Selecting likelihood weights by cross-validation. [1]

Xiaogang Wang and James V. Zidek [2]
Department of Mathematics and Statistics, York University
Department of Statistics, University of British Columbia,

Abstract

The (relevance) weighted likelihood was introduced to formally embrace a variety of statistical procedures that trade bias for precision. That likelihood unlike its classical counterpart, combines all relevant information while inheriting many of its desirable features including good asymptotic properties. However, to be effective, the weights involved in its construction need to be judiciously chosen. Choosing those weights is the subject of this paper in which we demonstrate the use of cross-validation. We prove the resulting weighted likelihood estimator (WLE) to be weakly consistent and asymptotically normal. Results of simulation studies have shown that WLE out-performs the traditional maximum likelihood estimator for small and moderate samples. An application to disease mapping data is also demonstrated.

1 Introduction

The weighted likelihood (WL for short) has been developed for a variety of purposes. Moreover, it shares its underlying purpose with other methods such as weighted least squares and kernel smoothers. Like the WL they can reduce an estimator's variance while increasing its bias to reduce mean-squared- error (MSE), *i.e.* increase its precision. However, the achievement of those gains depends on choosing the weights well, the subject of this paper. More specifically, we show they may be data dependent (*i. e.* "adaptive") and chosen by *cross-validation.* However, the idea of data dependent weights goes back at least to the celebrated James-Stein estimator, a WL estimator with adaptive weights that does successfully trades bias for variance (Hu and Zidek 2002).

To describe the WL, suppose we observe independent random response vectors $\mathbf{X_1}, ..., \mathbf{X_m}$ with probability density functions $f_1(.;\theta_1), ..., f_m(.;\theta_m)$, where $\mathbf{X_i} = (X_{i1}, ..., X_{in_i})^t$. Further suppose that only population 1, in particular θ_1, an unknown vector of parameters, is of inferential interest. Given data, $\mathbf{X} = \mathbf{x}$, the classical likelihood would be

$$L_1(\mathbf{x_1}, \theta_1) = \prod_{j=1}^{n_1} f(x_{1j}; \theta_1).$$

[1] This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.
AMS 2000 Classifications: 62F10; 62H12
STMA 2000 Classifications: 04:010; 01:180; 04:170
Keywords: Asymptotic normality; consistency; cross-validation; weighted likelihood
[2] *E-mail addresses:* stevenw@mathstat.yorku.ca (Steven Xiaogang Wang), jim@stat.ubc.ca (James V. Zidek).

An alternative, the WL, obtains when the remaining parameters, $\theta_2, ..., \theta_m$, are thought to resemble θ_1,

$$\mathrm{WL}(\mathbf{x}; \theta_1) = \prod_{i=1}^{m} \prod_{j=1}^{n_i} f_1(x_{ij}; \theta_1)^{\lambda_i},$$

where $\lambda = (\lambda_1, ..., \lambda_m)$, the " weights vector", must be specified. Notice that the parameters from the remaining populations, $\theta_2, ..., \theta_m$, unlike the data they generate, do not appear in the WL, since inferential interest focuses on θ_1. It follows that

$$log \mathrm{WL}(\mathbf{x}; \theta_1) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \lambda_i \; log \; f_1(x_{ij}; \theta_1).$$

The WL extends the local likelihood method of Tibshirani and Hastie (1987) for nonparametric regression, although the idea predates them (see Hu and Zidek 2002 for a review). Following Hu (1997), Hu and Zidek (1995, 2001, 2002) extend the local likelihood to a more general setting. However, the aim is the same, that of combining all relevant information in samples from populations thought to resemble the one of concern.

The maximum WL estimator (WLE) for θ_1 say $\tilde{\theta}_1$ is defined by

$$\tilde{\theta}_1 = arg \sup_{\theta_1 \in \Theta} \mathrm{WL}(\mathbf{x}; \boldsymbol{\theta_1}).$$

In many cases the WLE may be obtained by solving the *estimating equation:*

$$(\partial / \partial \theta_1) \; log \; \mathrm{WL}(\mathbf{x}; \theta_1) = 0.$$

Note that uniqueness of the WLE is not assumed.

Like the MLE, the WLE has a number of desirable properties (Hu and Zidek 2002), in particular consistency and asymptotic normality under reasonably regularity conditions (Hu 1997; Wang, van Eeden and Zidek 2002). However, those asymptotic properties have only been shown with fixed weights and hence need to be extended in this paper to cover the estimators we obtain using cross - validation.

In its most primitive but nevertheless useful form, cross validation consists of controlled and uncontrolled division of the data sample into two subsets. For example, the subsets can be derived by deleting one or more observations or it can be a random sample from the dataset. Stone (1974) began the systematic study of cross-validatory choice and assessment in statistical prediction. Both Stone (1974) and Geisser (1975) discuss its application to the *K-group* problem and use a linear combination of the sample means from the different groups to estimate a common mean. Breiman and Friedman (1997) also demonstrate the benefit of using cross-validation to obtain linear combinations of predictors that perform well in multivariate regression.

The paper is organized as follows. The adaptive weights are derived in Section 2. The asymptotic properties of the resulting WLE are presented in Section 3. Results of simulations studies are discussed in Section 4. In Section 5, an application to disease mapping data demonstrates the

benefits of using the proposed method in conjunction with the WLE when compared with traditional estimators.

2 Choosing Adaptive Weights

For cross-validation, there are many ways of dividing the entire sample into subsets such as a random selection technique. However, we use the simplest *leave-one-out* approach in this section since the analytic forms of the optimum weights are then completely tractable for the linear WLE.

Suppose that we have m populations which might be related to each other. The probability density functions or probability mass functions are of the form $f_i(x;\theta_i)$ with θ_i as the parameter for population i. Assume that

$$\begin{array}{llll} X_{11}, & X_{12}, & X_{13}, \ldots, & X_{1n_1} \overset{i.i.d}{\sim} f_1(x;\theta_1) \\ X_{21}, & X_{22}, & X_{23}, \ldots, & X_{2n_2} \overset{i.i.d}{\sim} f_2(x;\theta_2) \\ \vdots & & & \vdots \\ X_{m1}, & X_{m2}, & X_{m3}, \ldots, & X_{mn_m} \overset{i.i.d.}{\sim} f_m(x;\theta_m) \end{array}$$

where, for fixed i, the $\{X_{ij}\}$ are observations obtained from population i and so on. Assume that observations obtained within each population are independent and identically distributed. Further assume that observations from one population are independent of those from other populations except that $Corr(X_{ij}, X_{kj}) = \rho$. That is, observations obtained at the same time point or having the same second sub-scripts are not necessarily independent even though they are obtained from different populations. This would allow a spatial correlation structure but not a temporal one. We also assume that $E(X_{ij}) = \phi(\theta_i) = \phi_i$ say for $j = 1, 2, ..., n_i$. The population parameter of the first population, θ_1, is of inferential interest.

We denote the vector of parameters and the weight vector by $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)$ respectively. Let $\boldsymbol{\lambda}_e^{opt}$ and $\boldsymbol{\lambda}_u^{opt}$ denote optimum weight vectors determined in the sequel for samples with equal and unequal sizes, respectively. We require that $\sum_{i=1}^{m} \lambda_i = 1$ in this paper.

Our cross-validatory approach to estimating the weights for the WLE flows from taking prediction as our inferential objective. In other words we seek an estimator, $\hat{\theta}_1$, of θ_1 that enables us to predict accurately, in some sense, a randomly drawn element, X_1^*, from the first population. But how should the precision of $\hat{\theta}_1$ be assessed?

One answer is the expected log score. Denoting by 'E', expectation with respect to the conditional distribution of X_1 given θ_1, that score is $E[log\ f_1(X_1|\hat{\theta}_1)]$, an index of $\hat{\theta}_1$'s performance.

However, the complexity of that index makes its use impractical in applications such as that in Section 5. We therefore adopt an approximation as a compromise. To obtain that approximation, we assume a 1 - to - 1 mapping of θ_1 into (ϕ_1, τ_1) where the range of ϕ_1 contains that of X_1. In fact, with an abuse of notion we represent θ_1 by $\theta_1 = (\phi_1, \tau_1)$ and $\hat{\theta}_1$ in a similar way.

The approximation obtains under the assumptions,

$$\frac{\partial \log E[f_1(X_1|\hat{\theta}_1)]}{\partial \hat{\phi}_1}|_{\hat{\theta}_1=\theta_1} = 0,$$

and

$$\frac{\partial^2 \log E[f_1(X_1|\hat{\theta}_1)]}{\partial^2 \hat{\phi}_1}|_{\hat{\theta}_1=\theta_1} < 0$$

for all θ, all derivatives above along with the associated 3rd order derivative being assumed to exist. These assumptions obtain for the normal distribution, for example, and more importantly for our application in Section 5, the Poisson distribution. Under these assumptions, the first order term in a 3 term Taylor expansion of the expected log score vanishes. Therefore, ignoring irrelevant terms and factors, we obtain for any proposed estimator $\hat{\theta}_1$ as an approximation to its negative expected log score, $(\hat{\phi}_1 - \phi_1)^2$, as a measure of $\hat{\phi}_1$'s precision. Finally, for its empirical assessment, we estimate the unknown ϕ_1 in this measure by X_1. Moreover, we adopt that empirical measure in obtaining adaptive weights by cross-validation. To that end, we use $(-j)$ to indicate that the j-th item has been dropped from the sample, in a sense to be made more precise below.

Taking the usual path, we predict X_{1j} by $\phi(\tilde{\theta}_1^{(-j)})$, the WLE of its mean without using the X_{1j}. Note that $\phi(\tilde{\theta}_1^{(-j)})$ is a function of the weight vector $\boldsymbol{\lambda}$ by the construction of the WLE. A natural measure for the discrepancy of the WLE becomes:

$$D(\boldsymbol{\lambda}) = \sum_{j=1}^{n_1} \left(X_{1j} - \phi(\tilde{\theta}_1^{(-j)}) \right)^2. \tag{1}$$

The optimum weights are derived such that the minimum of $D(\boldsymbol{\lambda})$ is achieved for fixed sample sizes $n_1, n_2, ..., n_m$ and $\sum_{i=1}^{m} \lambda_i = 1$.

In the rest of this section, we will concentrate on linear WLE, i.e., linear combinations of maximum likelihood estimates. We remark that WLE takes the linear form for normal, Poisson and other members of the exponential family as shown by Wang (2001).

2.1 Linear WLEs for Equal Sample Sizes

Stone (1974) and Geisser (1975) discuss the application of the cross-validation approach to the so-called *K-group* problem. Suppose that the data set S consists of n observations in each of K groups. The mean predictor for the ith group is:

$$\hat{\mu}_i = (1-\alpha)\overline{X}_{i\cdot} + \alpha \overline{X}_{\cdot\cdot}.$$

where $\overline{X}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ and $\overline{X}_{\cdot\cdot} = \frac{1}{K}\sum_{i=1}^{m} \overline{X}_{i\cdot}$. If our interest focuses on group 1, the relevant predictor is

$$\hat{\mu}_1 = \left(1 - \frac{K-1}{K}\alpha\right)\overline{X}_{1\cdot} + \sum_{i=2}^{m} \frac{\alpha}{K}\overline{X}_{i\cdot}.$$

We remark that the above formula is just a particular linear combination of the sample means. Stone (1974) uses cross-validation to derive an optimal value of α.

We consider more general linear combinations and throughout this subsection assume $n_1 = n_2 = ... = n_m = n$. Let $\tilde{\theta}_1^{(e)}$ denote the WLE obtained through cross-validation. If $\phi(\theta) = \theta$, the linear WLE for θ_1 is then defined as

$$\tilde{\theta}_1^{(e)} = \sum_{i=1}^{m} \lambda_i \overline{X}_{i.}$$

where $\sum_{i=1}^{m} \lambda_i = 1$.

In this subsection, we will use cross-validation by simultaneously deleting $X_{1j}, X_{2j}, ..., X_{mj}$ for each fixed j. That is, we delete one data point from each sample at each step. This might be appropriate if these data points are obtained at the same time point and strong associations exist among these observations. By simultaneously deleting $X_{1j}, X_{2j}, ..., X_{mj}$ for each fixed j, we might achieve numerical stability of the cross-validation procedure. An alternative approach is to delete a data point from only the first sample at each step. That approach will be studied in the next subsection.

Let $\overline{X}_{i.}^{(-j)}$ be the sample mean of the ith sample with jth element in that sample excluded. A natural measure for the discrepancy of $\tilde{\theta}_1$ might be:

$$\begin{aligned} D_e^{(m)} &= \sum_{j=1}^{n} \left(X_{1j} - \sum_{i=1}^{m} \lambda_i \overline{X}_{i.}^{(-j)} \right)^2 \\ &= c(\underline{\boldsymbol{X}}) - 2\boldsymbol{\lambda}^t b_e(\underline{\boldsymbol{X}}) + \boldsymbol{\lambda}^t A_e(\underline{\boldsymbol{X}})\boldsymbol{\lambda} \end{aligned}$$

where $c(\underline{\boldsymbol{X}}) = \sum_{j=1}^{n} X_{1j}^2$, $(b_e(\underline{\boldsymbol{X}}))_i = \sum_{j=1}^{n} X_{1j}\overline{X}_{i.}^{(-j)}$, and $(A_e(\underline{X}))_{ik} = \sum_{j=1}^{n} \overline{X}_{i.}^{(-j)}\overline{X}_{k.}^{(-j)}, i = 1, 2, ..., n, k = 1, 2, ..., m$.

An optimum weight vector obtained by using the cross-validation rule is defined to be a weight vector which minimizes the objective function, $D_e^{(m)}$ and satisfies $\sum_{i=1}^{m} \lambda_i = 1$. For expository simplicity, let $b_e = b_e(\underline{X})$ and $A_e = A_e(\underline{X})$ in this paper.

2.1.1 Two Population Case

For simplicity, first consider the simple case of just two populations, *i.e.*

$$\begin{aligned} X_{11}, \quad X_{12}, \quad X_{13}, \quad \ldots, \quad X_{1n} &\overset{i.i.d.}{\sim} f_1(x;\theta_1) \\ X_{21}, \quad X_{22}, \quad X_{23}, \quad \ldots, \quad X_{2n} &\overset{i.i.d.}{\sim} f_2(x;\theta_2) \end{aligned}$$

with $E(X_{1j}) = \theta_1$ and $E(X_{2j}) = \theta_2$. Let σ_1^2 and σ_2^2 denote the variances of X_{1j} and X_{2j} respectively. Let $\rho = cor(X_{1j}, X_{2j})$. Thus observations obtained at the same time point are not necessarily independent of each other. Let $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0)$ where θ_1^0 and θ_2^0 are the true values for θ_1 and θ_2 respectively.

We seek the optimum weights , $\lambda_1,\ \lambda_2$ with $\lambda_1 + \lambda_2 = 1$ that minimize the objective function defined as follows:

$$D_e^{(2)} = \sum_{j=1}^{n} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \lambda_2 \overline{X}_{2.}^{(-j)} \right)^2 - \gamma(\lambda_1 + \lambda_2 - 1).$$

Differentiating $D_e^{(2)}$ with respect to λ_1 and λ_2, we have

$$\frac{\partial D_e^{(2)}}{\partial \lambda_1} = -\sum_{j=1}^{n} \overline{X}_{1.}^{(-j)} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \lambda_2 \overline{X}_{2.}^{(-j)} \right) - \gamma = 0,$$

$$\frac{\partial D_e^{(2)}}{\partial \lambda_2} = -\sum_{j=1}^{n} \overline{X}_{2.}^{(-j)} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \lambda_2 \overline{X}_{2.}^{(-j)} \right) - \gamma = 0.$$

It follows that

$$\begin{cases} \lambda_1^{opt}(\mathbf{X}) = 1 - \dfrac{\sum_{j=1}^{n} \left(\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)} \right) \left(\overline{X}_{1.}^{(-j)} - X_{1j} \right)}{\sum_{j=1}^{n} (\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)})^2}, \\ \lambda_2^{opt}(\mathbf{X}) = \dfrac{\sum_{j=1}^{n} \left(\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)} \right) \left(\overline{X}_{1.}^{(-j)} - X_{1j} \right)}{\sum_{j=1}^{n} \left(\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)} \right)^2}. \end{cases} \tag{2}$$

Lemma 2.1 *The following identity holds:*

$$\lambda_1^{opt} = 1 - \lambda_2^{opt} \quad and \quad \lambda_2^{opt} = S_2^e / S_1^e,$$

where

$$\begin{aligned} S_1^e &= \frac{n(n-2)}{(n-1)^2} \left(\overline{X}_{1.} - \overline{X}_{2.} \right)^2 + \frac{1}{n(n-1)^2} \sum_{j=1}^{n} \left(X_{1j} - X_{2j} \right)^2, \\ S_2^e &= \frac{n}{(n-1)^2} \left(\hat{\sigma}_1^2 - \widehat{cov} \right) \end{aligned}$$

where $\hat{\sigma_1}^2 = \frac{1}{n} \sum_{j=1}^{n} \left(X_{1j} - \overline{X}_{1.} \right)^2$ *and* $\widehat{cov} = \frac{1}{n} \sum_{j=1}^{n} \left(X_{1j} - \overline{X}_{1.} \right) \left(X_{2j} - \overline{X}_{2.} \right)$.

The value of the λ_2^{opt} serves as the measure of relevance between the two samples. If that measure is almost zero, it implies that there is no need to combine the samples from the two populations. From the above formula, it can be see that small value of λ_2^{opt} would happen if the difference between the two sample means is relatively large or the correlation between the two samples is close to 1. The weights chosen by the cross-validation rule will then guard against the undesirable scenario in which too much bias might be introduced into the estimation procedure. On the other hand, if the second sample does contain valuable information about the parameter of interest, the cross-validation procedure will recognize that by assigning a non-zero value to λ_2^{opt}. Note that the knowledge of the variances and the correlation are not used to construct the adaptive weights.

Proposition 2.1 *If* $\rho < \frac{\sigma_1}{\sigma_2}$, *then*

$$P_{\theta^0}(\lambda_2^{opt} > 0) \xrightarrow{P_{\theta^0}} 1.$$

We remark that the condition $\rho < \sigma_1/\sigma_2$ is satisfied if $\sigma_2 < \sigma_1$ or $\rho < 0$. If the condition $\rho < \sigma_1/\sigma_2$ is not satisfied, then λ_2^{opt} might have negative sign for small n. However, the value of λ_2^{opt} will converge to zero as shown in the next Proposition.

Proposition 2.2 *If* $\theta_1^0 \neq \theta_2^0$, *then, for any given* $\epsilon > 0$,

$$P_{\theta^0}(|\lambda_1^{opt} - 1| < \epsilon) \longrightarrow 1 \quad and \quad P_{\theta^0}(|\lambda_2^{opt}| < \epsilon) \longrightarrow 1.$$

The asymptotic limit of the weights will not exist if θ_1^0 equals θ_2^0. This is because the cross-validation procedure will not be able to detect the difference of the population means since there is none. This can be rectified by defining $\lambda_2^{opt} = \frac{S_2^e}{S_1^e + c}$ where $c > 0$. An alternative solution is to put a threshold on λ_2, say $\frac{M}{n}$ where M is some positive constant. We remark that the knowledge of the variances and covariance is not assumed.

2.1.2 Alternative Matrix Representation of the Optimum Weights

To study the case of more than two populations, it is necessary to derive an alternative matrix representation of λ^{opt}. Define $e_n = \frac{n}{n-1}$. It can be verified that

$$\begin{aligned} \overline{x}_{i.}^{(-j)}\overline{x}_{k.}^{(-j)} &= (e_n\overline{x}_{i.} - \frac{1}{n-1}x_{ij})(e_n\overline{x}_{k.} - \frac{1}{n-1}x_{kj}) \\ &= e_n^2\overline{x}_{i.}\overline{x}_{k.} - \frac{e_n}{n-1}x_{ij}\overline{x}_{k.} - \frac{e_n}{n-1}x_{kj}\overline{x}_{i.} + (\frac{1}{n-1})^2 x_{ij}x_{kj}. \end{aligned}$$

Thus, we have

$$\sum_{j=1}^{n} \overline{x}_{i.}^{(-j)}\overline{x}_{k.}^{(-j)} = \left(e_n^2(n-2) + \frac{e_n}{n-1}\right)\ \hat{\theta}_i\hat{\theta}_k \ + \ \frac{e_n}{n-1}\ \widehat{cov}_{ik}^2, \tag{3}$$

where

$$\begin{aligned} \hat{\theta}_i &= \overline{x}_{i.}, \quad i = 1, 2, ..., m; \\ \widehat{cov}_{ik} &= \frac{1}{n}\sum_{j=1}^{n}(x_{ij} - \overline{x}_{i.})(x_{kj} - \overline{x}_{k.}). \end{aligned}$$

Recall that, for $1 \leq i \leq m$ and $1 \leq k \leq m$,

$$A_{e(ik)} = \sum_{j=1}^{n} \overline{x}_{i.}^{(-j)}\overline{x}_{k.}^{(-j)}.$$

By equation (**??**), it follows that

$$A_e = \frac{e_n}{n-1}\ \hat{\Sigma} \ + \ \left(e_n^2(n-2) + \frac{e_n}{n-1}\right)\ \hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}^t \tag{4}$$

where $\Sigma_{ik} = \widehat{cov}_{ik}$ and $\hat{\boldsymbol{\theta}} = (\overline{x}_{1.}, ..., \overline{x}_{m.})$.
We also have

$$b_{e(i)}(\mathbf{x}) = A_{1i} - \frac{e_n}{n-1} \sum_{j=1}^{n} (x_{1j} - \overline{x}_{1.}) x_{ij}. \tag{5}$$

It then follows that

$$b_e(\mathbf{x}) = A_1 - e_n^2 \widehat{\Sigma}_1. \tag{6}$$

where A_1 is the first column of A_e and $\widehat{\Sigma}_1$ is the first column of the sample covariance matrix $\widehat{\Sigma}$.
We are now in a position to derive the optimum weights in the general case when sample sizes are equal.

Proposition 2.3 *The optimum weight vector which minimizes $D_e^{(m)}$ takes the following form*

$$\boldsymbol{\lambda}_e^{opt} = (1, 0, 0, ..., 0)^t - e_n^2 \left(A_e^{-1} \widehat{\Sigma}_1 - \frac{\mathbf{1}^t A_e^{-1} \widehat{\Sigma}_1}{\mathbf{1}^t A_e^{-1} \mathbf{1}} A_e^{-1} \mathbf{1} \right).$$

We remark that A_e is invertible since $\hat{\Sigma}$ is invertible. Note that the expression of the weight vector in the two population case can also be derived by using the matrix representation given as above. The detailed calculation is quite similar to that given in the previous subsection.

2.2 Linear WLE for Unequal Sample Sizes

In the previous subsection, we discussed choosing the optimum weights when the samples sizes are equal. In this section, we propose cross-validation methods for choosing adaptive weights for unequal sample sizes. If the sample sizes are not equal, it is not clear whether the *delete-one-column* approach is a reasonable one.

For example, suppose that there are 10 observations in the first sample and there are 5 observations in the second. Then there is no observation to delete for the second sample for half of the cross-validation steps. Furthermore, we might lose accuracy in prediction by deleting one column for small sample sizes. Therefore we propose an alternative method that deletes only one data point from the first sample and keeps all the data points from the rest of samples if the sample sizes are not equal.

2.2.1 Two Population Case

Let us again consider the two population case. The optimum weights $\boldsymbol{\lambda}_u^{opt}$ are obtained by minimizing the following objective function:

$$D_u^{(2)}(\boldsymbol{\lambda}) = \sum_{j=1}^{n_1} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \lambda_2 \overline{X}_{2.} \right)^2,$$

where $\sum_{i=1}^{m} \lambda_i = 1$ and $\overline{X}_{1.}^{(-j)} = \frac{1}{n_1 - 1} \sum_{k \neq j} X_{1k}$. We remark that the major difference between $D_e^{(2)}$ and $D_u^{(2)}$ is that only the jth data point of the first sample is left out for the jth term in $D_u^{(2)}$.

Under the condition that $\lambda_1 + \lambda_2 = 1$, we can rewrite $D_u^{(2)}$ as a function of λ_1:

$$\begin{aligned} D_u^{(2)} &= \sum_{j=1}^{n_1} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - (1-\lambda_1)\overline{X}_{2.} \right)^2 \\ &= \sum_{j=1}^{n_1} \left((X_{1j} - \overline{X}_{2.}) + \lambda_1 (\overline{X}_{2.} - \overline{X}_{1.}^{(-j)}) \right)^2. \end{aligned}$$

By differentiating $D_u^{(2)}$ with respect to λ_1 and using the fact that $\lambda_1 + \lambda_2 = 1$, we then have

$$\lambda_1^{opt} = \frac{n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 - \frac{n_1}{n_1-1}\hat{\sigma}_1^2}{n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{n_1}{(n_1-1)^2}\hat{\sigma}_1^2}; \quad \lambda_2 = 1 - \lambda_1^{opt}. \tag{7}$$

The adaptive optimum weights still converge to $(1,0)$ when the sample sizes are not equal.

Proposition 2.4 *If* $\theta_1^0 \neq \theta_2^0$, *then*
$\lambda_1^{opt} \xrightarrow{P_{\theta^0}} 1$ *and* $\lambda_2^{opt} \xrightarrow{P_{\theta^0}} 0$.

2.2.2 Optimum Weights By Cross-Validation

We now derive the general formula for the optimum weights by cross-validation when the sample sizes are not all equal.

The objective function is defined as follows:

$$\begin{aligned} D_u^{(m)} &= \sum_{j=1}^{n_1} \left(X_{1j} - \lambda_1 \overline{X}_{1.}^{(-j)} - \sum_{i=2}^{m} \lambda_i \overline{X}_{i.} \right)^2 \\ &= c(\underline{X}) - 2\boldsymbol{b}(\underline{X})\boldsymbol{\lambda}_u + \boldsymbol{\lambda}_u^t A(\underline{X})\boldsymbol{\lambda}_u \end{aligned}$$

where

$$\begin{aligned} b_1 &= \sum_{j=1}^{n_1} X_{1j} \left(\overline{X}_{1.} + \frac{1}{n_1 - 1}(\overline{X}_{1.} - X_{1j}) \right) = n_1 \overline{X}_{1.}^2 - \frac{n_1}{n_1 - 1}\hat{\sigma}_1^2; \\ b_i &= n_1 \overline{X}_{1.} \overline{X}_{i.}, \qquad i = 2, ..., m; \end{aligned}$$

and

$$\begin{aligned} a_{11} &= \sum_{j=1}^{n_1} \left(\overline{X}_{1.} + \frac{1}{n_1 - 1}(\overline{X}_{1.} - X_{1j}) \right)^2 = n_1 \overline{X}_{1.}^2 + \frac{n_1}{(n_1-1)^2}\hat{\sigma}_1^2 \\ a_{ij} &= n_1 \overline{X}_{i.} \overline{X}_{j.}, \qquad i \neq 1 \ \ or \ \ j \neq 1. \end{aligned}$$

It then follows that

$$A = n_1 \left(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m \right)^t \left(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m \right) + D$$

where

$$\begin{aligned} d_{11} &= \frac{n_1}{(n_1-1)^2}\hat{\sigma}_1^2; \\ d_{ij} &= 0, \qquad i \neq 1 \ \ or \ \ j \neq 1. \end{aligned}$$

By the elementary rank inequality, it follows

$$rank\,(A) \leq rank\big(\hat{\theta}^t\hat{\theta}\big) + rank(D) = 2.$$

Therefore, we have

$$rank(A) < m \quad if \quad m > 2.$$

It then follows that A is not invertible for $m > 2$. Therefore, the g-inverse of the matrix A should be considered in order to find the optimum weight vector.

3 Asymptotic Properties of the Adaptive Weights

In this section, we derive the asymptotic properties of the cross-validated weights. Let $\hat{\theta}_1^{(n_1)}$ be the MLE based on the first sample of size n_1. Let $\hat{\theta}_1^{(-j)}$ and $\tilde{\theta}_1^{(-j)}$ respectively be the MLE and WLE based on m samples without the jth data point from the first sample. This generalizes the two cases where either only the jth data point is deleted from the first sample or jth data point from each sample is deleted. Note that $\tilde{\theta}_1^{(-j)}$ is a function of the weight function $\boldsymbol{\lambda}$. Let $\frac{1}{n_1}D_{n_1}$ be the average discrepancy in the cross-validation which is defined as

$$\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda}) = \frac{1}{n_1}\sum_{j=1}^{n_1}\Big(X_{1j} - \phi(\tilde{\theta}_1^{(-j)})\Big)^2.$$

Let $\boldsymbol{\lambda}^{(cv)}$ be the optimum weights chosen by cross-validation. Let $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2, ..., \theta_m)$, where θ_1^0 is the true values of θ_1.

Theorem 3.1 *Assume that*

(1) $\frac{1}{n_1}D_{n_1}$ *has a unique minimum for any fixed* n_1*;*

(2) $\frac{1}{n_1}\sum\limits_{j=1}^{n_1}\Big(\phi(\hat{\theta}_1^{(-j)}) - \phi(\theta_1^0)\Big) \xrightarrow{P_{\theta^0}} 0$ *as* $n_1 \to \infty$*;*

(3) $P_{\theta^0}\left(\frac{1}{n_1}\sum\limits_{j=1}^{n_1}\Big(X_{1j} - \phi(\hat{\theta}_1^{(-j)})\Big)^2 < K\right) \xrightarrow{P_{\theta^0}} 1$ *for some constant* $0 < K < \infty$*;*

(4) $P_{\theta^0}\left(\Big|\phi(\hat{\theta}_1^{n_1}) - \phi(\tilde{\theta}_1^{n_1})\Big| > M\right) = o(\frac{1}{n_1})$ *for some constant* $0 < M < \infty$*;*

then

$$\boldsymbol{\lambda}^{(cv)} \xrightarrow{P_{\theta^0}} \boldsymbol{w}_0 = (1, 0, 0, ..., 0)^t. \tag{8}$$

To check the assumptions of the above theorem, let us consider the linear WLE for two samples with equal sample sizes. Assumption (1) is satisfied since $\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda})$ is a quadratic form in $\boldsymbol{\lambda}$ and its minimum is indeed unique for each fixed n_1. To check Assumption (2), consider

$$\begin{aligned}
\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\theta_1^0)\right) &= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\overline{X}_{1.}^{(-j)}-\theta_1^0\right)\\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\frac{1}{n_1-1}\sum_{l\neq j}X_{1l}-\theta_1^0\right)\\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\frac{n_1}{n_1-1}\overline{X}_{1.}-\frac{1}{n_1-1}X_{1j}\right)-\theta_1^0\\
&= \overline{X}_{1.}-\theta_1^0 \quad \xrightarrow{P_{\theta^0}} 0 \;\; as \;\; n_1\to\infty.
\end{aligned}$$

Next we consider

$$\begin{aligned}
\frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\hat{\theta}_1^{(-j)})\right)^2 &= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\overline{X}_{1.}^{(-j)}\right)^2\\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\left(\frac{n_1}{n_1-1}\overline{X}_{1.}-\frac{1}{n_1-1}X_{1j}\right)\right)^2\\
&= \left(\frac{n_1}{n_1-1}\right)^2\frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\overline{X}_{1.}\right)^2 \xrightarrow{P_{\theta^0}} var(X_{11})<\infty \quad as \;\; n_1\to\infty.
\end{aligned}$$

For the last assumption of the previous theorem, consider

$$\begin{aligned}
\left|\phi(\hat{\theta}_1^{(n_1)})-\phi(\tilde{\theta}_1^{(n_1)})\right| &= \left|\overline{X}_{1.}-(\lambda_1^{(cv)}\overline{X}_{1.}+\lambda_2^{(cv)}\overline{X}_{2.})\right|\\
&= \left|\lambda_2^{(cv)}\right|\left|\overline{X}_{1.}-\overline{X}_{2.}\right|.
\end{aligned}$$

It then follows from Lemma **??** that

$$\begin{aligned}
P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{(n_1)})-\phi(\tilde{\theta}_1^{(n_1)})\right|>\epsilon\right) &= P_{\theta^0}\left(\frac{1}{n-2}\left|\frac{(\hat{\sigma}_1^2-\widehat{cov})^2(\bar{X}_{1.}-\bar{X}_{2.})^2}{[(\bar{X}_{1.}-\bar{X}_{2.})^2+\frac{1}{n^2}\sum_{i=1}^{n}(X_{1j}-X_{2j})^2]^2}\right|>\epsilon\right)\\
&\leq \frac{1}{(n-2)^2\epsilon^2}E_{\theta^0}\left|\frac{(\hat{\sigma}_1^2-\widehat{cov})^2(\bar{X}_{1.}-\bar{X}_{2.})^2}{(\bar{X}_{1.}-\bar{X}_{2.})^4}\right|\\
&\leq \frac{1}{(n-2)^2\epsilon^2}E_{\theta^0}\left|\frac{\hat{\sigma}_1^2-\widehat{cov}}{\bar{X}_{1.}-\bar{X}_{2.}}\right|^2 = o(\frac{1}{n^2})
\end{aligned}$$

since $\bar{X}_{1.}-\bar{X}_{2.}\xrightarrow{a.s}\theta_1^0-\theta_2^0\neq 0$ and $\hat{\sigma}_1^2-\widehat{cov}\xrightarrow{a.s}\sigma_1^2-cov(X_{11},X_{21})$. Thus the assumptions of the theorem are all satisfied. We then have

$$\left|\lambda_2^{(cv)}\right|\xrightarrow{P_{\theta^0}}1; \qquad \left|\lambda_2^{(cv)}\right|\xrightarrow{P_{\theta^0}}0.$$

This is consistent with the result of Proposition **??**.

Wang, van Eeden and Zidek (2002) establish the asymptotic normality of the WLE for fixed weights. Under certain regularity conditions and by Theorem **??**, we then have the following asymptotic results for using adaptive weights.

Theorem 3.2 *For each θ_1^0, the true value of θ_1, and each $\theta_1 \neq \theta_1^0$,*

$$\lim_{n_1 \to \infty} P_{\theta^0} \left(\prod_{i=1}^{m} \prod_{j=1}^{n_i} f(X_{ij}; \theta_1^0)^{\lambda_i^{(cv)}(\mathbf{X})} > \prod_{i=1}^{m} \prod_{j=1}^{n_i} f(X_{ij}; \theta_1)^{\lambda_i^{(cv)}(\mathbf{X})} \right) = 1,$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

Theorem 3.3 *For any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ of θ_1 constructed with adaptive weights $\lambda_i^{(\mathbf{n})}(\mathbf{X})$, and for all $\epsilon > 0$,*

$$\lim_{n_1 \to \infty} P_{\theta^0} \left(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| > \epsilon \right) = 0,$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

Theorem 3.4 *For any sequence of maximum weighted likelihood estimates $\tilde{\theta}_1^{(n_1)}$ of θ_1 constructed with adaptive weights $\lambda_i(\mathbf{X})$, and for all $\epsilon > 0$*

$$\lim_{n_1 \to \infty} P_{\theta^0} \left(||\tilde{\theta}_1^{(n_1)} - \theta_1^0|| > \epsilon \right) = 0,$$

for any $\theta_2, \theta_3, ..., \theta_m, \theta_i \in \Theta, i = 2, 3, ..., m$.

We assume that the parameter space is an open subset of R^p. The asymptotic normality of the WLE constructed by cross-validated weights follows.

Theorem 3.5 *(Multi-dimensional) Then there exists a sequence of roots of the weighted likelihood function based on adaptive weights $\tilde{\theta}_1^{(n_1)}$ that is weakly consistent and*

$$\sqrt{n_1} \left(\tilde{\theta}_1^{(n_1)} - \theta_1^0 \right) \xrightarrow{D} N\left(0, I(\theta_1^0)\right), \quad as \ \ n_1 \to \infty.$$

4 Simulation Studies

To demonstrate the benefit of using cross-validation procedure described in previous sections, we perform simulations for normal and Poisson distributions based on *delete-one-column* approach. The algorithm is given as follows.

Step 1: Draw a random sample of size n from $f_1(x; \theta_1^0)$ and $f_2(x; \theta_2^0)$;

Step 2: Calculate the cross-validated optimum weights;

Step 3. Calculate (MLE- $\theta_1^0)^2$ and (WLE- $\theta_1^0)^2$;

Step 4: Repeat Step 1 - 3, 1000 times. Calculate the averages and standard deviations of the squared estimation error for both the MLE and WLE. Calculate the averages and standard deviations of the optimum weights.

n	MSE(MLE)	SD of (MLE-θ_1^0)^2	MSE(WLE)	SD of (WLE-θ_1^0)^2	$\frac{MSE(WLE)}{MSE(MLE)}$
10	0.100	0.145	0.079	0.122	0.798
20	0.047	0.062	0.040	0.054	0.849
30	0.032	0.044	0.028	0.039	0.872
40	0.026	0.036	0.023	0.033	0.906
50	0.018	0.025	0.016	0.024	0.923
60	0.016	0.022	0.015	0.021	0.935

Table 1: Averages and standard deviations of the MSE for the MLE and the WLE samples of equal size from $N(0,1)$ and $N(0.3,1)$.

We generate random samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, setting $\sigma_1 = \sigma_2 = 1$ for simplicity. For the purpose of the demonstration, we set $\mu_1 = 0$ and $\mu_2 = 0.3$. Table 1 shows the simulation results for this case.

n	AVE. of λ_1	AVE. of λ_2	SD of λ_1 and λ_2
10	0.8630	0.1369	0.0060
20	0.9331	0.0668	0.0013
30	0.9555	0.0444	0.0005
40	0.9656	0.0343	0.0003
50	0.9741	0.0258	0.0002
60	0.9792	0.0207	0.0001

Table 2: Optimum weights and their standard deviations for samples of equal size from $N(0,1)$ and $N(0.3,1)$.

It can be seen from Table 1 that the MSE of WLE is much smaller than that of the MLE for small and moderate sample sizes. The standard deviations of the squared differences for the WLE are uniformly smaller than those of the MLE. This suggests that not only does WLE achieve smaller MSE but also its MSE has less variation than that of MLE. Intuitively, as the sample size increases, the importance of the second sample diminishes. As indicated by Table 1, the cross-validation procedure assigns smaller values to λ_2, as the sample size increases. The optimum adaptive weights do converge towards the asymptotic limit $(1,0)$ as shown by Table 2.

We repeat the algorithm for Poisson distribution with $\mu = 2$ and $\mu = 2.5$. Simulation results are shown in Table 3 and Table 4. The results for the Poisson case demonstrate similar patterns for the ratio of the MSEs for WLE and MLE. This statement also applies to the behavior of the optimal adaptive weights as the sample size increases.

Other values for μ_1 and μ_2 were tried for the normal and Poisson distributions. In general, the larger the difference between the two means, the less improvement in the MSE. For the normal case, if we set $\sigma_1 = \sigma_2 = 1$ and $\mu_2 - \mu_1 = 1$, the ratio of the MSE for MLE and WLE will be almost 1.

This implies that the WLE will be almost identical to the MLE. This further implies that the weight for the second sample is almost zero. We also report that the reduction in MSE will disappear if we set $\mu_2 - \mu_1 = 1.5$ for the Poisson case. Thus, the cross-validation procedure will not combine the two samples if the second sample does not help for prediction.

n	MSE(mle)	SD of (MLE-θ_1^0)^2	MSE(wle)	SD of (WLE-θ_1^0)^2	$\frac{MSE(wle)}{MSE(mle)}$
10	0.192	0.081	0.155	0.056	0.808
20	0.108	0.022	0.093	0.017	0.862
30	0.068	0.010	0.061	0.008	0.899
40	0.051	0.005	0.047	0.005	0.932
50	0.040	0.003	0.038	0.003	0.943
60	0.036	0.002	0.034	0.002	0.947

Table 3: MSE of the MLE and the WLE and their standard deviations for samples with equal sizes from $\mathcal{P}(2)$ and $\mathcal{P}(2.5)$.

n	AVE. of λ_1	AVE. of λ_2	SD of λ_1 and λ_2
10	0.8667	0.1332	0.0064
20	0.9371	0.0628	0.0010
30	0.9588	0.0411	0.0004
40	0.9713	0.0286	0.0002
50	0.9771	0.0228	0.0001
60	0.9812	0.0187	0.0001

Table 4: Optimum weights and their standard deviations for samples with equal sizes from $\mathcal{P}(2)$ and $\mathcal{P}(2.5)$

We emphasize that the knowledge of the actual difference between the two populations and the values of the variance for the normal case is not used in any of these two simulations. We also remark that simulations of using the *delete-one-point* approach give very similar results.

5 Application to Disease Mapping

In this section, we address the problem of disease mapping. In particular, we demonstrate a weighted likelihood alternative to the hierarchical Bayes approach that has been used in references cited below. Our approach allows the data themselves to select the weights through cross- validation. We thereby avoid the need in prior modeling, to guess the latent patterns of environmental hazards that may lead to the adverse health effects being mapped. Such hazards include air pollution that have been associated with respiratory morbidity (see, for example, Burnett 1994 and Zidek *et al.* 1998).

Our demonstration involves parallel time series of weekly hospital admissions for respiratory disease in residents of 733 Census Sub-Divisions (CSD) in southern Ontario. The data derive from the May-to-August periods of the years, 1983 to 1988. For this demonstration, we confine attention to certain densely populated areas.

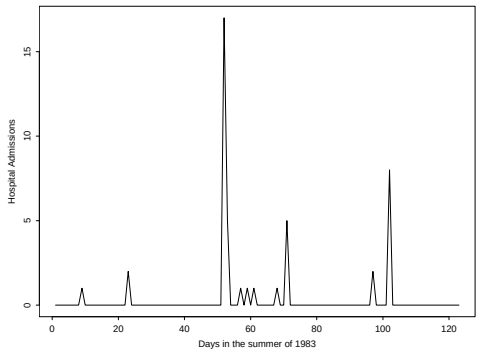

Figure 1: Daily hospital admissions for CSD No 380 in the summer of 1983.

Let us consider the problem of estimating the rate of weekly hospital admissions of CSD 380, the one with the largest annual hospital admissions total among all CSD's from 1983 to 1988. The daily data contain many 0's, representing no hospital admissions. For example, although CSD 380 has the largest number of hospital admissions overall among all the CSD's, no patients were admitted during 112 out of the 123 days in the summer of 1983. On certain of those days, however, quite a number of people sought treatment for acute respiratory disease, possibly due to high levels of air pollution in their region. Again referring to CSD 380, 17 patients were admitted on day 51.

A more graphical depiction of these irregularities in admission counts for this CSD are seen in Figure 1.

There daily counts are shown and the problem of data sparseness and high level of variation are extreme. In fact, in this demonstration we have chosen to avoid the complexities of modeling these daily count series and we turn instead to weekly counts. While the problems remain, they are not nearly so acute.

In total, each of the summers in the years covered by our study has 17 weeks. For simplicity, the data obtained in the last few days of each summer are dropped from the analysis since they do not constitute a whole week.

5.1 Weighted Likelihood Estimation

We assume the weekly number of hospital admissions for any given CSD follows a Poisson distribution, *i.e.*, for year q, CSD i and week j,

$$Y_{ij}^{q} \overset{ind.}{\sim} \mathcal{P}\left(\theta_{ij}^{q}\right), j = 1, 2, ..., 17; i = 1, 2, ..., 733; q = 1, 2, ..., 6.$$

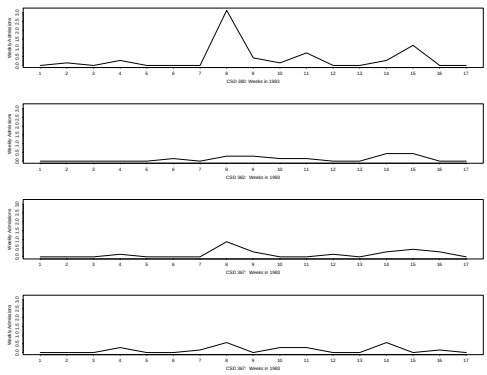

Figure 2: Hospital Admissions for CSD 380, 362, 366 and 367 in 1983.

The raw estimate of θ_{ij}^q, namely Y_{ij}^q is highly unreliable, the effective sample size in this case being but 1. Moreover, each CSD may contain only a small group of susceptible. These considerations point to the need to "borrow strength, a standard feature of disease mapping techniques (Lawson and Williams, 2001). That is, the information in neighboring CSD's is combined to produce more reliable estimates of a regions underlying propensity to generate adverse health outcomes while introducing only a small amount of bias.

For any given CSD, the "neighboring CSD's are defined to be CSD's in close proximity to that of interest, CSD 380 in our analysis. To estimate the rate of weekly hospital admissions in a particular CSD, we would expect that neighboring subdivisions to contain relevant information which might help us to derive a better estimate than the traditional sample average. Thus, the Euclidean distance between the target CSD and any other CSD in the dataset is calculated by using the longitudes and latitudes. We choose somewhat arbitrarily to count as neighbours, any CSD's whose Euclidean distances are less than 0.2 from the target CSD. For CSD 380, neighboring CSD's turn out to be CSD's 362, 366 and 367.

The time series plots of weekly hospital admissions for those CSD's in 1983 are shown in Figure 2. Hospital admissions of these CSD's indeed seem to be related since the major peaks in the time series plot occurred at roughly the same time points. However, as noted earlier the data from other CSD's may introduce bias. Thus the WLE's weights are needed to combine information effectively and control the bias at the same time.

To find cross-validatory choices for these weights we consider purely as a working assumption, that $\theta_{ij}^q = \theta_i^q$ for $j = 1, 2, ..., 17$. In fact, that assumption does not seem tenable since every year, week 8 always has the markedly larger numbers of hospital admissions for CSD 380 than the remaining weeks. For example, in 1983, there are 21 admissions in week 8 while the second largest weekly count is only 7 in week 15. Thus, we are forced to drop week 8 from our working assumption and instead, assume $\theta_{ij}^q = \theta_i^q$ for $j = 1, 2, ...7, 9, ..17$. In fact the sample means and variances of the weekly hospital admissions for those 16 weeks of CSD 380 are quite close to each other in support

of our assumption.

One alternative to assuming the constancy of weights over the whole summer that might yield stable estimates of weights would be the use of a moving window just a few weeks in width. We leave that option for future work.

For Poisson distributions, the MLE of θ_1^q is the sample average of the weekly admissions of CSD 380 while the WLE is a linear combination of the sample averages for each CSD. Thus, the *weighted likelihood estimate* of the population mean weekly hospital admission number for a CSD is θ_1^q, is

$$WLE^q = \sum_{i=1}^{4} \lambda_i^q \overline{Y}_{i.}^q, \quad q = 1, 2.., 16.$$

where $\overline{Y}_{i.}^q$ denotes be CSD i's overall sample average for the 16 weeks of year i.
For our analysis, the weights are selected by the cross-validation procedure proposed in Section 2.2.

5.2 Results of the Analysis

We assess the performance of the MLE and the WLE by comparing their MSE's. The MSE of the MLE and the WLE are defined as, for $q = 1, 2, .., 16$,

$$\begin{aligned} MSE_M^q(\theta_1^q) &= E_{\theta_1^q}\left(\overline{Y}_{1.}^q - \theta_1^q\right)^2 \\ MSE_W^q(\theta_1^q) &= E_{\theta_1^q}\left(\sum_{i=1}^{4} \lambda_i^q \overline{Y}_{i.}^q - \theta_1^q\right)^2. \end{aligned}$$

In fact, the θ_1^q are unknown. We then estimate the MSE_M and MSE_W by replacing θ_1^q by the MLE. Under the assumption of Poisson distributions, the estimated MSE for the MLE is given by:

$$MSE_M^q = \widehat{var(Y_{11})}/16, \quad q = 1, 2, ...6.$$

The estimated MSE for the WLE is give as following:

$$\begin{aligned} MSE_W^q &= E\left(\sum_{i=1}^{m} \lambda_i^q \overline{Y}_{i.}^q - \theta_1^q\right)^2 \\ &= Var\left(\sum_{i=1}^{m} \lambda_i^q \overline{Y}_{i.}^q\right) + \left(E\sum_{i=1}^{m} \lambda_i^q \overline{Y}_{i.}^q - \theta_1^q\right)^2 \\ &\approx \sum_{i=1}^{4}\sum_{k=1}^{4} \lambda_i^q \lambda_j^q \widehat{cov}(\overline{Y}_{i.}^q, \overline{Y}_{j.}^q) + \left(\sum_{i=1}^{m} \lambda_i^q \overline{Y}_{i.}^q - \overline{Y}_{1.}^q\right)^2. \end{aligned}$$

The estimated MSE for the MLE and the WLE are given in the following table. It can be seen from the table that the MSE for the WLE is much smaller than that of the MLE. In fact, the average reduction of the MSE by using WLE is about 25%.

Year	7 MLE	7 WLE	16 $\widehat{MSE}_M^q$	16 $\widehat{MSE}_W^q$	$\widehat{MSE}_W^q/\widehat{MSE}_M^q$
1	.185	.174	.101	.084	0.80
2	.328	.282	.241	.131	0.87
3	.227	.257	.286	.143	0.54
4	.151	.224	.159	.084	0.96
5	.303	.322	.298	.130	0.80
6	.378	.412	.410	.244	0.54

Table 5: Estimated MSE for the MLE and the WLE.

Combining information across these CSD's might also help us in the prediction since the patterns exhibited in one neighboring location in a particular year might manifest itself at the location of interest the next year. To assess the performance of the WLE, we also use the WLE derived from one particular year to predict the overall weekly average of the next year. The overall prediction error is defined as the average of those prediction errors. To be more specific, the overall prediction errors for the WLE and the traditional sample average are defined as follows:

$$\begin{aligned} PRED_M &= \sqrt{\frac{1}{5}\sum_{q=1}^{5}\left(\overline{Y}_{1.}^{q} - \overline{Y}_{1.}^{q+1}\right)^2}\ ; \\ PRED_W &= \sqrt{\frac{1}{5}\sum_{q=1}^{5}\left(WLE^q - \overline{Y}_{1.}^{q+1}\right)^2}\ . \end{aligned}$$

The average prediction error for the MLE, $Pred_M$, is 0.065 while the $Pred_W$, the average prediction error for the WLE, is 0.047 which is about 72% of that of the MLE.

Bayes methods are popular choices in the area of disease mapping. Manton *et al.* (1989) discuss the Empirical Bayes procedures for stabilizing maps of cancer mortality rates. Hierarchical Bayes generalized linear models are proposed for the analysis of disease mapping in Ghosh *et al.* (1999). But it is not obvious how one would specify a neighborhood which needs to be defined in these approaches. The numerical values of the weight functions can be used as a criterion to appropriately define the neighborhood in the Bayesian analysis. We will use the following example to demonstrate how a neighborhood can be defined by using the weight functions derived from the cross-validation procedure for the WLE.

From Table 6, we see that there is strong linear association between CSD 380 and CSD 366. However, the weight assigned to CSD 366 is the smallest one. It shows that CSD's with higher correlation contain less information for the prediction since they might have too similar a pattern to the target CSD for a given year to be helpful in the prediction for the next year. Thus CSD 366 which has the smallest weight should not be included the analysis. Therefore, the "neighborhood" of CSD 380 in the analysis should only include CSD 362 and CSD 367.

	CSD 380	CSD 362	CSD 366	CSD 367	Weights
CSD 380	1.000	0.421	0.906	0.572	0.455
CSD 362	0.421	1.000	0.400	0.634	0.202
CSD 366	0.906	0.400	1.000	0.553	0.128
CSD 367	0.572	0.634	0.553	1.000	0.215

Table 6: Correlation matrix and the weight function for 1984.

In general, we might examine those CSD which are in close proximity to the target CSD. We can calculate the weight for each CSD selected by using the cross-validation procedure. The CSD with small weights should be dropped from the analysis since they are not deemed to be helpful or relevant to our analysis according to the cross-validation procedure.

Year	CI_M	CI_W
1983	[0,3]	[0, 3]
1984	[0, 5]	[0, 4]
1985	[0, 4]	[0, 4]
1986	[0, 3]	[0, 4]
1987	[0, 4]	[0, 5]
1988	[0, 5]	[0, 6]

Table 7: MSE of the MLE and the WLE for CSD 380.

We remark that the weight function can also be helpful in selecting an appropriate distribution that takes into account the spatial structure. Ghosh *et al.* (1999) propose a very general hierarchical Bayes spatial generalized model that is considered broad enough to cover a large number of situations where a spatial structure needs to be incorporated. In particular, they propose the following:

$$\theta_i = q_i = x_i^t \mathbf{b} + u_i + v_i, i = 1, 2, ..., m$$

where the q_i are known constants, x_i are covariates, u_i and v_i are mutually independent with $v_i \overset{i.i.d.}{\sim} N(0, \sigma_v^2)$ and the u_i have joint pdf

$$f(u) \propto (\sigma_u)^2)^{-1/2m} exp\left(- \sum_{i=1}^{m} \sum_{j \neq i} (u_i - u_j)^2 w_{ij} / (2\sigma_u^2) \right)$$

where $w_{ij} \geq 0$ for all $1 \leq i \neq j \leq m$. The above distribution is designed to take into account the spatial structure. In their paper, they propose to use $w_i = 1$ if location i and j are neighbors. They also mention the possibility of using the inverse of the correlation matrix as the weight function. The weights function derived from the cross-validation procedure might be a better choice since it takes account of the spatial structure and underlying factors.

The predictive distribution for the weekly total is Poisson for WLE. We can then derive the 95% predictive interval for the weekly average hospital admissions. This might be criticized as failing to take into account the uncertainty of the unknown parameter. Smith (1998) argues that the traditional plug-in method has a small MSE compared to the posterior mean under certain circumstances. In particular, it has a smaller MSE when the true value of the parameter is not large. Let CI_W and CI_M be the 95% predictive intervals of the weekly averages calculated from the WLE and the MLE respectively. The results are shown in the following table.

We remark that this is merely a demonstration of the weighted likelihood method. Further analysis is needed if one wants to compare the performances of the WLE, the MLE and the Bayesian estimator in disease mapping.

Appendix

Proof of Lemma ?? : Observe that

$$\overline{X}_{i.}^{(-j)} = e_n \overline{X}_{i.} - \frac{1}{n-1} X_{ij}.$$

where $e_n = \frac{n}{n-1}$.
Let $\frac{1}{n} S_1^e = \frac{1}{n} \sum_{j=1}^{n} \left(\overline{X}_{1.}^{(-j)} - \overline{X}_{2.}^{(-j)} \right)^2$. It then follows that

$$\begin{aligned}
\frac{1}{n} S_1^e &= \frac{1}{n} \sum_{j=1}^{n} \left((e_n \overline{X}_{1.} - \frac{1}{n-1} X_{1j}) - (e_n \overline{X}_{2.} - \frac{1}{n-1} X_{2j}) \right)^2 \\
&= \frac{1}{n} \big(n e_n^2 (\overline{X}_{1.} - \overline{X}_{2.})^2 - 2 \frac{e_n}{n-1} (\overline{X}_{1.} - \overline{X}_{2.}) \sum_{j=1}^{n} (X_{1j} - X_{2j}) \\
&\qquad + (\frac{1}{n-1})^2 \sum_{j=1}^{n} (X_{1j} - X_{2j})^2 \big) \\
&= \frac{n(n-2)}{(n-1)^2} (\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{1}{n(n-1)^2} \sum_{j=1}^{n} (X_{1j} - X_{2j})^2.
\end{aligned}$$

Let $\frac{1}{n}S_2^e = \frac{1}{n}\sum_{j=1}^{n}(\overline{X}_{1.}^{(-j)} - X_{2.}^{(-j)})(\overline{X}_{1.}^{(-j)} - X_{1j})$. It follows that

$$\begin{aligned}
\frac{1}{n}S_2^e &= \frac{1}{n}\sum_{j=1}^{n}\left(e_n(\overline{X}_{1.} - \overline{X}_{2.}) - \frac{1}{n-1}(X_{1j} - X_{2j})\right)\left((e_n\overline{X}_{1.} - \frac{1}{n-1}X_{1j}) - X_{1j}\right) \\
&= \frac{1}{n}\sum_{j=1}^{n}\left(e_n(\overline{X}_{1.} - \overline{X}_{2.}) - \frac{1}{n-1}(X_{1j} - X_{2j})\right)(e_n\overline{X}_{1.} - e_nX_{1j}) \\
&= \frac{e_n^2}{n}(\overline{X}_{1.} - \overline{X}_{2.})\sum_{j=1}^{n}(\overline{X}_{1.} - X_{1j}) - \frac{e_n}{n(n-1)}\sum_{j=1}^{n}(X_{1j} - X_{2j})(\overline{X}_{1.} - X_{1j}) \\
&= -\frac{e_n}{n(n-1)}\sum_{j=1}^{n}(X_{1j} - X_{2j})(\overline{X}_{1.} - X_{1j}) \qquad (since \sum_{j=1}^{n}(\overline{X}_{1.} - X_{1j}) = 0) \\
&= -\frac{e_n}{n(n-1)}\left(\overline{X}_{1.}\sum_{j=1}^{n}(X_{1j} - X_{2j}) - \sum_{j=1}^{n}X_{1j}^2 + \sum_{j=1}^{n}X_{1j}X_{2j}\right) \\
&= -\frac{e_n}{n-1}\left(\overline{X}_{1.}(\overline{X}_{1.} - \overline{X}_{2.}) - \frac{1}{n}\sum_{j=1}^{n}X_{1j}^2 + \frac{1}{n}\sum_{j=1}^{n}X_{1j}X_{2j}\right) \\
&= \frac{n}{(n-1)^2}\left(\hat{\sigma}_1^2 - \widehat{cov}\right)
\end{aligned}$$

This completes the proof. $\diamond$

Proof of Proposition ?? : By the Weak Law of Large Numbers, it follows that

$$\hat{\sigma}_1^2 - \widehat{cov} \longrightarrow \sigma_1^2 - \rho\sigma_1\sigma_2.$$

Thus condition $\rho < \sigma_1/\sigma_2$ implies that $\hat{\sigma}_1^2 > \widehat{cov}$ for sufficiently large n. Thus, λ_2^{opt} eventually will be positive. $\diamond$

Proof of Proposition ??: From Lemma **??**, it follows that the second term of S_1 goes to zero in probability as n goes to infinity while the first term converges to $(\theta_1^0 - \theta_2^0)^2$ in probability.
Therefore we have

$$S_1^e \xrightarrow{P_{\theta^0}} (\theta_1^0 - \theta_2^0)^2 \ \ as \ \ n \to \infty,$$

where $(\theta_1^0 - \theta_2^0)^2 \neq 0$ by assumption.
Moreover, we see that $S_2^e = O_P(\frac{1}{n})$. By definition of λ_2^{opt}, it follows that

$$|\lambda_2^*| = |\frac{S_2^e}{S_1^e}| \xrightarrow{P_{\theta^0}} 0 \quad as \ \ n \to \infty.$$

This completes the proof. $\diamond$

Proof of Proposition ??:
By differentiating $D_e^{(m)} - \nu\,(\mathbf{1}^t\boldsymbol{\lambda} - 1)$ and setting the result to zero, it follows that

$$\frac{\partial D_e^{(m)} - \nu\,(\mathbf{1}^t\boldsymbol{\lambda} - 1)}{\partial\lambda} = -2b_e + 2A_e\boldsymbol{\lambda}_e^{opt} - \nu\mathbf{1} = 0.$$

It then follows that

$$\boldsymbol{\lambda}_e^{opt} = A_e^{-1}\left(b_e + \frac{\nu}{2}\mathbf{1}\right).$$

We then have

$$1 = \mathbf{1}^t\boldsymbol{\lambda}_e^{opt} = \mathbf{1}^t A_e^{-1}\left(b_e + \frac{\nu\mathbf{1}}{2}\right).$$

Thus,

$$\nu = \frac{2}{\mathbf{1}^t A_e^{-1}\mathbf{1}}\left(1 - \mathbf{1}^t A_e^{-1} b_e\right).$$

Therefore,

$$\boldsymbol{\lambda}_e^{opt} = A_e^{-1}\left(b_e + \frac{1 - \mathbf{1}^t A_e^{-1} b_e}{\mathbf{1}^t A_e^{-1}\mathbf{1}}\mathbf{1}\right).$$

Since $D_e^{(m)}$ is a quadratic function of $\boldsymbol{\lambda}$ and $A \geq 0$, the minimum is achieved at the point $\boldsymbol{\lambda}_e^{opt}$. Furthermore, by equation (**??**) and (**??**), we have

$$A_e^{-1} b_e = A_e^{-1}\left(A_1 - e_n^2\widehat{\Sigma}_1\right) = (1,0,0,...,0)^t - e_n^2 A_e^{-1}\widehat{\Sigma_1}.$$

Denote the optimum weight vector by λ^{opt}. It follows that

$$\boldsymbol{\lambda}_e^{opt} \quad = \quad (1,0,0,...,0)^t - e_n^2\left(A_e^{-1}\widehat{\Sigma}_1 - \frac{\mathbf{1}^t A_e^{-1}\widehat{\Sigma}_1}{\mathbf{1}^t A_e^{-1}\mathbf{1}} A_e^{-1}\mathbf{1}\right).$$

This completes the proof. $\diamond$

Proof of Proposition ??: ¿From equation (**??**), it follows that

$$\lambda_1^{opt} = 1 - \frac{\left(\frac{n_1}{n_1 - 1}\right)^2 \hat{\sigma}_1^2}{n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{1}{n_1 - 1}\hat{\sigma}_1^2}.$$

By the Weak Law of Large Numbers, we have

$$\begin{aligned} \hat{\sigma}_1^2 &\xrightarrow{P_{\theta^0}} \sigma_1^2 \\ (\overline{X}_{1.} - \overline{X}_{2.})^2 &\xrightarrow{P_{\theta^0}} (\theta_1^0 - \theta_2^0)^2 \neq 0. \end{aligned}$$

It then follows that

$$\frac{\left(\frac{n_1}{n_1 - 1}\right)^2 \hat{\sigma}_1^2}{n_1(\overline{X}_{1.} - \overline{X}_{2.})^2 + \frac{1}{n_1 - 1}\hat{\sigma}_1^2} \xrightarrow{P_{\theta^0}} 0.$$

We then have

$$\lambda_1^{opt} \xrightarrow{P_{\theta^0}} 1.$$

The last assertion of the theorem follows by the fact that $\lambda_1 + \lambda_2 = 1$. $\diamond$

Proof of Theorem ?? : Consider

$$\begin{aligned}
\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda}) &= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\tilde{\theta}_1^{(-j)})\right)^2 \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\left(X_{1j}-\phi(\hat{\theta}_1^{(-j)})\right)+\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right)\right)^2 \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\hat{\theta}_1^{(-j)})\right)^2+\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\hat{\theta}_1^{-j})-\phi(\tilde{\theta}_1^{(-j)})\right)^2 \\
&\quad +\frac{2}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\hat{\theta}_1^{(-j)})\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right).
\end{aligned}$$

Note that

$$\begin{aligned}
&\frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\hat{\theta}_1^{(-j)})\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right) \\
&= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\theta_1^0)\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right) \\
&\quad +\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\theta_1^0)-\phi(\hat{\theta}_1^{(-j)})\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right) \\
&= S_1+S_2
\end{aligned}$$

where

$$\begin{aligned}
S_1 &= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j}-\phi(\theta_1^0)\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right), \\
S_2 &= \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\theta_1^0)-\phi(\hat{\theta}_1^{(-j)})\right)\left(\phi(\hat{\theta}_1^{(-j)})-\phi(\tilde{\theta}_1^{(-j)})\right).
\end{aligned}$$

We first show that $S_1 \xrightarrow{P_{\theta^0}} 0$.

Consider

$$
\begin{aligned}
& P_{\theta^0}\left(|S_1| > \epsilon\right) \\
= \; & P_{\theta^0}\left(\epsilon < |S_1| \;\; and \;\; \left|\phi(\hat{\theta}_1^{(-j)}) - \phi(\tilde{\theta}_1^{(-j)})\right| < M \;\; for \;\; all \;\; j\right) \\
& + P_{\theta^0}\left(\epsilon < |S_1| \;\; and \;\; \left|\phi(\hat{\theta}_1^{(-l)}) - \phi(\tilde{\theta}_1^{(-l)})\right| \geq M \;\; for \;\; some \;\; l\right) \\
\leq \; & P_{\theta^0}\left(\epsilon < |S_1| < \frac{M}{n_1}\sum_{j=1}^{n_1}\left|X_{1j} - \phi(\theta_1^0)\right|\right) + \sum_{l=1}^{n_1} P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{(-l)}) - \phi(\tilde{\theta}_1^{(-l)})\right| \geq M\right) \\
\leq \; & P_{\theta^0}\left(\frac{\epsilon}{M} < \left|\frac{1}{n_1}\sum_{j=1}^{n_1}(X_{1j} - \phi(\theta_1^0))\right|\right) + n_1 P_{\theta^0}\left(|\phi(\hat{\theta}_1^{(-1)}) - \phi(\tilde{\theta}_1^{(-1)})| \geq M\right) \\
= \; & P_{\theta^0}\left(\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\theta_1^0)\right)\right| > \frac{1}{M}\epsilon\right) + n_1 P_{\theta^0}\left(|\phi(\hat{\theta}_1^{(n_1-1)}) - \phi(\tilde{\theta}_1^{(n_1-1)})| \geq M\right)
\end{aligned}
$$

The first term goes to zero by the Weak Law of Large Numbers. The second term also goes to zero by assumption (4). We then have

$$P_{\theta^0}\left(|S_1| > \epsilon\right) \longrightarrow 0 \quad as \;\; n_1 \;\; \to \infty. \tag{9}$$

We next show that $S_2 \xrightarrow{P_{\theta^0}} 0$ as $n_1 \to \infty$.
Consider

$$
\begin{aligned}
& P_{\theta^0}(|S_2| > \epsilon) \\
= \; & P_{\theta^0}\left(\epsilon < |S_2| \;\; and \;\; \left|\phi(\hat{\theta}_1^{(-j)}) - \phi(\tilde{\theta}_1^{(-j)})\right| < M \;\; for \;\; all \;\; j\right) \\
& + P_{\theta^0}\left(\epsilon < |S_2| \;\; and \;\; \left|\phi(\hat{\theta}_1^{(-l)}) - \phi(\tilde{\theta}_1^{(-l)})\right| \geq M \;\; for \;\; some \;\; l\right) \\
\leq \; & P_{\theta^0}\left(\epsilon < |S_2| < \frac{M}{n_1}\left|\sum_{j=1}^{n_1}(\phi(\hat{\theta}_1^{(-j)}) - \phi(\theta_1^0))\right|\right) + \sum_{l=1}^{n_1} P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{(-l)}) - \phi(\tilde{\theta}_1^{(-l)})\right| \geq M\right) \\
\leq \; & P_{\theta^0}\left(\frac{1}{M}\epsilon < \left|\frac{1}{n_1}\sum_{j=1}^{n_1}(\phi(\hat{\theta}_1^{(-j)}) - \phi(\theta_1^0))\right|\right) + n_1 P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{(-1)}) - \phi(\tilde{\theta}_1^{(-1)})\right| \geq M\right) \\
= \; & P_{\theta^0}\left(\left|\frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\hat{\theta}_1^{(-j)}) - \phi(\theta_1^0)\right)\right| > \frac{1}{M}\epsilon\right) + n_1 P_{\theta^0}\left(\left|\phi(\hat{\theta}_1^{(n_1-1)}) - \phi(\tilde{\theta}_1^{(n_1-1)})\right| \geq M\right)
\end{aligned}
$$

The first term goes to zero by assumption (2). The second term also goes to zero by assumption (4). We then have

$$P_{\theta^0}\left(|S_2| > \epsilon\right) \longrightarrow 0 \quad as \;\; n_1 \;\; \to \infty. \tag{10}$$

It then follows that

$$\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda}) = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\hat{\theta}_1^{(-j)})\right)^2 + \frac{1}{n_1}\sum_{j=1}^{n_1}\left(\phi(\hat{\theta}_1^{-j}) - \phi(\tilde{\theta}_1^{(-j)})\right)^2 + R_n \tag{11}$$

where $R_n \xrightarrow{P_{\theta^0}} 0$. Observe that the first term is independent of $\boldsymbol{\lambda}$. Therefore the second term must be minimized with respect to $\boldsymbol{\lambda}$ to obtain the minimum of $\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda})$. We see that the second term is always non-negative. It then follows that, with probability tending to 1,

$$\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda}) \geq \frac{1}{n_1}\sum_{j=1}^{n_1}\left(X_{1j} - \phi(\hat{\theta}_1^{(-j)})\right)^2 = \frac{1}{n_1}D_{n_1}(\boldsymbol{w}),$$

since $\phi(\hat{\theta}_1^{(-j)}) = \phi(\tilde{\theta}_1^{(-j)})$ for $\boldsymbol{\lambda}^{(cv)} = \boldsymbol{w}_0 = (1, 0, 0, ..., 0)^t$ for fixed n_1.
Finally, we will show that

$$\boldsymbol{\lambda}^{(cv)} \xrightarrow{P_{\theta^0}} \boldsymbol{w}_0, \quad as \;\; n_1 \to \infty.$$

Suppose to the contrary that $\boldsymbol{\lambda}^{(cv)} \xrightarrow{P_{\theta^0}} \boldsymbol{w}_0 + \boldsymbol{d}$ where d is a non-zero vector. Then there exists n_0 such that for $n_1 > n_0$,

$$\frac{1}{n_1}D_{n_1}\left(\boldsymbol{\lambda}^{(cv)}\right) \geq \frac{1}{n_1}D_{n_1}(\boldsymbol{w}).$$

This is a contradiction because $\boldsymbol{\lambda}^{(cv)}$ is the vector which minimizes $\frac{1}{n_1}D_{n_1}$ for any fixed n_1 and the minimum of $\frac{1}{n_1}D_{n_1}(\boldsymbol{\lambda})$ is unique by assumption. $\diamond$

The proofs of Theorem **??** - **??** are identical to the proofs for fixed weights as given by Wang, van Eeden and Zidek (2002) except that the fixed weights are replaced by adaptive weights and the utilization of Theorem **??**. Details can be found in Wang (2001).

References

Breiman, L. and Friedman, H.J. (1997). Predicting multivariate responses n multiple regression. *Journal of Royal Statistical Society: Series B*, **36**, 111-147.

Burnett, R. and Krewski, D. (1994). Air pollution effects on hospital admission rates: A random effects modeling approach. *The Canadian Journal of Statistics,* **22**, 441-458.

Hu, F. (1997). The asymptotic properties of the maximum-relevance weighted likelihood estimators, Canad. J. Statist., 25, 45-59.

Hu, F., Zidek, J.V. (1995). Incorporating relevant sample information using the likelihood. Technical Report No. 161 Dept. of Statistics, The University of British Columbia, Vancouver, B.C., Canada.

Hu, F., Zidek, J.V. (2001). The relevance weighted likelihood with applications. In: Ahmed, S.E., Reid, N. (Eds.), Empirical Bayes and Likelihood Inference, 211-235. Springer Verlag, New York.

Hu, F., Zidek, J.V. (2002). The weighted likelihood. Canadian J of Statist. *Canadian Journal of Statistics*, 30, 347-371.

Geisser, S. (1975). The predictive sample reuse method with applications, Journal of the American Statistical Association, 320-328.

Ghosh, M., Natarajan, K., Waller, L. A. and Kim, D. (1999). Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping. Journal of Statistical Planning and Inference, 305-318.

Lawson, A. W. and Williams, F.L.R. (2001). An introductory guide to disease mapping. John Wiley, New York.

Manton, K. G., Woodbury, M. A., Stallard, E. Riggan, W. B. Creason, J. P. and Pellon, A. C. (1989). Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. Journal of the American Statistical Association, 637-650.

Smith, R. L. (1998). Bayesian and frequentist approaches to parametric predictive inference. Bayesian Statistics, 589–612.

Stone, M. (1974). Cross-validation choice and assessment of statistical predictions. Journal of Royal Statistical Society: Series B, 1-147.

Tibshirani, R., Hastie, T. (1987). Local likelihood estimation. J. Amer. Statist. Assoc., 82, 559–567.

Wang, X. (2001). Maximum Weighted Likelihood Estimation. PhD Thesis. Department of Statistics, The University of British Columbia, Vancouver, B.C., Canada.

Wang, X., van Eeden, C. and Zidek, J.V. (2002). Asymptotic Properties of Maximum Weighted Likelihood Estimator. Journal of Statistical Inference and Planning. To appear.

Zidek, J.V., White, R. and Le, N.D. (1998). Using spatial data in assessing the association between air pollution episodes and respiratory morbidity. *Statistics for the Environment 4: Pollution Assessment of Control. 117-136.*