\pi(\theta)\pi(\delta|\gamma_0), \tag{2.4}$$

where γ_0 is a specified hyperparameter vector. A natural prior for δ would be a $Beta(a, b)$ distribution, or simply a uniform distribution, since $0 \leq \delta \leq 1$. The investigator may influence the prior weight on the historical data by adjusting the hyperparameters. Chen *et al.* ([10]) suggested using several sets of hyperparameters and conducting sensitivity analyses.

Furthermore, Ibrahim and Chen ([18]) generalized the prior defined in (2.4) to multiple historical data sets. Suppose there are m independent historical studies. Let D_{0j} to be the

historical data based on the jth study, $j = 1, ..., m$ and $D_0 = (D_{01}, ..., D_{0m})$. They suggested defining a different weight parameter δ_j for j^{th} historical study and taking the δ_j's to be i.i.d. *Beta* random variables with hyperparameters (a, b). Let $\underline{\delta} = (\delta_1, ..., \delta_m)$, then the power prior for multiple historical data set takes the form

$$\pi(\theta, \underline{\delta}|D_0) \propto \left(\prod_{j=1}^{m} [L(\theta|D_{0j})]^{\delta_j} \pi(\delta_j|a, b) \right) \pi(\theta).$$

This framework could accommodate potential heterogeneity among several historical studies, and hence the role of historical data can be more accurately evaluated.

2.3.2 Optimality Properties of the Power Prior

Ibrahim *et al.* ([19]) provided a formal justification of the power prior as an optimal class of informative priors. They first assume that δ is fixed, and then extend it to random δ. Let $K(f_0, f_1)$ denote the Kullback-Leibler (KL) divergence between two densities f_0 and f_1. Two extreme posterior densities are considered: one density is based on no incorporation of historical data and the other density is based on pooling the historical and current data. An attempt is made to find the posterior density $g(\theta)$ that minimizes a convex sum of KL divergences between the two posterior densities mentioned above

$$K_g = (1 - \delta)K(g, \pi(\theta|D, D_0, \delta = 0)) + \delta K(g, \pi(\theta|D, D_0, \delta = 1)),$$

where δ is a fixed scalar between 0 and 1. It turns out that the power prior $g_{opt} \propto L(\theta|D)L(\theta|D_0)^{\delta}\pi(\theta)$ achieves the desired minimum. Ibrahim *et al.* ([19]) stated this result as a theorem and gave a formal proof. Ibrahim *et al.* ([19]) also proved that K_g is convex in g, which implies that the minimum of K_g exists and the minimizer g_{opt} is unique. The optimality of the power prior is hence established. Furthermore, Ibrahim *et al.* ([19]) extended the optimality of the power prior to the case in which multiple historical datasets exist. Finally the assumption of δ being fixed is loosened. Ibrahim *et al.* ([19]) noted that when δ is random, the power prior minimizes $E(K_g)$, where the expectation is taken with respect to $\pi(\delta)$.

In addition, Ibrahim *et al.* ([19]) showed that following the optimal information processing rules (IPR) of Zellner ([43], [45]), the power prior is a 100% efficient IPR in the sense that the ratio of the output to input information is equal to 1.

Based on Zellner's theory of IPR, a weighted version of the information criterion function $\Delta[g(\theta)]$ is considered in our scenario.

$$\begin{aligned}\Delta[g(\theta)] &= \text{Output information} - \text{Input information} \\ &= \int g(\theta)\ln(g(\theta))\ d\theta + \int g(\theta)\ln(m(D, D_0))\ d\theta \\ &\quad - \left[\int g(\theta)\ln(\pi(\theta))\ d\theta + \int g(\theta)\ln(L(\theta|D))\ d\theta + \delta\int g(\theta)\ln(L(\theta|D_0))\ d\theta\right], \quad (2.5)\end{aligned}$$

where $g(\theta)$ is a proper posterior density $\pi(\theta|D, D_0)$ in our setting. Zellner defined a rule to be 100% efficient if the $g*$ that minimizes (2.5) yields $\Delta[g(\theta)] = 0$; that is, output information equals input information.

Ibrahim *et al.* ([19]) established the equivalence between the two criteria K_g and $\Delta[g(\theta)]$.

$$K_g = \Delta[g(\theta)] + C,$$

where C is a constant free of g. This relationship implies that K_g and $\Delta[g(\theta)]$ have the same minimizer, and therefore the power prior is a 100% IPR. This relationship was also shown to hold for the case of multiple historical datasets.

2.3.3 Power Priors for Regression Models

Ibrahim and Chen ([18]) and Chen *et al.* ([11]) examined the power prior for four commonly used classes of regression models, including generalized linear models, generalized linear mixed models, semiparametric proportional hazards models, and cure rate models for survival data. They discussed the construction of the power prior, propriety conditions, and its application on model selection. In the rest of this section, we let n_0 denote sample size for the historical data, y_0 be an $n_0 \times 1$ response vector for the historical data, and X_0 is an $n_0 \times k$ matrix of covariates corresponding to y_0. Also, let y_{0i} denote the ith component of y_0, and let $x'_{0i} = (x_{0i1}, x_{0i2}, ..., x_{0ik})$ be the ith row of X_0 with $x_{0i1} = 1$ corresponding to an intercept. Let β denote the vector of regression coefficients on covariates x'_{0i} in the linear predictor. For regression models, Ibrahim and Chen use $\pi(\beta) \propto 1$ as the initial prior for β, and $Beta(a, b)$ as the prior for δ.

Generalized linear models: For the generalized linear models, suppose the historical data is denoted by $D_0 = (n_0, y_0, X_0)$, and the linear predictor is denoted by $\eta_{0i} = x'_{0i}\beta$. Chen *et al.* ([10]) established some very general results concerning the propriety of the

joint prior distribution of (β, δ) for the generalized linear models. First they presented two sufficient conditions for the propriety. They pointed out that those conditions hold for many generalized linear models such as the normal, Poisson, and binomial models. However, neither condition will be satisfied when the y_{0i} are binary responses. Then Chen *et al.* ([10]) discussed the additional regularity conditions on the fixed covariates x_{0i} needed to establish the propriety of the power prior for binary responses. Part of the sufficient conditions are the conditions on hyperparametrs a and b. The sufficient condition on a is either $a > k$ or $a > k/2$, where k is the number of covariates plus 1. Chen *et al.* ([10]) also investigated whether this sufficient condition on a is also necessary. They first derive lower bounds for the normalizing constant of the conditional power prior density for β, which can be expressed as $\int [L(\beta|D_0)^\delta] d\beta$ based on (3.1). Then they use those lower bounds to show that a necessary condition for the propriety of the power prior distribution is either $a > k$ or $a > k/2$ for the logit model and probit model, and is $a > k$ for the log-log link model and Poisson regression model. Chen *et al.* ([10]) also extended the results concerning the propriety to the case of multiple historical data sets. They pointed out that the sufficient and necessary conditions on a for multiple historical data sets are weaker than those for the single historical data set because in the former case more prior information is incorporated into the analysis.

Generalized linear mixed model (GLMM): Suppose there exist historical data with N_0 subjects that yields the $n_{0i} \times 1$ response vector y_{0i} for subject i. The linear predictor is denoted by $\eta_{0it} = x'_{0it}\beta + z'_{0it}b_{0i}$, where b_{0i} is a $q \times 1$ vector of random effects and x'_{0it} and z'_{0it} are vectors of covariates. Chen *et al.* (2003) proposed to take the power prior for β given δ to be of the form

$$\pi(\beta|D_0, T, \delta) \propto \prod_{i=1}^{N_0} \left(\int_{R^q} \prod_{t=1}^{n_{0i}} [p(y_{0it}|\beta, b_{0it})]^\delta \pi(b_{0i}|T) db_{0i} \right) \pi(\beta), \tag{2.6}$$

where $p(y_{0it}|\beta, b_{0it})$ is the density of historical data y_{0it} based on GLMM, and $\pi(b_{0i}|T)$ is the normal distribution with mean 0 and covariance matrix T^{-1}. Note that this construction first exponentiates the historical data likelihood given the random effects, and then integrates the random effects out. It does not start by integrating out the random effects to get the marginal historical data likelihood and then raise this to a power δ. An obvious advantage of (2.6) is that its implementation is easier because the marginal historical data likelihood is computationally intractable. The power prior for GLMM is completed by specifying priors for δ and parameters in T. Chen *et al.* ([11]) assume that T is determined by two parameters σ_b^2 and ρ, and take an inverse gamma prior for σ_b^2 and a scaled beta prior for ρ.

Furthermore, they discussed the conditions for propriety of this joint power prior distribution $\pi(\beta, \delta, \sigma_b^2, \rho|D_0)$. The results are quite similar to those for generalized linear models.

Chen *et al.* ([11]) applied the proposed priors on the variable subset selection for the class of GLMM, and also developed efficient computational algorithms for implementation. The resulting posterior model probabilities can be evaluated directly without numerically computing the prior model probabilities. Due to the complexity of the GLMM, they adopted the importance-weighted marginal posterior density estimation (IWMDE) method of Chen ([9]) to estimate the marginal posterior densities of β. Chen *et al.* ([11]) pointed out that the power prior is especially attractive in variable selection because only very few hyperparameters need to be specified to carry out this elicitation method.

Proportional hazards models: Let $0 \leq s_0 < s_1 < ... < s_M$ denote a finite partition of the time axis constructed as in Ibrahim and Chen ([17]). Further, let δ_i denote the increment in the baseline hazard in the interval $(s_{i-1}, s_i], i = i, ..., M$, and let $\Delta = (\delta_1, ..., \delta_M)$. They use a piecewise-constant baseline hazard model to construct the likelihood function, and use only information about the interval where the failure times fall into. Let $\pi(\beta, \Delta)$ denote the initial prior distribution for (β, Δ). Ibrahim and Chen ([18]) suggested the following joint power prior for (β, Δ, δ) based on (2.4).

$$\pi(\beta, \Delta, \delta|D_0) \propto L(\beta, \Delta|D_0)^{\delta} \pi(\beta, \Delta) \delta^{a-1}(1-\delta)^{b-1}, \tag{2.7}$$

where $L(\beta, \Delta|D_0)$ is the likelihood function of (β, Δ) based on the historical data. To simplify the prior specification, $\pi(\beta, \Delta) = \pi(\beta)\pi(\Delta)$ was assumed. They suggested taking a p-dimensional multivariate normal distribution for $\pi(\beta)$, and a gamma density for $\pi(\Delta)$. If $\pi(\beta) \propto 1$, then (2.7) is proper if $\pi(\Delta)$ is proper and $\delta > p$. Also, Ibrahim and Chen ([17], [18]) applied the power prior approach on the variable selection problem for proportional hazards models.

Cure rate models: Let γ denote the indexing parameter, and let C_0 denote the unobserved vector of latent counts. Ibrahim and Chen ([18]) took the following form as the joint power prior for (β, γ, δ).

$$\pi(\beta, \gamma, \delta|D_{0,obs}) \propto \left[\sum_{C_0} L(\beta, \gamma|D_0) \right]^{\delta} \pi(\beta, \gamma) \delta^{a-1}(1-\delta)^{b-1}, \tag{2.8}$$

where $L(\beta, \gamma|D_0)$ is the complete historical data likelihood. To specify an initial prior $\pi(\beta, \gamma)$, where $\gamma = (\alpha, \lambda)$, they assumed independence among β, α, and λ. They suggested an improper uniform prior for β, a gamma prior for α, and a normal prior for λ. They also gave

the mild regularity condition needed to guarantee the propriety of the joint power prior in (2.8).

Ibrahim and Chen ([18]) extended the standard power prior elicitation to the situations where historical data y_0 are not available, or where the set of covariates measured in the previous study is a subset of the covariates measured in the current study. In the former situation, y_0 can be obtained through prediction based on a theoretical prediction model, expert opinion, or case-specific information. The latter issue can be addressed with some adjustments to the elicitation process. Let X_1 be the covariates in the current study that are common to the covariates in the previous study, and let X_2 be the new covariates in the current study which are not measured in the previous study. X_{01} and X_{02} are corresponding covariate components from the previous study. Also we partition the vector of parameters θ into θ_1 and θ_2 accordingly. Let $D_{0j} = (n_{0j}, y_{0j}, X_{0j})$, where y_{0j} is the historical data corresponding to $X_{0j}, j = 1, 2$. Then the construction of the power prior is completed by assuming the priori independence between θ_1 and θ_2.

$$\begin{aligned} \pi(\theta|D_0, \delta) &= \pi(\theta_1|D_{01}, \delta_1)\pi(\theta_2|D_{02}, \delta_2) \\ &= L(\theta_1|D_{01})^{\delta_1} L(\theta_2|D_{02})^{\delta_2} \pi(\theta_1, \theta_2). \end{aligned}$$

Obviously, X_{01} and y_{01} are the raw covariate matrix and raw response vector from the previous study, respectively. It is a common practice to take $X_{02} = X_2$ and to use predicted values to fill in y_{02}.

Ibrahim and Chen ([18]) investigated the relationships between the power prior approach and the maximum likelihood analysis, the maximum likelihood analysis using a random effects model, and a meta-analysis. The AIDS data were used to illustrate the comparison among different methods in terms of the estimates and standard errors of β. It turns out that the power prior method with $\delta = 0$ gives nearly identical results as a maximum likelihood analysis, and that the power prior method with $\delta = 1$ produces results similar to those of a maximum likelihood analysis based on pooling the historical and current data. Let $b_i \sim N(0, \sigma_b^2)$ denote the random effect. The analyses on AIDS data also show that the results from the random effect model with small σ_b^2 are very similar to those from the power prior method with $\delta = 1$. The estimates and standard errors of β are fairly robust to the increase of σ_b^2. In addition, Ibrahim and Chen noted that the estimates of a meta-analysis are quite comparable to the power prior method with $\delta = 1$, where the meta-analysis type estimate of β is constructed by weighting the maximum likelihood estimate of β from the historical and current studies.

The power priors were developed by Ibrahim and Chen as a general class of informative prior distributions for arbitrary regression models. The power parameter can be viewed as a precision parameter that quantifies the compatibility between historical and current data. The power priors are useful in a wide variety of applications, such as carcinogenicity studies and clinical trials, or in general situations when historical data are available. They are especially attractive in variable subset selection, since very few hyperparameters need to be elicited. In addition, the power prior has been shown to be optimal in the sense that it minimizes the convex sum of KL divergences between the posterior densities of θ when $\delta = 0$ and when $\delta = 1$ ([19]).

Chapter 3

A Modified Power Prior Elicitation

3.1 Introduction

In applying statistics to real experiments, it is common that the sample size in the current study is often inadequate to provide necessary precision for parameter estimation, while plenty of historical data or data from similar studies or research settings are available. For example, to assess violations of water quality standards, measurements of chemical constituents are typically collected on a monthly or quarterly basis at each monitoring station, and then analyzed to evaluate the percentage of samples exceeding the standard. Under the Clean Water Act, only observations over a two year period are allowed to be counted as current data in the assessment. The lack of sufficient data often leads to unacceptable levels of uncertainty. In a situation like this, "historical" data, a data set from previous time periods or from adjacent stations, can be very useful in interpreting the current status of water quality, if it can be combined with current data in some way.

Due to the nature of updating information sequentially, it is natural to use a Bayesian approach with an informative prior on the model parameters to incorporate the historical data into the current study. A traditional approach to incorporating historical data is to construct an informative prior using the historical data and such a prior is combined with the likelihood to yield the posterior distribution in statistical inference. This implies a simple pooling of current data and historical data together, since the two data sets are equally weighted. This approach can be well justified by assuming the current and historical data come from the same population. However, although the current and historical data

are usually assumed to follow distributions in the same family, the population parameters may change over time, or over different settings. If the sample size of the historical data is much larger than that of the current data and heterogeneity exists between these data sets, historical information could dominate the analysis and the data pooling may result in misleading conclusions.

To address this issue, Ibrahim and Chen ([18]) proposed the concept of *power priors*, based on the notion of the availability of historical data. The basic idea is to let a power parameter δ $(0 \leq \delta \leq 1)$ tell us how much historical data is to be used in the current study. However, in their approach, the power parameter always has a tendency to be close to zero, which suggests that much of a historical data set is not used. Here we propose a modified power prior Bayesian approach. In this modified approach, the power parameter quantifies the heterogeneity between current and historical data automatically, and hence controls the influence of historical data on the current study in a sensible way. In addition, the modified power prior needs little to ensure its propriety.

The rest of this chapter is organized as follows. In Section 3.2, the general development of the modified power prior approach is given and certain properties of the approach for the Bernoulli and normal families are discussed. In Section 3.3, optimality of the modified power prior approach will be investigated. In Section 3.4, the modified and original power priors are compared in terms of the behavior of the power parameter, the estimate of the parameter of interest, and its mean squared error (MSE).

3.2 A Modified Power Prior Approach

3.2.1 The Modified Power Prior

Suppose that θ is the parameter of interest, for instance, concentration of a chemical level in a water quality measurement. Assume that such a measurement follows a distribution and $L(\theta|D_0)$ is the likelihood function of θ based on the historical data, denoted by D_0. In this article, we assume that, given θ, historical data D_0 and current data, denoted by D, are independent random samples from an exponential family. Furthermore, denote by $\pi(\theta)$ the initial prior, which can be a noninformative prior. Given δ, the power parameter, Ibrahim and Chen ([18]) defined the power prior of θ for the current study as

$$\pi(\theta|D_0, \delta) \propto L(\theta|D_0)^{\delta}\pi(\theta). \tag{3.1}$$

The power parameter δ measures the portion of historical information needed in the current study and is described using the prior in (3.1).

The power prior $\pi(\theta|D_0, \delta)$ in (3.1) was initially elicited for fixed δ. However, since δ is not necessarily pre-determined and also because it is often difficult to specify in practice, we may extend it further to the case that δ is random. A random variable δ provides the researcher with more flexibility in weighting the historical data. A natural prior for δ would be a $Beta(\alpha, \beta)$ distribution, or simply a uniform distribution, since $0 \leq \delta \leq 1$. The elicitation of the power prior on (θ, δ) is then completed by specifying a prior distribution for δ. Ibrahim and Chen ([18]) constructed the joint power prior of (θ, δ) as

$$\pi(\theta, \delta|D_0) \propto L(\theta|D_0)^{\delta}\pi(\theta)\pi(\delta). \tag{3.2}$$

However, a problem of this approach arises as we investigate the application of power priors on Bernoulli and normal mean models. The influence of the historical data is generally small, i.e., δ is close to 0, no matter how compatible current and historical data are. In such a case, the inference on θ is not much different from the inference when the historical data is ignored (more discussions is referred to Section 3.4). Furthermore, this prior could also be improper. We feel that once the historical information is available, a prior elicited from such information would better be proper. Therefore, we propose a modified joint power prior distribution for (θ, δ) as

$$\pi(\theta, \delta|D_0) \propto \frac{L(\theta|D_0)^{\delta}\pi(\theta)\pi(\delta)}{\int_{\Theta} L(\theta|D_0)^{\delta}\pi(\theta)\ d\theta}, \tag{3.3}$$

in the region of δ such that the denominator in (3.3) is finite.

The difference in the form between (3.2) and (3.3) is that the prior distribution of (θ, δ) expressed in (3.3) is always proper given that $\pi(\delta)$ is proper, whereas it is not necessarily the case for that in (3.2). More importantly, the approach of (3.2) does not agree with the likelihood principle in the sense that any arbitrary positive number can be multiplied in (3.2). More discussion will be given in Section 3.4.

3.2.2 Development of the Modified Power Prior

The framework of the modified power prior method is built upon the initial idea of the power prior defined in (3.1), and an assumption that both current and historical data are needed to update the distribution of δ.

Development of $\pi(\theta, \delta|D_0)$

Consider the prior structure in (3.1) as the conditional power prior of θ given δ, which can be expressed as

$$\pi(\theta|D_0, \delta) = \frac{L(\theta|D_0)^\delta \pi(\theta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}, \tag{3.4}$$

where $0 \leq \delta \leq 1$.

Certain regularity conditions, including $L(\theta|D_0) \geq 0$, $\pi(\theta) \geq 0$, and $P_\theta(L(\theta|D_0) > 0) > 0$, are assumed to hold. Under those regularity conditions, we have

$$\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta > 0.$$

Define a set A as

$$A = \{\delta : 0 < \int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta\ < \infty\}, \tag{3.5}$$

and define A^- as the complement of A in $[0, 1]$. We assume that, given only the sampling model $L(\theta|D_0)$, $\pi(\theta)$ leads to a proper posterior of θ using a conventional Bayesian analysis, i.e. $\int_\Theta L(\theta|D_0)\pi(\theta)\ d\theta\ < \infty$. This assumption constrains our discussion to a sensible range, and also ensures that A is nonempty because $1 \in A$.

Furthermore, we define that $\pi(\delta|D_0) \propto \pi(\delta) I_A(\delta)$, where $\pi(\delta)$ is a *Beta* distribution, and $I_A(\delta) = 1$ if $\delta \in A$ and 0 otherwise. Later in Section 3.3.1, we show that δ may be interpreted as the probability that D_0 and D come from the same population. Roughly speaking, δ measures how similar D_0 is to D. Therefore, $\pi(\delta|D_0) \propto \pi(\delta)$ is a reasonable assumption because without being compared to the current data, the historical data alone do not provide any information on δ.

Multiplying $\pi(\delta|D_0)$ by $\pi(\theta|\delta, D_0)$ in (3.4) yields the following joint power prior distribution.

$$\pi(\theta, \delta|D_0) = \pi(\delta|D_0)\pi(\theta|\delta, D_0) \propto \frac{L(\theta|D_0)^\delta \pi(\theta)\pi(\delta) I_A(\delta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta) d\theta}. \tag{3.6}$$

It is straightforward to verify that the joint prior $\pi(\theta, \delta|D_0)$ is always proper, which also ensures the propriety of the joint posterior for (θ, δ).

Posterior distributions using the modified power prior

Using current data to update the prior distribution $\pi(\theta, \delta|D_0)$ in (3.6), we derive the joint posterior distribution for (θ, δ) as

$$\pi(\theta, \delta|D_0, D) \propto L(\theta|D)\pi(\theta, \delta|D_0) \propto \frac{L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\pi(\delta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta} I_A(\delta).$$

Integrating θ out of the expression above, the marginal posterior distribution of δ can be expressed as

$$\pi(\delta|D_0, D) \propto \pi(\delta)\frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta} I_A(\delta). \tag{3.7}$$

Similarly, the marginal posterior distribution of θ, $\pi(\theta|D_0, D)$, is obtained by integrating δ out. If our interest is only in θ, δ may be integrated out at an earlier stage. Then $\pi(\theta|D_0, D)$ may also be developed in the way described below.

An informative prior $\pi(\theta|D_0)$ and its posterior $\pi(\theta|D_0, D)$

Integrating δ out in $\pi(\theta, \delta|D_0)$ yields a new prior for θ, a prior updated by the historical information,

$$\pi(\theta|D_0) = \int_A \pi(\theta, \delta|D_0)d\delta \propto \pi(\theta) \int_A \frac{L(\theta|D_0)^\delta \pi(\delta) I_A(\delta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}\ d\delta. \tag{3.8}$$

With historical data appropriately incorporated, $\pi(\theta|D_0)$ can be viewed as an informative prior for the Bayesian analysis on the current data. Consequently, the posterior distribution of θ can be written as

$$\pi(\theta|D_0, D) \propto \pi(\theta|D_0)L(\theta|D_0, D) \propto \pi(\theta)L(\theta|D) \int_A \frac{L(\theta|D_0)^\delta \pi(\delta) I_A(\delta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}\ d\delta. \tag{3.9}$$

Extension to multiple historical data sets

Similar to the extension given by Ibrahim and Chen ([18]), the priors defined in (3.6) can easily be generalized to multiple historical data sets. Suppose there are m historical studies. "Historical" studies can be studies done previously as well as studies with settings similar to current study. Denote by D_{0j} the historical data for the jth study, $j = 1, ..., m$ and $D_0 = (D_{01}, ..., D_{0m})$. Different weight parameter δ_j for each historical study should be used.

Furthermore, δ_j's can be assumed i.i.d. *Beta* random variables with hyperparameters (α, β). Let $\underline{\delta} = (\delta_1, ..., \delta_m)$. The modified power prior in (3.6) can be generalized as

$$\pi(\theta, \underline{\delta}|D_0) \propto \frac{\left(\prod_{j=1}^{m} L(\theta|D_{0j})^{\delta_j} \pi(\delta_j|\alpha, \beta)\right)\pi(\theta)}{\int \left(\prod_{j=1}^{m} L(\theta|D_{0j})^{\delta_j}\right)\pi(\theta)\ d\theta} I_B(\underline{\delta}),$$

where $B = \{(\delta_1, ..., \delta_m) : 0 < \int \left(\prod_{j=1}^{m} L(\theta|D_{0j})^{\delta_j}\right)\pi(\theta)\ d\theta\ < \infty\}$.

Heterogeneity often exists among different studies but data collected at one study are relatively homogeneous. The framework introduced above would accommodate potential heterogeneity among data sets from different sources or collected at different times. For example, in water quality assessment, we could take data observed at neighboring sites as different "historical" data sets. Moreover, data collected over a long period may be divided into several historical data sets to ensure the homogeneity within each data set. In such a way, the role of historical data can be more accurately evaluated (Duan, Ye and Smith, [14]). In Chapter 5, we will discuss an example of implementing the modified power prior approach using multiple sites information.

3.2.3 Power Prior for the Bernoulli Population

Suppose we are interested in making inference on the probability of success p from a Bernoulli family. Denote by $D = (x_1, ..., x_n)$ the current data and $D_0 = (x_{01}, ..., x_{0n_0})$ the historical data. Define $y_0 = \sum_{i=1}^{n_0} x_{0i}$, and $y = \sum_{j=1}^{n} x_j$.

Theorem 1. Assume that the initial prior distribution of$p$ follows a $Beta(\alpha_p, \beta_p)$, and the prior distribution of δ follows a $Beta(\alpha_\delta, \beta_\delta)$ distribution, where the hyperparameters $\alpha_p, \beta_p, \alpha_\delta$ and β_δ are all known. The joint posterior distribution for (p, δ) is

$$\pi(p, \delta|D_0, D) \propto \frac{p^{\delta y_0 + y}(1-p)^{\delta(n_0 - y_0)+(n-y)}\delta^{\alpha_\delta - 1}(1-\delta)^{\beta_\delta - 1}}{B(\delta y_0 + \alpha_p, \delta(n_0 - y0) + \beta_p)},$$

where $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ stands for the *beta* function.

Proof. Following (3.6), the joint power prior of (p, δ) can be written as

$$\pi(p, \delta|D_0) = \frac{[p^{y_0}(1-p)^{n_0 - y_0}]^\delta}{B(\delta y_0 + \alpha_p, \delta(n_0 - y0) + \beta_p)} \frac{\delta^{\alpha_\delta - 1}(1-\delta)^{\beta_\delta - 1}}{B(\alpha_\delta, \beta_\delta)}.$$

It is easy to show that the set A defined in (3.5) is $[0,1]$ in the Bernoulli case.

Combining the joint power priors with the likelihood based on the current data $L(p|D)$, we obtain the joint posterior distribution of (p,δ) of the form in Theorem 1. □

Integrating p out in $\pi(p,\delta|D_0,D)$, the marginal posterior distribution of δ is given by

$$\pi(\delta|D_0,D) \propto \frac{B(\delta y_0 + y + \alpha_p, \delta(n_0-y_0)+n-y+\beta_p)}{B(\delta y_0+\alpha_p,\delta(n_0-y0)+\beta_p)}\delta^{\alpha_\delta-1}(1-\delta)^{\beta_\delta-1}.$$

The behavior of the power parameter δ can be examined from this marginal posterior distribution. Similarly, the marginal posterior distribution of p can be derived by integrating δ out in $\pi(p,\delta|D_0,D)$, but it does not have a closed form. Instead we may learn the characteristic of the marginal posterior of p by studying the conditional posterior distribution of p on δ, combined with $\pi(\delta|D_0,D)$. An implementation of the power prior for Bernoulli data can be found in Duan, Smith and Ye ([13]).

3.2.4 Normal Population

In this section we are interested in making inference on the normal mean with unknown variance, by incorporating both current and historical data. Suppose that current data $D=(x_1,...,x_n)$ come from a normal $N(\mu,\sigma^2)$ population with unknown mean μ and variance σ^2, and $D_0=(x_{01},...,x_{0n_0})$ is a historical data set. Define

$$\bar{x}_0=\frac{1}{n_0}\sum_{i=1}^{n_0}x_{0i},\quad \bar{x}=\frac{1}{n}\sum_{j=1}^{n}x_j,\quad \hat{\sigma}_0^2=\frac{1}{n_0}\sum_{i=1}^{n_0}(x_{0i}-\bar{x}_0)^2,\quad \text{and}\quad \hat{\sigma}^2=\frac{1}{n}\sum_{j=1}^{n}(x_i-\bar{x})^2.$$

Following (3.6), the modified power prior for the normal population with unknown variance is given in the following theorem.

Theorem 2. Let $\pi(\mu,\sigma^2)$ denote the initial prior distribution for (μ,σ^2). Assume that the prior distribution of δ is a $Beta(\alpha_\delta,\beta_\delta)$, where hyper-parameters α_δ and β_δ are known. Then the modified power prior distribution of (μ,σ^2,δ) is

$$\pi(\mu,\sigma^2,\delta|D_0) \propto \frac{(\sigma^2)^{-\frac{\delta n_0}{2}}\exp\big\{-\frac{\delta n_0}{2\sigma^2}[\hat{\sigma}_0^2+(\mu-\bar{x}_0)^2]\big\}\pi(\mu,\sigma^2)\delta^{\alpha_\delta-1}(1-\delta)^{\beta_\delta-1}}{\int_0^\infty\int_{-\infty}^{+\infty}(\sigma^2)^{-\frac{\delta n_0}{2}}\exp\big\{-\frac{\delta n_0}{2\sigma^2}[\hat{\sigma}_0^2+(\mu-\bar{x}_0)^2]\big\}\pi(\mu,\sigma^2)d\mu d\sigma^2}, \qquad (3.10)$$

in the region of δ such that the denominator in (3.10) is finite.

When considering a special form of $\pi(\mu, \sigma^2)$, we are led to Corollaries 2.1, 2.2, and 2.3 whose proofs are simple and thus omitted.

Corollary 2.1. Suppose that we use the prior $\pi(\mu, \sigma^2) \propto (\frac{1}{\sigma^2})^a$ as the initial prior of (μ, σ^2), where $a > 0$ is a pre-determined constant. The joint power prior distribution of (μ, σ^2, δ) can be expressed as

$$\pi(\mu, \sigma^2, \delta | D_0) \propto \frac{\delta^{\frac{\delta n_0}{2}+a+\alpha_\delta-2}(1-\delta)^{\beta_\delta-1}}{\big(\frac{2\sigma^2}{n_0\hat{\sigma}_0^2}\big)^{\frac{\delta n_0}{2}+a}\Gamma\big(\frac{\delta n_0-3}{2}+a\big)} \exp\big\{-\frac{\delta n_0}{2\sigma^2}[\hat{\sigma}_0^2+(\mu-\bar{x}_0)^2]\big\},$$

where the range of δ is $(b, 1]$ for $b \geq 0$ or $[0, 1]$ for $b < 0$. Here, $b = \frac{2}{n_0}\big(\frac{3}{2} - a\big)$.

Note that $a = 1$ corresponds to the reference prior (Berger and Bernardo, [6]), while $a = \frac{3}{2}$ results in the Jeffreys prior (Jeffreys, [20]). Hence, if the reference prior of θ is used, the set A defined in (3.5) is $(\frac{1}{n_0}, 1]$ for the normal mean model, while if the Jeffreys prior of θ is used, $A = (0, 1]$. The lower bound b suggests that the information in historical data is automatically taken into account to a certain extent, depending on the availability of historical data. However, such a case may be changed once the original prior is changed.

Corollary 2.2. Assume $\pi(\mu, \sigma^2) \propto (\frac{1}{\sigma^2})^a$. The marginal posterior distribution of δ is

$$\pi(\delta | D_0, D) \propto \frac{\delta^{\frac{\delta n_0}{2}+a+\alpha_\delta-2}(1-\delta)^{\beta_\delta-1}\Gamma(\frac{\delta n_0+n-3}{2}+a)}{\big[\frac{\delta n}{\delta n_0+n}\frac{(\bar{x}_0-\bar{x})^2}{\hat{\sigma}_0^2}+\delta+\frac{n}{n_0}\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\big]^{\frac{\delta n_0+n-3}{2}+a}\Gamma(\frac{\delta n_0-3}{2}+a)},$$

with the range described in Corollary 2.1.

Corollary 2.3. Assume $\pi(\mu, \sigma^2) \propto (\frac{1}{\sigma^2})^a$. The conditional posterior distribution of μ, given δ and data (D_0, D), follows a Student t-distribution with, respectively, the location parameter and the scale parameter

$$\frac{\delta n_0 \bar{x}_0 + n\bar{x}}{\delta n_0 + n}, \qquad \sqrt{\frac{2}{C(\delta)}\frac{1}{(\delta n_0 + n + 2a - 3)(\delta n_0 + n)}},$$

and degrees of freedom $\delta n_0 + n + 2a - 3$, where

$$C(\delta) = \frac{2}{\frac{\delta n_0 n(\bar{x}_0-\bar{x})^2}{\delta n_0+n} + \delta n_0\hat{\sigma}_0^2 + n\hat{\sigma}^2}.$$

Furthermore, the conditional posterior distribution of σ^2, given δ and the data, follows an inverse-gamma distribution with parameters $\frac{\delta n_0+n+2a-3}{2}$ and $C(\delta)^{-1}$.

Duan, Ye and Smith ([14]) provides an example of implementing the power prior for a normal population with unknown variance.

3.3 Optimality Properties of the Modified Power Prior

We now provide theoretical supports for the modified power prior in two steps. First we review the justification provided by Ibrahim *et al.* ([19]) for the class of power prior with δ being fixed, and give a new interpretation on the optimality results. Then the optimality property of the posterior for δ is discussed to complete the justification.

3.3.1 Optimality of the Conditional Posterior $\pi(\theta|D_0, D, \delta)$

The posterior distribution of θ conditional on δ based on the modified power prior is the same as that based on the original one, which can be written as

$$\pi(\theta|D_0, D, \delta) \propto L(\theta|D_0)^{\delta} L(\theta|D)\pi(\theta). \tag{3.11}$$

Ibrahim *et al.* ([19]) discussed the optimality properties of $\pi(\theta|D_0, D, \delta)$ from two aspects. First, the $\pi(\theta|D_0, D, \delta)$ in (3.11) minimizes a convex sum of Kullback-Leibler (KL) divergence between two extreme posterior densities. One density, denoted by f_0, is based on no incorporation of historical data and the other density, denoted by f_1, is based on pooling the historical and current data. Let $K(g, f)$ denote the KL divergence between two densities g and f. Ibrahim *et al.* ([19]) showed that the $\pi(\theta|D_0, D, \delta)$ in (3.11) is the unique minimizer g_{opt} for the convex sum

$$K_g = (1 - \delta)K(g, f_0) + \delta K(g, f_1),$$

where δ is a fixed scalar between 0 and 1.

We point out that K_g can be viewed as the expected loss of using the density g to estimate the true posterior distribution of θ, which is denoted by f. Suppose δ is interpreted as the probability that D_0 follow the distribution for D. Furthermore, the KL divergence is used as the loss function between the estimated density and true density. If the historical data do come from the population underlying the current data, two samples should be pooled and hence $f = f_1$. Otherwise, we should use $f = f_0$. It follows that $K_g = Pr(f = f_0)K(g, f_0) + Pr(f = f_1)K(g, f_1) = E(K(g, f))$. Therefore $g_{opt} = \pi(\theta|D_0, D, \delta)$ is optimal in terms of minimizing the expected loss.

Second, Ibrahim *et al.* ([19]) showed that following the optimal information processing rules (IPR) of Zellner ([43], [45]), the $\pi(\theta|D_0, D, \delta)$ in (3.11) yield a 100% efficient IPR in the sense that the ratio of the output to input information is equal to 1.

Based on Zellner's theory of IPR, a weighted version of the information criterion function $\Delta[g(\theta)]$ is considered in our scenario.

$$\begin{aligned}\Delta[g(\theta)] &= \text{Output information} - \text{Input information} \\ &= \int g(\theta)\ln g(\theta)\ d\theta + \int g(\theta)\ln m(D, D_0)\ d\theta \\ &\quad - \left[\int g(\theta)\ln \pi(\theta)\ d\theta + \int g(\theta)\ln L(\theta|D)\ d\theta + \delta\int g(\theta)\ln L(\theta|D_0)\ d\theta\right], \end{aligned} \tag{3.12}$$

where $g(\theta)$ denotes a proper posterior density $\pi(\theta|D, D_0)$ in our setting.

In our research setting, the sampling distribution for current data is assumed to be $L(D|\theta)$, with θ being unknown. It is our ultimate goal to make inference on θ. The true distribution underlying historical data is not of interest. Instead, we are interested in how to borrow information in historical data to gain knowledge on the population underlying current data. Therefore, when we pretend $L(D_0|\theta)$ is the true sampling distribution for D_0 and use it to extract the information in D_0 about θ, the quality of the information in historical data is not as good as that in the current data. Hence the information on θ provided by the historical data should be discounted with a fractional number δ.

Zellner defined a rule to be 100% efficient whenever $\Delta[g(\theta)] = 0$; that is, output information equals input information. It turns out that the $g^*(\theta) = \pi(\theta|D_0, D, \delta)$ obtained using power prior yields $\Delta[g^*(\theta)] = 0$. The optimality of the power prior is hence established for the case when δ is a fixed scaler. Meanwhile, to achieve $\Delta[g^*(\theta)] = 0$, the $m(D, D_0)$ in (3.12) has to be in the form of

$$m^*(D, D_0) = \int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta.$$

This can be easily verified by substituting $g(\theta)$ with $\pi(\theta|D_0, D, \delta)$ in (3.12). Notice that $m^*(D, D_0)$ depends on δ. However, it is not necessarily a proper probability density function with respect to D and D_0. The marginal density of (D, D_0) given δ can be derived by normalizing $m^*(D, D_0)$.

$$\begin{aligned} m(D, D_0|\delta) &= \frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int\int\left(\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)d\theta\right) dDdD_0} \\ &= \frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int\int_\Theta L(\theta|D_0)^\delta \pi(\theta)d\theta\ dD_0}. \end{aligned} \tag{3.13}$$

Similarly, if only historical data are considered in Zellner's IPR, i.e., no $\int g(\theta) \ln L(\theta|D) d\theta$ in (3.12), we have

$$m^*(D_0) = \int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta.$$

Consequently, we obtain the marginal density of D_0 given δ by normalizing $m^*(D_0)$.

$$m(D_0|\delta) = \frac{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int \int_\Theta L(\theta|D_0)^\delta \pi(\theta) d\theta\ dD_0}. \tag{3.14}$$

Following (3.13) and (3.14), $m(D|D_0, \delta)$ can be written as

$$m(D|D_0, \delta) = \frac{m(D, D_0|\delta)}{m(D_0|\delta)} = \frac{\int_\Theta L(\theta|D) L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}. \tag{3.15}$$

We will use (3.15) for our further investigation on optimality when δ is random.

3.3.2 Optimality of $\pi(\delta|D_0, D)$

In addition to the optimality properties discussed above, the modified power prior yields an optimal $\pi(\delta|D_0, D)$ when δ is considered as a random variable.

Shannon ([29]) used a mutual information function to measure the dependency between two variables X and Y. Shannon's mutual information is defined by the expected entropy difference,

$$\vartheta(Y \wedge X) \equiv H(Y) - E_x[H(Y|x)] = E_{(x,y)}\left[\ln \frac{f(x|y)}{f(x)}\right],$$

where $H(Y)$ is the entropy of $f(y)$ and $H(Y|x)$ is the entropy of the conditional distribution $f(y|x)$. The mutual information is a measure of the expected information about Y transmitted through a noisy channel, which is represented by X.

We borrow the concept of mutual information, and interpret $\ln \frac{m(D|D_0)}{m(D)}$ as the observed mutual information between two samples D_0 and D, where $m(D) = \int_\Theta L(\theta|D)\pi(\theta)\ d\theta$ is the marginal density of D, and $m(D|D_0)$ is the density of D given that D_0 is observed. The term $\ln \frac{m(D|D_0)}{m(D)}$ measures the amount of information in historical data that is useful in interpreting the current data.

Although it is independent of values of model parameters, $\ln \frac{m(D|D_0)}{m(D)}$ is model dependent. Both $m(D|D_0)$ and $m(D)$ depend on the type of the sampling distribution, because they are derived by weighting the sampling distribution with priors on model parameters. In

addition, $m(D|D_0)$ also depends on how historical data are incorporated. Such dependence is easier to be recognized if we break down $\ln m(D|D_0)$ into

$$\ln m(D|D_0) = \ln \frac{m(D|D_0,\delta)\pi(\delta|D_0)}{\pi(\delta|D_0,D)}. \tag{3.16}$$

As discussed in Section 3.2.2, we believe that historical data alone does not provide any information on δ, because δ is introduced to measure the probability that historical and current data come from the same population. Therefore, it would not compromise the implication of $\ln \frac{m(D|D_0)}{m(D)}$ if the prior of δ, $\pi(\delta)$, is used to substitute the $\pi(\delta|D_0)$ in (3.16). Then the observed mutual information between D and D_0 becomes $\ln \frac{m(D|D_0,\delta)\pi(\delta)}{\pi(\delta|D_0,D)} - \ln m(D)$ in this setting. Furthermore we may take its expectation with respect to the posterior distribution of δ, and define

$$\varpi(D \wedge D_0) \equiv E_{\pi(\delta|D,D_0)}\left[\ln \frac{m(D|D_0,\delta)\pi(\delta)}{\pi(\delta|D_0,D)} - \ln m(D)\right]$$

as the weighted mutual information between historical and current data.

The goal of our research is to find a suitable method for incorporating historical data into current study. The power prior method with a fixed δ has been well justified as an optimal method for this purpose in section 3.3.1. So our search can be constrained to finding a optimal power prior method with a random δ.

In the context of power priors, $\varpi(D_0 \wedge D)$ measures the expected information in historical data transmitted through a power prior model for understanding the sampling distribution of current data. Therefore, it is desirable to find an optimal power prior method that maximizes $\varpi(D_0 \wedge D)$, which further comes down to finding an optimal $\pi(\delta|D_0,D)$. The result is stated as a formal theorem.

Theorem 3. The density $\pi(\delta|D_0,D)$ that maximizes $\varpi(D_0 \wedge D)$ is

$$\pi^*(\delta|D_0,D) \propto \pi(\delta)\frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}.$$

Proof. We have

$$\begin{aligned}
\varpi(D_0 \wedge D) &\equiv E_{\pi(\delta|D,D_0)}\left[\ln \frac{m(D|D_0,\delta)\pi(\delta)}{\pi(\delta|D_0,D)} - \ln m(D)\right] \\
&= \int \pi(\delta|D_0,D) \ln \frac{m(D|D_0,\delta)\pi(\delta)}{\pi(\delta|D_0,D)} d\delta - \ln m(D) \\
&= -K\left(\pi(\delta|D_0,D), \frac{m(D|D_0,\delta)\pi(\delta)}{M}\right) + \ln M - \ln m(D),
\end{aligned}$$

where $M = \int m(D|D_0, \delta)\pi(\delta)\, d\delta$ is the normalizing constant of $m(D|D_0, \delta)\pi(\delta)$. Now clearly $-K\left(\pi^*(\delta|D_0, D), \frac{m(D|D_0,\delta)\pi(\delta)}{M}\right)$ is maximized and equal to 0 when

$$\pi^*(\delta|D_0, D) = \frac{m(D|D_0, \delta)\pi(\delta)}{M} \propto m(D|D_0, \delta)\pi(\delta).$$

Combined with (3.15), it leads to

$$\pi^*(\delta|D_0, D) \propto \pi(\delta)\frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta\pi(\theta)\ d\theta}{\int_\Theta L(\theta|D_0)^\delta\pi(\theta)\ d\theta}.$$

□

This optimal $\pi^*(\delta|D_0, D)$ is precisely the marginal posterior of δ based on the modified power prior (see (3.7)).

3.4 Comparisons of Two Power Prior Approaches

Ibrahim and Chen ([18]) constructed the joint power prior of (θ, δ) as

$$\pi(\theta, \delta|D_0) \propto (L(\theta|D_0))^\delta\pi(\theta)\pi(\delta).$$

Notice that if we multiply the likelihood function $L(\theta|D_0)$ with a positive constant k, the joint prior distribution of (θ, δ) becomes

$$\pi(\theta, \delta|D_0) \propto k^\delta(L(\theta|D_0))^\delta\pi(\theta)\pi(\delta).$$

The joint prior of (θ, δ) and consequently the posterior will change by k^δ. Therefore, in the original power prior approach, proportional likelihood does not produce the same posterior distribution. This result is not consistent with the likelihood principle.

On the other hand, multiplying the likelihood function $L(\theta|D_0)$ by a positive constant k wouldn't change results in the modified approach, which is shown as follows.

$$\pi(\theta, \delta|D_0) = M\frac{k^\delta(L(\theta|D_0))^\delta\pi(\theta)\pi(\delta)}{\int_\Theta k^\delta(L(\theta|D_0))^\delta\pi(\theta)\ d\theta}I_A(\delta) = M\frac{(L(\theta|D_0))^\delta\pi(\theta)\pi(\delta)}{\int_\Theta(L(\theta|D_0))^\delta\pi(\theta)\ d\theta}I_A(\delta),$$

Another feature of the modified power prior approach is that the only condition needed to ensure the propriety of the joint prior for (θ, δ) is $\int_\Theta L(\theta|D_0)\pi(\theta)\ d\theta\ < \infty$. Since any prior $\pi(\theta)$ used in the regular Bayesian updating scheme ($\delta = 1$) has to satisfy this condition

to produce a proper posterior, this is an appropriate assumption. Hence no additional effort is needed in checking the propriety of $\pi(\theta, \delta)$ under the modified approach. For the original power prior approach, certain conditions are required to achieve a proper power prior. Ibrahim and Chen ([18]) and Chen *et al.* ([10]) examined the propriety conditions for four commonly used classes of regression models.

Although the joint power priors of (θ, δ) are different, the conditional power prior $\pi(\theta|D_0, \delta)$ in (3.1) and the conditional posterior $\pi(\theta|D_0, D, \delta)$ in (3.11) are the same for both approaches. This feature indicates that the two approaches are equivalent for a fixed δ, which is expected because both approaches are rooted in the same idea presented by the definition of $\pi(\theta|D_0, \delta)$. This also implies that the differences in results between two approaches come from their difference in the posterior marginal distributions of δ. Therefore we may examine their differences in $\pi(\theta, \delta|D_0, D)$ by comparing $\pi(\delta|D_0, D)$ between two approaches.

3.4.1 Comparison of Posteriors

The marginal posterior distributions of θ and δ are of interest here, because the former leads to the final inference on θ and the latter describes the characteristics of the power prior approach. The marginal posterior means of θ and δ and the marginal posterior mode of δ will be used to compare the posterior distributions of the two approaches. The marginal posterior mode of δ represents the most likely value of δ given by the historical and current data. Since $\pi(\delta|D_0, D)$ is often asymmetric, the marginal posterior mode of δ is an important statistic for studying the marginal posterior distribution of δ.

Compatibility statistics

To discuss how well the marginal posterior mode of δ responses to the compatibility between current and historical data, the notion of "compatibility statistic" is defined. Let $x_1, ..., x_n$ be i.i.d observations from an exponential family with probability density function or probability mass function of the form

$$f(x|\theta) = h(x) \exp\left(\sum_{i=1}^{k} w_i(\theta) t_i(x) + \tau(\theta) \right),$$

where the dimension of θ is no larger than k. Clearly,

$$T(\underline{x}) = (\sum_{j=1}^{n} t_1(x_j), ..., \sum_{j=1}^{n} t_k(x_j))$$

is a sufficient statistic for θ (Casella and Berger, [8]). One underlying assumption of this sufficiency is that the sample size n is fixed when the experiment is performed repeatedly. However, the current and historical data often have different sample sizes. This then raises the question of how to measure the difference between two samples with unequal sizes in terms of their information about θ. Note that in this section $\underline{x}$ represents an arbitrary sample in general, including current and historical sample.

Define

$$C(\underline{x}) = \Big(c_1(\underline{x}), ..., c_k(\underline{x})\Big) = \Big(\sum_{j=1}^{n} \frac{t_1(x_j)}{n}, ..., \sum_{j=1}^{n} \frac{t_k(x_j)}{n}\Big)$$

as the compatibility statistic of a sample $\underline{x} = (x_1, ..., x_n)$ for θ. For example, $C(\underline{x}) = \frac{y}{n} = \bar{x}$ for the Bernoulli case, and $C(\underline{x}) = (\bar{x}, \hat{\sigma}^2)$ for the normal case, where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the maximum likelihood estimator of σ^2. The density function of the sample $\underline{x}$ may be expressed as

$$f(\underline{x}|\theta) = \Big(\prod_{j=1}^{n} h(x_j)\Big) g(\theta|C(\underline{x}))^n,$$

where $g(\theta|C(\underline{x}))$ depends on the sample $\underline{x}$, including the sample size n, only through $C(\underline{x})$. Since $\prod_{j=1}^{n} h(x_j)$ does not contain θ, the log likelihood function of θ based on the sample can be written as $\ln L(\theta|\underline{x}) \propto n \ln g(\theta|C(\underline{x}))$. For any specific distribution, $C(\underline{x})$ determines the averaged log likelihood, and hence characterizes a sample in terms of the averaged information about θ it carries. Most of information measures are functions of log likelihood. For example, suppose θ is a one-dimensional parameter, then the Fisher information about θ provided by the sample $\underline{x}$ is

$$J_n(\theta) = \sum_{j=1}^{n} -\frac{\partial^2}{\partial\theta^2} \ln f(x_j|\theta) = -n\Big[\sum_{i=1}^{k}\Big(w_i''(\theta)\sum_{j=1}^{n}\frac{t_i(x_j)}{n}\Big) + \tau''(\theta)\Big]. \tag{3.17}$$

It is easy to check that for high dimensional θ, each element of $J_n(\theta)$ is of the similar form as in (3.17). Therefore the average Fisher information $J_n(\theta)/n$ carried by the sample depends on the data only through $C(\underline{x})$.

Notice that in either current or previous study, $C(\underline{x})$ is also a sufficient statistic for θ when n is fixed. Therefore $C(\underline{x})$ captures all the information about θ contained in a sample within the scope of each study. On the other hand, it is comparable among samples across studies with the same distributional assumption but different sample sizes. Note that $C(\underline{x})$ is not sufficient anymore if current and previous studies are combined as a single experiment with random n, since n becomes part of the data.

Applying the concept of the compatibility statistic on our investigation of power priors, we have the following result whose proof is in the Appendix.

Result 3.1 *Suppose that historical data D_0 and current data D are two independent random samples from an exponential family. Furthermore, suppose that $\pi(\delta) = 1$ is used as the prior for δ. The compatibility statistic for the historical data and current data are $C(D_0)$ and $C(D)$ respectively, as defined earlier. If two samples are fully compatible, i.e. $C(D_0) = C(D)$, the marginal posterior mode of δ is always 1 under the modified power prior approach, for any n_0 and n (no matter how high the ratio n_0/n is).*

This is very rational since when the historical data contribute necessary information into the current study, we would like to use it as much as possible to achieve higher precision. As long as the difference between $C(D_0)$ and $C(D)$ is negligible from a practical point of view, it is appropriate to view the historical and current samples as fully compatible, and hence the marginal posterior mode of δ would be 1 or very close to 1 under the modified power prior approach. This property is also supported by the numerical results that will be presented later in this section.

However, in the original power prior approach, the posterior mode of δ changes if we multiply the likelihood function by a constant, because the likelihood principle doesn't hold there. Furthermore, for any historical and current data, we can always find a positive constant k_0 such that the marginal posterior mode of δ becomes 0 after the likelihood function is multiplied with k_0.

Result 3.2 *Suppose that current data D are from a population with a density function $f(x|\theta)$, and D_0 is a related historical data set. Furthermore, suppose that $\pi(\delta) = 1$ is used as the prior for δ and the conditional posterior distribution of θ on δ is proper for any δ. Then for any D_0 and D, there exists at least one positive constant k_0 such that $\pi(\delta|D_0, D)$ has mode at $\delta = 0$ under the original power prior approach, where $L(\theta|x) = k_0 f(x|\theta)$.*

The proofs of these results are given in the Appendix. For a normal or Bernoulli population, our research reveals that $\pi(\delta|D_0, D)$ has mode at $\delta = 0$ even when $k_0 = 1$. This strong tendency of δ toward 0 in the original approach compromises the flexibility of using a random δ. Also, the role of historical data is underestimated.

Change of posteriors in terms of compatibility between D and D_0

The compatibility between historical and current data may be measured by their difference in each element of the compatibility statistic, e.g. $|c_i(D) - c_i(D_0)|$ or $c_i(D)/c_i(D_0)$ for $i = 1, ..., k$. For example, in the case of the normal mean model, the compatibility between the historical and current data can be measured by the differences in the sample mean and MLE of the variance ($\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_i - \bar{x})^2$). In the Bernoulli case, the compatibility between two data sets is measured by the difference in sample proportion. Since the power parameter is initiated to quantify the heterogeneity between two samples, the magnitude of δ should depend on samples' difference in the compatibility statistic. In addition, the relative sample size of the historical data to the current data, denoted by n_0/n, may also influence the amount of penalty that should be put on the historical data. This is because when two samples are not perfectly homogeneous, the dominance of historical data in current analysis should be prevented.

Since the analytical forms of the marginal posterior mean and mode of δ as well as the marginal posterior mean of θ are intractable, we investigate their properties by computing numerical results under various situations. In this article, normal and Bernoulli populations are extensively studied using current and historical samples with a wide range of sample sizes, sample means, and sample variances. In the rest of this section, some illustrative figures will be presented (see Figures 3.1, 3.2, 3.3, 3.4, and 3.5), and several remarks will be discussed.

Figures 3.1 and 3.2 compare two power prior methods in posterior means and mode for the Bernoulli population; Figures 3.3, 3.4, and 3.5 are for the normal population. Those figures show that with the modified power prior approach, the posterior mode of δ is sensitive to the compatibility between historical and current data. The posterior mode of δ goes to zero very fast when the compatibility decreases. In addition, when the two samples are not perfectly homogeneous, the posterior mode of δ decreases with n_0/n, and attains 1 with very small n_0/n. How quick the mode reaches 1 also depends on the extent of compatibility between historical and current data. However, the posterior mode of δ would never be zero under the modified approach. This is reasonable because the historical population is subjectively believed to have similarity with the current population, more or less. These trends imply that the random δ responds to data in a sensible way in the modified approach.

For both normal and Bernoulli populations, the posterior mean of θ is much more sensitive to the change of compatibility or n_0/n under the modified approach than under the

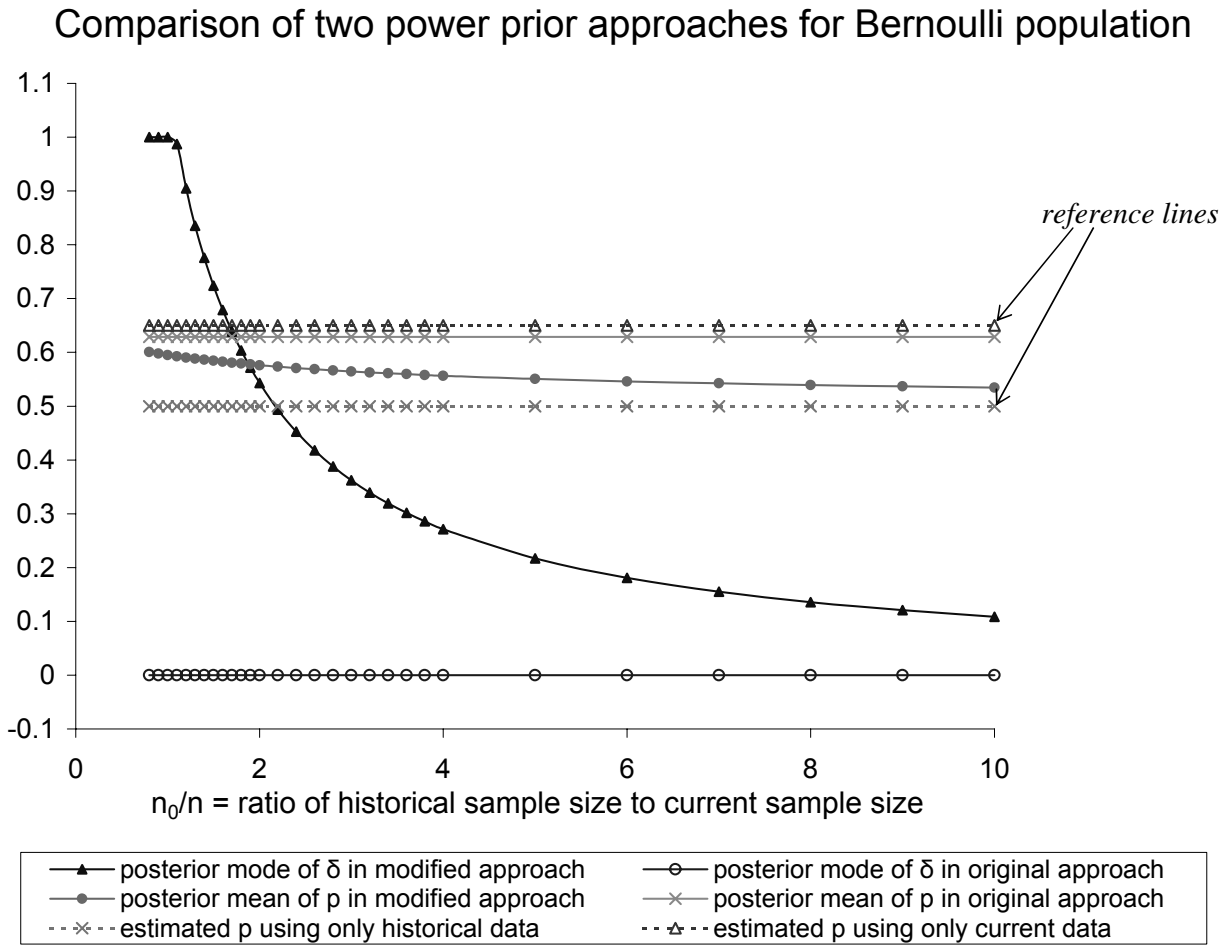

Figure 3.1: Marginal posterior mode of δ or marginal posterior mean of p using different ratios of historical sample size to current sample size, when $n = 20, \bar{x} = 0.65, \bar{x}_0 = 0.5$.

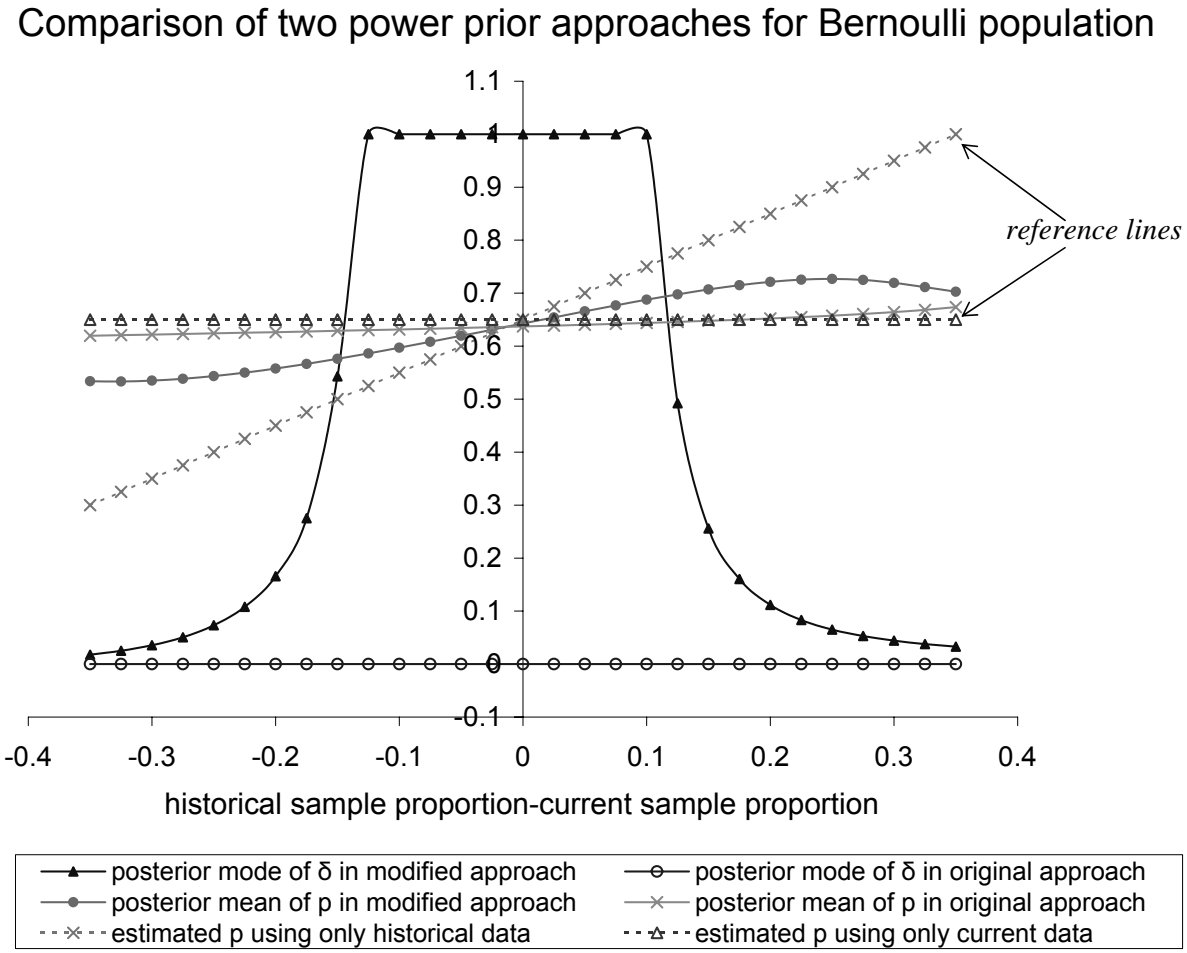

Figure 3.2: Marginal posterior mode of δ or marginal posterior mean of p considering different divergence in sample proportion between historical and current data, when $n = 20, \bar{x} = 0.65, n_0 = 40$.

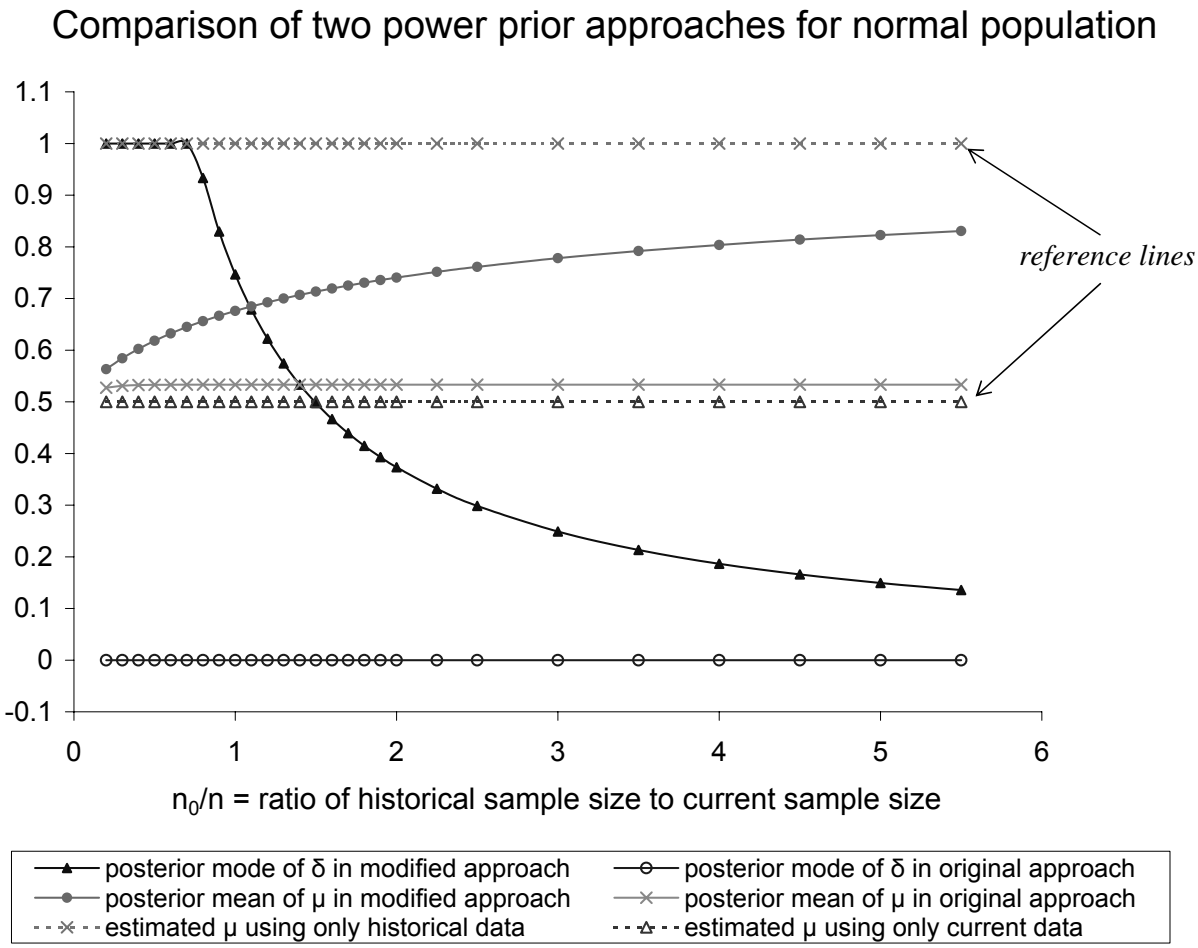

Figure 3.3: Marginal posterior mode of δ or marginal posterior mean of μ using different ratios of historical sample size to current sample size, when $n = 20, \bar{x} = 0.5, \hat{\sigma}^2 = 0.8, \bar{x}_0 = 1, \hat{\sigma}_0^2 = 1$.

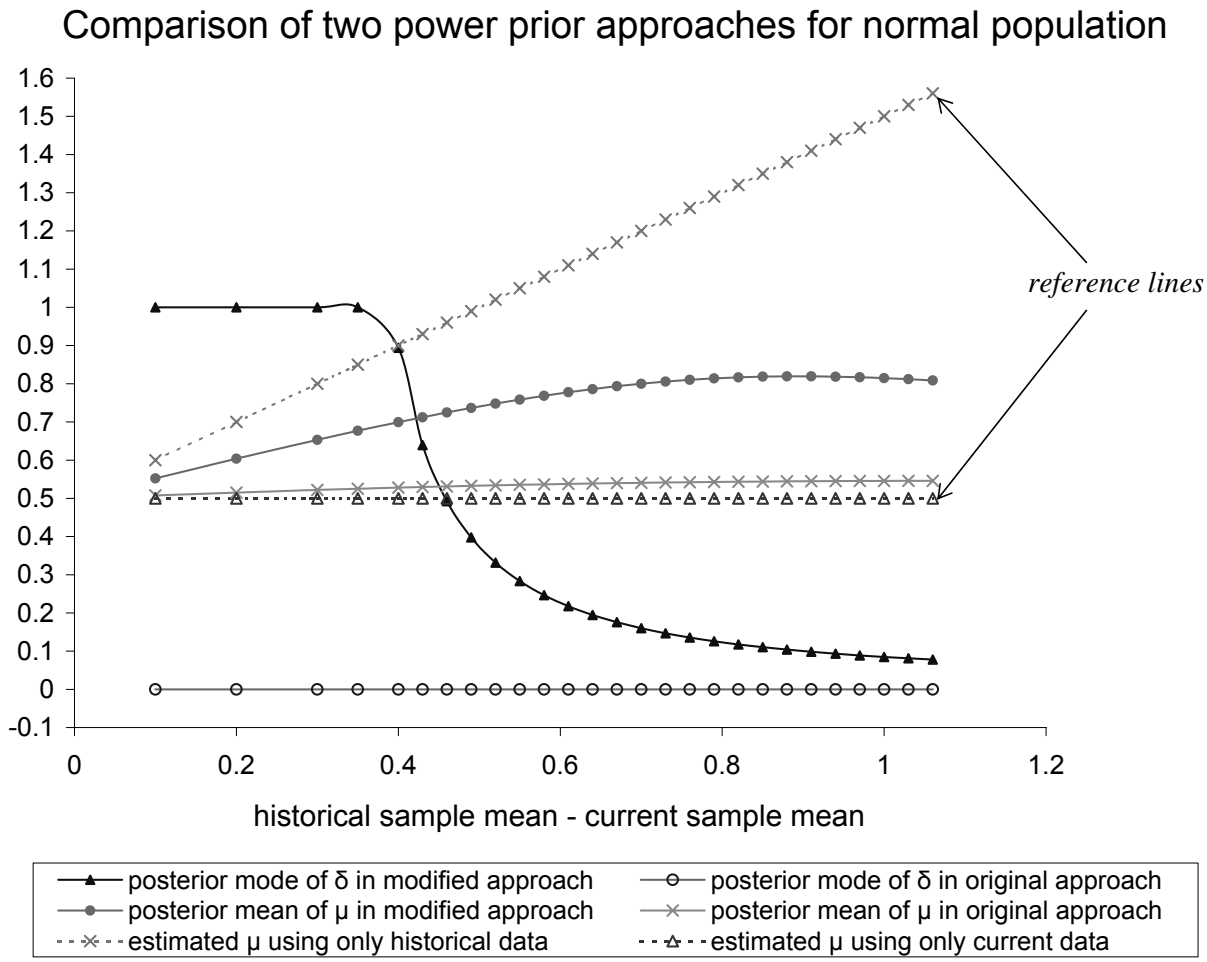

Figure 3.4: Marginal posterior mode of δ or marginal posterior mean of μ considering different divergence in sample mean between historical and current data, when $n = 20, \bar{x} = 0.5, \hat{\sigma}^2 = 0.8, n_0 = 40, \hat{\sigma}_0^2 = 1$.

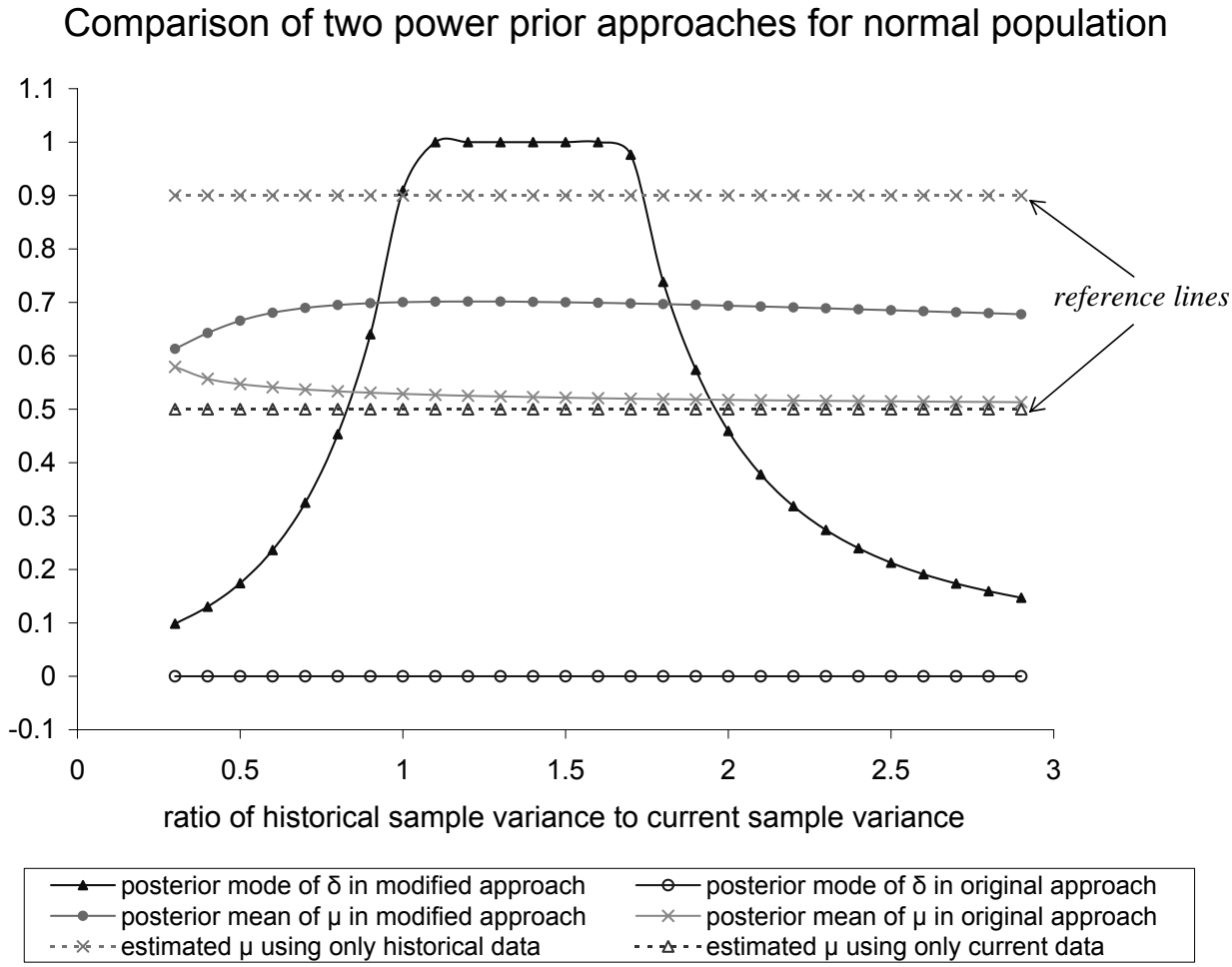

Figure 3.5: Marginal posterior mode of δ or marginal posterior mean of μ using different ratios of historical sample variance to current sample variance, when $n = 20, \bar{x} = 0.5, \hat{\sigma}^2 = 1, n_0 = 40, \bar{x}_0 = 1$. Note that here we use $\hat{\sigma}^2$, the MLE of population variance, to measure the sample variance.

original approach. Since two approaches have the same conditional posterior $\pi(\theta|D_0, D, \delta)$, their difference in $\pi(\delta|D_0, D)$ will explain the difference in the marginal posterior of θ. In the modified approach, the observed trends of posterior mean of θ are consistent with the trends of posterior mean or mode of δ. With n_0/n increasing, the posterior mean of θ is getting closer to the estimated θ derived based on the historical data alone, and it is getting closer to the estimated θ using only the current data when the compatibility between two samples is decreasing. However, no trend was found for the posterior mean of θ under the original approach.

In the original approach, the power parameter δ always has a tendency to be close to zero, which suggests that much of historical information is not used. The posterior mode of δ is always zero, and the posterior mean of δ is consistently lower than that in the modified approach. The original approach always puts a very light weight on the historical data, even when the historical and current data are perfectly homogeneous or when n_0 is much smaller than n. This suggests that the original approach may underestimate the power parameter in general.

Based on the empirical results represented by five figures, the trends of the posterior mean and mode of δ using both original and modified power prior approaches are summarized in Table 3.1; the trends of the posterior mean of θ in various situations are presented in Table 3.2.

The role of the power parameter δ is to control the influence of the historical data on the current study. The empirical results show that the original approach underestimates δ in general and hence the influence of historical data is small no matter how compatible current and historical data are. One consequence is that in the original approach, the posterior mean and variance of θ are very close to those without considering historical data. Therefore the historical data turn out not to be very helpful in parameter estimation, and this questions the necessity of incorporating the historical data. On the other hand, in the modified approach, this power control parameter is adjusted automatically based on the compatibility between the historical and current data, and also based on the sample sizes of the two studies. Therefore the modified power prior approach may be recommended over the original one.

Table 3.1: *Comparison of the posterior mode and mean of δ under two power prior approaches for normal and Bernoulli populations. The trends of posterior mode and mean of δ with respect to the ratio of two sample sizes and the compatibility between historical and current data are summarized based on empirical results.*

Incompatibility measure	Marginal Posterior Mode of δ		Marginal Posterior Mean of δ	
	original method	modified method	original method	modified method
normal:				
$\left\lvert\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}-1\right\rvert$ increases	0	decreases	decreases	decreases
$\lvert\bar{x}_0-\bar{x}\rvert$ increases	0	decreases	decreases	decreases
$\frac{n_0}{n}$ increases	0	decreases	decreases	decreases
Bernoulli:				
$\lvert\bar{x}_0-\bar{x}\rvert$ increases	0	decreases	decreases	decreases
$\frac{n_0}{n}$ increases	0	decreases	decreases	decreases

Table 3.2: *Comparison of the posterior mean of θ under two power prior approaches for normal and Bernoulli populations. The trends of posterior mean of θ with respect to the ratio of two sample sizes and the compatibility between historical and current data are summarized based on empirical results. "original" refers to the original power prior approach; "modified" refers to the modified power prior approach.*

Incompatibility measure	Marginal Posterior Mean of θ			
	normal population		Bernoulli population	
	original	modified	original	modified
$\left\lvert\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}-1\right\rvert$ increases	no trend	goes to the estimate by current data alone		
$\lvert\bar{x}_0-\bar{x}\rvert$ increases	no trend	goes to the estimate by current data alone	no trend	goes to the estimate by current data alone
$\frac{n_0}{n}$ increases	no trend	goes to the estimate by historical data alone	no trend	goes to the estimate by historical data alone

3.4.2 Comparison in Mean Squared Error (MSE)

Normal population

First, we compare MSE of the estimated μ in a normal population with unknown variance under the original and modified power prior approaches. Simulation is conducted to compare

the performance of two approaches in terms of MSE. A Monte-Carlo estimate of the MSE for estimating θ, $\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}_i-\theta)^2$, is used for comparisons, where m is the number of total runs, θ is the true parameter, and $\hat{\theta}_i$ is the estimate of parameter in the ith run.

Suppose that the current sample is from a normal $N(\mu,\sigma^2)$ population and the historical sample is from a normal $N(\mu_0,\sigma_0^2)$ population, with both mean and variance unknown. Furthermore, suppose the population mean of current sample, μ, is the parameter of interest. Denote by $\hat{\mu}$ the marginal posterior mean of μ to be used as the estimate of μ. The MSE of $\hat{\mu}$ is $E(\hat{\mu}-\mu)^2$. After simplifying $E(\hat{\mu}-\mu)^2$ using the posterior distributions derived in Section 3.2.4, we find that $E(\hat{\mu}-\mu)^2$ depends on $n, n_0/n, \sigma, \sigma_0/\sigma$, and $(\mu_0-\mu)/\sigma$ under both power prior approaches. Therefore, we consider different combinations of $n, n_0/n, \sigma, \sigma_0/\sigma$, and $(\mu_0-\mu)/\sigma$ when investigating the MSE. In each simulation scenario, i.e., a combination of $n, n_0/n, \sigma, \sigma_0/\sigma$, and $(\mu_0-\mu)/\sigma$, five thousand runs are performed and then the averaged squared error of $\hat{\mu}$ is calculated as an estimate of MSE. In this section, we use the reference prior $\pi(\mu,\sigma^2)\propto 1/\sigma^2$ as the initial prior for (μ,σ^2).

Figures 3.6 and 3.7 present the simulated MSE for n_0/n from 0.5 to 5 from the original and modified power prior approaches in the scenarios where $n=10$, $\sigma=\sigma_0=1$, and $(\mu_0-\mu)/\sigma=0.1$ to 0.5. It shows that the MSE from the modified approach decreases with n_0/n increasing when $0<(\mu_0-\mu)/\sigma\leq 0.3$, decreases first and then increases in n_0/n when $0<(\mu_0-\mu)/\sigma=0.4$, and increases in n_0/n when $(\mu_0-\mu)/\sigma>0.4$. This implies when the divergency between the current and historical populations is mild, incorporating more historical data would substantially decrease the MSE using the modified power prior. In the modified approach not only the MSE significantly increases with $(\mu_0-\mu)/\sigma$, but also its trend changes with n_0/n. If the two populations are quite heterogeneous, the increase in the bias of estimate is faster than the decrease in the variability of estimate when more historical data is available, and therefore the MSE increases. In contrast, the MSE from the original approach seems stable with different historical sample sizes and only slightly increases with the standardized difference in population mean.

The modified approach gives consistently smaller MSEs than those from the original approach if the standardized difference in mean, $(\mu_0-\mu)/\sigma$, is no greater than 0.3; the MSE of two approaches are comparable if $(\mu_0-\mu)/\sigma=0.4$; and the original approach performs better than the modified one if $(\mu_0-\mu)/\sigma\geq 0.5$. If the current and historical populations are significantly heterogeneous, we would expect large variability in the data and consequently in parameter estimates when incorporating samples from both populations

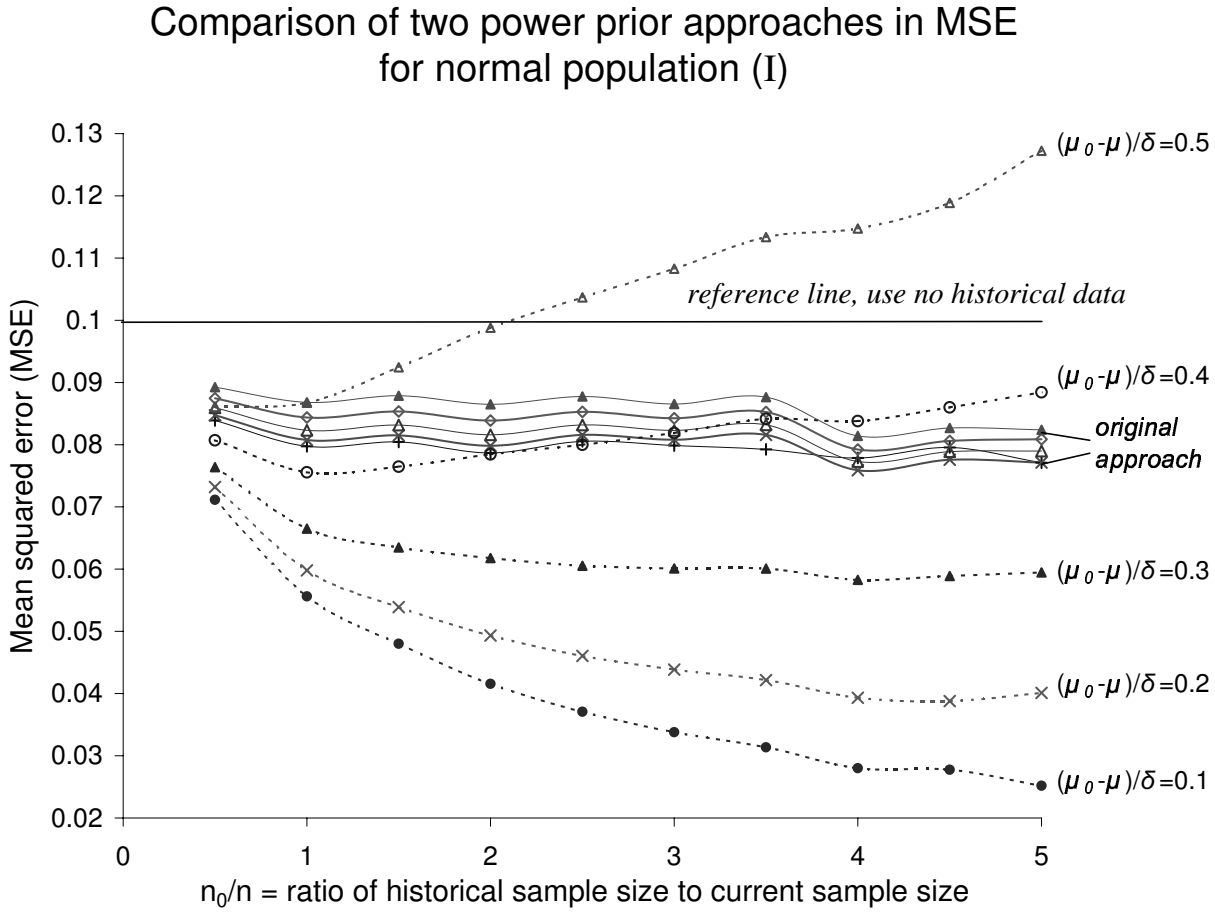

Figure 3.6: MSE of $\hat{\mu}$ using two power prior approaches (part I), where $\hat{\mu}$ is the marginal posterior mean of μ. Solid lines represent results from the original approach with $(\mu_0-\mu)/\sigma = 0.1$ to 0.5; dashed lines represent results from the modified approach with $(\mu_0 - \mu)/\sigma = 0.1$ to 0.5. $n = 10$, $n_0 = 5$ to 50, and $\sigma = \sigma_0 = 1$ are used.

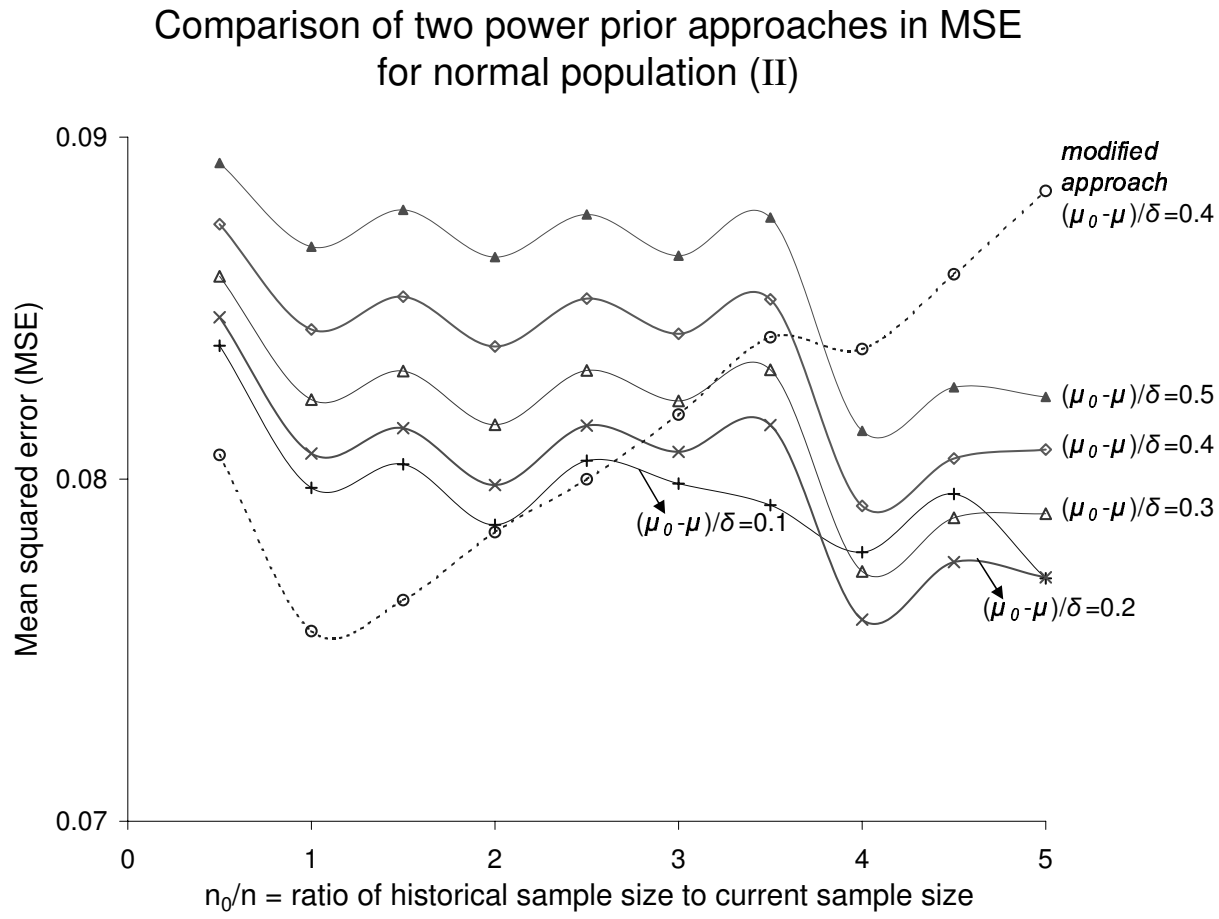

Figure 3.7: MSE of $\hat{\mu}$ using two power prior approaches (part II), where $\hat{\mu}$ is the marginal posterior mean of μ. Solid lines represent results from the original approach with $(\mu_0-\mu)/\sigma = 0.1$ to 0.5; dashed lines represent results from the modified approach with $(\mu_0-\mu)/\sigma = 0.4$. $n = 10$, $n_0 = 5$ to 50, and $\sigma = \sigma_0 = 1$ are used.

into one analysis. Therefore it is expected to observe a relatively large MSE in the modified approach. The original approach always puts very light weight on historical data, so its MSE is not affected much by the heterogeneity between current and historical populations. Besides those presented in Figure 3.6, simulations were run using different n and unequal variances for historical and current populations. In all simulation scenarios, similar trends are found as those in Figure 3.6. Table 3.3 summarizes the simulation results using the "trend turning point" and "MSE change point".

In Figure 3.7, the scenario with $(\mu_0 - \mu)/\sigma = 0.4$ and $n_0/n = 1$ appears to be a "trend turning point" for the situation when $n = 10$ and $\sigma = \sigma_0 = 1$. If $(\mu_0 - \mu)/\sigma > 0.4$, or $(\mu_0 - \mu)/\sigma = 0.4$ and $n_0/n > 1$, the performance of the modified power prior start to slip down. When $n_0/n \geq 4$ under $(\mu_0 - \mu)/\sigma = 0.4$, the MSE from the modified approach becomes larger than that from the original approach, which is also observed when $(\mu_0 - \mu)/\sigma > 0.4$ with any n_0. We refer to the scenario with $(\mu_0 - \mu)/\sigma = 0.4$ and $n_0/n = 4$ as the "MSE change point" in Figure 3.7.

Furthermore, we found that such a trend turning point and an MSE change point exist for other combinations of n, σ and σ_0, which is illustrated in Table 3.3. For each combination of n, σ and σ_0 in Table 3.3, the MSE from the two power prior approaches has similar trend as in Figure 3.6. With fixed n, σ and σ_0, We may evaluate the performance of two power prior approaches with a sequence of increasing $(\mu_0 - \mu)/\sigma$ first and then increasing n_0/n for each level of $(\mu_0 - \mu)/\sigma$. In this process usually we encounter the trend turning point first and then the MSE change point. Finally the MSE from the modified approach may even exceed the MSE derived using only the current sample, when $(\mu_0 - \mu)/\sigma$ and n_0/n are large. Therefore, the modified power prior approach is suitable when the divergence between current and historical populations is mild, but may be dangerous when the divergence is large. This is no surprise because the idea of borrowing information from historical data is established on the belief that the current and historical populations are quite similar. The performance of the original approach is not affected much by the divergence between two populations and neither by the availability of historical data. This property is an advantage when the divergence is large but a disadvantage when the divergence is small.

Bernoulli population

We also compared the MSE of estimated p in a Bernoulli population, where p is the true probability of success in the current sample. Suppose the current sample is from a Bernoulli(n,p)

Table 3.3: *The "trend turning point" and "MSE change point" for different combinations of* $\sigma, \sigma_0/\sigma$ *and* n.

Setting			Trend turning point		MSE change point	
σ	σ_0/σ	n	$(\mu_0-\mu)/\sigma$	n_0/n	$(\mu_0-\mu)/\sigma$	n_0/n
1	1	10	0.4	1	0.4	4
1	1	20	0.3	1	0.3	3.5
1	1	50	0.2	0.5	0.2	3.5
1	1.5	10	0.4	1.5	0.4	2
1	1.5	20	0.3	1	0.3	4
1	1.5	50	0.2	1.5	0.3	1

population and the historical sample is from a Bernoulli(n_0,p_0) population, with both p and p_0 unknown. Since the posterior distribution of p is not symmetric, the marginal posterior mean may not be a suitable estimate of p. In this investigation, we use $E_\delta((\delta y_0+y)/(\delta n_0+n))$ as $\hat{p}$, the estimate of p, where the expectation is taken over the marginal posterior distribution of δ. Here y and y_0 are the total number of successes in the current and historical samples, as defined in section 3.2.3. It is straightforward to show that when a $uniform(0,1)$ is used as the initial prior for p, $(\delta y_0 + y)/(\delta n_0 + n)$ is the posterior mode of p conditional on δ, and furthermore $E_\delta((\delta y_0 + y)/(\delta n_0 + n))$ is a mode of marginal posterior distribution of p.

Tables 3.4 and 3.5 empirically summarize the situations under which the modified power prior approach leads to a smaller MSE than the original one. When the divergence between current and historical populations, represented by $|p - p_0|$, is small or mild, the modified approach gives lower MSE than the original one. This result is consistent with what is found in the normal case and can be interpreted similarly. As stated in the previous section, the justification of using historical data is built on the substantial similarity between the current and historical populations.

Furthermore, when the current sample size is 50, the modified approach shows much superior performance (smaller MSE most of time) compared to the original approach in terms of MSE. But we need to be cautious that the MSE from the modified approach may blow up when the current and historical populations are substantially heterogeneous.

Table 3.4: *The range of p_0 where MSE from the modified approach is smaller than that from the original approach, under different combination of n, n_0, and p.*

n	n_0/n	p				
		0.1	0.2	0.3	0.4	0.5
10	0.5	$0 \sim 0.2$	$0 \sim 0.4$	$0 \sim 0.6$	$0 \sim 0.7$	$0.2 \sim 0.8$
10	1.0	$0 \sim 0.2$	$0 \sim 0.4$	$0 \sim 0.5$	$0.2 \sim 0.6$	$0.3 \sim 0.8$
10	1.5	$0 \sim 0.2$	$0 \sim 0.3$	$0 \sim 0.5$	$0.2 \sim 0.6$	$0.3 \sim 0.7$
10	2.0	$0 \sim 0.2$	$0 \sim 0.3$	$0 \sim 0.5$	$0.2 \sim 0.6$	$0.3 \sim 0.7$
10	2.5	$0 \sim 0.2$	$0 \sim 0.3$	$0 \sim 0.5$	$0.2 \sim 0.6$	$0.3 \sim 0.7$
10	3.0	$0 \sim 0.2$	$0 \sim 0.3$	$0.2 \sim 0.4$	$0.2 \sim 0.6$	$0.3 \sim 0.7$
20	0.5	$0 \sim 0.2$	$0 \sim 0.3$	$0 \sim 0.5$	$0.2 \sim 0.6$	$0.3 \sim 0.7$
20	1.0	$0 \sim 0.2$	$0 \sim 0.3$	$0.2 \sim 0.4$	$0.2 \sim 0.6$	$0.4 \sim 0.7$
20	1.5	$0 \sim 0.19$	$0 \sim 0.3$	$0.2 \sim 0.4$	$0.3 \sim 0.5$	$0.4 \sim 0.7$
20	2.0	$0 \sim 0.19$	$0 \sim 0.3$	$0.2 \sim 0.4$	$0.3 \sim 0.5$	$0.4 \sim 0.7$
20	2.5	$0 \sim 0.18$	$0 \sim 0.3$	$0.2 \sim 0.4$	$0.3 \sim 0.5$	$0.4 \sim 0.7$
20	3.0	$0 \sim 0.17$	$0.1 \sim 0.2$	$0.2 \sim 0.4$	$0.3 \sim 0.65$	$0.3 \sim 0.7$
50	0.5	$0 \sim 0.1$	$0 \sim 0.3$	$0.2 \sim 0.4$	$0.3 \sim 0.5$	$0.4 \sim 0.6$
50	1.0	$0 \sim 0.1, 0.8 \sim 1$	$0 \sim 0.3, 0.7 \sim 1$	$0.2 \sim 1$	$0.3 \sim 0.9$	$0.3 \sim 0.9$
50	1.5	$0 \sim 0.1, 0.6 \sim 1$	$0.2 \sim 1$	$0.2 \sim 1$	$0.2 \sim 1$	$0 \sim 0.9$
50	2.0	$0 \sim 0.1, 0.6 \sim 1$	$0.2 \sim 1$	$0.2 \sim 1$	$0 \sim 1$	$0 \sim 1$
50	2.5	$0 \sim 0.1, 0.5 \sim 1$	$0.1 \sim 1$	$0.1 \sim 1$	$0 \sim 1$	$0 \sim 1$
50	3.0	$0 \sim 0.1, 0.5 \sim 1$	$0.1 \sim 1$	$0.1 \sim 1$	$0 \sim 1$	$0 \sim 1$

Table 3.5: *The range of p_0 where MSE from the modified approach is smaller than that from the original approach, under different combination of n, n_0, and p. (cont')*

n	n_0/n	p			
		0.6	0.7	0.8	0.9
10	0.5	$0.3 \sim 0.9$	$0.4 \sim 1$	$0.6 \sim 1$	$0.8 \sim 1$
10	1.0	$0.4 \sim 0.8$	$0.5 \sim 0.9$	$0.6 \sim 1$	$0.8 \sim 1$
10	1.5	$0.4 \sim 0.8$	$0.5 \sim 0.9$	$0.7 \sim 1$	$0.8 \sim 1$
10	2.0	$0.4 \sim 0.8$	$0.5 \sim 0.9$	$0.7 \sim 0.9$	$0.8 \sim 1$
10	2.5	$0.4 \sim 0.8$	$0.5 \sim 0.9$	$0.7 \sim 0.9$	$0.8 \sim 1$
10	3.0	$0.4 \sim 0.8$	$0.5 \sim 0.9$	$0.7 \sim 0.9$	$0.8 \sim 1$
20	0.5	$0.4 \sim 0.8$	$0.5 \sim 0.9$	$0.7 \sim 1$	$0.8 \sim 1$
20	1.0	$0.5 \sim 0.8$	$0.6 \sim 0.9$	$0 \sim 0.1, 0.7 \sim 0.9$	$0 \sim 0.3, 0.8 \sim 1$
20	1.5	$0.5 \sim 0.8$	$0 \sim 0.1, 0.6 \sim 0.8$	$0 \sim 0.3, 0.7 \sim 0.9$	$0 \sim 0.4, 0.8 \sim 1$
20	2.0	$0.5 \sim 0.8$	$0 \sim 0.3, 0.5 \sim 0.8$	$0 \sim 0.9$	$0 \sim 0.5, 0.8 \sim 1$
20	2.5	$0 \sim 0.2, 0.4 \sim 0.8$	$0 \sim 0.9$	$0 \sim 0.9$	$0 \sim 1$
20	3.0	$0 \sim 0.9$	$0 \sim 1$	$0 \sim 1$	$0 \sim 1$
50	0.5	$0.5 \sim 0.7$	$0.6 \sim 0.8$	$0 \sim 0.3, 0.7 \sim 0.9$	$0 \sim 0.6, 0.9 \sim 1$
50	1.0	$0 \sim 1$	$0 \sim 1$	$0 \sim 0.9$	$0 \sim 0.9$
50	1.5	$0 \sim 0.9$	$0 \sim 0.9$	$0 \sim 1$	$0 \sim 0.9$
50	2.0	$0 \sim 0.9$	$0 \sim 0.9$	$0 \sim 0.9$	$0 \sim 1$
50	2.5	$0 \sim 0.9$	$0 \sim 0.9$	$0 \sim 0.9$	$0 \sim 1$
50	3.0	$0 \sim 1$	$0 \sim 1$	$0 \sim 1$	$0 \sim 1$

Appendix: Proofs

Result 3.1.

Proof. The likelihood functions $L(\theta|D_0)$ and $L(\theta|D)$ can be written as $L(\theta|D_0) \propto g(\theta|C(D_0))^{n_0}$ and $L(\theta|D) \propto g(\theta|C(D))^n$ respectively. If $C(D_0) = C(D)$, then

$$L(\theta|D_0) \propto g(\theta)^{n_0} \text{ and } L(\theta|D) \propto g(\theta)^n, \tag{3.18}$$

where $g(\theta) = g(\theta|C(D_0)) = g(\theta|C(D))$. To prove that the marginal posterior mode of δ is 1, it is sufficient to show that $\frac{\partial \pi(\delta|D_0,D)}{\partial \delta} \geq 0$ for any $\delta \in A$, where

$$\pi(\delta|D_0, D) \propto \pi(\delta) \frac{\int L(\theta|D) L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int L(\theta|D_0)^\delta \pi(\theta)\ d\theta} I_A(\delta). \tag{3.19}$$

Denote by $u(\delta) = \int g(\theta)^n g(\theta)^{\delta n_0} \pi(\theta)\ d\theta$ and $v(\delta) = \int g(\theta)^{\delta n_0} \pi(\theta)\ d\theta$. Then (3.19) may be written as

$$\pi(\delta|D_0, D) \propto \pi(\delta) \frac{u(\delta)}{v(\delta)} I_A(\delta).$$

Using $\pi(\delta) = 1$ and equation (3.18) to simplify $\pi(\delta|D_0, D)$, and furthermore exchanging differentiation and integration under stated regularity conditions, we have

$$\begin{aligned}
&\frac{\partial \pi(\delta|D_0, D)}{\partial \delta} \geq 0 \Longleftrightarrow u'(\delta)v(\delta) - u(\delta)v'(\delta) \geq 0 \Longleftrightarrow \\
&\int g(\theta)^{\delta n_0} g(\theta)^n \pi(\theta) \ln g(\theta)^{n_0} d\theta \int g(\theta)^{\delta n_0} \pi(\theta) d\theta \\
&- \int g(\theta)^{\delta n_0} \pi(\theta) \ln g(\theta)^{n_0} d\theta \int g(\theta)^{\delta n_0} g(\theta)^n \pi(\theta) d\theta \geq 0 \\
&\Longleftrightarrow \\
&\int g(\theta)^{\delta n_0} g(\theta)^n \pi(\theta) \ln g(\theta) d\theta \int g(\theta)^{\delta n_0} \pi(\theta) d\theta \\
&- \int g(\theta)^{\delta n_0} \pi(\theta) \ln g(\theta) d\theta \int g(\theta)^{\delta n_0} g(\theta)^n \pi(\theta) d\theta \geq 0,
\end{aligned} \tag{3.20}$$

for any $\delta \in A$.

Applying the property of the *Kullback-Leibler* function between two distributions, which is

$$K(f_1 : f_2) = \int f_1(x) \ln \frac{f_1(x)}{f_2(x)} dx \geq 0,$$

we obtain

$$\begin{aligned}
&\int \ln \frac{\pi(\theta|D_0, D, \delta)}{\pi(\theta|D_0, \delta)}[\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)]\ d\theta \\
&= \int \ln \frac{\pi(\theta|D_0, D, \delta)}{\pi(\theta|D_0, \delta)}\pi(\theta|D_0, D, \delta)\ d\theta + \int \ln \frac{\pi(\theta|D_0, \delta)}{\pi(\theta|D_0, D, \delta)}\pi(\theta|D_0, \delta)\ d\theta \\
&\geq 0. \qquad (3.21)
\end{aligned}$$

This further leads to

$$\begin{aligned}
&n \int \ln g(\theta) \left[\frac{g(\theta)^{\delta n_0+n}\pi(\theta)}{\int g(\theta)^{\delta n_0+n}\pi(\theta)d\theta} - \frac{g(\theta)^{\delta n_0}\pi(\theta)}{\int g(\theta)^{\delta n_0}\pi(\theta)d\theta}\right] d\theta \\
&= \int \ln g(\theta)^n\ [\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)]d\theta \qquad \text{by equation (3.18)} \\
&= \int \ln \left[\frac{\pi(\theta|D_0, D, \delta)}{\pi(\theta|D_0, \delta)}M(\delta)\right][\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)]\ d\theta \\
&= \int \ln \frac{\pi(\theta|D_0, D, \delta)}{\pi(\theta|D_0, \delta)}[\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)]\ d\theta \\
&+ \int \ln M(\delta)\ [\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)]\ d\theta,
\end{aligned}$$

where $M(\delta) = \int g(\theta)^{\delta n_0+n}\pi(\theta)d\theta / \int g(\theta)^{\delta n_0}\pi(\theta)d\theta$. Notice that

$$\begin{aligned}
&\int \ln M(\delta)\ [\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)]\ d\theta \\
&= \ln M(\delta) \left[\int \pi(\theta|D_0, D, \delta)d\theta - \int \pi(\theta|D_0, \delta)d\theta\right] = \ln M(\delta)\ (1-1)\ = 0,
\end{aligned}$$

combining it with inequality (3.21), we get

$$n \int \ln g(\theta) \left[\frac{g(\theta)^{\delta n_0+n}\pi(\theta)}{\int g(\theta)^{\delta n_0+n}\pi(\theta)d\theta} - \frac{g(\theta)^{\delta n_0}\pi(\theta)}{\int g(\theta)^{\delta n_0}\pi(\theta)d\theta}\right] d\theta \geq 0. \qquad (3.22)$$

Finally, by multiplying both sides of inequality (3.22) by $\frac{1}{n}\int g(\theta)^{\delta n_0+n}\pi(\theta)d\theta \int g(\theta)^{\delta n_0}\pi(\theta)d\theta$, it follows that the sufficient condition in (3.20) for the marginal posterior mode of δ being 1 is met for any $\delta \in A$. □

Result 3.2.

Proof. Suppose that k is an arbitrary positive constant. We take the likelihood function of the form $L(\theta|x) = kf(x|\theta)$, then $L(\theta|D) = k^n f(D|\theta)$ and $L(\theta|D_0) = k^{n_0} f(D_0|\theta)$. Then the marginal posterior distribution of δ can be rewritten as

$$\pi(\delta|D_0, D) \propto \pi(\delta) \int L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta \propto \pi(\delta) \int f(D|\theta)[k^{n_0} f(D_0|\theta)]^\delta \pi(\theta)\ d\theta.$$

To prove that the marginal posterior mode of δ is 0, it is sufficient to show that $\frac{\partial \pi(\delta|D_0,D)}{\partial \delta} \leq 0$ for any $\delta \in [0,1]$.

Using $\pi(\delta) = 1$ and the exchange of differentiation and integration, we obtain

$$\begin{aligned}
&\frac{\partial \pi(\delta|D_0,D)}{\partial \delta} \leq 0 \Longleftrightarrow \int f(D|\theta)\frac{\partial [k^{n_0} f(D_0|\theta)]^\delta}{\partial \delta}\pi(\theta)\ d\theta \leq 0 \\
&\Longleftrightarrow k^{n_0\delta}\int \pi(\theta) f(D|\theta) f(D_0|\theta)^\delta (n_0 \ln k + \ln f(D_0|\theta)) d\theta \leq 0 \\
&\Longleftrightarrow \frac{\int \pi(\theta) f(D|\theta) f(D_0|\theta)^\delta \ln f(D_0|\theta) d\theta}{\int \pi(\theta) f(D|\theta) f(D_0|\theta)^\delta d\theta} \leq n_0 \ln \frac{1}{k}. \qquad (3.23)
\end{aligned}$$

Since $f(D_0|\theta) \geq 0$, then $\ln f(D_0|\theta) \leq f(D_0|\theta)$. Furthermore, it is reasonable to assume that the conditional posterior distribution of θ on δ is proper for any $\delta \in [0,1]$, where $\pi(\theta|D_0, D, \delta) \propto \int \pi(\theta) f(D|\theta) f(D_0|\theta)^\delta d\theta$. So $0 < \int \pi(\theta) f(D|\theta) f(D_0|\theta)^\delta d\theta < \infty$ for any $\delta \in [0,1]$. Suppose that a new data set D_1 is obtained by pooling D and D_0 together. If we use D_1 as the current data and D_0 as the historical data, the conditional posterior $\pi(\theta|D_0, D_1, \delta) \propto \int \pi(\theta) f(D|\theta) f(D_0|\theta) f(D_0|\theta)^\delta d\theta$. Therefore, we have

$$\int \pi(\theta) f(D|\theta) f(D_0|\theta)^\delta \ln f(D_0|\theta) d\theta \leq \int \pi(\theta) f(D|\theta) f(D_0|\theta) f(D_0|\theta)^\delta d\theta < \infty,$$

for any $\delta \in [0,1]$.

If we take

$$k_0 = \exp\left(- \frac{\max_\delta [\frac{\int \pi(\theta) f(D|\theta) f(D_0|\theta)^\delta \ln f(D_0|\theta) d\theta}{\int \pi(\theta) f(D|\theta) f(D_0|\theta)^\delta d\theta}]}{n_0} \right),$$

then the sufficient condition in (3.23) for the marginal posterior mode of δ being 0 is met for any δ. □

Chapter 4

Modified Power Priors with Multiple Historical Data Sets

4.1 Introduction

In this chapter, we further investigate how to incorporate multiple historical data sets under the framework of modified power priors. This issue is tackled from several aspects. First, we propose three power prior methods for incorporating multiple historical data sets. Those three methods yield the same conditional posterior $\pi(\theta|\delta, \underline{D}_0, D)$, but different $\pi(\underline{\delta}|\underline{D}_0, D)$. One method stands out in the comparison of MSE among the three.

Second, the modified power prior approach is compared with the random effects model for the accommodation of potential heterogeneity between current and historical samples. Furthermore, we combine two methods together and apply the modified power prior on top of a random effects model.

At the end, we relax the assumption that multiple power parameters are independent, and discuss the implementation of time weighted power priors.

4.2 Three Methods in Incorporating Multiple Historical Data Sets

The modified power priors defined in (3.6) can easily be generalized to multiple historical data sets. Suppose there are k historical studies. We define D_{0j} to be the historical data based on the jth study, $j = 1, ..., k$ and $\underline{D}_0 = (D_{01}, ..., D_{0k})$. Chen *et al.* ([10]) suggested defining a different weight parameter δ_j for each historical study and taking the δ_j's to be i.i.d. *Beta* random variables with hyperparameters (α, β). Let $\underline{\delta} = (\delta_1, ..., \delta_k)$. The modified power prior in (3.6) can be generalized as

$$\pi(\theta, \underline{\delta}|\underline{D}_0) \propto \frac{\left(\prod_{j=1}^{k} L(\theta|D_{0j})^{\delta_j}\pi(\delta_j|\alpha,\beta)\right)\pi(\theta)}{\int\left(\prod_{j=1}^{k} L(\theta|D_{0j})^{\delta_j}\right)\pi(\theta)\ d\theta} I_B(\underline{\delta}), \tag{4.1}$$

where $B = \{(\delta_1, ..., \delta_k) : 0 < \int\left(\prod_{j=1}^{k} L(\theta|D_{0j})^{\delta_j}\right)\pi(\theta)\ d\theta\ < \infty\}$.

Incorporating the current data D, it follows that the posterior distribution of θ conditional on $\underline{\delta}$ is

$$\pi(\theta|\underline{\delta}, \underline{D}_0, D) \propto \left(\prod_{j=1}^{k} L(\theta|D_{0j})^{\delta_j}\pi(\delta_j|\alpha,\beta)\right)L(D)\pi(\theta), \tag{4.2}$$

and the marginal posterior for $\underline{\delta}$ is

$$\pi(\underline{\delta}|\underline{D}_0, D) \propto \frac{\int\left(\prod_{j=1}^{k} L(\theta|D_{0j})^{\delta_j}\pi(\delta_j|\alpha,\beta)\right)L(D)\pi(\theta)\ d\theta}{\int\left(\prod_{j=1}^{k} L(\theta|D_{0j})^{\delta_j}\right)\pi(\theta)\ d\theta} I_B(\underline{\delta}). \tag{4.3}$$

We refer to the method defined above for incorporating multiple historical data sets as "method A". Note that in method A each power parameter δ_j is determined not only by the relative availability of D_{0j} and its discrepancy from D, but also by the discrepancy among D and all $\underline{D}_0$. Equation (4.3) implies that all δs interact with each other. There is not much debate on the conditional posterior $\pi(\theta|\underline{\delta}, \underline{D}_0, D)$ defined in (4.2), however other possibilities may be considered to define the behavior of $\underline{\delta}$ for the case when multiple historical data sets are available. For example, consider a method C, which has the same $\pi(\theta|\underline{\delta}, \underline{D}_0, D)$ as method A but leads to a different $\pi(\underline{\delta}|\underline{D}_0, D)$.

$$\pi(\underline{\delta}|\underline{D}_0, D) \propto \prod_{j=1}^{k} \frac{\int L(\theta|D_{0j})^{\delta_j}\pi(\delta_j|\alpha,\beta)L(D)\pi(\theta)\ d\theta}{\int L(\theta|D_{0j})^{\delta_j}\pi(\theta)\ d\theta} I_{A_j}(\delta_j), \tag{4.4}$$

where where $A_j = \{\delta_j : 0 < \int L(\theta|D_{0j})^{\delta_j}\pi(\delta_j|\alpha,\beta)\pi(\theta)\ d\theta\ < \infty\}$ for $j = 1, ..., k$.

Method C says that each power parameter δ_j is controlled only by D_{0j} and D. So historical data sets do not interact with each other; the role of each D_{0j} in the current study is determined independently.

Similarly, we may define an intermediate method between methods A and C, referred to as method B, to control the influence of multiple historical data sets. In method B, the marginal posterior of $\underline{\delta}$ is defined as

$$\pi(\underline{\delta}|\underline{D}_0, D) \propto \frac{\int \left(\prod_{j=1}^{k} L(\theta|D_{0j})^{\delta_j}\pi(\delta_j|\alpha,\beta)\right) L(D)\pi(\theta)\ d\theta}{\prod_{j=1}^{k}\int L(\theta|D_{0j})^{\delta_j}\pi(\theta)\ d\theta} \prod_{j=1}^{k} I_{A_j}(\delta_j). \tag{4.5}$$

$\pi(\theta|\underline{\delta}, \underline{D}_0, D)$ is the same in methods A, B, and C, but $\underline{\delta}$ behave differently. The question is under which method the random power parameters $\underline{\delta}$ function well. To investigate this issue, simulations are performed to compare the three methods in terms of MSE for estimate of θ.

Normal populations are considered in the simulations. More specifically, we would like to make inference on the normal mean with unknown variance, by incorporating both current and historical data. Suppose that current data $D = (x_1, ..., x_n)$ come from a normal population with unknown mean μ and variance σ^2, and

$$D_{01} = (x_{011}, ..., x_{01m_1}) \quad \text{and} \quad D_{02} = (x_{021}, ..., x_{02m_2})$$

are two historical data sets. Let

$$\bar{x}_{01} = \frac{1}{m_1}\sum_{i=1}^{m_1} x_{01i}, \qquad \bar{x}_{02} = \frac{1}{m_2}\sum_{i=1}^{m_2} x_{02i}, \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\hat{\sigma}^2_{01} = \frac{1}{m_1}\sum_{i=1}^{m_1}(x_{01i} - \bar{x}_{01})^2, \quad \hat{\sigma}^2_{02} = \frac{1}{m_2}\sum_{i=1}^{m_2}(x_{02i} - \bar{x}_{02})^2, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

We use the Jeffreys prior ([20]) as the initial prior for (μ, σ^2), i.e. $\pi(\mu, \sigma^2) \propto (\sigma^2)^{-1.5}$. Furthermore a uniform distribution on $(0, 1)$ is used as $\pi(\delta_j)$. Following the equation (4.2),

the posterior distributions for (μ, σ^2) are derived as follows.

$$\mu|\sigma^2, \underline{\delta}, \underline{D}_0, D \sim \text{Normal}\left(\frac{\delta_1 m_1 \bar{x}_{01} + \delta_2 m_2 \bar{x}_{02} + n\bar{x}}{\delta_1 m_1 + \delta_2 m_2 + n}, \frac{\sigma^2}{\delta_1 m_1 + \delta_2 m_2 + n}\right) \tag{4.6}$$

$$\sigma^2|\underline{\delta}, \underline{D}_0, D \sim \text{Inverse-Gamma}\left(\frac{\delta_1 m_1 + \delta_2 m_2 + n}{2}, \beta_1\right),$$

$$\text{where } \beta_1 = \frac{2}{n\hat{\sigma}^2 + \delta_1 m_1 \hat{\sigma}_{01}^2 + \delta_2 m_2 \hat{\sigma}_{02}^2 + \frac{n\delta_1 m_1 (\bar{x}-\bar{x}_{01})^2 + n\delta_2 m_2 (\bar{x}-\bar{x}_{02})^2 + \delta_1 m_1 \delta_2 m_2 (\bar{x}_{01}-\bar{x}_{02})^2}{\delta_1 m_1 + \delta_2 m_2 + n}}. \tag{4.7}$$

Following equations (4.3), (4.5), and (4.4), the marginal posterior of $\underline{\delta}$ is

$$\pi(\delta_1, \delta_2|\underline{D}_0, D) \propto \frac{\Gamma(\frac{\delta_1 m_1 + \delta_2 m_2 + n}{2})(\delta_1 m_1 + \delta_2 m_2 + n)^{-0.5} \beta_1^{\frac{\delta_1 m_1 + \delta_2 m_2 + n}{2}}}{\Gamma(\frac{\delta_1 m_1 + \delta_2 m_2}{2})(\delta_1 m_1 + \delta_2 m_2)^{-0.5} \beta_2^{\frac{\delta_1 m_1 + \delta_2 m_2}{2}}} \text{ in method A,}$$

$$\text{where } \beta_2 = \frac{2}{\delta_1 m_1 \hat{\sigma}_{01}^2 + \delta_2 m_2 \hat{\sigma}_{02}^2 + \frac{\delta_1 m_1 \delta_2 m_2 (\bar{x}_{01}-\bar{x}_{02})^2}{\delta_1 m_1 + \delta_2 m_2}}; \tag{4.8}$$

$$\pi(\delta_1, \delta_2|\underline{D}_0, D) \propto \frac{\Gamma(\frac{\delta_1 m_1 + \delta_2 m_2 + n}{2})(\delta_1 m_1 + \delta_2 m_2 + n)^{-0.5} \beta_1^{\frac{\delta_1 m_1 + \delta_2 m_2 + n}{2}}}{\Gamma(\frac{\delta_1 m_1}{2})(\delta_1 m_1)^{-0.5} \left(\frac{2}{\delta_1 m_1 \hat{\sigma}_{01}^2}\right)^{\frac{\delta_1 m_1}{2}} \Gamma(\frac{\delta_2 m_2}{2})(\delta_2 m_2)^{-0.5} \left(\frac{2}{\delta_2 m_2 \hat{\sigma}_{02}^2}\right)^{\frac{\delta_2 m_2}{2}}}$$

in method B;

$$\pi(\delta_1, \delta_2|\underline{D}_0, D) = \pi(\delta_1|\underline{D}_0, D)\pi(\delta_2|\underline{D}_0, D) \text{ in method C,}$$

$$\text{where } \pi(\delta_1|\underline{D}_0, D) \propto \frac{\Gamma(\frac{\delta_1 m_1 + n}{2})(\delta_1 m_1 + n)^{-0.5} \left(\frac{2}{\delta_1 m_1 \hat{\sigma}_{01}^2 + n\hat{\sigma}^2 + \frac{\delta_1 m_1 n (\bar{x}_{01}-\bar{x})^2}{\delta_1 m_1 + n}}\right)^{\frac{\delta_1 m_1 + n}{2}}}{\Gamma(\frac{\delta_1 m_1}{2})(\delta_1 m_1)^{-0.5} \left(\frac{2}{\delta_1 m_1 \hat{\sigma}_{01}^2}\right)^{\frac{\delta_1 m_1}{2}}},$$

$$\text{and } \pi(\delta_2|\underline{D}_0, D) \propto \frac{\Gamma(\frac{\delta_2 m_2 + n}{2})(\delta_2 m_2 + n)^{-0.5} \left(\frac{2}{\delta_2 m_2 \hat{\sigma}_{02}^2 + n\hat{\sigma}^2 + \frac{\delta_2 m_2 n (\bar{x}_{02}-\bar{x})^2}{\delta_2 m_2 + n}}\right)^{\frac{\delta_2 m_2 + n}{2}}}{\Gamma(\frac{\delta_2 m_2}{2})(\delta_2 m_2)^{-0.5} \left(\frac{2}{\delta_2 m_2 \hat{\sigma}_{02}^2}\right)^{\frac{\delta_2 m_2}{2}}}.$$

In this simulation study, suppose that the current sample is from a normal $N(\mu, \sigma^2)$ population and each historical sample D_{0j} is from a normal $N(\mu_{0j}, \sigma_{0j}^2)$ population, for $j = 1, 2$, with both mean and variance unknown. Furthermore, suppose the population mean of current sample, μ, is the parameter of interest. In a Bayesian analysis, the marginal posterior mean of μ is usually used as $\hat{\mu}$, the estimate of μ. Then the MSE of $\hat{\mu}$ is $E(\hat{\mu} - \mu)^2$.

We consider eight combinations of μ_{01}, μ_{02}, m_1, and m_2 to cover different sample sizes and different degree of divergence between current and historical populations. $n = 20, \mu =$

$0, \sigma = 1$, and $\sigma_{01} = \sigma_{02} = 1.5$ are used for all simulation scenarios. For each simulation scenario, three thousand sets of current and historical samples are generated and then the averaged squared error of $\hat{\mu}$ is calculated as an estimate of MSE. The simulation results obtained using *MATLAB* 6.1 are presented in Table 4.1.

Table 4.1: *Comparison of the MSE of $\hat{\mu}$ in a normal population using three methods for incorporating multiple historical data sets. $n = 20, \mu = 0, \sigma = 1$, and $\sigma_{01} = \sigma_{02} = 1.5$.*

Setting		Method A	Method B	Method C	Use no historical data
$\mu_{01} = 0.1, \mu_{02} = 0.2$					
m_1	m_2				
20	30	0.0302	0.0312	0.0314	0.05
40	50	0.0275	0.0288	0.0344	0.05
60	70	0.0250	0.0267	0.1503	0.05
80	90	0.0239	0.0256	1.0312	0.05
$\mu_{01} = 0.2, \mu_{02} = 0.3$					
m_1	m_2				
20	30	0.0396	0.0444	0.0413	0.05
40	50	0.0425	0.0471	0.0516	0.05
60	70	0.0441	0.0492	0.3182	0.05
80	90	0.0458	0.0508	1.2562	0.05

As shown in Table 4.1, method A leads to consistently smaller MSE compared to methods B and C, and its MSE is also smaller than the MSE of $\hat{\mu}$ estimated without using historical data. Hence, method A improves the MSE of $\hat{\mu}$ by incorporating historical information, and furthermore it gives the greatest improvement amongst the methods. Therefore method A will be adopted in the rest of this dissertation to incorporate multiple historical data sets.

Heterogeneity often exists among different studies but data collected in one study are relatively homogeneous. The framework defined in (4.1) would accommodate potential heterogeneity among data sets from different sources or collected at different times. For example, in water quality assessment, we could take data observed at neighboring sites as different "historical" data sets. Moreover, data collected over a long period may be divided into several historical data sets to ensure the homogeneity within each data set. In such a way, the role of historical data can be more accurately evaluated (Duan, Ye and Smith, [14]). In

Chapter 5, we will discuss an example of implementing the modified power prior approach using multiple sites information.

4.3 Comparing the Modified Power Prior Approach with the Random Effects Model

The random effects model is often used to accommodate potential differences in means among several populations. So it is of interest to compare the modified power prior approach defined in (4.1) with the random effects model in incorporating historical data sets. A simulation study is performed to compare their performances in terms of MSE and coverage probability of 95% confidence region.

Again we are interested in a normal population, and focus on the estimation of the normal mean with unknown variance, by incorporating both current and historical data. Suppose that the current data $D = (x_1, ..., x_n)$ come from a normal population with unknown mean μ and variance σ^2, and $D_{01} = (x_{011}, ..., x_{01m_1})$ and $D_{02} = (x_{021}, ..., x_{02m_2})$ are two historical data sets. Let $\bar{x}_{01}$, $\bar{x}_{02}$, and $\bar{x}$ denote the sample means as defined in Section 4.2. In the power prior approach, a Jeffreys prior is used as the initial prior for (μ, σ^2), i.e. $\pi(\mu, \sigma^2) \propto (\sigma^2)^{-1.5}$, and a uniform distribution on $(0, 1)$ is used as $\pi(\delta_j)$. Then the posterior distributions using a modified power prior are given in (4.6), (4.7), and (4.8).

Alternatively, we use a one-way random effects model to incorporate both current and historical data, with D, D_{01}, and D_{02} being treated as three groups. The model is expressed as

$$y_{ij} = \mu_0 + \alpha_i + e_{ij}, \text{ with } i = 1, 2, 3;$$
$$\alpha_i \sim \text{i.i.d. Normal}(0, \sigma_\alpha^2), e_{ij} \sim \text{i.i.d. Normal}(0, \sigma_e^2);$$

where $\underline{y}_1$ is the current sample $\underline{x}$ and $\underline{y}_2, \underline{y}_3$ are the historical samples, i.e. $\underline{y}_2 = \underline{x}_{01}$ and $\underline{y}_3 = \underline{x}_{02}$. Then the following results can be obtained (see Searle *et al.*, [27]).

The Best linear unbiased estimator (BLUE) of μ_0 is $\hat{\mu}_0 = \dfrac{\frac{n\bar{x}}{\hat{\sigma}_e^2+n\hat{\sigma}_\alpha^2} + \frac{m_1\bar{x}_{01}}{\hat{\sigma}_e^2+m_1\hat{\sigma}_\alpha^2} + \frac{m_2\bar{x}_{02}}{\hat{\sigma}_e^2+m_2\hat{\sigma}_\alpha^2}}{\frac{n}{\hat{\sigma}_e^2+n\hat{\sigma}_\alpha^2} + \frac{m_1}{\hat{\sigma}_e^2+m_1\hat{\sigma}_\alpha^2} + \frac{m_2}{\hat{\sigma}_e^2+m_2\hat{\sigma}_\alpha^2}}$,

with $Var(\hat{\mu}_0) = \dfrac{1}{\frac{n}{\hat{\sigma}_e^2+n\hat{\sigma}_\alpha^2} + \frac{m_1}{\hat{\sigma}_e^2+m_1\hat{\sigma}_\alpha^2} + \frac{m_2}{\hat{\sigma}_e^2+m_2\hat{\sigma}_\alpha^2}}$;

The Best linear unbiased predictor (BLUP) of $(\mu_0 + \alpha_1) = \hat{\mu}_0 + \dfrac{n\hat{\sigma}_\alpha^2}{\hat{\sigma}_e^2 + n\hat{\sigma}_\alpha^2}(\bar{x} - \hat{\mu}_0)$,

where $\hat{\sigma}_e^2$ and $\hat{\sigma}_\alpha^2$ are ANOVA estimators for σ_e^2 and σ_α^2.

In the simulations, suppose the current data sample is from a normal population $N(\mu, \sigma^2)$ and each historical sample $D_{0j}, j = 1, 2$ is from a $N(\mu_{0j}, \sigma_{0j}^2)$ population, with both mean and variance unknown. Furthermore, suppose the population mean of current sample, μ, is the parameter of interest. In the power prior analysis, the marginal posterior mean of μ is used as $\hat{\mu}$, the estimate of μ; in the analysis of random effects model, both the BLUE of μ_0 and BLUP of $(\mu_0 + \alpha_1)$ are considered for estimating the population mean of the current data.

We consider various sample sizes, means and variances for historical populations, but use the same population to generate current data, i.e. $n = 20, \mu = 0, \sigma = 1$ for all simulation scenarios. For each simulation scenario, three thousand sets of current and historical samples are generated and then $\hat{MSE} = \frac{1}{3000}\sum_{k=1}^{3000}(\hat{\mu} - \mu)^2$ is calculated. *MATLAB* 6.1 and *SAS* 8.2 are used to run the simulations, whose results are presented in Tables 4.2, 4.3, 4.4 and 4.5.

For most simulation scenarios described in Tables 4.2 and 4.3, the modified power prior method yields smaller MSE for estimation of μ than the random effects model regardless of whether BLUE or BLUP is used to estimate μ. Exceptions are indicated with italicized entries in the table. For example, when the historical and current populations have homogenous variability ($\sigma = \sigma_{01} = \sigma_{02} = 1$), the random effects model gives smaller MSE than the power prior method if the difference in mean between current and historical populations is only 0.1 and historical sample sizes are no larger than 40. In such situations, current and historical populations are almost homogenous in terms of both variance and mean, and hence the random effects model works like a normal means model. Since the simulation model is very close to a normal means model, the random effects model gives better performance than the power prior method. When the divergence in mean between current and historical populations is large, i.e. $\mu_{01} - \mu = \mu_{01} - \mu = 0.4$, we have shown in Chapter 3 that the modified power prior does not perform well. So it is no surprise that the modified power prior method leads to larger MSE than the random effects model when $m_1 = m_2 \geq 40$ with $\sigma_{01} = \sigma_{02} = 1$ and when $m_1 = m_2 \geq 80$ with $\sigma_{01} = \sigma_{02} = 1.5$.

When the current and historical populations have different variances (see Table 4.3), i.e. $\sigma_{01} = \sigma_{02} = 1.5$ but $\sigma = 1$, the advantage of the power prior method over the random effects model in MSE grows. The MSE using the power prior method is smaller than that using the random effects model even when $\mu_{01} = \mu_{02} = \mu$ (see Table 4.4). This is because a

Table 4.2: *Comparison of the MSE of $\hat{\mu}$ in a normal population using the modified power prior method and random effects model. $n = 20, \mu = 0, \sigma = 1$, and $\sigma_{01} = \sigma_{02} = 1$.*

Setting		Modified Power Prior	Random Effects Model		Means Model
			BLUE(μ_0)	BLUP($\mu_0 + \alpha_1$)	use only D
$\mu_{01} = \mu_{02} = 0.1$					
m_1	m_2				
20	20	*0.02353*	*0.02229*	0.02749	0.05
40	40	*0.01669*	*0.01659*	0.02193	0.05
60	60	0.01400	0.01457	0.01750	0.05
80	80	0.01193	0.01273	0.01649	0.05
100	100	0.01152	0.01226	0.01531	0.05
$\mu_{01} = \mu_{02} = 0.2$					
m_1	m_2				
20	20	0.03037	0.03625	0.03633	0.05
40	40	0.02804	0.03405	0.03305	0.05
60	60	0.02880	0.03438	0.03335	0.05
80	80	0.02833	0.03335	0.03225	0.05
100	100	0.02987	0.03396	0.03230	0.05
$\mu_{01} = \mu_{02} = 0.3$					
m_1	m_2				
20	20	0.04109	0.05910	0.04827	0.05
40	40	0.04524	0.06131	0.04989	0.05
60	60	0.04963	0.06255	0.05223	0.05
80	80	0.05486	0.06560	0.05252	0.05
100	100	0.05976	0.06850	0.05523	0.05
$\mu_{01} = \mu_{02} = 0.4$					
m_1	m_2				
20	20	0.05509	0.09084	0.05960	0.05
40	40	*0.07029*	0.09857	*0.06406*	0.05
60	60	*0.08363*	0.10559	*0.06793*	0.05
80	80	*0.09080*	0.10749	*0.07413*	0.05
100	100	*0.09950*	0.11157	*0.07708*	0.05

random effects model with the assumption of homogenous variance is used here and therefore it can only take into account the divergence in means among several populations. The power

Table 4.3: *Comparison of the MSE of $\hat{\mu}$ in a normal population using the modified power prior method and random effects model. $n = 20, \mu = 0, \sigma = 1$, and $\sigma_{01} = \sigma_{02} = 1.5$.*

Setting		Modified Power Prior	Random Effects Model		Means Model
			BLUE(μ_0)	BLUP($\mu_0 + \alpha_1$)	use only D
$\mu_{01} = \mu_{02} = 0.1$					
m_1	m_2				
20	20	0.02898	0.03543	0.03344	0.05
40	40	0.02241	0.02750	0.02582	0.05
60	60	0.01776	0.02157	0.02031	0.05
80	80	0.01717	0.02042	0.01879	0.05
100	100	0.01581	0.01835	0.01712	0.05
$\mu_{01} = \mu_{02} = 0.2$					
m_1	m_2				
20	20	0.03646	0.05082	0.04031	0.05
40	40	0.03262	0.04509	0.03700	0.05
60	60	0.03296	0.04393	0.03702	0.05
80	80	0.03163	0.04142	0.03699	0.05
100	100	0.03277	0.04118	0.03582	0.05
$\mu_{01} = \mu_{02} = 0.3$					
m_1	m_2				
20	20	0.04236	0.07047	0.05194	0.05
40	40	0.04830	0.07376	0.05528	0.05
60	60	0.05167	0.07404	0.06012	0.05
80	80	0.05752	0.07749	0.06112	0.05
100	100	0.06194	0.07967	0.06529	0.05
$\mu_{01} = \mu_{02} = 0.4$					
m_1	m_2				
20	20	0.05479	0.10249	0.06411	0.05
40	40	0.06919	0.11135	0.07377	0.05
60	60	0.08348	0.11948	0.08388	0.05
80	80	*0.09566*	0.12702	*0.09136*	0.05
100	100	*0.10136*	0.12747	*0.09605*	0.05

prior method is initiated to address the potential heterogeneity among several populations in general, including differences in both means and variances.

Table 4.4 presents the situations where means are the same but variances are different between current and historical populations. A special case is that $\sigma_{01} = \sigma_{02} = \sigma = 1$, which is a reduced random effects model with $\sigma_\alpha^2 = 0$ and $\sigma_e^2 = 1$. Since the true simulation model is a special case of the random effects model, this situation favors the random effects model over the power prior method. When $\sigma_{01} = \sigma_{02} = 0.5$, the variability in data is dragged down by the relatively low variance of historical samples because the random effects model does not discount the role of historical data. As a consequence, the variability of BLUE(μ_0) or BLUP($\mu_0 + \alpha_1$) appears to be lower than it should be. On the other hand, the power prior method controls the influence of historical data and hence the decrease in variance due to the its discrepancy from current data. Therefore the random effects model yields smaller MSE than the power prior method. We speculate that the estimation on σ using the power prior method would be more accurate than that from the random effects model. When $\sigma_{01} = \sigma_{02} = 1.5$ or when $\sigma_{01} = \sigma_{02} = 2$ and $m_1 = m_2 \leq 60$, the power prior method performs better in terms of MSE than the random effects model.

Table 4.5 compares the two methods in terms of the coverage probability of 95% confidence regions for $\hat{\mu}$. For the modified power prior method, we calculate a 95% credible region of the posterior distribution for μ in each iteration. For the random effects model, the 95% confidence interval is used for BLUE(μ_0) and the 95% prediction interval is used for BLUP($\mu_0 + \alpha_1$). Only the balanced case ($n = m_1 = m_1 = 20$) is investigated in our simulations. In Table 4.5, an obvious trend is that the modified power prior method yields consistently larger coverage probability compared to the random effects model.

Criticism may arise on the fairness of using $D \sim N(\mu, \sigma^2)$ and $D_{0j} \sim N(\mu_{0j}, \sigma_{0j}^2)$ to simulate current and historical samples. However, this setup mimics the situations encountered in practice. It is fair in a sense that it is neither the true model for the power prior method nor a random effects model. Our interest is to find an appropriate method to incorporate historical data, whose underlying population may be different from the current population. The simulation scenarios used precisely address our research interest.

4.4 Power Priors on the Random Effects Model

In the previous sections, power priors are used on the normal mean model. As a general method, the modified power prior approach can be applied to various models, including the random effects model. In this section, we introduce the implementation of the modified

Table 4.4: *Comparison of the MSE of $\hat{\mu}$ in a normal population using the modified power prior method and random effects model. $n = 20, \mu = 0, \sigma = 1$, and $\mu_{01} = \mu_{02} = 0$.*

Setting		Modified Power Prior	Random Effects Model		Means Model
			BLUE(μ_0)	BLUP($\mu_0 + \alpha_1$)	use only D
$\sigma_{01} = \sigma_{02} = 0.5$					
m_1	m_2				
20	20	0.02504	0.00865	0.02326	0.05
40	40	0.02274	0.00550	0.02278	0.05
60	60	0.02263	0.00426	0.02065	0.05
80	80	0.02324	0.00342	0.01772	0.05
100	100	0.02744	0.00320	0.01818	0.05
$\sigma_{01} = \sigma_{02} = 1$					
m_1	m_2				
20	20	0.02094	0.01722	0.02427	0.05
40	40	0.01283	0.01066	0.01810	0.05
60	60	0.00887	0.00765	0.01354	0.05
80	80	0.00669	0.00596	0.01056	0.05
100	100	0.00554	0.00502	0.00927	0.05
$\sigma_{01} = \sigma_{02} = 1.5$					
m_1	m_2				
20	20	0.02835	0.03132	0.03088	0.05
40	40	0.01921	0.02098	0.02204	0.05
60	60	0.01405	0.01509	0.01570	0.05
80	80	0.01106	0.01189	0.01262	0.05
100	100	0.00926	0.00954	0.01041	0.05
$\sigma_{01} = \sigma_{02} = 2$					
m_1	m_2				
20	20	0.03642	0.05096	0.04207	0.05
40	40	0.02770	0.03646	0.03204	0.05
60	60	0.02259	0.02784	0.02366	0.05
80	80	0.02056	0.02138	0.01976	0.05
100	100	0.01665	0.01768	0.01574	0.05

power priors on a random effects model and demonstrate its use.

Consider the following one-way random effects model to accommodate current and his-

Table 4.5: *Comparison of the coverage probability of* 95% *confidence regions for* $\hat{\mu}$ *in a normal population using the modified power prior method and random effects model.* $n = 20, \mu = 0, \sigma = 1,$ *and* $m_1 = m_2 = 20$.

Setting		Modified Power Prior	Random Effects Model	
			BLUE(μ_0)	BLUP($\mu_0 + \alpha_1$)
$\sigma_{01} = \sigma_{02} = 1$				
μ_1	μ_2			
0	0	0.97233	0.94200	0.9546
0.1	0.1	0.96000	0.93733	0.9452
0.2	0.2	0.93733	0.91500	0.8994
0.3	0.3	0.90733	0.89967	0.8530
0.4	0.4	0.86133	0.89967	0.8160
$\sigma_{01} = \sigma_{02} = 1.5$				
μ_1	μ_2			
0	0	0.98500	0.94600	0.9718
0.1	0.1	0.98433	0.94667	0.9648
0.2	0.2	0.97133	0.94500	0.9508
0.3	0.3	0.96267	0.93667	0.9284
0.4	0.4	0.93836	0.93770	0.8998

torical data.

$$y_{ij} = \mu_0 + \alpha_i + e_{ij}, \text{ with } i = 1, ..., k+1 \text{ and } j = 1, ..., n_i;$$
$$\alpha_i \sim \text{i.i.d. Normal}(0, \sigma_\alpha^2), e_{ij} \sim \text{i.i.d. Normal}(0, \sigma_e^2);$$

where $\underline{y}_1$ is the current data and $\underline{y}_2, ..., \underline{y}_{k+1}$ are k historical data sets. Let $\underline{\delta} = (\delta_1, ..., \delta_k)$ and $\phi = \sigma_\alpha^2/\sigma_e^2$. The modified power prior on a random effects model is defined as

$$\pi(\mu, \sigma_e^2, \phi, \underline{\delta}|\underline{y}_2, ..., \underline{y}_{k+1}) \propto \frac{\left(\prod_{i=2}^{k+1} \int_{-\infty}^{\infty} f(\underline{y}_i|\underline{\alpha}_i)^{\delta_i} f(\underline{\alpha}_i) d\underline{\alpha}_i\right) \pi(\mu, \sigma_e^2, \phi)\pi(\underline{\delta})}{\int \left(\prod_{i=2}^{k+1} \int_{-\infty}^{\infty} f(\underline{y}_i|\underline{\alpha}_i)^{\delta_i} f(\underline{\alpha}_i) d\underline{\alpha}_i\right) \pi(\mu, \sigma_e^2, \phi) d\mu d\sigma_e^2 d\phi} I_B(\underline{\delta}),$$

where $B = \{(\delta_1, ..., \delta_k) : 0 < \int \left(\prod_{i=2}^{k+1} \int_{-\infty}^{\infty} f(\underline{y}_i|\underline{\alpha}_i)^{\delta_i} f(\underline{\alpha}_i) d\underline{\alpha}_i\right) \pi(\mu, \sigma_e^2, \phi) d\mu d\sigma_e^2 d\phi < \infty\}$.

For illustration, we consider a data set describing yields (in bushels per acre) of three varieties of corn (presented in Table 4.6). Three methods are implemented in *Winbugs 4* to

make inference on μ_i (the population mean of each variety), and their results are compared in Table 4.7. For the random effects model with a power prior or with a Jeffreys prior, the BLUP is used to predict each population mean. When the normal means model with a power prior is applied, we use the posterior mean of μ_i as the $\hat{\mu}_i$ for each variety.

Table 4.6: *Yields (in bushels per acre) of three varieties of corn.*

Varieties of Corn		
Four country	Lodent	Lancaster
7.3	6.9	9.6
4.5	6.8	7.8
7.4	7.6	9.6
7.4	8.1	7.7
5.0	9.4	8.2
5.9	12.0	7.3
6.4	15.9	11.3
6.3	7.4	9.5
5.0	9.0	8.8
6.1	5.2	8.4
7.9	9.2	6.8
5.7	8.6	5.0

Table 4.7 shows that the normal means model with a power prior tends to pull all estimates together compared to the other two methods. This is because the other two methods use a random effects model as the base model, which assumes the divergence in mean among groups. By contrast, the random effects model with a power prior recognizes the characteristics of each individual variety the most among three methods. For example, estimates are closest to sample means, and the standard deviation (s.d.) for the mean estimate of a variety is less affected by other varieties' variances. For a variety ("Four country") with small sample variance, the random effects model with a power prior yields relatively smaller estimates for variance and therefore a smaller s.d. for the mean estimate, compared to the other two methods. We examined several other examples and similar results were found. This feature of the random effects model with a power prior can be explained by the presence of two-level structure built-in to model the heterogeneity among groups. This feature also leads to a concern of over-parameterizing, which needs to be further investigated.

Table 4.7: *Application of three methods to estimate the population mean for each variety of corn.*

	Varieties of Corn		
	Four country	Lodent	Lancaster
Sample mean	6.24	8.84	8.33
Sample variance	1.19	**7.79**	2.62
Random effects model with power prior:			
BLUP	6.50	8.61	8.22
s.d. of BLUP	0.54	0.64	0.52
Normal mean model with power prior:			
$\hat{\mu}$	7.00	8.32	8.01
s.d. of $\hat{\mu}$	0.56	0.60	0.51
Random effects model with Jeffreys prior:			
BLUP	6.63	8.58	8.19
s.d. of BLUP	0.62	0.57	0.54

4.5 Time-Weighted Power Priors

In environmental studies, historical data over the past twenty years or even longer is often available. It is desirable to divide all the historical data into several data sets based on time to ensure the homogeneity within each data set. The power prior with multiple historical data sets should be applied in this case and a different power parameter will be put on each historical data set. In the framework of power priors with multiple historical data sets, it is assumed that those different power parameters are independent and follow the same prior distribution. However, the power parameters are dependent on each other in the setting mentioned above. A question arises then of how to make the prior time-weighted so that there is more weight assigned to more recent times. One solution is to impose a constraint on the power parameters, i.e. $0 \leq \delta_k \leq ... \leq \delta_2 \leq \delta_1 \leq 1$. For the rest of this section, we will demonstrate this method with a real set of water quality data.

In this example, we use measurements of pH collected during years of 1991 to 2000 at a monitoring station on Chopawamsic creek in Virginia. Of interest in these data is the population mean of pH over the period of $1999 - 2000$. Therefore pH data collected in years of 1999 and 2000 are treated as the current data, while pH data collected from 1991 to 1998

are divided into four samples of historical data. D_{01} represents the pH data collected in years of 1997 and 1998; D_{02} denotes the historical data collected in years of 1995 and 1996; D_{03} is data collected in years of 1993 and 1994; and D_{04} is data collected in years of 1991 and 1992. Table 4.8 presents the summary statistics of current and historical samples.

Table 4.8: *pH data collected during* 1991 *to* 2000 *at a water monitoring station.*

	Current data	Historical data			
	D	D_{01}	D_{02}	D_{03}	D_{04}
Sample size	20	19	18	21	21
Sample mean	6.37	6.95	7.03	7.15	6.74
Sample variance	0.90	0.25	0.22	0.15	0.28

Suppose that current data D come from a normal population with unknown mean μ and variance σ^2. Using the Jeffrey's prior as the initial prior on (μ, σ^2) and an uniform distribution as $\pi(\delta_j)$ for $j = 1, 2, 3, 4$, the time weighted power prior can be expressed as

$$\pi(\mu, \sigma^2, \underline{\delta}|\underline{D}_0) \propto \frac{\left(\prod_{j=1}^{4} L(\mu, \sigma^2|D_{0j})^{\delta_j}\pi(\delta_j)\right)\pi(\mu, \sigma^2)}{\int\left(\prod_{j=1}^{4} L(\mu, \sigma^2|D_{0j})^{\delta_j}\right)\pi(\mu, \sigma^2)\ d\mu d\sigma^2}, \tag{4.9}$$

where $\pi(\delta_1)$ is an uniform distribution on $[\delta_2, 1]$, $\pi(\delta_2)$ is an uniform distribution on $[\delta_3, \delta_1]$, $\pi(\delta_3)$ is an uniform distribution on $[\delta_4, \delta_2]$, and $\pi(\delta_4)$ is an uniform distribution on $[0, \delta_3]$.

The power prior in (4.9) is a modification of (4.1) by imposing a constraint $0 \leq \delta_4 \leq \delta_3 \leq \delta_2 \leq \delta_1 \leq 1$, and it can be easily implemented in *Winbugs 4*. The results are presented in Table 4.9 along with the results using a power prior with i.i.d. δ's, using all historical data and using no historical data.

Table 4.9 says that D_{04} would have a stronger impact on the estimation without the constraint $0 \leq \delta_4 \leq \delta_3 \leq \delta_2 \leq \delta_1 \leq 1$. When i.i.d. δ's are used, the influence of historical data on estimation is only determined by objective information, i.e. availability of historical data and its discrepancy from current data. The constraint $0 \leq \delta_4 \leq \delta_3 \leq \delta_2 \leq \delta_1 \leq 1$ is in fact subjective information, based on the belief that data collected at more recent time are more helpful in explaining current water quality. The special case where $0 \leq \delta_4 \leq \delta_3 \leq \delta_2 \leq \delta_1 \leq 1$ may be applied, including modelling $\text{logit}(\delta_j)$ as a linear function of time. However we experienced some difficulties in implementing this model due to its complexity. In addition, whether or not $\text{logit}(\delta_j)$ has a linear relationship with time is another question.

Table 4.9: *Analysis of pH data using the power prior method with different specifications on power parameters.*

Methods			Posterior mean			
	$\hat{\mu}$	s.d. of $\hat{\mu}$	δ_1	δ_2	δ_3	δ_4
Time-weighted power prior	6.63	0.17	0.42	0.22	0.12	0.06
Power prior with i.i.d. δ's	6.62	0.16	0.23	0.20	0.13	0.39
Use all historical data (all δ's $= 1$)	6.85	0.07				
Use no historical data (all δ's $= 0$)	6.37	0.22				

In practice, often more than one "historical" data sets are available. These could be data collected previously, at the same region, or from similar research settings. The "historical" data set is a vague concept and has no strict definition. Therefore, how many and which historical data sets should be incorporated are of particular interest to practitioners. A related issue is, for example, how many pieces a large historical data set collected over a long time period should be divided into. In addition, researchers may have prior knowledge on relative importance of each historical data set in the current study, and also on the relationship among several historical data sets. So how to build such subjective information into the power prior may not always be solved by adjusting the $\pi(\underline{\delta})$. Further research is needed on incorporation of multiple historical data sets, and this is essential for the power prior method to be widely used in real applications.

Chapter 5

Evaluating Water Quality: Using Power Priors to Incorporate Historical Information

5.1 Introduction

One important problem in environmental statistics is the evaluation of air or water quality standards. Issues include the definition of standards (Barnett and O'Hagan, [1]), trend assessment (Hirsch *et al.*, [16]) and the evaluation of data from locations to determine compliance. A standard for a chemical or pollutant is a qualitative or quantitative description of expectation for the chemical or pollutant. To implement such a standard a numerical criterion is often required. The numerical criteria might be different for a lake used for drinking water than a lake used for fishing. Also associated with the standard are expectations related to frequency, magnitude and duration. Air quality evaluation often involves the expected frequency of violation (for example the ozone standard, Thompson *et al.*, [35]). However, evaluation of water quality standards often involves a percentile view. For example, for dissolved oxygen, a site is expected to have 10% or fewer samples in violation.

To assess water quality standards, measurements of water quality under the Clean Water Act (e.g. pH, dissolved oxygen, biological oxygen demand) are collected on a regular basis (e.g. quarterly) over a two year period and analyzed to evaluate the percentage of samples in violation of the standard. A common approach accepted by the *US Environmental*

Protection Agency (USEPA) is the raw score approach that simply calculates the proportion of violations and declares the water segment impaired if this proportion exceeds 10%. Smith *et al.* (2001) noted that this is essentially a statistical hypothesis testing problem assuming a *binomial* population without controlling for error rates. They suggested to use a binomial test and discussed both Type I and Type II error probabilities. They showed that the tests using a statistical method have greater power than the raw score approach by USEPA. Furthermore, Ye and Smith (2002) proposed to use a Bayesian test on the desired percentile of the distribution (e.g., 10th percentile) that is of interest, to check for impairment of the monitoring site (see also McBride and Ellis, [24]). By using Bayesian methodology, the quantity of interest (e.g., the percentile of the measurement distribution) can be naturally treated as a parameter and thus its posterior distribution can be used to make needed decisions.

However, because all the approaches mentioned above do not fully use all the information provided by the data in the sense that only the data with "standard violation" or not is used in the analysis, Smith *et al.* ([31]) suggested an approach using a tolerance limit, pointing out that this would reflect the magnitude of violation (see also Smith, [32]).

Suppose a water quality measurement follows a certain distribution and a small value indicates a violation. Then the raw data can be used to test a hypotheses

$$H_0 : L \geq L_0 \text{ (no violation) versus } H_1 : L < L_0 \text{ (violation)}, \tag{5.1}$$

where L denotes the true lower percentile of the population distribution, and L_0 is the standard. To test the hypotheses in (5.1), one may consider rejecting the null hypothesis and hence declaring impairment when the posterior probability of the null hypothesis given the data is small (e.g. < 0.05).

Unfortunately, because decisions are based on data from a limited time period (for water monitoring data, many current samples have only two years of data), the sample size is often inadequate to provide necessary precision in parameter estimates. In such situations, "historical" data, a data set from similar studies or a data set from previous time periods, can be very helpful in interpreting the current status of water quality. Due to the nature of updating information sequentially, it is straightforward to use a Bayesian approach with an informative prior on the model parameters to incorporate the historical data into the current study. A traditional approach to incorporating historical data is to construct an informative prior using the historical data and then to combine the prior with the likelihood to yield the posterior distribution for statistical inference. This implies a simple pooling of

current data and historical data, since the two data sets are equally weighted, and can be well justified by assuming current and historical data are from the same population. However, the population parameters may change over time, or over different sites, although current and historical data are usually assumed to follow distributions in the same family. If the sample size of the historical data is much larger than that of current data and heterogeneity exists between the current and previous studies, historical data would dominate the analysis and the data pooling may result in misleading conclusions.

To address this issue, Ibrahim and Chen ([18]) proposed the concept of the *power prior*, based on the notion of the availability of historical data. The basic idea is to let a power parameter δ $(0 \leq \delta \leq 1)$ tell us how much historical data are to be used in the current study. Ibrahim and Chen ([18]) and Chen *et al.* ([10]) demonstrated how to construct power priors and discussed the general conditions for propriety. They also examined the power prior approach for generalized linear models, generalized linear mixed models, semiparametric proportional hazards models, and cure rate models with real data examples. Ibrahim *et al.* ([19]) gave a formal justification of the power prior as an optimal class of informative priors, and showed that the power prior is a 100% efficient information-processing rule in the sense that the ratio of the output to input information is equal to 1. However, in their approach, the power parameter δ always has a tendency to be close to zero, which suggests that much of a historical data set is not used. Here we propose a modified power prior approach with applications in water quality assessment. In Section 2, the general development of a modified power prior approach is given and certain properties of the approach under a normal population are discussed. More development of this modified approach can be found in Chapter 3. In Section 3, we will demonstrate the application of power priors in water quality assessment with two examples.

5.2 Power Prior Bayesian Analysis

5.2.1 Power Prior Approach

Suppose that θ is the parameter of interest in a water quality measurement, for instance, concentration of a chemical. Suppose that such a measurement follows a distribution and $L(\theta|D_0)$ is the likelihood function of θ based on the historical data, denoted by D_0. In this article we assume that, given θ, historical data (D_0) and current data (D) are independent random samples from an exponential family. $\pi(\theta)$ is taken as the initial prior before any

historical information is gathered and usually it is a noninformative prior. Given δ, Ibrahim and Chen ([18]) define the power prior of θ for the current study as

$$\pi(\theta|D_0,\delta) \propto L(\theta|D_0)^{\delta}\pi(\theta). \tag{5.2}$$

The parameter δ measures the portion of historical information needed in the current study and is described using the prior in (5.2). The case $\delta = 0$ means that no historical data should be used, while $\delta = 1$ gives equal weight to $L(\theta|D_0)$ and the likelihood of the current study $L(\theta|D)$, resulting in full incorporation of the historical data. Therefore, (5.2) can be viewed as a generalization of the usual Bayesian update of $\pi(\theta)$ (see discussion in Ibrahim and Chen, [18]). The power parameter δ can be interpreted as a precision parameter. For example, consider the case of a normal sample with known variance. Suppose that D_0 consists of n_0 observations, $X_1, X_2, \ldots, X_{n_0}$, from the normal population with unknown mean parameter θ and known variance σ^2. If the prior $\pi(\theta)$ is assumed to be uniform (non-informative), (5.2) implies a prior distribution of θ for the current data set D, $\theta|D_0,\delta \sim N(\bar{x}_0, \sigma^2/\delta n_0)$, where $\bar{x}_0$ is the sample mean of the historical data. Hence, δ can be viewed as part of the precision parameter, because smaller δ implies larger power prior variance while larger δ means smaller power prior variance.

The power prior $\pi(\theta|D_0,\delta)$ in (5.2) was initially elicited for fixed δ. However, since δ is not necessarily pre-determined, we may extend it further to the case that δ is random. Thus the power prior $\pi(\theta|D_0,\delta)$ in (5.2) is treated as a distribution of θ conditional on δ and historical information. The power prior specification on (θ,δ) is then completed by specifying a prior distribution for δ. A random δ gives the investigator more flexibility in weighting the historical data. A natural prior for δ would be a $Beta(\alpha,\beta)$ distribution, or simply a uniform distribution, since $0 \leq \delta \leq 1$.

We prefer a modification of the original approach of Ibrahim and Chen ([18]) for two reasons. First, in their original approach, Ibrahim and Chen ([18]) constructed the joint power prior of (θ,δ) as

$$\pi(\theta,\delta|D_0) \propto L(\theta|D_0)^{\delta}\pi(\theta)\pi(\delta). \tag{5.3}$$

Notice that multiplying the likelihood function $L(\theta|D_0)$ by a positive constant would change the joint prior of (θ,δ) and consequently the posterior, which does not agree with the likelihood principle. Another problem with the original approach comes up when we investigated the application of power priors on normal mean models. The influence of historical data is small no matter how compatible current and historical data are, and the inference on θ is not much different from that without considering historical data. Therefore, we propose the

following power prior:

$$\pi(\theta, \delta|D_0) = \frac{L(\theta|D_0)^\delta \pi(\theta)\pi(\delta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}, \tag{5.4}$$

in the region of δ such that the denominator in (5.4) is not infinity. In this article, Θ is used to denote the range of parameter θ. This new approach is free of the above mentioned problems and it needs little to ensure the propriety of the joint power prior for (θ, δ). More discussion of this prior will be given in next section.

5.2.2 General Development of the Power Prior

For ease of exposition, first we develop a power prior and the consequent posterior with only one historical data set, then follow with the extension to multiple historical data sets in Section 5.2.4.

The modified power prior distribution of (θ, δ) is given in (5.4). The assumptions and conditions of this prior are $0 \leq \delta \leq 1$, $L(\theta|D_0) \geq 0$, $\pi(\theta) \geq 0$, and $P_\theta(L(\theta|D_0) > 0) > 0$. Consequently, we have $\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta > 0$. Define A as the following

$$A = \{\delta : 0 < \int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta\ < \infty\}. \tag{5.5}$$

Furthermore, define A^- as the complement of A in [0,1]. The assumption

$$\int_\Theta L(\theta|D_0)\pi(\theta)\ d\theta\ < \infty$$

guarantees that $\pi(\theta)$ can be used as a prior in a coherent Bayesian updating scheme due to the consequent proper posterior of θ. This assumption constrains our discussion to a sensible range, and implies $1 \in A$.

We propose a joint power prior distribution for (θ, δ) of the form

$$\pi(\theta, \delta|D_0) = \begin{cases} M \frac{L(\theta|D_0)^\delta \pi(\theta)\pi(\delta)}{\int L(\theta|D_0)^\delta \pi(\theta)\ d\theta}, & \text{if } \delta \in A; \\ 0, & \text{if } \delta \in A^-. \end{cases} \tag{5.6}$$

where M is a normalization constant. Consequently

$$\pi(\delta|D_0) = \begin{cases} M\ \pi(\delta), & \text{if } \delta \in A; \\ 0, & \text{if } \delta \in A^-. \end{cases}$$

It is easy to check that the joint prior $\pi(\theta, \delta|D_0)$ defined above is always proper, which also ensures the propriety of the joint posterior for (θ, δ). Such a joint posterior distribution for (θ, δ) can be written as

$$\pi(\theta, \delta|D_0, D) \propto L(\theta|D)\pi(\theta, \delta|D_0) \propto \frac{L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\pi(\delta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta} I_A(\delta),$$

where $I_A(\delta) = 1$ if $\delta \in A$ and 0 otherwise.

The marginal posterior distributions of δ and θ can be derived as follows.

$$\pi(\delta|D_0, D) \propto \pi(\delta)\frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta} I_A(\delta), \tag{5.7}$$

and

$$\pi(\theta|D_0, D) \propto \pi(\theta)L(\theta|D) \int_A \frac{L(\theta|D_0)^\delta \pi(\delta) I_A(\delta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}\ d\delta. \tag{5.8}$$

We can make inferences on θ, e.g. in hypothesis testing, Bayesian interval calculation, or model selection, based on the marginal posterior distribution of θ in equation (5.8). On the other hand, using the marginal posterior distribution of δ shown in (5.7), the properties of the power parameter δ can be studied.

We next discuss the case of a normal population. Then in Section 5.3 environmental applications are demonstrated.

5.2.3 Normal Population

Suppose we are interested in a normal population and inference for the normal mean with unknown variance, by incorporating both current and historical data. Although there may be many observations available for study, not all are used. For instance, in one location there may be more than 50 years of data available, however, it is not reasonable to treat observations 50 years ago as providing the same information as just one prior year of information about the water quality at the same location. If the data set can be assumed approximately normal, we can use the approach discussed below, the power prior analysis.

Suppose that current data $D = (x_1, ..., x_n)$ come from a normal $n(\mu, \sigma^2)$ population with unknown μ and σ^2, and $D_0 = (x_{01}, ..., x_{0n_0})$ is a historical data set. Let $\bar{x}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} x_{0i}$, $\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$, $\hat{\sigma}_0^2 = \frac{1}{n_0}\sum_{i=1}^{n_0}(x_{0i} - \bar{x}_0)^2$, and $\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_i - \bar{x})^2$. Furthermore, suppose that we use the prior $\pi(\mu, \sigma^2) \propto (1/\sigma^2)^a$ as the initial prior of (μ, σ^2), where $a > 0$ is a

pre-determined constant. Note that $a = 1$ corresponds to the reference prior (Berger and Bernardo, [6]), while $a = \frac{3}{2}$ results in the Jeffreys prior (Jeffrey, 1946). Also, the prior distribution of δ is a $Beta(\alpha, \beta)$, where hyper-parameters α and β are all known. Following (5.6), the joint power prior distribution of (μ, σ^2, δ) can be expressed as

$$\pi(\mu, \sigma^2, \delta|D_0) \propto \frac{\delta^{\frac{\delta n_0}{2}+a+\alpha-2}(1-\delta)^{\beta-1}}{\left(\frac{2\sigma^2}{n_0\hat{\sigma}_0^2}\right)^{\frac{\delta n_0}{2}+a}\Gamma\left(\frac{\delta n_0-3}{2}+a\right)} \exp\Big\{-\frac{\delta n_0}{2\sigma^2}[\hat{\sigma}_0^2+(\mu-\bar{x}_0)^2]\Big\},$$

where the range of δ is $(b, 1]$ for $b \geq 0$ or $[0, 1]$ for $b < 0$ and $b = \frac{2}{n_0}\left(\frac{3}{2} - a\right)$. Hence, if the reference prior of θ is used, the set A defined in (5.5) is $(\frac{1}{n_0}, 1]$ for the normal mean model, while if the Jeffreys prior of θ is used, $A = (0, 1]$. The lower bound b suggests that the information in historical data is automatically taken into account to a certain extent, depending on the availability of historical data. However, such a case may be changed once the original prior is changed.

Combining the joint power priors with the likelihood based on the current data, we obtain the joint posterior distribution of (μ, σ^2, δ). Integrating μ and σ^2 out of $\pi(\mu, \sigma^2, \delta|D_0, D)$ then leads to the marginal posterior distribution of δ.

$$\pi(\delta|D_0, D) \propto \frac{\delta^{\frac{\delta n_0}{2}+a+\alpha-2}(1-\delta)^{\beta-1}\Gamma(\frac{\delta n_0+n-3}{2}+a)}{\left[\frac{\delta n}{\delta n_0+n}\frac{(\bar{x}_0-\bar{x})^2}{\hat{\sigma}_0^2}+\delta+\frac{n}{n_0}\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right]^{\frac{\delta n_0+n-3}{2}+a}\Gamma(\frac{\delta n_0-3}{2}+a)},$$

with the range described above. The behavior of the power parameter δ can be studied from this marginal posterior distribution.

Similarly the marginal posterior distribution of (μ, σ^2) can be derived by integrating δ out of $\pi(\mu, \sigma^2, \delta|D_0, D)$, but it does not have a closed form. Combined with $\pi(\delta|D_0, D)$, the conditional posterior distribution of (μ, σ^2) given δ indirectly reflects the behavior of $\pi(\mu, \sigma^2|D_0, D)$. Both the posterior distribution of μ conditional on δ and the posterior of σ^2 conditional on δ turn out to be commonly used distributions. From standard calculations of Bayesian analysis using a normal population (see Gelman *et al.*, [15]), we find that the conditional posterior distribution of μ, given δ and data, follows a *Student-t* distribution with location parameter $(\delta n_0\bar{x}_0 + n\bar{x})/(\delta n_0 + n)$, scale parameter $\sqrt{\frac{2}{C(\delta)}\frac{1}{(\delta n_0+n+2a-3)(\delta n_0+n)}}$ and degrees of freedom $\delta n_0 + n + 2a - 3$, where

$$C(\delta) = \frac{2}{\frac{\delta n_0 n(\bar{x}_0-\bar{x})^2}{\delta n_0+n}+\delta n_0\hat{\sigma}_0^2+n\hat{\sigma}^2}.$$

Furthermore, the conditional posterior distribution of σ^2, given δ and the data, follows an

inverse-gamma distribution with parameter $(\delta n_0 + n + 2a - 3)/2$ and $C(\delta)^{-1}$ (again, see Gelman *et al.*, [15]).

The role of the power parameter δ is to control the influence of the historical data on the current study. Using simulation studies, we found that this power control parameter is adjusted automatically based on the compatibility between the historical and current data, and also based on the sample sizes of the two studies. In the case of the normal mean model, the compatibility between the historical and current data can be measured by differences in sample means and sample variances. If historical and current data are fully compatible, i.e. two data sets have the same sample mean and sample variance, the marginal posterior mode of δ is always 1, no matter how high the ratio n_0/n is. This seems very reasonable since when the historical data can contribute necessary information into the current study, we would like to use it as much as possible to achieve higher precision. How quickly the mode reaches 1 also depends on the extent of compatibility between the historical and current data. When the historical data population is very different from the current data population, the posterior mode of δ goes to zero very fast. The posterior mode of δ decreases with n_0/n, and attains 1 with very small n_0/n. These trends imply that the random δ responds to data in a sensitive and desirable way.

5.2.4 Extension to Multiple Historical Data Sets

The priors defined in (5.6) can easily be generalized to multiple historical data sets. Suppose there are m historical studies. We define D_{0j} to be the historical data based on the jth study, $j = 1, ..., m$ and $D_0 = (D_{01}, ..., D_{0m})$. Chen *et al.* ([10]) suggested defining a different weight parameter δ_j for j^{th} historical study and taking the δ_j's to be i.i.d. *Beta* random variables with hyperparameters (α, β). Let $\underline{\delta} = (\delta_1, ..., \delta_m)$. Under the new approach, the power prior in (5.6) can be generalized as

$$\pi(\theta, \underline{\delta}|D_0) \propto \frac{\left(\prod_{j=1}^{m} L(\theta|D_{0j})^{\delta_j} \pi(\delta_j|\alpha, \beta)\right)\pi(\theta)}{\int_\Theta \left(\prod_{j=1}^{m} L(\theta|D_{0j})^{\delta_j}\right)\pi(\theta)\ d\theta} I_B(\underline{\delta}),$$

where $B = \{(\delta_1, ..., \delta_m) : 0 < \int_\Theta \left(\prod_{j=1}^{m} L(\theta|D_{0j})^{\delta_j}\right)\pi(\theta)\ d\theta\ < \infty\}$.

This framework can accommodate potential compatibility among different sites or different time periods. For example, we could take data collected over time at different but

adjacent sites as different historical data sets. Moreover, data collected over a long period may be divided into several historical data sets instead of being treated as one big historical data set. In such a way, the role of historical data can be more accurately evaluated. In Section 5.3.2, we will discuss an example implementing the modified power prior approach using multiple site information.

5.3 Power Prior Applications in Evaluating Site Impairment

When applying Bayesian analysis with power priors to water quality data, two kinds of additional information could be incorporated: past information or information from adjacent sites. We compare the modified power prior approach to other methods including the USEPA's raw score method, the binomial test, and a traditional Bayesian approach using the reference prior. Suppose that the measurements of water quality follow a normal distribution, and for ease of comparison, the normal model with a simple mean is considered. Based on the equations derived in Section 5.2.3, we use *Winbugs* (Release 1.4) to simulate the marginal posterior distribution for (μ, σ^2), and then the percentile of interest.

5.3.1 Using Past Information to Build the Prior

In this example, we use measurements of pH to evaluate impairment of four sites in Virginia individually. Of interest in these data sets is the determination of whether the pH values at a site indicate that the site violates a (lower) standard of 6.0 more than 10% of the time. For each site, larger sample size is associated with the historical and smaller with the current data. In this example, pH data collected over a two-year or three-year period are treated as the current data, while pH data collected over the previous nine years represents one single historical data set. The current data and historical data are plotted side by side for each site in Figure 1. In the power prior approach, a violation is evaluated using a Bayesian test of

$$H_0 : L \geq 6.0 \text{ (no impairment, don't list)},$$
$$H_1 : L < 6.0 \text{ (impairment, list)},$$

where L is the lower 10th percentile of the distribution for pH. Comparison of results from different methods is presented in Table 5.1.

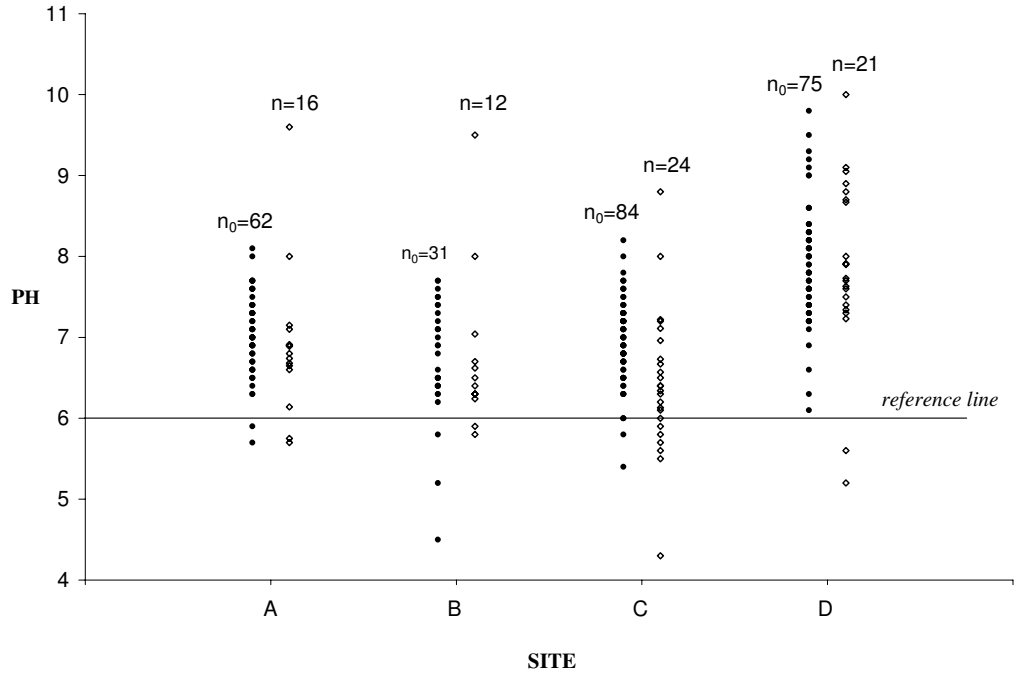

Figure 5.1: pH data collected at four stations. For each site, historical data are on the left (circle) and current data on the right (diamond).

In Table 5.1, the summaries of the current and historical data are given, along with the percentage of samples below the standard (raw score values). Furthermore, the test results using a binomial population as well as the test results using both the reference prior analysis (without incorporating historical data) and power prior analysis (with reference prior, i.e, $a = 1$ in Section 5.2.3) are presented. For sites C and D, the raw data are used in the Bayesian analysis, while for sites A and B, a log transformation is applied to the data, because the lognormal distribution fits the data better than normal.

All four sites have greater than 10% of observations below the standard of 6.0. Therefore, we would declare that all sites are impaired using EPA's raw score approach. Note that this raw score approach results in higher type I error probability (see Smith *et al.*, 2001) which means that it would declare more impaired sites than it should have. On the other hand, if the 0.05 significance level is used, the binomial test would only indicate site C as impaired. The Bayesian test using the reference prior results in a similar conclusion although the p-values are smaller, compared to the binomial tests, in all cases but one. Here we use the posterior probability of H_0 as equivalent to the p-value (Berger, [2]) for testing a one-sided

Table 5.1: *Comparison of the power prior method with alternative methods in evaluating site impairment when one historical data set is available. In the table, n and n_0 are sample sizes, mean (s.d.) refers to sample mean (sample standard deviation), and s.d. of L is the posterior standard deviation of L. % below is the percentage of samples below the EPA standard (6 for pH).*

Site	Current data		Historical data		% below	Binomial P-value	Posterior probability of H_0 (s.d. of L)	
	n	mean (s.d.)	n_0	mean (s.d.)			reference prior	power prior
A	16	6.91 (0.90)	62	7.05 (0.47)	0.13	0.4853	0.2074 (0.27)	0.6027 (0.21)
B	12	6.78 (1.03)	31	6.73 (0.71)	0.17	0.3410	0.0627 (0.34)	0.0294 (0.19)
C	24	6.43 (0.88)	84	6.95 (0.49)	0.25	0.0277	0.0003 (0.26)	0.0017 (0.24)
D	21	7.87 (1.11)	75	7.88 (0.67)	0.10	0.6353	0.8673 (0.36)	0.9831 (0.25)

hypothesis. Using historical data does lead to different conclusions for site B. The test using a power prior results in significance for sites B & C. In the case of site B, there are around 10% of historical observations below 6.0. Hence our prior opinion of the site is suggestive of impairment. Less information is therefore required to declare impairment relative to a reference prior and the result is a smaller p-value.

Another notable advantage of the power prior method is that it improves the estimation of L by using past information. This can be shown by the consistently smaller posterior standard deviation of L with the power prior than with the reference prior for all four sites.

5.3.2 Borrowing Information from Adjacent Sites

Alternate sources of information from other locations can also be used to aid inference on environmental quantity at the site of interest. In this section, measurements of dissolved oxygen (DO) are used to evaluate impairment of four sites in the Philpott reservoir in Virginia. Of interest here is to determine whether DO values violate a (lower) standard of 5.0 more than 10% of the time at each site. DO values are plotted in Figure 2.

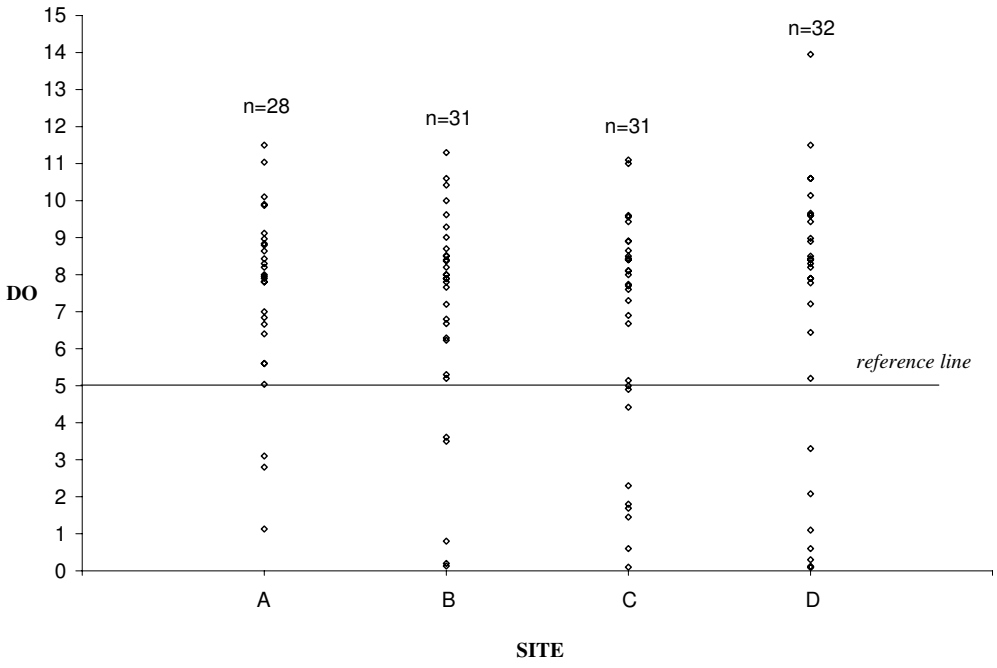

Figure 5.2: DO data collected at four stations on Philpott reservoir (years 2001, 2002, and 2003).

Since the four sites are on the same reservoir, the geographical connection among the four sites suggests similarity of sites due to spatial correlation among their data. So data collected at one site should partially reflect the water quality at the other three. For each site, DO data collected over a three-year period are treated as the current data, while DO data collected over the same period but at the other three sites are referred to three "historical" data sets. This section illustrates the utility of power priors with multiple historical data sets that is discussed in Section 5.3.2. In the power prior approach, the problem is set up as

a Bayesian test of

$$H_0 : L \geq 5.0 \text{ (no impairment, don't list)},$$
$$H_1 : L < 5.0 \text{ (impairment, list)},$$

where L is the lower 10th percentile of the distribution for DO. Comparison of results from different methods is presented in Table 5.2.

Table 5.2: *Comparison of the power prior method with alternative methods for evaluating site impairment when multiple historical data sets are available. In the table, n is sample size, mean (s.d.) refers to sample mean (sample standard deviation), and s.d. of L is the posterior standard deviation of L. % below is the percentage of samples below the EPA standard (5 for DO).*

Site	n	mean (s.d.)	% below	Binomial P-value	Posterior probability of H_0 (s.d. of L)	
					reference prior	power prior
A	28	7.55 (2.41)	0.11	0.5406	0.1640 (0.65)	0 (0.55)
B	31	7.10 (2.88)	0.16	0.1932	0.0038 (0.73)	0 (0.54)
C	31	6.66 (3.09)	0.26	0.0096	0 (0.79)	0 (0.54)
D	32	6.67 (4.02)	0.28	0.0033	0 (1.01)	0 (0.62)

Both binomial and reference prior Bayesian tests, which are only based on each individual site's data, showed the same results on sites A, C and D, and different results on site B. However, the power prior Bayesian analysis results in significance of all tests and hence a decision to declare all sites as impaired. Coincidentally, this conclusion consistently matches the EPA's raw data approach.

Clearly, the power prior approach improves the estimation of L by borrowing information from adjacent sites, which can be seen by the consistently smaller posterior standard deviation of L using the power prior relative to the use of the reference prior for all four sites. Other sites' information does affect the inference substantially. For example, based on each site's data individually, site A seems to have higher water quality than the other three. However, when we use the power prior, the low water quality at site B, C, and D drags down the estimate of L for site A. So we would not declare site A impaired based on the posterior probability of H_0 without considering the other three sites' information, but we would do so with a power prior.

Chapter 6

Using Power Priors to Improve the Binomial Test of Water Quality

6.1 Introduction

Section 303(d) of the Clean Water Act requires states to assess and report the condition of their waters based on *US Environmental Protection Agency* (USEPA) guidelines. Reports describe the condition of water segments based on measurements of chemical constituents collected at a large number of monitoring sites or segments, with each segment typically associated with a single site. Each water sample at a site is assumed to represent a background population of water quality conditions. From a sample of water quality measurements collected over time the assessor must decide if the site should be listed as impaired, resulting in a listing decision process. A common approach accepted by the USEPA is the *raw score* approach that declares a violation if the proportion of violations exceeds 10%. Smith, Ye, Hughes, and Shabman ([30]) suggest that the water quality assessment process can be viewed as a statistical decision problem, and one may use hypothesis testing to help with the decision process.

If a statistical approach is used, the null hypothesis is that the site is not impaired, while the alternative hypothesis is that the site is impaired. The hypotheses may be framed in terms of a parameter p describing the true degree or probability of impairment and p^*, the "safe level" or hypothesized probability of impairment under safe conditions. The impairment decision is based on the test $H_0 : p \leq p^*$ versus $H_1 : p > p^*$ where p^* is a constant

between 0 and 1 (in our discussion in this manuscript, it is 0.10).

The hypothesis testing scenario suggests evaluation of tests based on error rates. From the environmental perspective, the assessor needs to be concerned about falsely declaring a healthy segment as impaired (Type I error or false positive) and failing to declare a segment impaired when in fact it is impaired (Type II error or false negative). A false declaration of standards violation may trigger a costly process, the Total Maximum Daily Load (TMDL) process, and incur unnecessary constraints on agriculture or industry. On the other hand, a false declaration of no violation may pose a risk to human and ecological health. The error rates are bounded between 0 and 1, with 0 indicating no error.

In the frequentist binomial test, the probability of exceeding the standard is treated as fixed and the binary data (whether the measurement is exceeding the standard or not) are treated as random. A Bayesian approach computes the probability that the site exceeds the standard by treating the impairment probability as a random variable that has an associated distribution. Initially the form of this distribution is based on previous information that is used to define a prior distribution. After data are collected, the prior is updated with data to form the posterior distribution of the impairment probability using Bayes rule. Based on the posterior distribution, a decision can be made using either a cutoff approach or an odds-ratio approach (Bayes factor). McBride and Ellis ([24]) point out that using a Bayesian approach can address some issues presented in a frequentist analysis. For example, switching null and alternative hypotheses will not lead to a different decision because the use of Type I and Type II error probabilities are less relevant in a Bayesian framework. And more importantly, a Bayesian approach appears to provide simpler and more direct statements about the probability of compliance than frequentist hypothesis testing (McBride and Ellis, [24]). Smith *et al.* ([30]) and McBride and Ellis ([24]) discussed the implementation of a Bayesian approach to the binomial test using a noninformative prior for evaluating site impairment.

Since listing decisions are typically required to be made on a site-by-site basis over a two year period, all the methods mentioned above face the challenge resulting from using a limited amount of data to determine if the stream is violating standards. The lack of sufficient data may lead to unacceptable levels of uncertainty. One way to tackle this issue is to use the information from surrounding sites or from previous reports to elicit an informative prior. For ease of discussion, we refer to data from surrounding sites as well as from previous reports as historical data. A traditional approach to incorporating historical data is to construct the informative prior as the posterior distribution updated by the historical data.

This actually implies a simple pooling of current data and historical data together, since the two data sets are equally weighted. This traditional approach can only be justified by assuming that the current and historical data are from the same population.

It is often reasonable to assume that the distributions underlying current and historical samples belong to the same family. However, the population parameters and hence the probability of impairment may change over time, or over different locations. If the sample size of historical data is much larger than that of current data or two data sets are not compatible, historical data would dominate the analysis and the data pooling may result in misleading conclusions. To address this issue, Ibrahim and Chen ([18]) propose the concept of *power priors*, which are based on the notion of the availability of historical data. The basic idea is to let a power parameter δ $(0 \leq \delta \leq 1)$ tell us how much historical data is used in the current study. Since δ is not necessarily pre-determined, we may treat it as random and specify a prior distribution $\pi(\delta)$ for it. However, in their implementation, the power parameter δ always has a tendency to be close to zero, which suggests that much of a historical data set is usually not used. Here we propose to use a modified power prior approach to help with water quality assessment. The general development of this modified approach and more discussions can be found in Chapter 3. In Section 6.2, the modified power prior approach for a Bernoulli population is proposed and the Bayesian binomial test using the power prior is examined. In Section 6.3, error probabilities using the frequentist binomial test, and the Bayesian binomial test with uniform prior and power prior will be compared under different scenarios. In Section 6.4, we will demonstrate the application of power priors in water quality assessment with two real data sets.

6.2 Bayesian Binomial Test with the Power Prior

Assume that the current sample observations are independent of each other, and that the water quality parameter of interest has a distribution that does not change over the period of collection. Suppose that this assumption also holds for the historical sample, and the distributions underlying the two samples belong to the same family possibly with different parameters. The number exceeding the standard among current data, denoted by x, may be modelled as a binomial random variable with associated current sample size n and probability of violation p. Similarly the number in violation of the standard among historical data, denoted by x_0, can be modelled as a binomial random variable with associated historical sample size n_0. However, the probability of violation associated with the historical sample

may be different from that associated with the current sample. Given the initial prior for p, $\pi(p)$, the joint power prior distribution of (p, δ) is constructed as

$$\pi(p, \delta|D_0) = \pi(p|\delta, D_0)\pi(\delta) = \frac{f(D_0|p)^\delta \pi(p)\pi(\delta)}{\int_0^1 f(D_0|p)^\delta \pi(p)\ dp},$$

where D_0 is the historical data and $f(D_0|p)$ is the density of the historical data given p. A natural prior for δ, $\pi(\delta)$, would be a $Beta(\alpha, \beta)$ distribution, or possibly a uniform distribution, since $0 \leq \delta \leq 1$. The uniform distribution is the reference prior for p, one of the most commonly used noninformative priors. Suppose now both $\pi(p)$ and $\pi(\delta)$ are uniform distributions on $[0, 1]$. Then the joint power prior distribution of (p, δ) becomes

$$\pi(p, \delta|D_0) = \frac{p^{\delta x_0}(1-p)^{\delta(n_0-x_0)}}{B(\delta x_0 + 1, \delta(n_0 - x_0) + 1)}, \tag{6.1}$$

where B stands for the $Beta$ function, $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Combining the power prior $\pi(p, \delta|D_0)$ with the density of the current data D, the joint posterior distribution of (p, δ) is

$$\pi(p, \delta|D_0, D) \propto f(D|p)\pi(p, \delta|D_0) \propto \frac{p^{\delta x_0 + x}(1-p)^{\delta(n_0-x_0)+(n-x)}}{B(\delta x_0 + 1, \delta(n_0 - x_0) + 1)}.$$

The marginal posterior distribution of δ can be derived as follows by integrating p out in $\pi(p, \delta|D_0, D)$,

$$\pi(\delta|D_0, D) \propto \frac{B(\delta x_0 + x + 1, \delta(n_0 - x_0) + (n - x) + 1)}{B(\delta x_0 + 1, \delta(n_0 - x_0) + 1)}. \tag{6.2}$$

This distribution can be used to investigate the behavior of the power parameter δ.

On the other hand, integrating δ out of $\pi(p, \delta|D_0, D)$ leads to the marginal posterior distribution of p, which is used to make inference about p.

$$\pi(p|D_0, D) \propto \int_0^1 \frac{p^{\delta x_0 + x}(1-p)^{\delta(n_0-x_0)+(n-x)}}{B(\delta x_0 + 1, \delta(n_0 - x_0) + 1)} d\delta. \tag{6.3}$$

This posterior distribution represents the current knowledge about the probability of a violation found by updating the prior information. Using the above distribution, the posterior probability of the null and alternative hypotheses may be calculated. For the null hypothesis (H_0) that the site is not exceeding a standard, the probability is computed as $\alpha_0 = P(H_0|\text{data}) = P(p \leq p^*|D_0, D)$, where p^* is 0.1 in our case. For the alternative (H_1)

that the site is exceeding standards, the posterior may be calculated as $\alpha_1 = P(H_1|\text{data}) = P(p > p^*|D_0, D)$. Two approaches are available to make decisions based on these probabilities: the cutoff method and the odds-ratio method (see Smith *et al.* 2001).

The cutoff method uses the posterior probability to determine the rejection rule. If $P(H_0|\text{data}) < q$, then we reject the null hypothesis and conclude that the water is impaired. The quantity q is the posterior cutoff and is analogous to the binomial method Type I error rate. The odds-ratio method uses the Bayes factor to determine the rejection rule. The Bayes factor of H_1 against H_0 can be expressed as

$$B_{10} = \frac{P(H_1|D_0, D)/P(H_0|D_0, D)}{P(H_1)/P(H_0)}$$

A large value of the Bayes factor would indicate that the null hypothesis is not correct. Since the uniform distribution is used as the initial prior for p throughout our discussions, here $P(H_1)/P(H_0)$ is a fixed number. Consequently any Bayes factor cutoff can be converted to a corresponding posterior cutoff. Therefore, in our examples we only adopted the posterior cutoff method as the decision rule for Bayesian binomial tests.

6.3 Error Probabilities

For the listing decision process, Smith *et al.* (2001) considered alternatives to the EPA's raw score approach, including the frequentist binomial test and the Bayesian approach to binomial test with a noninformative prior (specifically, uniform prior on p). We will use "binomial method" and "Bayesian method" to refer to these two approaches for the rest of this chapter. Both methods have error rates that may be controlled through sample size and selection of nominal Type I error or posterior cutoff. The Type II error rates are reasonable when sample sizes are large (e.g. more than 20), but power is low for smaller sample sizes. The lack of power is a problem since sample sizes for assessment of site impairment are frequently small. For example, the assessment is often required to be conducted on two-year observations, and some measurements (e.g. pH, dissolved oxygen) are collected quarterly. However, the power prior approach may be used to legitimately incorporate more data by constructing an informative prior, and therefore should improve the power. Our interest is to investigate the benefit gained by using the power prior approach under various scenarios, when the current sample has 20 or fewer observations. Three sample sizes $n = 8, 12$, and 20 are considered for illustration.

For the binomial method, the Type I error rate is preset at α, which is an upper bound on the error. The actual error rate for the binomial method is determined by computing the cumulative probability of getting less than "x" samples exceeding the standard. The actual Type I error rate is calculated as the greatest cumulative probability that does not exceed α. To illustrate the improvement brought by the power prior approach over the binomial and Bayesian methods, α and q (described in Section 6.2) are chosen to achieve the same Type I and Type II error probabilities using binomial and Bayesian methods. Specifically, α and q are set at 0.05 when $n = 8$ or 20, and at 0.03 when $n = 12$, for ease of exposition. This setup puts frequentist and Bayesian methods on the same base for comparisons, and therefore emphasizes the changes in decision making brought by incorporating historical information.

In the following subsections, we give the expressions for the error probabilities from the power prior approach. Furthermore, exact error probabilities are computed and compared for the tests using the power prior, binomial and Bayesian methods.

6.3.1 Error Probabilities under Various p_0

We denote by p_0 the true probability of violation for a water segment at an earlier period or for a neighboring segment. This indicates that x_0 (the number exceeding the standard among historical data) follows a binomial distribution with parameters p_0 and n_0.

Using the posterior cutoff method, the Type I error probability using Bayesian methods for a specified p_0 is computed as

$$\begin{aligned} &P(P(H_0|D_0, D) < q | x \sim \text{binomial}(n, p^*), x_0 \sim \text{binomial}(n_0, p_0)) \\ &\quad = \sum_{x_0=0}^{n_0} \sum_{x=0}^{n} I(P(p \le p^*|D_0, D) < q) f(x|n, p^*) f(x_0|n_0, p_0), \end{aligned} \tag{6.4}$$

where $I(\cdot)$ is an indicator function, and $f(\cdot)$ is the probability mass function of a binomial distribution. To compare the error rates, the acceptable probability of violation is set at 10%, i.e. $p^* = 0.1$. The listing decision process may be viewed as a test of the null hypothesis that the probability of violation is less than or equal to 10% versus the alternative that it is greater than 10%.

To compute a Type II error rate (given that the site is impaired, how likely that we do not detect impairment), the true probability of exceeding the standard in the current data set must be specified; this percentage is denoted by p^{**}. Then the Type II error probability

for a certain p_0 is

$$P(P(H_0|D_0, D) \geq q|x \sim \text{binomial}(n, p^{**}), x_0 \sim \text{binomial}(n_0, p_0))$$
$$= \sum_{x_0=0}^{n_0} \sum_{x=0}^{n} I(P(p \leq p^*|D_0, D) \geq q) f(x|n, p^{**}) f(x_0|n_0, p_0). \quad (6.5)$$

For illustration in our situation, p^{**} is set at 25%. This value was selected as indicating severe problems and represents the minimum violation percentage we would almost always want to detect.

Four types of scenarios are considered to accommodate some situations possibly encountered in water quality assessment:

1. water is currently healthy and was also healthy when historical data were collected
2. water is currently impaired but was healthy before
3. water is currently healthy but was impaired before
4. water is currently impaired and was also impaired before

Here a water segment is viewed as healthy if the probability of violation is no greater than 0.1, with 0.1 being marginally healthy. Corresponding to the four scenarios, the following setups of p and p_0 are used to investigate the error probabilities for various methods, where p and p_0 are the true probability of violation underlying current and historical samples, respectively.

- $p = 0.1, p_0 = 0.05$ and 0.1 for scenario 1
- $p = 0.25, p_0 = 0.05$ and 0.1 for scenario 2
- $p = 0.1, p_0 = 0.15$ to 0.5 for scenario 3
- $p = 0.25, p_0 = 0.15$ to 0.5 for scenario 4

In Figures 6.1-6.4, exact error probabilities of the test using the power prior are compared to those derived using binomial and Bayesian methods under the above four situations. In Figures 6.1 and 6.3, the Type I error rate represents the probability that a site is declared as impaired when in fact its violation rate is 0.1 (false positive); in Figures 6.2 and 6.4, the

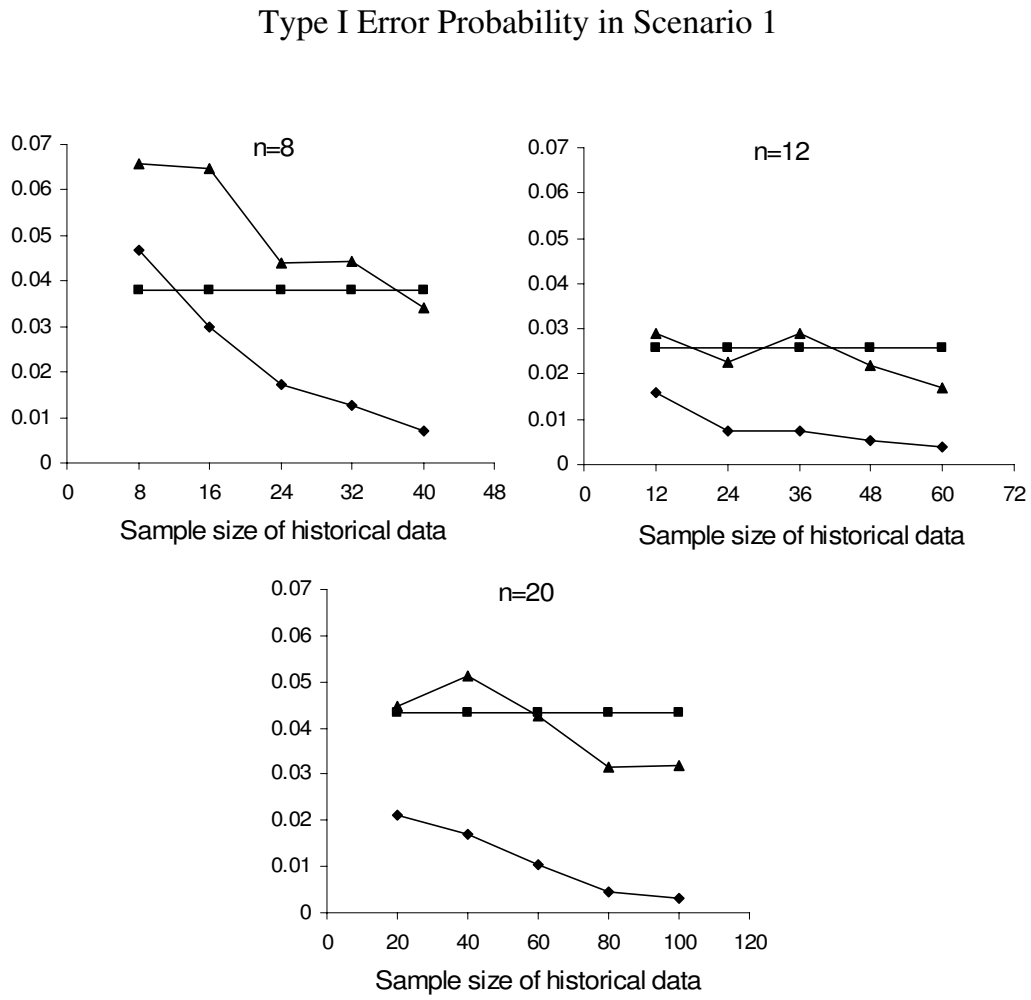

Figure 6.1: Comparisons of Type I error probability using three procedures for the situation in which both the true current and prior status of water are healthy. The three graphs are for $n = 8, 12$, and 20. In each graph, ■ is for the binomial and Bayesian methods; ♦ is for the power prior method with $p_0 = 0.05$; ▲ is for the power prior method with $p_0 = 0.1$, where p_0 is the true probability of violation for the past status of water.

Type II error rate is interpreted as the probability of failing to declare a site as impaired when in fact its violation rate is 0.25 (false negative).

When the true impairment status of a water segment does not change over time period (Figures 6.1 and 6.4), using the power prior method does reduce the Type II error as expected, but does not have advantages in terms of the Type I error. This is no surprise since the Type

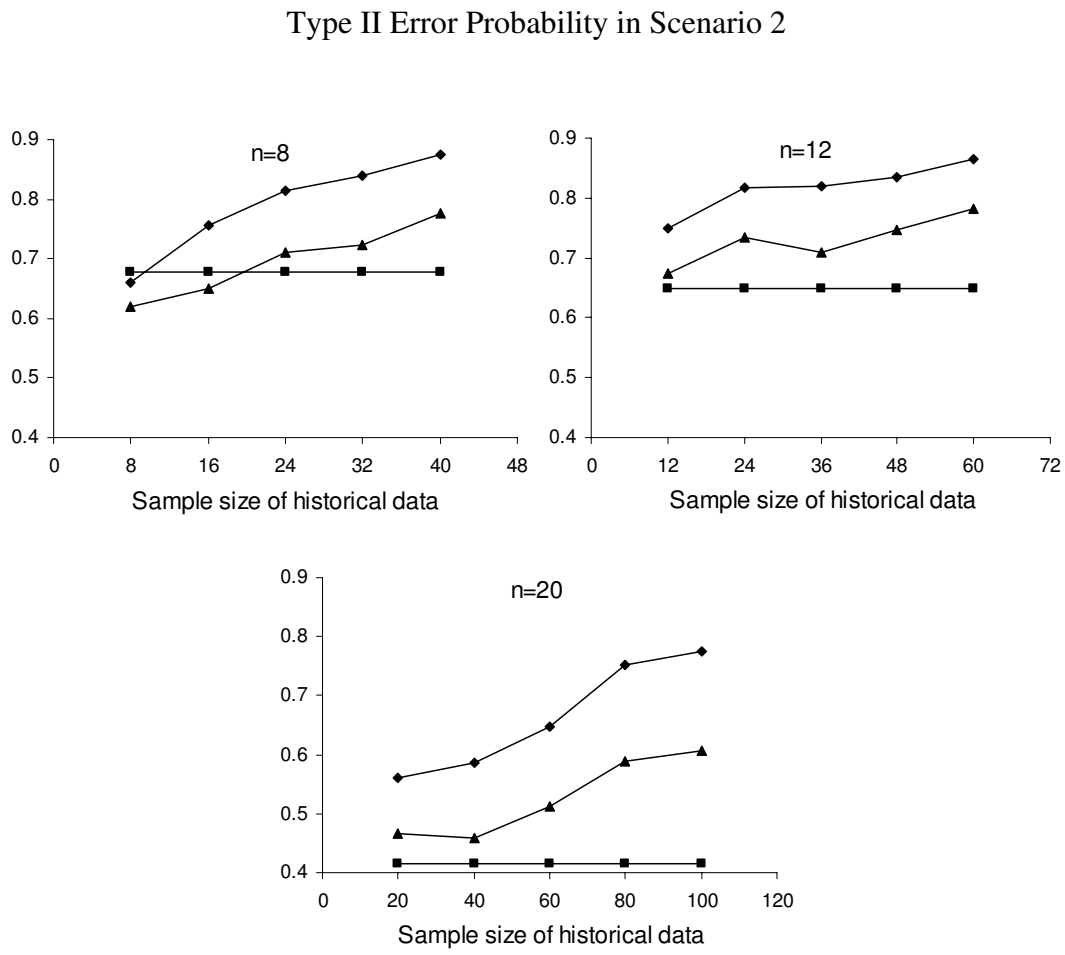

Figure 6.2: Comparisons of Type II error probability using three procedures for the situation in which water was healthy before but is currently impaired. The three graphs are for $n = 8, 12$, and 20. In each graph, ■ is for the binomial and Bayesian methods; ♦ is for the power prior method with $p_0 = 0.05$; ▲ is for the power prior method with $p_0 = 0.1$, where p_0 is the true probability of violation for the past status of water.

I error is controlled by the cutoff, and hence is generally not reduced by increasing the sample size. On the other hand, when the water was healthy during the past period (true probability of violation is 0.05), the Type I error probability for the power prior method becomes smaller than that for the other two methods, and it decreases with increasing historical sample size.

Figure 6.4 shows that the Type II error probability for the power prior method is quite low relative to the Bayesian and binomial methods. Furthermore, the more historical data available, the lower the Type II error. This implies that the power prior may be adopted to

resolve the issue of lack of power for decisions with small sample sizes.

Figures 6.2 and 6.3 illustrate the results under the situations when the true impairment status of a water segment changes over time. When water was previously healthy but becomes currently impaired (Figure 6.2), the power prior method is prone to Type II error relative to the Bayesian and binomial methods. The Type II error probability using the power prior method increases in n_0 and decreases in p_0 (true probability of violation for the past condition of water). The results indicate that incorporating historical data from previously healthy water leads to making fewer decisions that there is impairment, so stronger evidence is needed to show that water is currently impaired. With more historical data or healthier water, it is harder to declare the current water impaired. This conclusion is also supported by the trend of Type I probability in Figure 6.1 (when water was previously healthy). Note that in Figures 6.1 and 6.2 the lines appear to be bumpy due to the discreteness of binomial distribution, since error probabilities are plotted against historical sample size.

When water condition that was previously declared impaired is restored to a healthy state (Figure 6.3), the Type I error probability for the power prior method is higher than that for the Bayesian or binomial method, and it increases with n_0. This means that when the water condition was previously impaired, we are more likely to continue to declare the site as impaired, so stronger evidence is needed to unlist the site. Using more historical data leads to more decisions that there is still impairment. The same effect can be seen in Figure 6.4 (when water was previously impaired).

When water was previously impaired and the impairment continues (Figure 6.4), it is interesting to note that although the Type II error probability decreases with p_0 in generally, it starts to increase when $n = 20$, $n_0/n \geq 3$ and $p_0 > 0.3$. This is because when the discrepancy between historical and current data is large and at the same time a lot of historical data are incorporated, the power parameter automatically becomes smaller to control the influence of historical data on inference. A similar trend shows up in Figure 6.3. However, this trend does not appear when $n = 8$ or 12 in Figures 6.3 and 6.4 because of the flexible adjustment made through the power parameter. The need for additional (historical) information is stronger with smaller current sample size, and therefore the influence of historical current diminishes with its discrepancy from current data in a slower rate.

This investigation reveals that the use of power priors is beneficial in improving the power given that historical and current data are from water segments with similar quality. On the other hand, when the water quality changes over time, the history of impairment

influences the listing process and it is more difficult to change a decision than if the historical data were ignored. It is easy to see that similar conclusions hold when the "historical" data refer to the measurements observed at adjacent stations.

6.3.2 Error Probabilities under Various x_0/n_0

Although the investigation of error probabilities under various p_0 is theoretically important for establishing the advantages of the power prior method, in practice we would never know the true probability of violation for a water segment. Instead, with a set of historical data at hand, practitioners are more interested in whether and how much the error probabilities would be improved by switching to the power prior method.

x_0/n_0 is the observed proportion of violations in the historical sample. Let p^* and p^{**} be defined the same as in Section 6.3.1. Using the posterior cutoff method, the Type I error probability when x_0 violations are observed in the historical sample is computed as

$$\begin{aligned} &P(P(H_0|D_0, D) < q|x \sim \text{binomial}(n, p^*)) \\ &= \sum_{x=0}^{n} I(P(p \leq p^*|D_0, D) < q) f(x|n, p^*), \end{aligned} \tag{6.6}$$

and the Type II error probability is

$$\begin{aligned} &P(P(H_0|D_0, D) \geq q|x \sim \text{binomial}(n, p^{**})) \\ &= \sum_{x=0}^{n} I(P(p \leq p^*|D_0, D) \geq q) f(x|n, p^{**}). \end{aligned} \tag{6.7}$$

The exact Type I and Type II error probabilities for the power prior method were calculated using equations (6.6) and (6.7) for various historical sample proportion of violation. These error probabilities along with those for Bayesian and binomial methods are presented in Figures 6.5 and 6.6. As in Section 6.3.1, for estimation of Type I error probabilities in Figure 6.5 the current water quality is assumed to meet standards ($p^* = 0.1$), while for the estimation of Type II error probabilities in Figure 6.6 the current water quality is assumed not to meet standards ($p^{**} = 0.25$).

Figures 6.5 and 6.6 show that if the proportion of violations in historical data is less than 0.13, the use of power prior would reduce Type I error but increase Type II error. Therefore incorporating a historical sample with a low rate of violations from a site would

make it harder to declare it impaired. On the other hand, if the historical sample proportion is greater than 0.13, the use of power prior would reduce the Type II error but increase the Type I error. This implies that we are more likely to list a site after using a historical sample with a high violation rate. Figures 6.5 and 6.6 graphically display the gain and loss in error probabilities from incorporating historical data. The graphs provide a useful reference for practitioners to decide whether to use a specific set of historical data. Using equations 6.6 and 6.7, we can easily compute error probabilities for any situation not listed in the graphs (any combination of n, n_0, x_0, p^*, and p^{**}). Note that the more historical data available, the more rapidly the error probabilities change with historical sample proportion, and therefore the historical information has a greater influence on the listing decision.

The investigation of error probabilities reveals that the binary methods currently in use may be improved by adopting power priors.

6.4 Applications

We present two examples to demonstrate the influence of historical information on the listing decision through a power prior. As the first example we consider dissolved oxygen (DO) data collected at site A, with data in years of 2000 and 2001 as the current sample and those in years of 1992 to 1999 as the historical sample. The values are plotted in Figure 6.7. Of interest in these data is the determination of whether the DO values indicate that site A violates a (lower) standard of 5.0 more than 10% of the time (i.e. $p^* = 0.1$). As the second example we consider pH data collected at site B, with data in year of 1999 and 2000 as the current sample and those in year of 1992 to 1998 as the historical sample. The values are plotted in Figure 6.8. Of interest in these data is the determination of whether the pH values indicate that site B violates a (lower) standard of 6.0 or an (upper) standard of 9.0 more than 10% of the time (i.e. $p^* = 0.1$). Site A is located on William creek and site B is on Chopawamsic creek, both of which are part of Potomac and Shenandoah river basin in Virginia.

Table 6.1 summarizes the information about the two sites. In the historical sample, there are 20.6% violations for site A and only 2.8% violations for site B. Thus site A is an example of the scenario in which a water segment was previously impaired; site B is an example of the scenario in which a water segment would be considered healthy.

Table 6.2 provides the number of violations needed to list the site using EPA's raw score

Table 6.1: *Summary of data collected at sites A and B.*

	Site A	Site B
measurement of interest	DO	pH
standard	$5 \leq$ DO	$6 \leq$ pH ≤ 9
size of current sample	8	20
No. of violations in current sample	1	6
size of historical sample	34	71
No. of violations in historical sample	7	2
proportion of violation in historical sample	20.6%	2.8%

Table 6.2: *Number of violations needed to list the site using EPA's raw score method, binomial method, Bayesian method, and power prior method.* $\alpha = q = 0.05$.

	Site A	Site B
EPA's raw score method	1	2
Binomial method	3	5
Bayesian method	3	5
Power prior method	2	7

method, binomial method, Bayesian method, and power prior method. For the binomial method, the nominal type I error α is set at 0.05; for the Bayesian and power prior method, the posterior cutoff q is set at 0.05. As mentioned in Section 6.3.1, binomial and Bayesian methods lead to the same decision and the same Type I and Type II error probabilities when $n = 8$ or 20, with $\alpha = q = 0.05$. Table 6.2 shows that it is most likely to declare a site impaired with EPA's raw score method, which is expected according to the discussion made in Smith *et al.* (2001). In the site A example, the power prior method takes into account the previously high violation rate at site A, so less violations are needed in the current sample to list site A (or keep the site listed) using the power prior than using the binomial and Bayesian methods. On the other hand, site B was quite healthy, and therefore more evidence is needed in the current sample to prove this site impaired.

Only one violation is observed in the current sample of site A, so with any method except the raw score method the declaration of impairment would not be made. Six violations are present in the current sample of site B. Hence, site B would be listed using the raw score,

binomial and Bayesian methods, but not listed using the power prior method. In Figure 6.8, note that among the six violations observed in the current data at site B, five violations happened in the first year of the two-year period, but only one violation happened in the second year. Therefore further investigation of the reason for the dramatic decrease in violations may help the assessor to make a more informative decision.

In our power prior approach, the power parameter, δ, is viewed as a random variable that is integrated out to make the inference independent of the choice of δ. Because of the dependence on prior data, it may be useful to monitor the influence of the historical sample on the decision by fixing δ at a series of values between 0 and 1. Figure 6.9 illustrates how the posterior probability of H_0 (the site is not impaired) is affected by the weight put on the historical data, represented by δ. For site A, the violation rate is high in the historical sample, so the more historical data used in the analysis, the lower the probability that site A is not impaired. At site B, the historical data show that this water segment had good quality previously, so we are more confident that site B is not impaired when more historical data is taken into account.

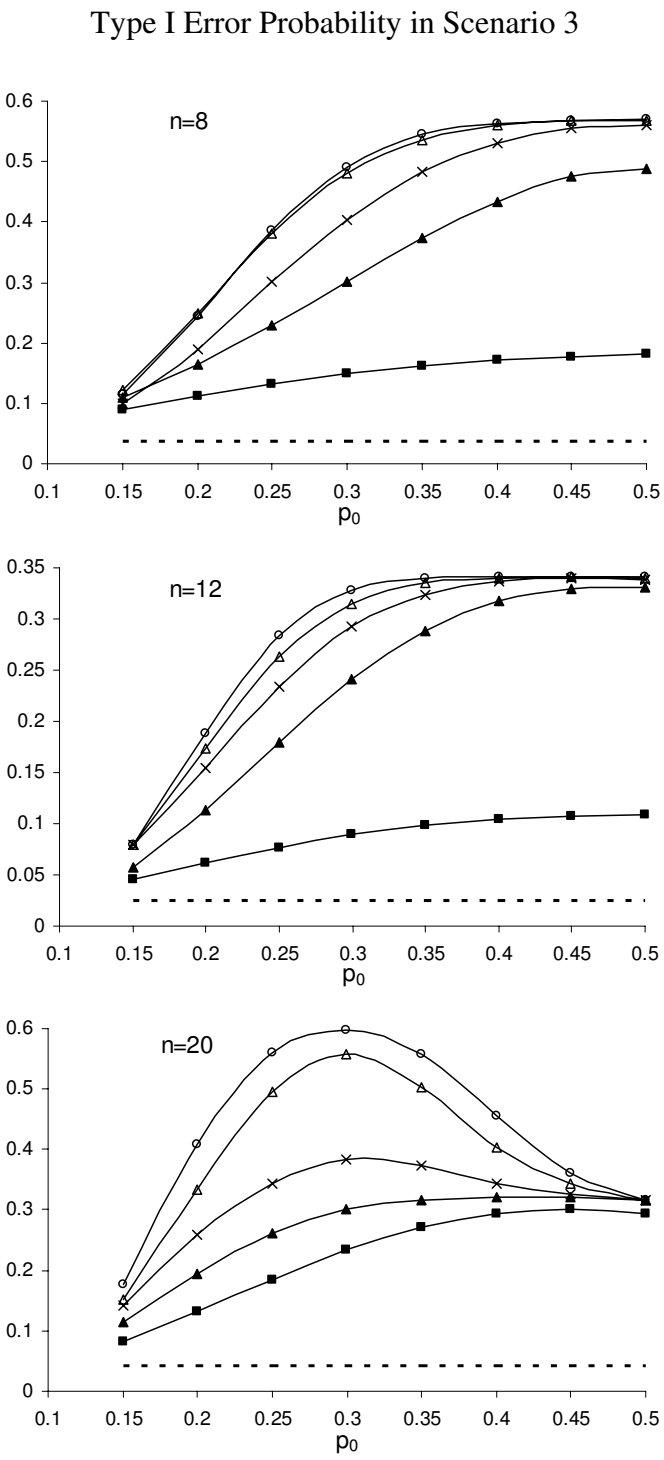

Figure 6.3: Comparisons of Type I error probability using the binomial and Bayesian methods (dotted line) and the power prior method (solid lines) for the situation in which water was impaired before but is currently healthy. The three graphs are for $n = 8, 12$, and 20. In each graph, the Type I error probabilities derived from the power prior method with different sample size ratios (n_0/n) are shown with different symbols. (■ $n_0/n = 1$, ▲ $n_0/n = 2$, × $n_0/n = 3$, △ $n_0/n = 4$, and ⊙ $n_0/n = 5$.)

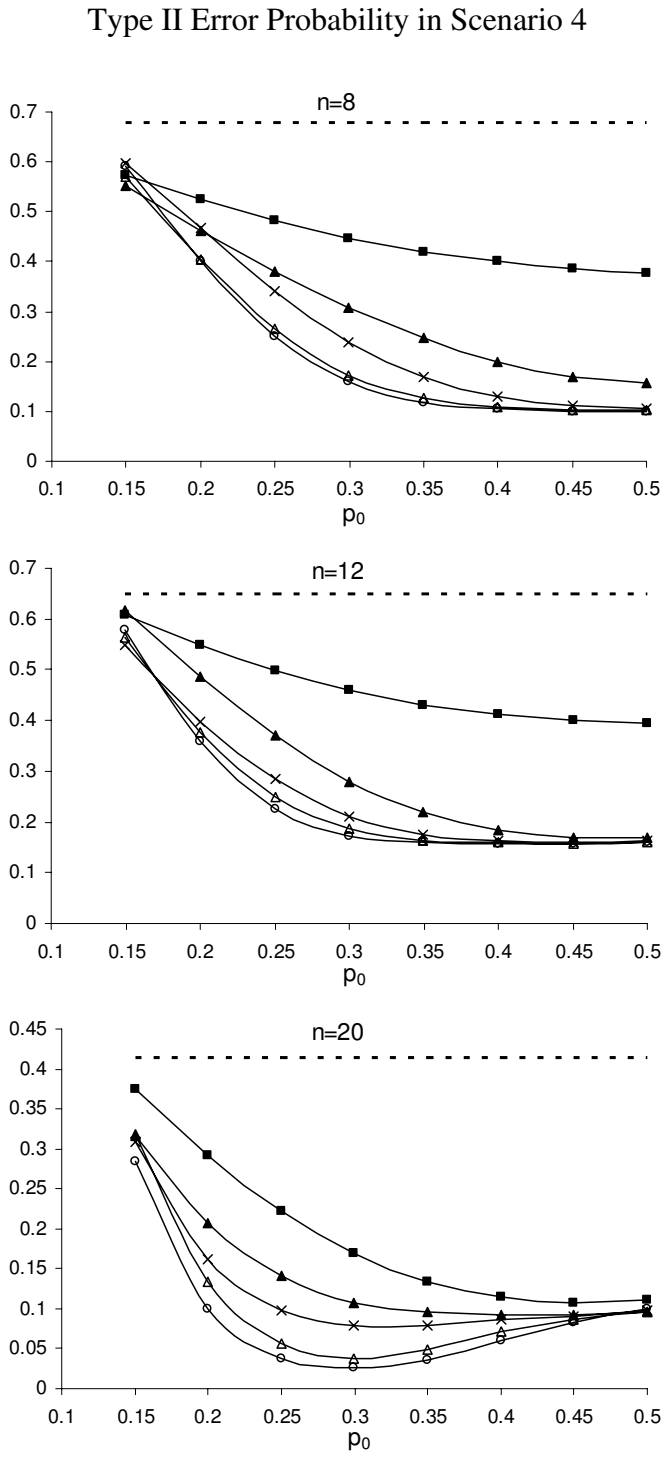

Figure 6.4: Comparisons of Type II error probability using the binomial and Bayesian methods (dotted line) and the power prior method (solid lines) for the situation in which water was impaired and is still impaired. The three graphs are for $n = 8, 12$, and 20. In each graph, the Type II error probabilities derived from the power prior method with different sample size ratios (n_0/n) are shown with different symbols. ($\blacksquare$ $n_0/n = 1$, $\blacktriangle$ $n_0/n = 2$, $\times$ $n_0/n = 3$, $\triangle$ $n_0/n = 4$, and $\odot$ $n_0/n = 5$.)

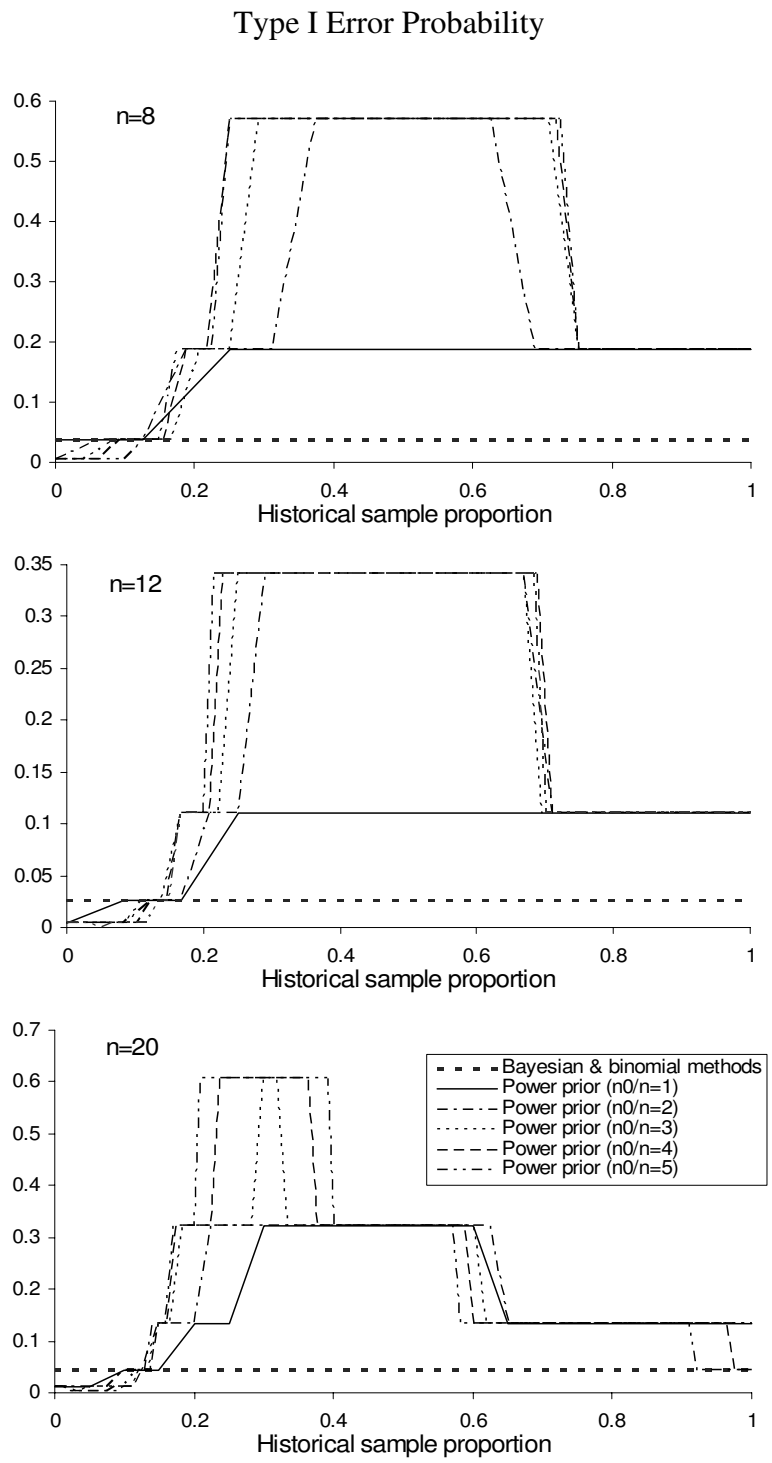

Figure 6.5: Type I error probabilities using the binomial and Bayesian methods and the power prior method with various sample proportion of violations in historical data. The three graphs are for $n = 8, 12$, and 20. In each graph, the Type I error probabilities derived from the power prior method with different sample size ratios (n_0/n) are shown with different types of lines (see the legend).

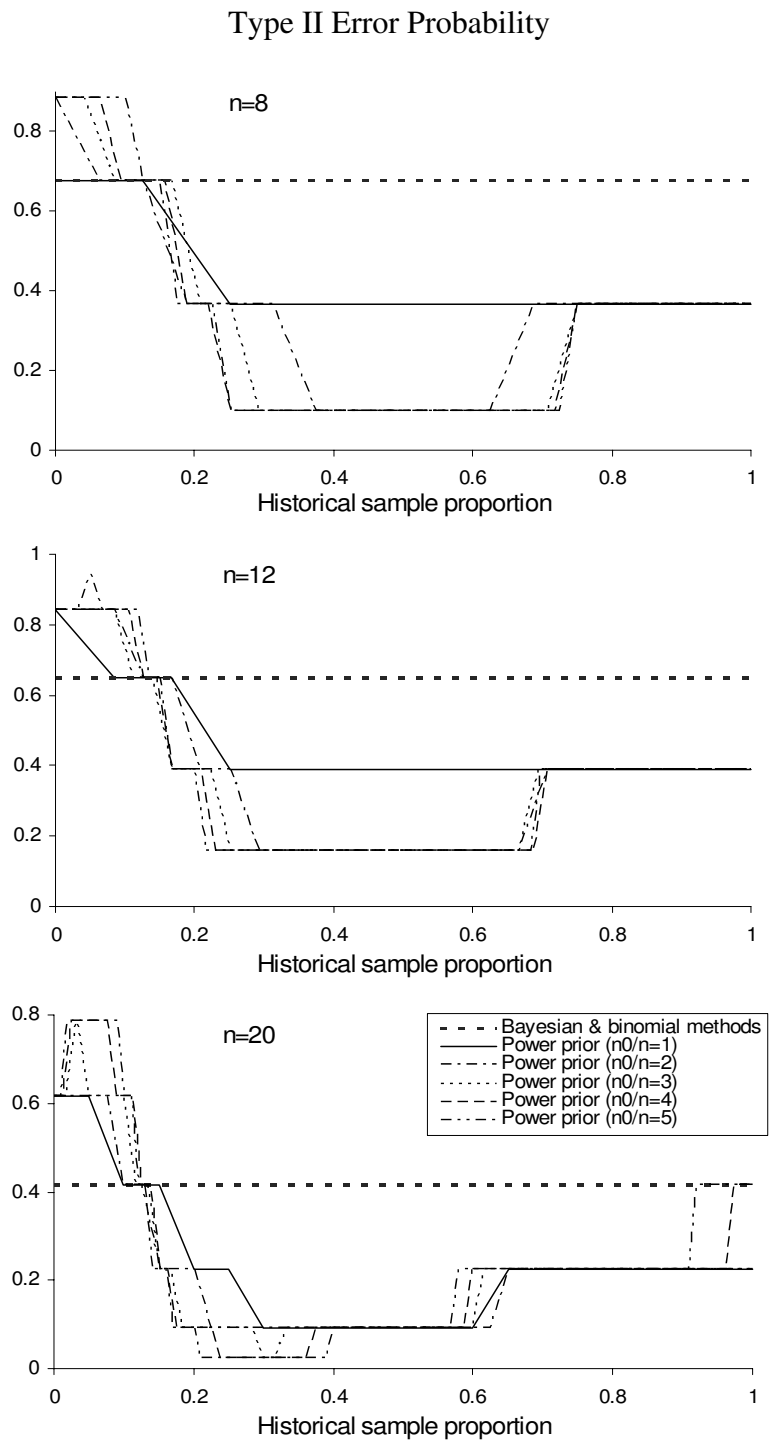

Figure 6.6: Type II error probabilities using the binomial and Bayesian methods and the power prior method with various sample proportion of violations in historical data. The three graphs are for $n = 8, 12$, and 20. In each graph, the Type I error probabilities derived from the power prior method with different sample size ratios (n_0/n) are shown with different types of lines (see the legend).

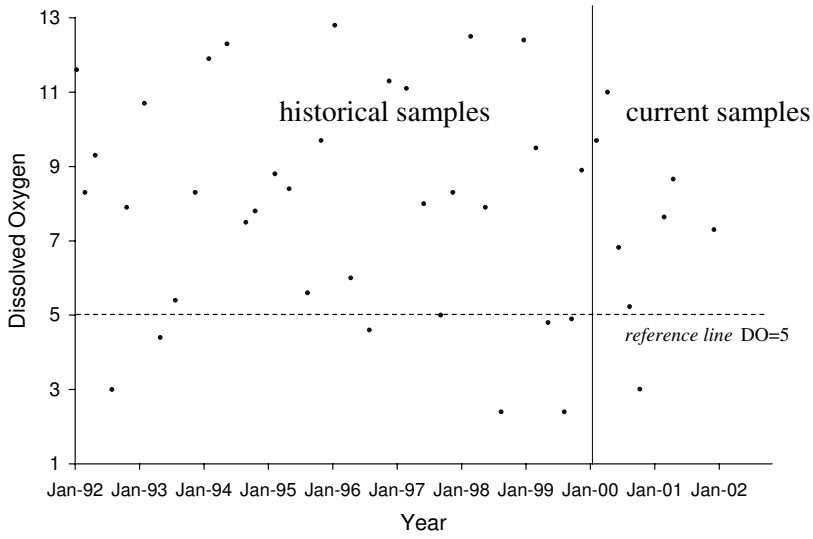

Figure 6.7: Plot of observations on dissolved oxygen collected at site A from 1992 to 2001.

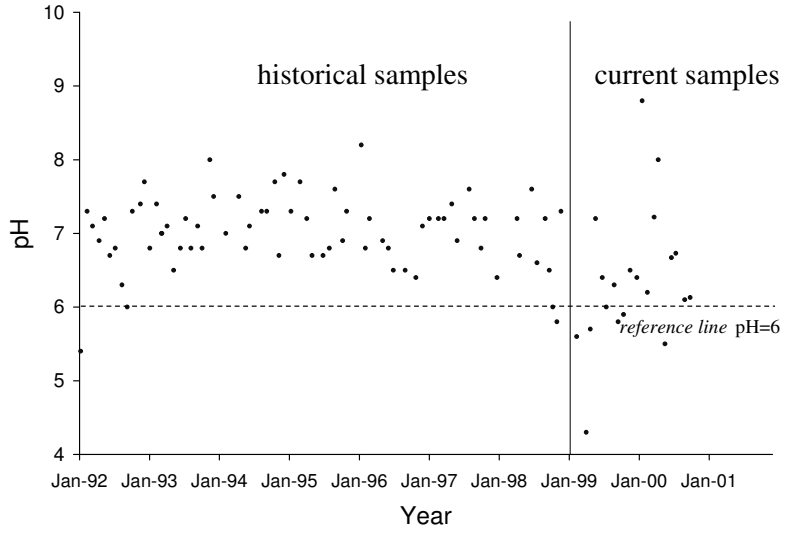

Figure 6.8: Plot of observations on pH collected at site B from 1992 to 2000.

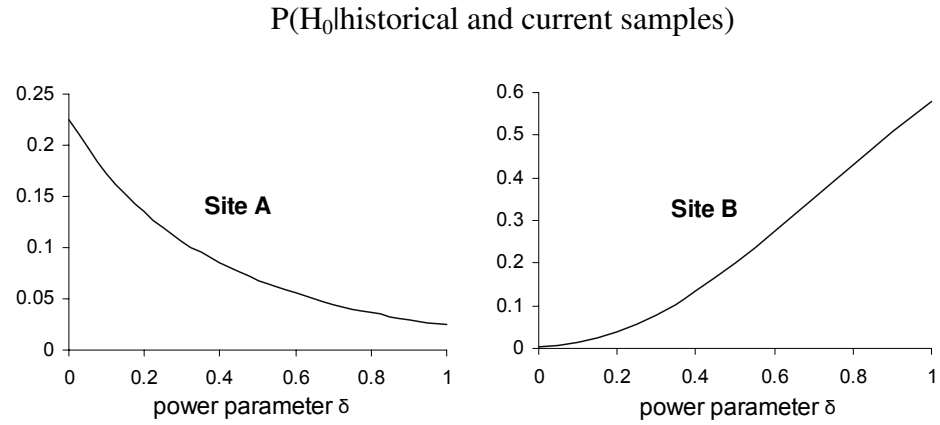

Figure 6.9: Posterior probability of H_0 (the site is not impaired) conditional on different values of δ (the power parameter) for sites A and B.

Chapter 7

Summary and Future Research

The modified power prior method provides a framework to incorporate data from alternative sources, whose influence on current study is automatically adjusted according to its availability and discrepancy from current data. As consequence of using more data, the power prior method has advantages in terms of power and estimation precision for decisions with small sample sizes. In water quality evaluations, the investigation of error probabilities reveals that the binary methods currently in use may be improved by adopting power prior approach. The power prior method may also be implemented on the numerical values of water quality measurements (e.g. pH values), resulting in more precise estimation.

On the other hand, the power prior may be viewed as a general class of informative priors in Bayesian inference. The power prior is elicited to take into account the heterogeneity between historical and current data when we are not able to describe or adequately model such heterogeneity explicitly. The power priors are semi-automatic, in a sense that they always take the form of raising the likelihood function $L(D_0|\theta)$ to a fractional power, regardless of the specific form of heterogeneity between D_0 and D. The fact that we often do not have enough knowledge to model such heterogeneity or to specify a fixed power makes the power prior with a random power parameter δ especially attractive in practice.

The power prior with a random power parameter is very flexible in determining the role of historical data. The subjective information about the difference in two populations is incorporated by adjusting the hyperparameters in the prior for δ; and the discrepancy between two samples is automatically taken into account through a random δ. An extreme case would be when researchers believe that the current and historical data are from the same population, in which case $\delta = 1$ should be used even when two samples seem incompatible.

In this dissertation we propose a modified joint power prior distribution for (θ, δ) to address the inconsistency of the original power prior with the likelihood principle, where θ is the parameter of interest. The modified and original approaches are essentially the same when the power parameter is fixed. Therefore the modified power prior shares all the nice properties of the original one discussed in a series of papers by Ibrahim and Chen ([18], [19], and [10]), such as generality of this methodology, optimality from the aspect of information processing, flexibility in expressing the uncertainty about the power parameter, and broad applications. In addition, the degree in which the historical data affect the posterior distributions via the power parameter is adjusted automatically in the modified approach, based on the compatibility between the historical and current samples, and also based on their sample sizes. With the modified power prior, the power parameter behaves in a sensible and desirable way. However, the original power prior approach underestimates the influence of historical data on the current study in general and therefore little benefits are gained from incorporation of historical data.

Furthermore, empirical evidence shows that the modified power prior leads to smaller MSE for estimated θ than the original one, when the divergence between historical and current populations is small to moderate. When such divergence is large, the MSE from the modified power prior is larger than that from the original one, sometimes even larger than the MSE obtained using no historical data. This limitation of the modified power prior method actually prevents researchers from abusing this method. The performance of estimation will be penalized if we incorporate a "historical" data set picked arbitrarily. This result indicates that the modified power prior approach is more liberal compared to the original one, so it needs to be implemented with caution, especially when the populations underlying current and historical data truly have little similarity.

We need to point out the following two situations when the power prior method is not preferred. First, to increase the estimating efficiency, we should model the heterogeneity between historical and current samples explicitly whenever possible. Second, the power prior is most useful when the size of current data set is small but ample historical data are available. If the current data set is large enough to achieve a satisfactory precision for parameter estimates, the benefit from historical data is diminished. On the other hand, the accuracy of estimates lost would be more than the precision gained if the historical data are used in such a situation.

Our research shows that the modified power prior approach has improved performance compared to the original approach. However, further research is needed to investigate pos-

sible modifications, e.g. a mixture power prior, to adapt with different levels of divergence between current and historical populations. This dissertation focuses on point estimation in comparison of two methods. An investigation of the role of power priors on interval estimation is warranted.

This dissertation builds a basic framework for the modified power prior method. There is plenty of opportunity for future research on its applications. Since the power prior is a general class of prior distributions, it could be quite useful in a wide variety of applications, including many kinds of regression and survival models. The power priors can also be used in model selection context since they automate the prior elicitation procedure. In addition, more research can be done on modelling the power parameter as a function of time, which is required for the implementation of time-weighted power priors.

Bibliography

[1] Barnett, V., and O'Hagan, A. *Setting Environmental Standards*. Chapman and Hall: London, 1997.

[2] Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer Verlag: New York, 1985.

[3] Berger, J. O., and Bernardo, J. M. Estimating a product of means: Bayesian analysis with reference priors. *J. Amer. Statist. Assoc. 84* (1989), 200–207.

[4] Berger, J. O., and Bernardo, J. M. Reference priors in a variance components problem. In *Bayesian Analysis in Statistics and Econometrics*, P. Goel and N. Iyengar, Eds. New York: Springer Verlag, 1992a.

[5] Berger, J. O., and Bernardo, J. M. Ordered group reference priors with application to a multinomial problem. *Biometrika 79* (1992b), 25–37.

[6] Berger, J. O., and Bernardo, J. M. On the development of the reference prior method. In *Bayesian Analysis*, J.M.Bernardo, J.O.Berger, D.V.Lindley, and A.F.M.Smith, Eds., vol. 4. London: Oxford University Press, 1992c.

[7] Bernardo, J. M. Reference posterior distributions for Bayes inference. *J. Roy. Statist. Soc. Ser B 41* (1979), 113–147.

[8] Casella, G., and Berger, R. L. *Ststistical Inference*, 2nd ed. Duxbury, 2001.

[9] Chen, M.-H. Importance-weighted marginal Bayesian posterior density estimation. *Journal of the American Statistical Association 89* (1994), 818–824.

[10] Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference 84* (2000), 121–137.

[11] Chen, M.-H., Ibrahim, J. G., Shao, Q.-M., and Weiss, R. E. Prior elicitation for model selection and estimation in generalized linear mixed models. *Journal of Statistical Planning and Inference 111* (2003), 57–76.

[12] Diaconis, P., and Ylvisaker, D. Conjugate priors for exponential families. *Ann. Statist. 7* (1979), 269–281.

[13] Duan, Y., Smith, E. P., and Ye, K. Power prior approach to the binomial test in water quality assessment. *Journal of Agricultural, Biological, and Environmental Statistics* (2005), under revision.

[14] Duan, Y., Ye, K., and Smith, E. P. Evaluating water quality: Using power priors to incorporate historical information. *Environmetrics* (2005), to appear.

[15] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. *Bayesian Data Analysis*, 2nd ed. Chapman Hall and CRC Press, Boca Raton, Florida, 1995.

[16] Hirsch, R. M., Slack, J. R., and Smith, R. A. Techniques for trend analysis for monthly water quality data. *Water Resources Research 18* (1982), 107–121.

[17] Ibrahim, J. G., and Chen, M.-H. Prior distributions and bayesian computation for proportional hazards models. *Sankhya Ser. B 60* (1998), 48–64.

[18] Ibrahim, J. G., and Chen, M.-H. Power prior distributions for regression models. *Statistical Science 15* (2000), 46–60.

[19] Ibrahim, J. G., Chen, M.-H., and Sinha, D. On optimality properties of the power prior. *Journal of the American Statistical Association 98* (2003), 204–213.

[20] Jeffreys, H. An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Statistical Society of London* (1946), vol. 186 of *A*, pp. 453–461.

[21] Jeffreys, H. *Theory of Probability*. London: Oxford University Press, 1961.

[22] Kass, R. E., and Wasserman, L. The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc. 91* (1996), 1343–1370.

[23] Laplace, P. S. *Theorie Analytique des Probabilities*. Courcier, Paris, 1812.

[24] McBride, G. B., and Ellis, J. C. Confidence of compliance: A bayesian approach for percentile standards. *Water Research 35* (2001), 1117–1124.

[25] Morris, C. N. Natural exponential families with quadratic variance functions: statistical theory. *Ann. Statist. 11* (1983), 515–529.

[26] Peers, H. W. On confidence points and bayesian probability points in the case of several parameters. *J. Roy. Statist. Soc. Ser B 27* (1965), 9–16.

[27] Searle, S. R., Casella, G., and McCulloch, C. E. *Variance Components*. New York: John Wiley & Sons, 1992.

[28] Severini, T. A. On the relationship between bayesian and nonbayesian interval estimates. *J. Roy. Statist. Soc. Ser B 53* (1991), 611–618.

[29] Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal 27* (1948), 379–423.

[30] Smith, E. P., Ye, K., Hughes, C., and Shabman, L. Statistical assessment of violations of water quality standards under section 303(d) of the clean water act. *nvironmental Science and Technology 35* (2001), 606–612.

[31] Smith, E. P., Zahran, A., Mahmoud, M., and Ye, K. Evaluation of water quality using acceptance sampling by variables. *Environmetrics 14* (2003), 373–386.

[32] Smith, R. W. The use of random-model tolerance intervals in environmental monitoring and regulation. *Journal of Agricultural, Biological and Environmental Statistics 7* (2002), 74–94.

[33] Stein, C. On the coverage probability of confidence sets based on a prior distribution. In *In Sequential Methods in Statistics*, vol. 16 of *Banach Center Publications*. Warsaw: PWN-Polish Scientific Publishers, 1985, pp. 485–514.

[34] Sweeting, T. J. On the implementation of local probability matching priors for interest parameters. Tech. rep., University College London, WC1E 6BT, UK, 2004.

[35] Thompson, M. L., Cox, L. H., Sampson, P. D., and Caccia, D. C. Statitical hypothesis testing for U.S. environmental regulatory standards for ozone. *Environmental and Ecological Statistics 9* (2002), 321–339.

[36] Tibshirani, R. Noninformative priors for one parameter of many. *Biometrika 76* (1989), 604–608.

[37] Wallis, W. A. Use of variables in acceptance inspection for percent defective. In *Selected Techniques of Statistical Analysis for Scientific and Industrial Research and Production and Management Engineering*. McGraw Hill: New York, 1947, pp. 3–93.

[38] Welch, B. L., and Peers, H. W. On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. Ser B 25* (1963), 318–329.

[39] Ye, K., and Berger, J. O. Noninformative priors for inferences in exponential regression models. *Biometrika 78* (1991), 645–656.

[40] Ye, K., and Smith, E. P. A bayesian approach to evaluating site impairment. *Environmental and Ecological Statistics 9* (2002), 379–392.

[41] Zellner, A. *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley, 1971.

[42] Zellner, A. Maximal data information prior distributions. In *New Developments in the Applications of Bayesian Methods*, A. Aykac and C. Brumat, Eds. Amsterdam: North-Holland, 1977, pp. 201–215.

[43] Zellner, A. Optimal information processing and bayes's theorem. *The American Statistician 42* (1988), 278–284.

[44] Zellner, A. Bayesian methods and entropy in economics and econometrics. In *Maximum Entropy and Bayesian Methods*, W. Grandy and L. Schick, Eds. Boston: Kluwer, 1991, pp. 17–31.

[45] Zellner, A., and Min, C. Bayesian analysis, model selection and prediction. In *Physics and Probability: Essays in Honor of Edwin T. Jaynes*, W. J. Grandy and P. Milonmi, Eds. U.K.: Cambridge University Press, 1993, pp. 195–206.

Vita

The author, Yuyan Duan, was born on May 12, 1975 in Yichang, P. R. China. She received her Bachelor degree of Economics, in 1995 from Renmin University of China. From 1995 to 1999 she worked in China State Shipbuilding Corporation as an operational research analyst. In 2001 she began her studies in statistics in the Department of Statistics of Virginia Tech, where she received her M.S. degree in 2002 and Ph.D. degree in 2005.