THREE ESSAYS ON MIXTURE MODEL AND GAUSSIAN PROCESSES

A Dissertation

by

WENBIN WU

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ximing Wu
Committee Members,	Yu Zhang
	David Leatham
	Qi Li
Head of Department,	Mark L. Waller

May 2019

Major Subject: Agricultural Economics

ABSTRACT

This dissertation includes three essays. In the first essay I study the problem of density estimation using normal mixture models. Instead of selecting the 'right' number of components in a normal mixture model, I propose an Averaged Normal Mixture (ANM) model to estimate the underlying densities based on model averaging methods, combining normal mixture models with different number of components. I use two methods to estimate the mixing weights of the proposed Averaged Normal Mixture model, one is based on likelihood cross validation and the other is based on Bayesian information criterion (BIC) weights. I also establish the theoretical properties of the proposed estimator and the simulation results demonstrate its good performance in estimating different types of underlying densities. The proposed method is also employed to a real world data set, empirical evidence demonstrates the efficiency of this estimator.

The second essay studies short term electricity demand forecasting using Gaussian Processes and different forecast strategies. I propose a hybrid forecasting strategy that combines the strength of different forecasting schemes to predict 24 hourly electricity demand for the next day. This method is shown to provide superior point and overall probabilistic forecasts. I demonstrate the economic utility of the proposed method by illustrating how the Gaussian Process probabilistic forecasts can be used to reduce the expected cost of electricity supply relative to conventional regression methods, and in a decision-theoretic framework to derive an optimal bidding strategy under a stylized asymmetric loss function for electricity suppliers.

The third essay studies a non-stationary modeling approach based on the method of Gaussian process regression for crop yields modeling and crop insurance rate estimation. Our approach departs from the conventional two-step estimation procedure and allows the yield distributions to vary over time. I further develop a performance weighted model averaging method to construct densities as mixture of Gaussian processes. This method not only facilitates information pooling but greatly improves the flexibility of the resultant predictive density of crop yields. The simulation results on corp insurance premium rates show that the proposed method compares favorably to conventional

two stage estimators, especially when the underlying distributions are non-stationary. I illustrate the efficacy of the proposed method with an application to crop insurance policy selection by insurance companies. I adopt a decision theoretic framework in this exploration and demonstrate that insurance companies can use the proposed method to effectively identify profitable policies under symmetric or asymmetric loss functions.

DEDICATION

To my mother, my father, and my grandparents.

ACKNOWLEDGMENTS

First and foremost, I want to thank my committee chair, Dr. Ximing Wu. He made it possible for me to write this dissertation and stood by my side while I began my research in applied economics and while I was entering the world of econometrics. I want to thank Dr. Wu for all the time we spent in his office discussing all the questions I had, correcting my mistakes, making improvements and sharing his ideas with me. I am very grateful for all the opportunities he gave me during the past five years. Without his supervision, this dissertation would not have been completed.

I also want to express my gratitude to Dr. Yu Zhang, Dr. David Leatham and Dr. Qi Li for their continuous support and enlightenment. I thank them for reading previous drafts of this dissertation and providing valuable comments that improved the contents of this dissertation. Last but not least, I thank my family for their support during my Ph.D. study. This dissertation would not have been completed without their encouragement.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Ximing Wu, Professor Yu Zhang and Professor David Leatham of the Department of Agricultural Economics and Professor Qi Li of the Department of Economics.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a departmental teaching assistance scholarship from the Department of Agricultural Economics at Texas A&M University.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Finite mixture of distributions especially normal mixture models have always been a powerful tool to statistical modeling of a wide variety of phenomena. It has been widely applied in economics, biology, engineering and social sciences. Despite its wide adoption in density estimation and clustering, there remain some issues to overcome. One important problem is that it is usually difficult to determine the number of components in a mixture model since the discrete choice of components number and non-nested structure in incremental model building. Due to this problem, normal mixture models based density estimation may change significantly if different number of components is used in the model, it could be unstable with respect to change of component numbers or to small perturbations of the data. It is important to estimate the appropriate number of components of mixture model if researchers are interested in the underlying heterogeneity of the distribution, but when approximation is the goal, we do not necessarily need to know the correct number of components. In the first essay, I choose to use a model averaging approach to tackle the density estimation problem based on normal mixture models. Since it is hard to choose the appropriate number of components in a normal mixture model, I first estimate a series of normal mixture models with different number of components, then I take these estimated normal mixture models as given, and find different ways to mix all these models.I propose two methods to find the mixing weights, one is based on likelihood cross validation and the other one is based on BIC weights. Simulation and empirical results demonstrate the efficiency of our model.

Electricity demand forecasting plays a vital role in power system planning, operations, transmission design, and financial risk management. Since electricity is difficult to store, supply and demand have to be balanced at every point in time. Consequently, overestimation of electricity demand may result in excessive purchase and an unnecessary waste of energy while underestimation may cause disturbance in the power system. There is a vast literature on the forecasting of electricity demand, ranging from long, medium to short term demand. However, most of the current literature focuses on point forecasts of electricity demand, there only exists a small literature

on probabilistic forecasting that predicts quantities such as the quantiles, intervals, or distribution/density functions. In the second essay, I develop a Gaussian Process regression model based on a hybrid forecasting strategy to estimate short term electricity demand, which is shown to provide superior point and overall probabilistic forecasts. In addition to statistical investigation, I further illustrate how the probabilistic forecasts obtained from the Gaussian Process models can be used in a decision-theoretic framework to optimize economic decision making and risk management in the electricity industry.

The federal crop insurance program has been an important part of U.S. agricultural policy to stabilize farmers' income and protect against unpredictable risks for several decades. It covers more than 100 crops with a variety of yield-based, revenue-based and area-based policies. An actuarially sound premium is critical to the effectiveness and robustness of crop yield insurances. Since the calculation of this parameter requires the knowledge of the future distribution of yields, one needs a reliable predictive yield distribution. In the third essay, I propose a new estimation approach for crop yields based on the method of Gaussian process regression. This modeling approach is probabilistic in nature and yields not only point estimates but entire predictive distributions. This is particularly appealing to one of the primary focuses on the corp yield estimation, which is to obtain reliable predictive yield distribution. I illustrate the efficacy of the proposed method with an application to crop insurance policy selection by insurance companies. I adopt a decision theoretic framework in this exploration and demonstrate that insurance companies can use the proposed method to effectively identify profitable policies under symmetric or asymmetric loss functions.

2. AVERAGED NORMAL MIXTURE MODEL FOR DENSITY ESTIMATION

2.1 Introduction

Finite mixture of distributions especially normal mixture models have always been a powerful tool to statistical modeling of a wide variety of phenomena (McLachlan and Basford [1988]; McLachlan and Peel [2004]). It has been widely applied in economics, biology, engineering and social sciences. Also as it is well known that any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities (McLachlan and Peel [2004]), normal mixture models have provided a convenient semi-parametric framework to model unknown distributions. In fact, normal mixture models based density estimation and clustering have shown great performance in many applications (McLachlan and Peel [2004]; Fraley and Raftery [2002]).

In the density estimation problem, we are given an i.i.d sample $S = (x_1, \cdots, x_n)$ drawn from an unknown density f of a $p-$dimensional random variable x, and the goal is to estimate this density function f from the realizations x_i of x. In fitting a finite normal mixtures of k components to these data, it is assumed that the probability density function of sample S can be represented in the form

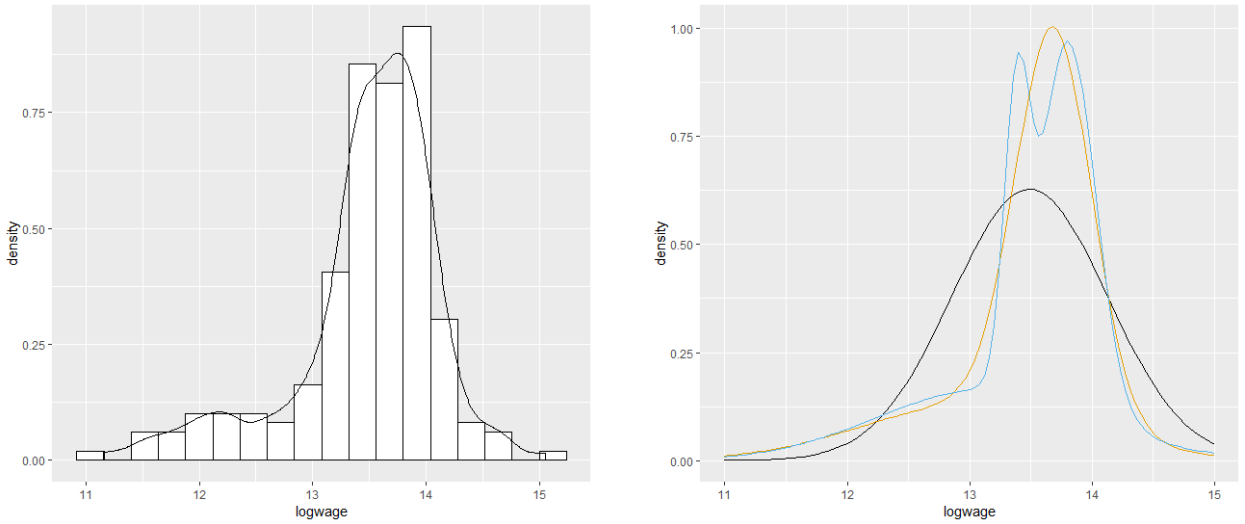
$$f(x) = \sum_{i=1}^{k} \lambda_i \phi(x; \mu_i, \Sigma_i) \tag{2.1}$$

where $\phi(x; \mu_i, \Sigma_i)$ is the normal density function with mean vector μ_i and covariance matrix Σ_i corresponding to the ith mixture component. $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_k)$ is the vector of mixing weights which sums to 1. We usually use maximize likelihood methods to estimate these unknown parameters.

Despite normal mixture models based density estimation and clustering have been proved very useful, there remain some issues to overcome. One important problem is that it is usually difficult to determine the number of components in a mixture model since the discrete choice of components number and non-nested structure in incremental model building. Due to this problem, the density estimated by normal mixture models may change significantly if we choose different num-

ber of components, it could be unstable with respect to change of component numbers or to small perturbations of the data.

In this paper, we start by an example to illustrate the instability of density estimation using a normal mixture model. We use “cps71" data set in R “np" package, which consists a random sample taken from the 1971 Canadian Census Public Use Tapes for male individuals having common education level (grade 13). There are 205 observations in total. We plot the kernel density and histogram for logarithm of wage data in Figure 2.1 (a), in Figure 2.1 (b) we plot the corresponding normal mixture models estimated densities using components number from $k = 1$ to $k = 3$.



(a) (b)

Figure 2.1: Density estimation of “cps71" data set
(a)kernel density estimator and histogram for logarithm of data. (b)normal mixture density estimators using components number from $k = 1$ to $k = 3$, which corresponds to black, orange and blue curve.

As we can see from figure 2.1 (b), normal mixture density estimator will change a lot if we use different number of components in the model. If we use only one component, the density estimator will just be a Gaussian distribution as shown in black curve. When we use two components, the normal mixture density estimator will be similar to the kernel density estimator in Figure 2.1 (a), but it is relatively smooth and does not yield enough features. Then if we use three components

in normal mixture density estimator, the shape of this new estimator will change dramatically, and have a clear tendency of overfitting the underlying density. Actually in practice, with too many components, the normal mixture models tend to overfit the data and yield poor interpretations, while with too few components, the normal mixture models may not be flexible enough to approximate the true underlying densities. So as we mentioned above, it is difficult to determine the number of components in a normal mixture model, also due to the non-nested structure of the model, change between normal mixture models with different components can be significant.

In this article we propose a new method of density estimation based on model averaging. It is an important issue to estimate the appropriate number of components in a mixture model if we are interested in the underlying heterogeneity of the distribution, but when approximation is the goal, we do not necessarily need to know the correct number of components, that is why we choose to use a model averaging approach to solve the problem. Since it is hard to choose the appropriate number of components in a normal mixture model, we first estimate a series of normal mixture models with different number of components, then we take these estimated normal mixture models as given, and we find different ways to give appropriate weights to these models and combine all these models. We proposed two methods to find the appropriate weights, one is based on likelihood cross validation and the other one is based on BIC weights. The presentation of this article goes as follows. Section 2.2 gives a brief overview of previous research. In Section 2.3, we propose our averaged normal mixture (ANM) model, along with an investigation of the asymptotic properties of the proposed methods. Section 2.4 reports results of Monte Carlo studies on normal mixture densities and non-normal mixture densities. Section 2.5 discusses some real world applications of proposed methods and our conclusions are presented in Section 2.6. Proofs of results are contained in the Appendix.

2.2 Literature reviews

2.2.1 Difficulties of normal mixture model specification

As we mentioned above, determining the number of components k in a mixture model could be very difficult and it has not been completely solved. For decades, researchers have been strove to develop an optimal way to find the appropriate number of components in a mixture model, in the existing literature, there are several ways to estimate the number of components in a mixture model.

One way is to use information based criteria such as the Akaike Information Criterion (AIC, Akaike [1974]), the Bayesian Information Criterion (BIC, Schwarz et al. [1978]) and the consistent AIC Information Criterion (CAIC, Bozdogan [1987]). Different information based criteria are essentially likelihood functions with distinct penalties. Leroux et al. [1992] systematically studied the use of AIC and BIC criteria to select the number of components in finite mixture models, he argued that the estimated number of components selected by these criteria is at least as large as the true parameter in large samples. Roeder and Wasserman [1997] showed the consistency of selecting the number of components in a mixture model using BIC criterion. There are also other similar criteria such as Integrated Classification Likelihood criterion (ICL, Biernacki et al. [2000]), Normalized Entropy Criterion (NEC, Biernacki et al. [1999]) and Minimum Information Ratio criterion (MIR, Windham and Cutler [1992]). Studies have also shown that BIC type criteria tend to underestimate the number of components when sample sizes are small. On the contrary, the AIC type criteria typically overestimate the number of components substantially.

Another way to select the number of components in a mixture model is to use the Bayesian framework. For instance, variational inference can be used to determine the number of the components in a fully Bayesian way (Corduneanu and Bishop [2001] or Bishop [2006] Chapter 10.2), which is an approximation of Bayesian inference. Also by choosing appropriate priors, the maximum a posteriori (MAP) estimator can be used for model selection purpose (Ormoneit and Tresp [1998] and Zivkovic and van der Heijden [2004]).

Some researchers also use likelihood ratio test techniques to select the component of the mixture model, such as McLachlan [1987], Dacunha-Castelle et al. [1999], Chen et al. [2004b], Kasahara and Shimotsu [2015]. However, in most cases, these tests will suffer from the boundary problems and difficult to determine the asymptotic distribution.

Moreover, there are many other ways to select the number of components in a mixture model, like the well adapted gap statistics (Tibshirani et al. [2001]), and distance measure like penalized minimum-distance method (Chen and Kalbfleisch [1996]), the Kullback-Leibler distance method (James et al. [2001]) and the Hellinger distance method (Woo and Sriram [2007]), which evaluates the distance between the fitted model and nonparametric estimation of underlying distribution

As we can see, determining the number of components in a mixture model has always been an arguing topic. It is an important issue if researchers are interested in the underlying heterogeneity of the distribution, but when approximation is the goal, we do not necessarily need to know the correct number of components, that is the reason why we introduce our Averaged Normal Mixture model.

2.2.2 Model averaging

Model selection has always been an integral part of statistical modeling. The goal of model selection is to choose the best model among all candidate models considered in the framework. The procedure of selecting the most "suitable" model and conducting analysis and inference on this "suitable" model is well adapted, but also has been criticized since this procedure usually leads to too optimistic tests and confidence intervals, and generally to biased inference statements. An alternative to selecting one model and basing all further work on this chosen model is model averaging. Model averaging exploits information from all candidate models and incorporates model uncertainty into the estimation. Like statistical estimation, model selection is subject to stochastic errors due to sample variation. In contrast, combining the strength of multiple models/estimators can often lead to better performance in practice.

There are two major framework for model averaging: Bayesian model averaging and frequentist model averaging (FMA). Bayesian model averaging provides a coherent mechanism for ac-

counting for model uncertainties. Reviews of Bayesian literature can be found in the works of Hoeting et al. [1999], Raftery et al. [1997]. For FMA strategies, the most widely used methods are weighting strategies based on the AIC or BIC values proposed by Buckland et al. [1997]. Also there are other researchers proposed different mixing strategies based on different framework. Yang [2000, 2001] proposed adaptive mixing strategies for density estimation and regression. Hjort and Claeskens [2003] studied some results on the large sample behavior of likelihood based model average estimators under the assumption of local model misspecification. Leung and Barron [2006] proposed a mixture least squares estimator with weights depending on the estimator's risk characteristics, they also derived a finite sample risk bound for this mixture estimator. More recently, there has been increasing interests in asymptotically optimal model averaging including Mallows model averaging (MMA, Hansen [2007]), optimal mean squared error averaging (Liang et al. [2012]), jackknife model averaging (JMA, Hansen and Racine [2012]), heteroskedasticity robust Cp (Liu and Okui [2013]), and so on.

2.3 Model setup and estimators

In this paper we assume that given number of components in a mixture model k, the normal mixture model is identifiable and estimable. In this case identifiability means that suppose we have two mixture models, given by

$$f(x) = \sum_{i=1}^{k} \lambda_i \phi(x; \mu_i, \Sigma_i) \qquad f'(x) = \sum_{i=1}^{k'} \lambda'_i \phi(x; \mu'_i, \Sigma'_i) \tag{2.2}$$

and that $f(x) \equiv f'(x)$ if and only if $k = k'$ and we can order the summations such that $\lambda_i = \lambda'_i$, $\mu_i = \mu'_i$ and $\Sigma_i = \Sigma'_i$ for $i = 1, \cdots, k$. Then we say $f(x)$ is identifiable. If the model is identifiable, We can use the widely known EM algorithm to estimate the parameters of mixture models. In this paper, we always assume the normal mixture models are identifiable and estimable, once they have been estimated, instead of selecting a best one among them, we take them as given and find a way to mix them.

As we mentioned before, our goal is to find appropriate ways to mix normal mixture models

with different number of components. Here we introduce our Averaged Normal Mixture (ANM) model, the estimation procedures for our proposed ANM model are described as follows:

1. Given any i.i.d sample $S = (x_1, \cdots, x_n)$, we can fit these data to a series of normal mixture models denoted as $\hat{f}_1, \cdots, \hat{f}_k, \cdots, \hat{f_K}$, which $\hat{f}_k$ is

$$\hat{f}_k = \sum_{i=1}^{k} \hat{\lambda}_{ik} \phi(x; \hat{\mu}_{ik}, \hat{\Sigma}_{ik}), \tag{2.3}$$

   and the components number of $\hat{f}_1, \cdots, \hat{f_K}$ is $k = 1, k = 2, \cdots, k = K$. The reason we have subscript $\cdot k$ in $\hat{\lambda}_{ik}$, $\hat{\mu}_{ik}$ and $\hat{\Sigma}_{ik}$ is because for each normal mixture model, we can have different parametrization for individual component.

2. We take $\hat{f}_1, \cdots, \hat{f_K}$ as given, and our proposed ANM model can be written as

$$\hat{f}(x) = \sum_{i=1}^{K} \hat{\omega}_i \hat{f}_i(x), \tag{2.4}$$

   A crucial issue of this method is how to find the mixture weights $(\hat{\omega}_1, \cdots, \hat{\omega}_K)$ for the averaged normal mixture model, here we propose two methods to find the appropriate weights. The first strategy is based on the likelihood cross validation, which is similar to the JMA estimator (Hansen and Racine [2012]), but instead of minimizing cross validation squared errors, our weights are based on maximization of cross validation likelihood. The second strategy we use the well-known Bayesian information criterion (BIC) weights.

2.3.1 Weights based on likelihood cross validation

In this section, we study the first strategy which is based on the likelihood cross validation. We first introduce the notion of Kullback-Leibler (KL) distance to evaluate the estimation accuracy of the density estimate. The KL distance is defined as the discrepancy between two distributions f and g as

$$D(f||g) = \int f(x) \log \frac{f(x)}{g(x)} dx = E \log \frac{f(x)}{g(x)}. \tag{2.5}$$

Note that the KL distance is not really a metric, it does not satisfy triangle inequality and it is not symmetric.

Our first method is to use likelihood cross validation to find the weights. Since we have an i.i.d. sample $S = (x_1, \cdots, x_n)$, we write $\hat{f}^{(-j)}(x)$ as the estimator of $f(x)$ with the jth data removed from the sample, $j = 1, \cdots, n$. We then define the log likelihood of data point x_j evaluated by the model estimated using data $\{x_1, \cdots, x_{j-1}, x_{j+1}, \cdots, x_n\}$ as $\log \hat{f}^{(-j)}(x_j)$, then our likelihood cross validation criterion is formulated to be

$$CV(w) = \sum_{j=1}^{n} \log \hat{f}^{(-j)}(x_j) \tag{2.6}$$

where

$$\hat{f}^{(-j)}(x_j) = \sum_{i=1}^{K} \omega_i \hat{f}_i^{(-j)}(x_j)$$

ω_i is the weight for normal mixture model with component number i and $\hat{f}_i^{(-j)}(x_j)$ is the estimator of normal mixtures with component number i evaluated at point x_j with the jth data removed from the sample. K is the largest number of components we use in the normal mixtures. Then weight ω is then selected via

$$\hat{w} = \text{agrmax}_{w \in \mathcal{W}} CV(w). \tag{2.7}$$

We try to maximize the likelihood of the sample using leave one out cross validation estimators, which is equal to minimize the KL distance between the underlying true model and our proposed ANM model.

Theorem 1. *Under assumptions $A.1 - A.3$ presented in Appendix A, we have*

$$\frac{D(f||\sum_j \hat{\omega_j} \hat{f}_j)}{\inf_{\omega} D(f||\sum_j \omega_j f_j)} \to 1 \tag{2.8}$$

in probability as $n \to \infty$.

Note $D(\cdot\,||\cdot\,)$ in Theorem 1 stands for KL distance between two distributions. Theorem 1 states

that our ANM estimator is asymptotically optimal in the sense that the KL distance is asymptotically identical to that between true density and the infeasible best possible model average estimator. The detailed proof for Theorem 1 is given in the Appendix.

2.3.2 Weights based on smooth Bayesian information criterion (BIC)

The Bayesian information criterion (BIC) of Schwarz et al. [1978] takes the form of a penalised log-likelihood function. In detail,

$$BIC_k = -2\log(L_k) + \log(n)\dim(k), \tag{2.9}$$

for each candidate model k, L_k is the maximized value of the likelihood function for the estimated model, $\dim(k)$ is the number of parameters estimated in the model, and n is the sample size of the data. Best model is usually chosen by minimizing corresponding BIC value of the model. It has been proved that when sample size is large enough, the BIC criterion will choose the true model with probability tending to 1.

Our second method is to use the BIC weights (Buckland et al. [1997]) to combine different mixture models. Suppose there are K underlying models, BIC weights are defined as

$$P(f_k) = \frac{\exp\{-\frac{1}{2}\Delta BIC_k\}}{\sum_{j=1}^{K}\exp\{-\frac{1}{2}\Delta BIC_j\}}, \tag{2.10}$$

where $\Delta BIC_k = BIC_k - BIC_{min}$, BIC_{min} is the minimum BIC_k over the K models. We then prove that the proposed weights are consistent in selecting the true model (if the true model is in the candidates set) or the quasi-true model (if the true model is not in the candidates set).

Here we define the quasi-true model(Buckland et al. [1997]) as follows:

For a set of K models, the Kullback-Leibler distance of model g_i from the true density f is denoted by $D(f||g_i)$. We assume the models are indexed from worst (g_1) to best (g_K), so that $D(f||g_1) \geq D(f||g_2) \geq \cdots \geq D(f||g_K)$. Let T be the tail subset of the models defined by $\{g_r, r \geq t, 1 \leq t \leq K | D(f||g_{t-1}) > D(f||g_t) = \cdots = D(f||g_K)\}$. When $t = K$, Set T only

contains the best model which minimizes the KL distance from the true density f. For the case when T contains more than one model (i.e., $1 \leq t < K$), we assume the models g_t to g_K are ordered such that $\dim(t) < \dim(t+1) \leq \cdots \leq \dim(K)$. The set T contains models that are all equally good approximations by KL distance to truth f. However, we can further distinguish them by their parameter space dimension, and we prefer the smallest dimension model. If $t < K$, and $\dim(t) < \dim(t+1)$ holds, then model g_t is the unique quasi-true model of the K models. With the definition of quasi-true model, we can prove the following theorem.

Theorem 2. *If there exists a true model f_i in the candidates set, then $P(f_i) \to 1 \, as \, n \to \infty$; if there does not exist a true model in the candidates set, then when $n \to \infty$, $with \; probablity \; P(f_i) \to 1$ the corresponding BIC weights will select the quasi-true model f_i.*

The details of the proof of Theorem 2 will also be given in the Appendix.

2.4 Simulations

In this section, we present Monte Carlo simulations of the proposed Averaged Normal Mixture(ANM) model. We use the ANM estimator to approximate various kinds of densities. We start with densities actually generated by normal mixture models, then we estimate some smooth densities. We set the sample size to be 50 and 100, all specifications repeat 1000 times. We report mean integrated square error (MISE) and mean absolute error (MAE) of ANM model based on two proposed estimation strategies, we also present the corresponding model selection results as comparison, which is the best model selected by likelihood cross validation criterion and BIC criterion.

Before we can really estimate the proposed ANM model, we need to choose the value of K in ANM model, which is the largest number of components we use in the normal mixtures f_k, theoretically we can set K to be arbitrarily large, here we choose $K = 5$, it should be large enough for a modest sample size. Also we can use a screening method, we know that one easy way to select the number of components in normal mixture models is to use information theoretic approaches based on penalized likelihood, such as AIC criterion. Studies have shown that the AIC

type criterion typically overestimates the number of components substantially, so we can use the number of components chosen by AIC criterion as our maximum number of components in the ANM model. The simulation results of ANM estimator using screening method and fixed K are very similar, here we just report the results of using fixed $K = 5$.

2.4.1 Simulation type I: densities generated by normal mixtures

As the family of normal mixture is extremely flexible, Marron and Wand [1992] used it to represent a wide range of densities in their study of the mean integrated squared error of the kernel density estimators. We select 8 examples out of their univariate normal mixture densities, whose coefficients are presented as follows,

Case 1:Gaussian	$N(0,1)$
Case 2:Skewed	$\frac{1}{5}N(0,1) + \frac{1}{5}N(\frac{1}{2},(\frac{2}{3})^2) + \frac{3}{5}N(\frac{13}{15},(\frac{5}{9})^2)$
Case 3:Strongly Skewed	$\sum_{i=0}^{7}\frac{1}{8}N(3\{(\frac{2}{3})^i - 1\},(\frac{2}{3})^{2i})$
Case 4:Kurtotic	$\frac{2}{3}N(0,1) + \frac{1}{3}N(0,(\frac{1}{10})^2)$
Case 5:Bimodal	$\frac{1}{2}N(-1,(\frac{2}{3})^2) + \frac{1}{2}N(1,(\frac{2}{3})^2)$
Case 6:Separated Bimodal	$\frac{1}{2}N(-1.5,(\frac{1}{2})^2) + \frac{1}{2}N(1.5,(\frac{1}{2})^2)$
Case 7:Asymmetric Bimodal	$\frac{3}{4}N(0,1) + \frac{1}{4}N(1.5,(\frac{1}{3})^2)$
Case 8:Trimodal	$\frac{9}{20}N(-\frac{5}{6},(\frac{3}{5})^2) + \frac{9}{20}N(\frac{5}{6},(\frac{3}{5})^2) + \frac{1}{10}N(0,(\frac{1}{4})^2)$

Table 2.1 and Table 2.2 displays how close the estimated density is to the true density in terms of mean integrated squared estimation error (MISE) and mean absolute error (MAE) with sample size 50 and 100 across the four methods we mentioned at the beginning of this section: ANM model based on likelihood cross validation weights (mixcv) and BIC weights (mixbic), as well as the best model selected by likelihood cross validation criterion (mscv) and BIC criterion (msbic). Except for MISE and MAE, we also report the ratios between ANM estimator and model selection estimator. Therefore, our proposed ANM estimator is superior to its corresponding model selection estimator if the ratio is larger than 1.

Table 2.1: Simulation results on Marron and Wand densities, sample size:50

		mise_mixcv	**mise_mscv**	**mise_mixbic**	**mise_msbic**	**R1**	**R2**
Case 1	mean_mise	0.3576	0.3505	0.2460	0.2795	0.9802	1.1365
	mean_mae	3.9530	3.6091	3.2975	3.3164	0.9130	1.0057
	med_mise	0.1876	0.1092	0.1121	0.1034	0.5819	0.9224
	med_mae	3.6058	2.9265	2.9520	2.8530	0.8116	0.9665
	sd_mise	0.4937	0.9450	0.6444	0.8359		
	sd_mae	2.1880	2.7124	2.0145	2.2932		
Case 2	mean_mise	1.2682	1.7862	1.3363	1.5772	1.4085	1.1803
	mean_mae	7.2119	7.8373	6.8129	7.4378	1.0867	1.0917
	med_mise	0.7216	0.6701	0.5337	0.6365	0.9286	1.1926
	med_mae	6.8014	7.0230	6.2475	6.9090	1.0326	1.1059
	sd_mise	4.1281	9.8779	8.9903	9.8628		
	sd_mae	3.0780	4.1293	3.3683	3.6851		
Case 3	mean_mise	8.5930	11.0700	9.5299	11.2080	1.2883	1.1761
	mean_mae	18.9294	21.2735	20.2297	22.0077	1.1238	1.0879
	med_mise	7.2620	8.6946	7.7990	9.1814	1.1973	1.1773
	med_mae	18.4970	20.7176	19.8498	21.3667	1.1201	1.0764
	sd_mise	6.5200	9.8527	7.4492	9.5252		
	sd_mae	5.2362	5.6407	5.4450	5.2564		
Case 4	mean_mise	9.4843	13.1749	10.6850	11.5854	1.3891	1.0843
	mean_mae	17.5476	18.6224	19.1525	19.4749	1.0613	1.0168
	med_mise	6.4263	5.7273	7.0773	6.4784	0.8912	0.9154
	med_mae	16.7878	16.0692	17.4430	16.7802	0.9572	0.9620
	sd_mise	18.9931	61.4051	10.8671	13.1412		
	sd_mae	7.2696	9.7777	8.9838	10.2575		
Case 5	mean_mise	0.5489	0.7272	0.5815	0.7395	1.3249	1.2716
	mean_mae	5.5127	6.5072	5.9253	6.7718	1.1804	1.1429
	med_mise	0.4557	0.5828	0.4985	0.6126	1.2789	1.2289
	med_mae	5.5579	6.4094	5.9069	6.5791	1.1532	1.1138
	sd_mise	0.4263	0.6206	0.4456	0.5736		
	sd_mae	1.8550	1.9294	1.8192	1.5653		
Case 6	mean_mise	0.6847	0.6974	0.7398	0.7639	1.0186	1.0325
	mean_mae	6.0417	5.9374	5.9947	6.0106	0.9827	1.0026
	med_mise	0.5102	0.4444	0.4576	0.4477	0.8710	0.9785
	med_mae	5.7753	5.5062	5.6090	5.5360	0.9534	0.9870
	sd_mise	0.8592	1.3666	2.2287	2.4472		
	sd_mae	2.2904	2.6797	2.5037	2.6405		
Case 7	mean_mise	0.7623	1.0371	0.8819	1.0784	1.3606	1.2229
	mean_mae	6.2717	7.2648	6.6857	7.5347	1.1584	1.1270
	med_mise	0.6676	0.8027	0.7052	0.8473	1.2023	1.2016
	med_mae	6.2244	7.0315	6.4740	7.2056	1.1297	1.1130
	sd_mise	0.5600	1.0729	41.9631	0.9813		
	sd_mae	2.1038	2.2637	2.5103	1.9194		
Case 8	mean_mise	0.5651	0.7913	0.6445	0.7855	1.4003	1.2188
	mean_mae	5.7308	6.8385	6.3582	7.1397	1.1933	1.1229
	med_mise	0.4790	0.6954	0.5768	0.7176	1.4520	1.2439
	med_mae	5.5690	7.0234	6.4707	7.3539	1.2612	1.1365
	sd_mise	0.3740	0.6354	0.4536	0.5565		
	sd_mae	1.7161	1.9109	1.8226	1.7356		

Table 2.2: Simulation results on Marron and Wand densities, sample size:100

		mise_mixcv	**mise_mscv**	**mise_mixbic**	**mise_msbic**	**R1**	**R2**
Case 1	mean_mise	0.2041	0.2705	0.0784	0.0837	1.3255	1.0677
	mean_mae	2.9372	2.8405	2.0777	2.0879	0.9671	1.0049
	med_mise	0.1062	0.0544	0.0457	0.0438	0.5123	0.9585
	med_mae	2.6509	2.0793	1.8670	1.8543	0.7844	0.9932
	sd_mise	0.3032	0.7223	0.1377	0.1857		
	sd_mae	1.6593	2.3952	1.1841	1.2864		
Case 2	mean_mise	0.6010	0.8483	0.4793	0.5761	1.4116	1.2019
	mean_mae	5.4666	6.2010	5.2806	5.8340	1.1343	1.1048
	med_mise	0.4179	0.4455	0.3503	0.4228	1.0661	1.2072
	med_mae	5.2579	5.6688	5.0288	5.5893	1.0782	1.1115
	sd_mise	0.6531	1.2759	0.6045	0.7113		
	sd_mae	2.2849	3.1998	2.2396	2.3331		
Case 3	mean_mise	4.7650	5.9730	5.5357	6.2453	1.2535	1.1282
	mean_mae	14.2884	15.8239	15.6298	16.6705	1.1075	1.0666
	med_mise	4.0194	4.6977	4.7155	5.2184	1.1688	1.1066
	med_mae	13.9162	15.3746	15.2745	16.2692	1.1048	1.0651
	sd_mise	3.3254	4.6813	4.0518	4.5029		
	sd_mae	3.8427	4.2668	4.2128	4.3426		
Case 4	mean_mise	4.2378	4.5464	4.0355	4.0841	1.0728	1.0120
	mean_mae	11.9597	11.7755	11.4730	11.4491	0.9846	0.9979
	med_mise	2.8228	2.4220	2.3090	2.2698	0.8580	0.9830
	med_mae	11.0363	10.4572	10.1034	10.0493	0.9475	0.9946
	sd_mise	4.1765	6.5871	4.8932	5.3708		
	sd_mae	5.3054	6.1341	6.1766	6.4383		
Case 5	mean_mise	0.3072	0.4003	0.3294	0.3974	1.3030	1.2062
	mean_mae	4.0769	4.4492	4.5072	4.9572	1.0913	1.0998
	med_mise	0.2431	0.2606	0.2957	0.4180	1.0720	1.4137
	med_mae	3.9548	4.1661	4.5541	5.2844	1.0534	1.1604
	sd_mise	0.2512	0.4886	0.2215	0.2670		
	sd_mae	1.4661	1.9723	1.5639	1.7829		
Case 6	mean_mise	0.3059	0.3324	0.2385	0.2404	1.0870	1.0080
	mean_mae	4.0855	4.0775	3.7394	3.7380	0.9981	0.9996
	med_mise	0.2300	0.2014	0.1789	0.1749	0.8756	0.9778
	med_mae	3.9498	3.7221	3.4728	3.4527	0.9423	0.9942
	sd_mise	0.2767	0.4596	0.2068	0.2205		
	sd_mae	1.6044	1.9261	1.5443	1.5709		
Case 7	mean_mise	0.4179	0.5434	0.5105	0.6239	1.3003	1.2222
	mean_mae	4.6125	5.0757	5.1633	5.7203	1.1004	1.1079
	med_mise	0.3488	0.3812	0.4547	0.6473	1.0930	1.4237
	med_mae	4.4694	4.8012	5.2101	6.0116	1.0742	1.1538
	sd_mise	0.3181	0.5519	0.3887	0.4329		
	sd_mae	1.6237	2.1290	1.7743	1.8911		
Case 8	mean_mise	0.3438	0.4773	0.3725	0.4223	1.3884	1.1337
	mean_mae	4.3834	4.8194	4.8073	5.0760	1.0995	1.0559
	med_mise	0.2794	0.3208	0.3134	0.3252	1.1480	1.0378
	med_mae	4.2351	4.5054	4.6219	4.7176	1.0638	1.0207
	sd_mise	0.3133	1.3227	0.2358	0.2881		
	sd_mae	1.3193	1.6877	1.5298	1.8316		

As the simulation results show, all of the estimators improve with the sample size increases, which means with larger sample size, we have smaller MISE and MAE value for all estimators. We also notice that the ratio R1 (mean of MISE (or MAE) for ANM model using likelihood cross validation weights/mean of MISE (or MAE) for best selected model by likelihood cross validation criterion) and R2 (mean of MISE (or MAE) for ANM model using BIC weights/mean of MISE (or MAE) for best selected model by BIC criterion) are basically the same when the sample size increases, so we focus on the case when sample size is 50.

For both of ANM estimators with likelihood cross validation weights and BIC weights, they basically behave better or similar compared to their corresponding best selected model. In Case 1 and Case 6, the ANM estimators yield similar results with their corresponding best selected model, it is reasonable since in Case 1, the underlying true distribution is a simple Gaussian distribution, even though the ANM estimators put most of weights on the normal mixture model with one component, we still have a lot of parameters to estimate and it will introduce a lot of noises in the model, also in this case it is not hard for model selection algorithm to identify a Gaussian distribution. Case 6 is Separated Bimodal, in this case it is relatively easy for the model selection algorithm to select the appropriate number of components, since it is two separate Gaussian distributions with same shape but different locations that relatively far from each other. In other cases like Case 2 and Case 4, which are Skewed and Kurtotic distributions, our proposed ANM estimators behave slightly better than the best selected models, especially for the mean of the MISE, which means our proposed estimator is more robust to outliers. In other cases like Case 3, Case 5, Case 7 and Case 8, our proposed ANM estimators behave significantly better than the best selected models, for the ANM estimator with likelihood cross validation weights, it improve the results by 12% to 45% according to different criterion, the ANM estimator with BIC weights also improve the results by 7% to 27%. Since in all these cases, the underlying distributions are either highly asymmetric or generated by normal mixtures hard to separate. Also we notice that almost all the standard deviation of the ANM estimators are smaller than its corresponding selection models, which means our proposed estimators are more robust and stable.

2.4.2 Simulation type II: smooth densities

In this part we run experiments on some densities not generated by normal mixture models. Bai and Ng [2005] used six symmetric and eight skewed distributions in their paper, these densities include t distribution, log-normal distribution and chi-squared distribution, as well as some other distributions generated from the generalized lambda family. This family encompasses a range of symmetric and asymmetric distributions. The coefficients of these densities are listed as follows,

$S1$ t_5

$S2$ $e_1 I(z \leq .5) + e_2 I(z > .5), where\ z \sim U(0,1), e_1 \sim N(-1,1), and\ e_2 \sim N(1,1)$

$S3$ $F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1-u)^{\lambda_4}]/\lambda_2, \lambda_1 = 0, \lambda_2 = .19754, \lambda_3 = .134915, \lambda_4 = .134915$

$S4$ $F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1-u)^{\lambda_4}]/\lambda_2, \lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -.08, \lambda_4 = -.08$

$S5$ $F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1-u)^{\lambda_4}]/\lambda_2, \lambda_1 = 0, \lambda_2 = -.397912, \lambda_3 = -.16, \lambda_4 = -.16$

$S6$ $F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1-u)^{\lambda_4}]/\lambda_2, \lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -.24, \lambda_4 = -.24$

$A1$ $lognormal : exp(e), e \sim N(0,1)$

$A2$ $\mathcal{X}_2^2$

$A3$ $exponential : -\log(e), e \sim N(0,1)$

$A4$ $F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1-u)^{\lambda_4}]/\lambda_2, \lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 1.4, \lambda_4 = .25$

$A5$ $F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1-u)^{\lambda_4}]/\lambda_2, \lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -.0075, \lambda_4 = -.03$

$A6$ $F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1-u)^{\lambda_4}]/\lambda_2, \lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -.1, \lambda_4 = -.18$

$A7$ $F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1-u)^{\lambda_4}]/\lambda_2, \lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -.001, \lambda_4 = -.13$

$A8$ $F^{-1}(u) = \lambda_1 + [u^{\lambda_3} - (1-u)^{\lambda_4}]/\lambda_2, \lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -.0001, \lambda_4 = -.17$

Table 2.3 and Table 2.4 display how close estimated density is to the true density in terms of MISE and MAE with sample size 50 and 100 across the four methods we mentioned above. We also report the boxplots for readers to visually compare the results. Since several distributions yield similar results, for the ease of presentation, we just report the simulation results of $S1, S3, S5, S6, A1, A2, A3$ and $A5$.

Table 2.3: Simulation results on Bai and Ng densities, sample size:50

		mise_mixcv	mise_mscv	mise_mixbic	mise_msbic	R1	R2
S1	mean_mise	0.3494	0.4534	0.2997	0.3552	1.2976	1.1852
	mean_mae	4.0383	4.2993	3.7131	3.9123	1.0646	1.0536
	med_mise	0.2229	0.1824	0.1628	0.1652	0.8180	1.0143
	med_mae	3.7962	3.6716	3.3733	3.4864	0.9672	1.0335
	sd_mise	0.4448	0.8952	0.5635	0.7719		
	sd_mae	1.8362	2.6096	1.9965	2.2573		
S3	mean_mise	0.4029	0.4061	0.3105	0.3444	1.0081	1.1091
	mean_mae	4.1992	3.8375	3.4601	3.4725	0.9139	1.0036
	med_mise	0.2034	0.1151	0.1179	0.1104	0.5661	0.9361
	med_mae	3.6977	2.9952	3.0054	2.9534	0.8100	0.9827
	sd_mise	0.6085	1.0755	1.1081	1.3344		
	sd_mae	2.3593	3.0405	2.2108	2.4221		
S5	mean_mise	0.8692	1.2455	1.0259	1.1973	1.4329	1.1670
	mean_mae	6.4247	7.3188	6.2894	6.7534	1.1392	1.0738
	med_mise	0.5793	0.6427	0.4923	0.5495	1.1095	1.1163
	med_mae	6.1265	6.7240	5.8000	6.2140	1.0975	1.0714
	sd_mise	0.9719	2.0938	2.9420	3.2275		
	sd_mae	2.7254	3.8082	3.1559	3.4498		
S6	mean_mise	2.3789	3.3349	2.7313	3.3942	1.4018	1.2427
	mean_mae	10.5402	11.9698	10.5768	11.3904	1.1356	1.0769
	med_mise	1.6269	1.7433	1.4115	1.6090	1.0715	1.1400
	med_mae	10.3045	10.9532	9.7898	10.5761	1.0630	1.0803
	sd_mise	3.3098	7.3877	5.8626	8.6260		
	sd_mae	4.4366	6.0786	5.1067	5.6657		
A1	mean_mise	2.7447	3.9344	3.0229	3.4887	1.4335	1.1541
	mean_mae	11.4409	13.6121	12.2048	13.2546	1.1898	1.0860
	med_mise	2.3018	3.0158	2.3526	2.7975	1.3102	1.1891
	med_mae	11.1123	13.2008	11.8453	12.9476	1.1879	1.0931
	sd_mise	1.9047	3.3200	2.4722	2.6923		
	sd_mae	3.1294	3.6157	3.4841	3.3007		
A2	mean_mise	1.7175	2.3164	1.9320	2.2476	1.3487	1.1633
	mean_mae	8.8003	10.1606	9.3219	10.1663	1.1546	1.0906
	med_mise	1.4603	1.8415	1.5572	1.7697	1.2610	1.1364
	med_mae	8.6824	9.8705	9.1278	9.8558	1.1368	1.0798
	sd_mise	1.0164	1.6730	1.5077	1.6423		
	sd_mae	2.0001	2.1972	2.2573	2.0897		
A3	mean_mise	6.9267	9.2566	7.5839	8.9737	1.3364	1.1833
	mean_mae	17.5792	20.3233	18.6510	20.2825	1.1561	1.0875
	med_mise	5.8884	7.3694	6.3373	7.2086	1.2515	1.1375
	med_mae	17.1729	19.6042	18.1731	19.5726	1.1416	1.0770
	sd_mise	4.7894	8.8557	6.0535	9.0155		
	sd_mae	3.9132	4.4376	4.2852	4.1072		
A5	mean_mise	915.1130	1254.9829	1020.2936	1261.4159	1.3714	1.2363
	mean_mae	212.9856	252.9258	221.7764	243.6075	1.1875	1.0984
	med_mise	661.5490	847.1154	635.6542	770.7268	1.2805	1.2125
	med_mae	204.7086	241.2889	209.9631	232.5765	1.1787	1.1077
	sd_mise	976.8954	1461.3492	2020.5386	3383.1201		
	sd_mae	81.6599	103.6218	89.7911	94.1933		

Table 2.4: Simulation results on Bai and Ng densities, sample size:100

		mise_mixcv	**mise_mscv**	**mise_mixbic**	**mise_msbic**	**R1**	**R2**
S1	mean_mise	0.2128	0.3067	0.1591	0.1864	1.4415	1.1718
	mean_mae	3.2324	3.6139	2.9548	3.1799	1.1180	1.0762
	med_mise	0.1445	0.1374	0.1038	0.1173	0.9511	1.1295
	med_mae	3.0668	3.1196	2.7478	2.9369	1.0172	1.0688
	sd_mise	0.2461	0.5340	0.1859	0.2406		
	sd_mae	1.4775	2.0721	1.4047	1.5263		
S3	mean_mise	0.2094	0.2769	0.0908	0.0977	1.3220	1.0766
	mean_mae	2.9736	2.8482	2.1105	2.1077	0.9578	0.9987
	med_mise	0.1085	0.0626	0.0489	0.0467	0.5774	0.9534
	med_mae	2.7150	2.2116	1.9629	1.9311	0.8146	0.9838
	sd_mise	0.3882	1.3561	0.2828	0.4568		
	sd_mae	1.5943	2.2889	1.1711	1.2484		
S5	mean_mise	0.5201	0.7825	0.4850	0.6217	1.5044	1.2819
	mean_mae	5.0311	5.6599	4.8519	5.2186	1.1250	1.0756
	med_mise	0.3783	0.3815	0.2983	0.3460	1.0087	1.1600
	med_mae	4.9241	5.1604	4.5781	4.9069	1.0480	1.0718
	sd_mise	0.5856	1.4539	1.2330	2.7664		
	sd_mae	2.1157	2.9667	2.2329	2.4342		
S6	mean_mise	1.3081	1.8922	1.1257	1.2847	1.4465	1.1412
	mean_mae	7.9234	8.4994	7.5346	7.8935	1.0727	1.0476
	med_mise	0.9486	0.7811	0.7427	0.7710	0.8234	1.0380
	med_mae	7.6314	7.3535	7.0661	7.2283	0.9636	1.0230
	sd_mise	1.3532	3.2685	1.6317	2.0373		
	sd_mae	3.2101	4.6671	3.3855	3.7799		
A1	mean_mise	1.6519	2.1787	1.7642	2.0202	1.3189	1.1451
	mean_mae	8.9507	10.2403	9.5257	10.2128	1.1441	1.0721
	med_mise	1.4180	1.8197	1.5493	1.7598	1.2833	1.1359
	med_mae	8.7870	9.9937	9.3931	10.1008	1.1373	1.0753
	sd_mise	1.0012	1.4673	1.2688	1.5154		
	sd_mae	2.2853	2.6855	2.2910	2.3037		
A2	mean_mise	1.1494	1.4301	1.2771	1.4430	1.2442	1.1300
	mean_mae	7.1076	7.9586	7.6371	8.2284	1.1197	1.0774
	med_mise	1.0283	1.2667	1.1644	1.3026	1.2319	1.1187
	med_mae	7.0331	7.8419	7.5366	8.1326	1.1150	1.0791
	sd_mise	0.5642	0.7291	0.6220	0.6720		
	sd_mae	1.4526	1.5441	1.5085	1.3609		
A3	mean_mise	4.6559	5.8048	5.1064	5.7516	1.2468	1.1264
	mean_mae	14.3232	15.9969	15.2523	16.4402	1.1169	1.0779
	med_mise	4.2012	4.9889	4.5707	5.1596	1.1875	1.1288
	med_mae	14.1406	15.7244	14.9917	16.2461	1.1120	1.0837
	sd_mise	2.3365	3.5213	2.6160	2.6791		
	sd_mae	2.9759	3.2019	3.1208	2.8539		
A5	mean_mise	507.5066	712.5088	502.1480	647.2115	1.4039	1.2889
	mean_mae	161.5284	183.7813	161.2635	169.0281	1.1378	1.0481
	med_mise	402.5623	431.1228	341.3955	361.8346	1.0709	1.0599
	med_mae	158.7149	171.8432	152.8234	158.0199	1.0827	1.0340
	sd_mise	458.3201	1009.8079	1714.6962	4761.0670		
	sd_mae	59.2563	79.9294	61.3866	67.1398		

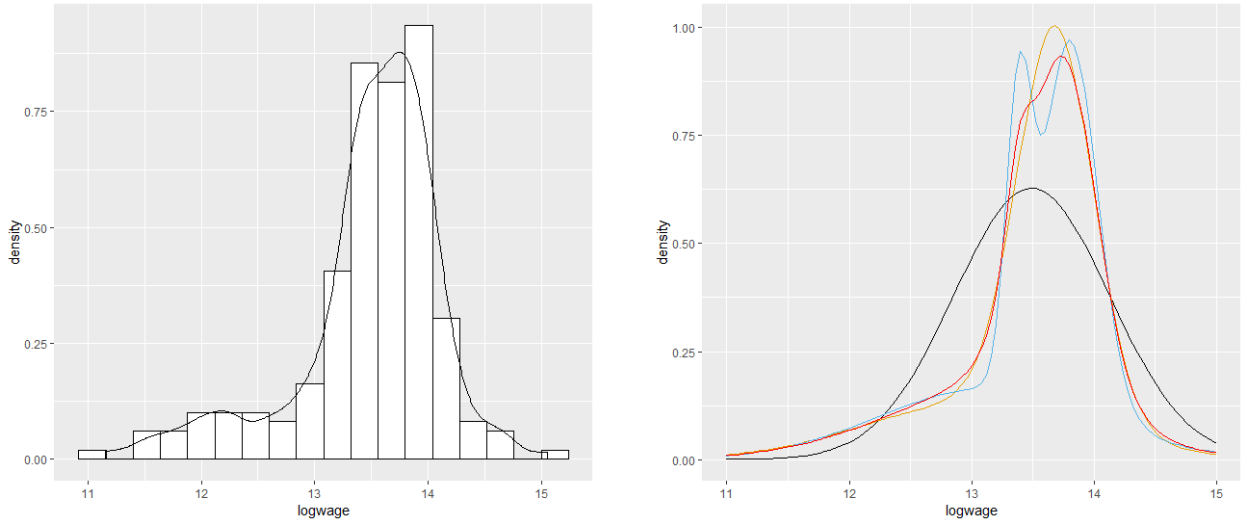
The simulation results show that our proposed ANM estimators also yield good performance on the densities not generated by normal mixture models. Similar to the simulation results of previous subsection, all of the estimators improve with the sample size increases, the ratio R1 and R2 are basically stay unchanged when the sample size increases, so we focus on when sample size is 50. For cases like $S1, S3, S5$ and $S6$, the ANM estimators yield similar results with their corresponding best selected model, since in all these cases the underlying densities are symmetric and similar to Gaussian distributions, they are relatively easy for the model selection algorithm to identify, or even if the model selection algorithm could not correctly identify the underlying distribution structure, the misspecification loss is relatively small. Also notice that in almost all of these cases, mean of the MISE of the proposed ANM estimators are much smaller than their corresponding best selected estimators, which means our proposed estimators are more robust to outliers. For cases like $A1, A2, A3$ and $A5$, our proposed ANM estimators behave significantly better than the best selected models, for the ANM estimator with likelihood cross validation weights, it improve the results by 14% to 43% according to different criterion, the ANM estimator with BIC weights also improve the results by 8% to 23%. From the simulations results we can tell that our proposed ANM estimators behave better if the underlying densities are asymmetric. Also the standard deviation of the ANM estimators are smaller than its corresponding selection models, so our proposed estimators are more stable.

2.5 Empirical applications

In this section we apply the proposed ANM estimator to the real world example we showed in the introduction. The dataset contains a random sample taken from the 1971 Canadian Census Public Use Tapes for male individuals having common education level (grade 13) and there are 205 observations in total. For the ease of presentation, we just show the ANM estimator using cross validation weights.

As we discussed in the introduction, normal mixture density estimator is unstable if we change the number of components used in the model, it tends to underfit the underlying density if we choose a small value for the number of components and overfit it if we choose a large value for

the number of components. We show our proposed ANM density estimator in red line in Figure 2.2 (b), it contains more information compared to the normal mixtures density estimator with 2 components which is shown in the orange line, and its shape is more similar to the kernel density estimator in Figure 2.2 (a). Also the ANM density estimator is more "smooth" compared to the normal mixture density estimator with 3 components and shows no tendency of overfitting. For the best selected model, the model selection algorithm based on likelihood cross validation chooses normal mixtures with 3 components. Clearly it is the most wiggly one and overfit the underlying distribution.



(a) (b)

Figure 2.2: Density estimation of "cps71" data set, empirical study
(a)kernel density estimator and histogram for logarithm of data. (b)normal mixture density estimators using components number from $k = 1$ to $k = 3$, which corresponds to black, orange and blue curve, and we denote our proposed ANM estimator as red line.

2.6 Conclusion

In this study, we propose an Averaged Normal Mixture model for density estimation based on normal mixture models. Instead of selecting the appropriate number of components in a normal mixture model, we first estimate a series of normal mixture models with different number of com-

ponents, then we take these estimated normal mixture models as given, and we find different ways to give appropriate weights to these models and combine all these models. This new method is more stable and generally more accurate than the best selected normal mixture models. We propose two methods to find the appropriate weights in the Averaged Normal Mixture model, one is based on likelihood cross validation and the other one is based on BIC weights. We have established the theoretical properties of the proposed estimator and the simulation results demonstrate its good performance on different kind of densities. Finally, we illustrate that our proposed estimator behaves well on a real world data set. For future studies, we can extend the univariate cases to multivariate cases and explore the properties of the proposed estimator in high dimension.

3. GAUSSIAN PROCESS MODELS OF ELECTRICITY DEMAND FORECASTING

3.1 Introduction

During the past few decades, electricity demand forecasting has played an increasingly important role in the electric power industry. It is vital to many aspects of the electricity industry such as power system planning and operation, transmission design, and financial risk management. Since electricity is difficult to store, supply and demand have to be balanced at every point in time. Consequently, overestimation of electricity demand may result in excessive purchase and an unnecessary waste of energy while underestimation may cause disturbance in the power system. There is a vast literature on the forecasting of electricity demand, ranging from long, medium to short term demand. See e.g. Hong and Fan [2016] for an overall review. In this paper we focus on the short term demand forecasting. More specifically, given historical hourly electricity demand data, we shall predict the 24 hourly electricity demand for the next day. This is of particular importance to many electricity suppliers as they customarily submit daily bid schedules 24 hours ahead of time.

Many methods have been applied to forecast electricity demand, largely focusing on point forecasts. For an overview of common methods see Weron [2007] and Taylor and McSharry [2007]. Most widely used methods include linear regression models (Weron [2007]; Bianco et al. [2009]), exponential smoothing (Taylor [2003]) and ARIMA models (Huang and Shih [2003]; Erdogdu [2007]). More recently, machine learning techniques have been adopted for this purpose; see e.g. support vector regression (Chen et al. [2004a]; Kavaklioglu [2011]) and artificial neural networks (Hippert et al. [2001]; Taylor and Buizza [2002]). There also exists a small literature on probabilistic forecasting that predicts quantities such as the quantiles, intervals, or distribution/density functions (Fan and Hyndman [2012]; Hong and Fan [2016]). The probabilistic approach is advantageous in that it provides not only point estimates but other inferential information pertinent to forecasting uncertainty and therefore facilitates risk assessment and management.

Following the probabilistic approach, in this study we adopt the method of Gaussian Process modeling for electricity demand forecasting. Under the framework of Gaussian Process, each marginal distribution retains Gaussianity. We thus naturally obtain probabilistic forecasts, based on which point forecasts and other quantities of interest can be easily inferred. The Gaussian Process is a powerful nonparametric machine learning method for regression analysis and it has been widely used in time series forecasting. The key feature of Gaussian Process modeling is the construction of covariance matrix, through which the covariates influence the outcome of interest. Leith et al. [2004] first used Gaussian Process to forecast weekly Irish electricity demand, employing an exponential squared covariance function of smooth time trend and a seasonal component. Mori and Ohmi [2005] and Lourenco and Santos [2012] used similar Gaussian Process models with a variety of covariates. Blum and Riedmiller [2013] extended the model by Leith et al. [2004] by incorporating weather information. Alamaniotis et al. [2014] compared different covariance functions in forecasting electricity demand of the New England region.

In this study, we use the Gaussian Process models to predict short term electricity demand in the state of Texas. We focus our investigation on probabilistic forecast, which is indispensable to the electricity industry.

The main contribution of this study is a novel hybrid Gaussian Process forecasting strategy that combines point forecasts from the DirRec strategy and variance forecasts from the direct strategy (more on these strategies below). We show that the proposed hybrid forecasting approach outperforms other forecasting methods considerably. We also show that the Gaussian Process models provide superior forecasts relative to conventional regression models.

In addition to statistical investigation, we further illustrate how the probabilistic forecasts obtained from the Gaussian Process models can be used in a decision-theoretic framework to optimize economic decision making and risk management in the electricity industry. In particular, we apply the proposed method to derive the optimal bidding strategy for electricity suppliers with a stylized asymmetric loss function. Our examples suggest that the proposed hybrid Gaussian Process models provide reliable and valuable probabilistic forecasts that inform and help facilitate operation

planning and risk management in the electricity industry.

The rest of this article is organized as follows. Section 3.2 gives a brief introduction of the Gaussian Process regression model. In Section 3.3, we describe the electricity market in Texas and the data used in this study. Section 3.4 presents the hybrid forecasting approach and section 3.5 reports the forecasting results. Section 3.6 provides some economic applications of our proposed method. The last section concludes.

3.2 Preliminaries on Gaussian Process models

In this section, we provide a brief introduction to the Gaussian Process regression models. Interested readers are referred to Rasmussen and Williams [2006] for an illuminating treatment of this subject. Gaussian process (GP) is a powerful nonparametric machine learning approach for regression and classification. It has also been widely used in time series analysis. The Gaussian process extends multivariate Gaussian distributions to infinite dimensionality. Formally, a Gaussian process generates data from a certain domain such that any finite subset within the range follows a multivariate Gaussian distribution and each single element of the set follows a Gaussian distribution. A GP regression model is formulated as follows. Consider a training set $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, N\}$ of N pairs of input x_i and output y_i from an underlying relationship f. Here f is typically assumed to be a zero-mean Gaussian process with a covariance (kernel) function $k(\cdot, \cdot)$, and the observations y_i are given by

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, N, \tag{3.1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are white noises independent of $f(x_i)$.

Let $x = [x_1, x_2, \ldots, x_N]'$ and $y = [y_1, y_2, \ldots, y_N]'$. Denote the predictive distribution of outcome at a test location x_* by $f_* = f(x_*)$. The joint distribution of (f_*, y) is then given by

$$\begin{bmatrix} f_* \\ y \end{bmatrix} = \mathcal{N}\left(0, \begin{bmatrix} k_{**} & k'_{x_*} \\ k_{x_*} & \sigma^2 I + K_{xx} \end{bmatrix}\right), \tag{3.2}$$

where $k_{**} = k(x_*, x_*)$, $k_{x_*} = (k(x_1, x_*), \ldots, k(x_N, x_*))'$, K_{xx} is an $N \times N$ matrix with the $(i,j)^{th}$ entity $k(x_i, x_j)$ and I is a N-dimensional identity matrix. The predictive distribution of f_* given y is

$$f_*|y = \mathcal{N}(k'_{x_*}(\sigma^2 I + K_{xx})^{-1} y, k_{**} - k'_{x_*}(\sigma^2 I + K_{xx})^{-1} k_{x_*}). \tag{3.3}$$

The predictive mean $k'_{x_*}(\sigma^2 I + K_{xx})^{-1} y$ gives the point forecast of $f(x)$ at location x_*, whose uncertainty is measured by the predictive variance $k_{**} - k'_{x_*}(\sigma^2 I + K_{xx})^{-1} k_{x_*}$. Note that the point forecast at location x_* depends on y and the various variance and covariance components, and is usually non-zero. The covariates influence the predictive outcome through the covariance. In this sense, the covariance is the determining factor of a GP predictor as it encodes our assumptions about the underlying relationship we wish to learn.

The most popular choice of the covariance function in GP models is the Squared Exponential (SE) covariance given by

$$k_{SE}(x_i, x_j) = \sigma_f^2 \exp\left[-\frac{(x_i - x_j)^2}{2l^2}\right], \tag{3.4}$$

where σ_f^2 reflects the maximum allowed covariance that usually increases with the variation of y. The so-called length scale l determines the relevancy of input x to the outcome y. To see this, note that the covariance between x_i and x_j vanishes under a sufficiently large length scale l, effectively removing it from the inference. A covariance function with this feature is called an Automatic Relevance Determination (ARD) covariance function. There exist a large selection of kernel functions that are suitable to model various functional relationship; see Chapter 4 of Rasmussen and Williams [2006] for details.

The performance of GP models hinges on the configuration of the covariance function and its tuning parameters, which are referred to as the hyperparameters in the machine learning literature. Given model (3.1) and the SE kernel (3.4), the covariance function for the training set takes the form

$$k(x_i, x_j) = \sigma_f^2 \exp\left[-\frac{(x_i - x_j)^2}{2l^2}\right] + \sigma^2 \delta(x_i, x_j), \tag{3.5}$$

where $\delta(x_i, x_j)$ is the Kronecker delta function. The hyperparameter of this model then consists of

$\theta = (\sigma_f^2, \sigma^2, l)$. One possibility of hyperparameter selection is via Maximizing A Posteriori (MAP) likelihood $p(\theta|x, y)$ of the GP model given the observed data. The log likelihood $\log p(y|x, \theta)$ is given by

$$\log p(y|x, \theta) = -\frac{1}{2} y' K_{xx}^{-1} y - \frac{1}{2} \log |K_{xx}| - \frac{N}{2} \log(2\pi), \tag{3.6}$$

where K_{xx} is an $N \times N$ matrix with the $(i, j)^{th}$ entity $k(x_i, x_j)$ given in (3.5). Owing to its close connection to the Bayesian analysis, this likelihood function $\log p(y|x, \theta)$ for a GP model is often referred to as the marginal likelihood function. The first part of the likelihood function $-\frac{1}{2} y' K_{xx}^{-1} y$ reflects the goodness of fit. The second part $-\frac{1}{2} \log |K_{xx}|$ can be viewed as a complexity penalty that depends on the covariance function and the inputs. The third part is a normalization constant. The presence of the complexity penalty, which is partially controled by the hyperparamters, in the objective function can effectively prevent overfitting.

3.3 ERCOT electricity market and data description

In this section, we provide a brief introduction to the electricity market in the state of Texas and discuss the data used in our analysis.

3.3.1 ERCOT electricity market

We focus on the electricity market of Texas in our study of electricity demand forecasting. The Electric Reliability Council of Texas (ERCOT) manages the flow of electric power to 24 million Texas customers, representing about 90 percent of the state's electric load. As the independent system operator for the region, ERCOT schedules power on an electric grid that connects more than 46,500 miles of transmission lines and 570+ generation units. It also performs financial settlement for the competitive wholesale bulk-power market and administers retail switching for 7 million premises in competitive choice areas. Participants in this electricity market mainly include generation companies, retail electric providers, consumers and transmission and distribution utilities.

In the ERCOT electricity market, a retail electric provider buys electricity from electricity generation companies and sells the electricity to consumers. Between generation companies and

retail electric providers, most of the electricity is traded via bilateral agreements one day ahead of the planned transactions. In addition to this bilateral market, ERCOT, as the system operator, administers a market to balance the real time electricity supply and demand. There exists two types of bilateral markets: the "day ahead market" and the "real time market." In the ERCOT electricity market, although most of the electricity is traded in the day ahead market, it is the short term fluctuations of electricity usage in the real time market that expose the participants in the electricity industry to substantial financial risk.

Typically in the day ahead market, firms submit to ERCOT an hourly schedule of electricity supply to inject and withdraw at specific locations and times for the next day. The actual electricity usage frequently deviates from the scheduled supply and demand due to a myriad of reasons. Whenever this occurs, firms have to increase or decrease their scheduled electricity supply accordingly. These real time adjustments are often costly. When the actual demand exceeds the scheduled supply, a retail electricity provider has to pay a premium on top of the spot price to acquire extra electricity from the producers. On the contrary, when the actual demand falls below the scheduled supply, the retail electricity provider can only unload the excessive supply at a price below the spot price. Therefore one-day-ahead forecasting of hourly electricity demand is of tremendous importance in this market. Reliable forecasts can effectively improve the overall efficiency of this market, reduce energy waste and increase overall social welfare.

3.3.2 Data description

ERCOT has made publicly available the electricity load data for its eight weather zones. This study focuses on electricity demand of the southern weather zone, one of the major weatherzones of Texas. Specifically we shall conduct one day ahead forecasts of electricity demands, which are most useful for the markets' participants. Each day is divided into 24 hourly periods, which correspond to ERCOT's daily market settlement periods.

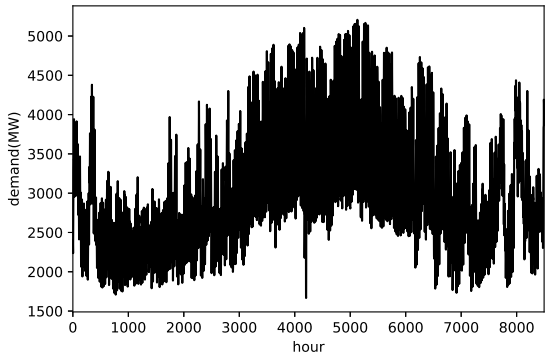
We use the hourly electricity data of year 2013 obtained from ERCOT website in our investigation. As we can see from Table 3.1, the hourly electricity load, in the unit of million watts per hour (MW/h), ranges from 1665.98 MW/h to 5206.73 MW/h, with an average demand of 3070.14

MW/h for the south weatherzone of Texas. Hourly weather data are obtained from the Local Climatological Database of National Oceanic and Atmospheric Administration (NOAA). Our hourly weather data contain temperature (in whole degrees Fahrenheit (F)), relative humidity (given to the nearest whole percentage) and wind speed (in miles per hour (mph)). We obtain these weather data from 3 weather stations in the southern regions of Texas and use their averages in our analysis. Since our investigation suggests little influence of humidity and wind speed on electricity usage, we only include temperature in this study. After removing national holidays and days with missing data, we end up with 354 days for year 2013, with a total of 8,496 hours. Summary statistics of the data are reported in Table 3.1. The hourly demand data for year 2013 are illustrated in Figure 3.1 (a), which shows clear seasonal and weekly patterns in electricity usage.

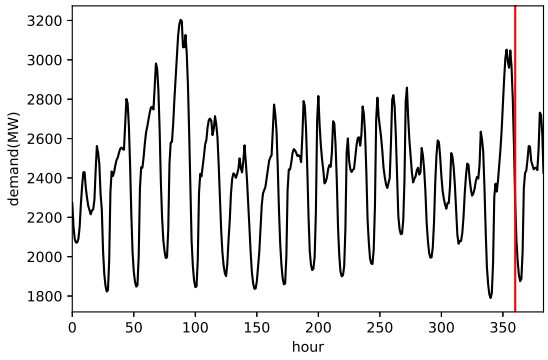
Table 3.1: Data summary statistics

	Load(MW)	Temperature(F)
Mean	3070.14	72.14
Min	1665.98	33.00
Median	2911.04	74.50
Max	5206.73	108.00
StDev	750.96	14.67

Given the objective of forecasting electricity demand one day ahead, we opt to use the previous 15 days' data in the prediction for any given day, which amounts to using the previous 360 hours' data to predict the next 24 hours' electricity demand. We have experimented with longer and shorter length of historical data in this investigation. The results are not sensitive to the choice of window length. To save space we only report forecasting results using 15 days' historical data. An illustration of the (360+24) hour window is given in Figure 3.1(b) for an arbitrarily selected day (March 5, 2013) in our sample. In this plot, we use a vertical line to separate the data used for estimation and those to be forecasted. There is an evident intra-day pattern that peaks in the afternoon. Below we shall randomly choose 120 days from the year of 2013 for forecasting and



(a) Electricity usage of the southern Texas in year 2013



(b) Electricity usage of March 5, 2013 and 15 days prior

Figure 3.1: Electricity usage of southern Texas

out-of-sample evaluation.

3.4 GP model of electricity demand

3.4.1 Design of covariance function

The key modeling component in the GP approach is the covariance matrix. There are a lot of covariance functions we can choose from besides the Squared Exponential form, such as the Matern, Periodic and linear kernels. Further flexibility is afforded as one can construct a composite covariance matrix with sum or product of kernels. See Rasmussen and Williams [2006] for a detailed discussion of covariance matrix for GP models. As noted above, the length scales of the kernels have the appealing property of automatic relevance determination such that irrelevant covariates are effectively removing from the modeling via data-driving selection of the hyperparameters. Our preferred model employs the following covariance function for electricity demand forecasting:

$$C_{forecast} = k_T + k_{temperature} + k_{day} + k_{hour} + k_{load_{-1}} \tag{3.7}$$

where k takes the form of Squared Exponential covariance function. This covariance function is constructed as a sum of univariate kernels of individual covariates. The input variables included in the model are time (T), temperature, day of the week (day), hour of the day ($hour$) and the

electricity usage of the previous hour ($load_{-1}$). In particular, we use the time variable to capture the long term trend of the electricity usage, and day of the week and hour of the day to model the intra-week and intra-day seasonality. We use temperature and the lagged electricity usage to help explain short term variations. We have also experimented with additional weather conditions such as relative humidity and wind speed. These variables turn out to have little effects on the estimation and forecasting and therefore are not included in our model. We use the MAP approach, discussed in the previous section, to select the hyperparameters.

3.4.2 Forecasting strategy

Our task of predicting 24 hourly electricity demands entails forecasts up to 24 steps ahead. Various strategies have been proposed in the literature to tackle the multiple-step-ahead forecast problem. Two commonly strategies are the Recursive strategy and the Direct strategy; see e.g. Taieb et al. [2012] and Xiong et al. [2013]. In the Recursive strategy, a single model is trained to perform a one-step ahead forecast; for multiple-step-ahead forecast, the previously forecasted values are used as input in subsequent forecasts (using the same one-step ahead model). In the Direct strategy, different models are constructed for each forecasting horizon separately; both one-step and multi-step forecasts use only historical observations up to the time of forecast.

In this paper, we propose a hybrid Gaussian Process model. In particular we adopt the DirRec strategy proposed by Sorjamaa and Lendasse [2006] for point forecasts. This approach is a combination of the Direct strategy and the Recursive strategy and hence the name DirRec. At every forecast time step, the DirRec strategy uses a different model (same as the Direct strategy) and incorporates the forecasted values from previous steps into the input set (same as Recursive strategy). At the same time, we use the Direct strategy proposed by Cox [1961] for variance forecasts.

Point Forecasts Based on the DirRec Strategy

We first describe our DirRec strategy for Gaussian Process point forecasts. For a given time t, we aim to forecast the electricity demand of the next 24 hours, namely $t+1, t+2, \cdots, t+24$. Our

forecast model can be written as:

$$y_t = G(T_t, temperature_t, day_t, hour_t, load_{t-1}) + \varepsilon_t \tag{3.8}$$

where G is a Gaussian processes model with zero mean and covariance function $C_{forecast}$. This model is trained using the previous 360 hours' data. Following the common practice in GP modeling, we standardize the independent variable in our estimation to improve its numerical stability.

We then use the following procedure to forecast one-day-ahead hourly electricity demand for the next 24 hours.

- We first predict the next period demand y_{t+1}, using information $T_{t-360}, \ldots,$ $hour_{t-360}$, $load_{t-361}, \ldots, T_t, hour_t, load_{t-1}$ to train the Gaussian process model G. Denote the estimated model by $\hat{G}_1$. We use $\hat{G}_1$ to predict the next period demand $\hat{y}_{t+1} = \hat{G}_1(T_{t+1}, \ldots, hour_{t+1}, load_t)$.

- Similarly we use $T_{t-360}, ..., hour_{t-360}$,$load_{t-361}, \ldots, T_{t+1}, hour_{t+1}, load_t$ to obtain an updated model $\hat{G}_2$. We then predict the electricity demand at time $t+2$ with $\hat{y}_{t+2} = \hat{G}_2(T_{t+2}, \ldots,$ $hour_{t+2}, \hat{y}_{t+1})$. Note that since $load_{t+1}$ is not known to the forecaster at the time of forecasting, we replace it with the prediction $\hat{y}_{t+1}$ from the previous step.

- We next use $T_{t-360}, ..., hour_{t-360}$,$load_{t-361}, \ldots, T_{t+2}, hour_{t+2}, \hat{y}_{t+1}$ to obtain an updated model $\hat{G}_3$ and predict the next period demand with $\hat{y}_{t+3} = \hat{G}_3(T_{t+3}, \ldots, hour_{t+3}, \hat{y}_{t+2})$. Similarly to the previous step, we use $\hat{y}_{t+2}$ in the place of $load_{t+2}$ in the forecasting step. This procedure is repeated to obtain the subsequent forecasts $\hat{y}_{t+4}, \cdots, \hat{y}_{t+24}$.

In the above forecast strategy, for every step we train a Gaussian Process model that includes previously forecasted results as part of its input. The advantage of this approach is that rather than only using observations up to the point of forecasting, it also incorporates proxy of more recent outputs, which may help improve multi-step-ahead predictions. However the forecasted outcomes are bound to differ from actual outcomes and along the course of this incremental incorporation of

previous forecasts, prediction errors tend to accumulate. Since the DirRec approach takes the previously forecasted outcomes as given and ignores forecasting uncertainty, its variance predictions tend to underestimate the true variance of the forecast and the degree of under-estimation generally increases with the forecasting horizon.

Variance Forecasts Based on the Direct Strategy

As noted above, the 'naive' variance forecasts from the DirRec approach tend to underestimate the true forecasting variation. One possible alternative to construct variance forecasts is to use Monte Carlo simulations. However in our case this method is going to be increasingly expensive as we proceed along the forecasting time path. We therefore choose to use the Direct strategy to construct the variance forecasts.

In particular, we use the following procedure to generate a second set of hourly electricity demand probabilistic forecasts for the next 24 hours. Given an arbitrary period t, we aim to forecast the electricity demand of periods $t+1, t+2, \cdots, t+24$. For an h-step ahead forecast, we use

$$y_{t+h} = G'(T_{t+h}, temperature_{t+h}, day_{t+h}, hour_{t+h}, y_t) + \varepsilon_{t,h} \tag{3.9}$$

where $h = 1, \ldots, H$, and

$$\hat{y}_{t+h} = \hat{G'}(T_{t+h}, temperature_{t+h}, day_{t+h}, hour_{t+h}, y_t) \tag{3.10}$$

In this procedure, for every forecast step we train a Gaussian Process model based only on past information of electricity demand up to y_t. Different from the DirRec approach, previously predicted outcomes are not incorporated as input in multi-step ahead forecasting. Although the point forecasts under this approach is generally not as accurate as those from the DirRec approach, the Direct approach tends to produce more reliable variance estimates for multi-step ahead forecasts. This is confirmed by our numerical experiments below.

Probabilistic Forecasts Based on the Hybrid Strategy

To combine the strength of the DirRec and Direct strategies and eschew their respective weakness, we propose to use the point estimates from the DirRec strategy and variance estimates from the Direct strategy. We term the resultant distribution a hybrid Gaussian Process forecast to reflect that it combines Gaussian Process point and variance forecasts from two complementary forecasting strategies. Since a Gaussian distribution is uniquely determined by its mean and variance, the resultant probabilistic prediction remains a well-defined Gaussian Process forecast. We note that combining point and variation estimates from different procedures is not uncommon in practice. For instance Mallow's C_p is commonly used for the purpose of model comparison/selection. This criterion, an estimate of the mean squared prediction error, is given by $C_p = 1/n(RSS + 2d\hat{\sigma}^2)$, where RSS and d are the residual sum of squares and number of covariates for a particular model in the candidate set, while $\hat{\sigma}^2$ is an estimate of the residual variance, which is customarily estimated based on the full model. Under this procedure, except for the full model, the point estimates (and the subsequent RSS) and the variance estimate $\hat{\sigma}^2$ are obtained from two different models. Another example of combining point and variance estimates from different models can be found in semi- and non-parametric estimations, wherein the optimal tuning parameters (such as the bandwidth for kernel-based estimation and the penalty parameter for penalized spline estimation) depend on the subject of estimation and often differ between the point and variance estimates.

3.5 Forecasting results

3.5.1 Comparison of point forecasts

We estimate our GP model on 120 randomly selected days in year 2013. For comparison, we also consider a benchmark linear regression model as considered in Hong [2010]. This model is

given by

$$\begin{aligned} y_t =& \beta_0 + \beta_1 * T_t + \beta_2 * T_t^2 \\ &+ \beta_3 * temperature_t + \beta_4 * temperature_t^2 \\ &+ \beta_5 * load_{t-1} + \beta_6 * hour_t + \beta_7 * day_t \\ &+ \beta_8 * T_t * hour_t + \beta_9 * T_t * day_t + u_t \end{aligned} \tag{3.11}$$

where u_t is an error term with mean zero and finite variance. Note here β_i denotes a single coefficient when the corresponding covariate is quantitative and a vector of coefficients when the covariate is categorical. To ensure direct compatibility, we use the same information input and forecasting strategies, as are used in the GP models, to obtain 24 hours' predictions.

We use the DirRec strategy, as described in the previous section, to obtain point forecasts for both the GP and linear regression models. We use the mean absolute percentage error (MAPE), mean absolute error (MAE) and root mean square error (RMSE) to evaluate the forecast performance. The results are reported in Table 3.2. For each hour of the day, we report the average performance across 120 randomly selected days used for forecasting. It is seen that the GP model outperforms the linear regression model consistently, often by substantial margins. The average results across all 24 hours are reported at the bottom of Table 3.2. On average, the GP models reduce the MAE by 34% and the RMSE by 29% relative to the linear regression models. Figure 3.2 illustrates the 24-hour-ahead demand point and interval forecasts averaged across the 120 randomly chosen days used in our forecasting. The forecasts closely track the actual electricity usage; at the same time, the precision of forecasts generally deteriorates as the forecast horizon increases. Also reported in Table 3.2 are results from the Direct forecasting strategy. Clearly in terms of point forecasts, the DirRec strategy is preferred to the Direct strategy.

In practice if temperature information is incorporated in the forecasting process, the forecasters have to resort to temperature forecasts as actual hourly temperature is not available at the time of forecasting. Since data on historical hourly temperature forecasts are not available to this study, we conduct our forecasting using the actual temperature. To check the sensitivity of our results

Table 3.2: Average performance of out-of-sample forecasts for 120 randomly selected days

	GP Model (DirRec)			GP Model (Direct)			Regression Model (DirRec)			Regression Model (Direct)		
Period	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE
1	0.0167	44.14	65.79	0.0167	44.14	65.79	0.0306	80.76	135.76	0.0306	80.76	135.76
2	0.0322	75.11	170.88	0.0356	83.13	176.64	0.0495	117.34	201.60	0.0714	169.41	308.44
3	0.0301	68.64	126.50	0.0429	95.68	181.26	0.0421	95.38	144.68	0.0974	220.06	415.78
4	0.0278	61.00	88.40	0.0490	107.39	174.10	0.0366	79.81	107.29	0.1187	259.18	505.97
5	0.0305	66.16	122.40	0.0553	120.23	175.56	0.0402	87.90	123.34	0.1347	288.28	582.32
6	0.0253	55.43	98.80	0.0614	134.00	202.50	0.0419	93.79	150.23	0.1461	313.09	660.97
7	0.0294	67.55	88.53	0.0668	153.43	208.98	0.0564	131.56	186.47	0.1481	336.10	739.83
8	0.0601	152.49	187.28	0.0824	204.99	275.74	0.0909	228.64	281.03	0.1464	368.44	843.25
9	0.0557	146.39	190.29	0.0844	217.81	280.71	0.0848	221.91	268.61	0.1533	398.94	954.94
10	0.0386	106.35	158.25	0.0758	202.87	267.82	0.0489	134.38	178.82	0.1547	411.47	1051.95
11	0.0338	97.20	155.92	0.0696	196.12	259.66	0.0342	99.02	142.62	0.1613	443.49	1133.40
12	0.0359	106.76	162.36	0.0706	209.62	280.04	0.0389	114.72	161.10	0.1657	468.16	1185.39
13	0.0358	108.04	143.58	0.0776	236.25	307.48	0.0474	141.17	184.22	0.1719	495.67	1215.37
14	0.0323	99.73	128.20	0.0808	252.81	324.16	0.0510	154.01	193.90	0.1774	519.50	1238.19
15	0.0307	97.21	127.35	0.0916	295.82	371.62	0.0487	151.02	184.15	0.1811	540.73	1273.00
16	0.0261	84.81	113.56	0.0975	318.87	394.56	0.0459	145.24	174.49	0.1843	558.08	1318.67
17	0.0261	85.00	113.14	0.0939	313.76	391.69	0.0407	131.48	162.58	0.1846	563.01	1355.12
18	0.0222	74.64	100.43	0.0943	318.40	399.82	0.0364	119.85	155.35	0.1843	566.06	1351.03
19	0.0229	77.37	104.56	0.0866	291.38	370.59	0.0387	127.55	168.36	0.1793	548.66	1295.94
20	0.0281	95.07	128.75	0.0775	261.04	335.27	0.0425	140.06	189.88	0.1636	510.13	1183.32
21	0.0260	84.47	116.40	0.0734	244.68	310.90	0.0456	146.94	198.70	0.1484	466.13	1037.43
22	0.0249	79.94	106.13	0.0681	224.98	288.23	0.0462	147.66	196.24	0.1368	429.07	913.13
23	0.0254	79.18	104.91	0.0711	226.36	284.42	0.0470	145.69	184.20	0.1354	412.99	859.21
24	0.0277	81.90	105.35	0.0594	177.17	236.53	0.0508	147.35	177.91	0.1404	401.09	862.50
Average	0.0310	87.27	129.16	0.0701	205.45	285.74	0.0473	132.63	181.18	0.1465	407.02	995.57

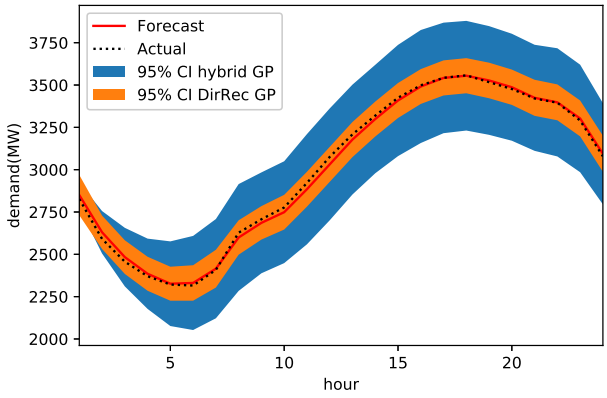


Figure 3.2: Averaged actual load vs predicted load from GP models

to this choice, we construct our own hourly temperature forecasts via a simple average of those of the previous three days. We then conduct the same forecasting exercise using instead these predicted hourly temperatures. The results are virtually identical to those obtained under the true temperature data, suggesting that our forecasting models are not favorably affected by the use of true weather information. This result is hardly surprising given the high precision of modern temperature forecasting.

3.5.2 Comparison of probabilistic forecasts

Comparison of distribution behavior

In addition to the point forecasts, probabilistic forecasts also play a vital role in decision making under uncertainty. For Gaussian Process models, the probabilistic forecasts depend entirely on the point and variance estimates. The results reported in Table 3.2 show that compared with the Direct strategy, the DirRec strategy produces superior point estimates. At the same time, it tends to underestimate the estimation variation as it does not take into account the prediction error of previously forecasted results used as input in subsequent forecasting. Figure 3.2 shows the confidence bands, centered at DirRec point estimates, from the DirRec (orange) and Direct strategy (blue). Clearly the former underestimates the forecast variation. We therefore use the proposed hybrid strategy to construct probabilistic forecast for our Gaussian Process models. For a given

point of time, denote the predicted mean and variance from the DirRec strategy by $\hat{\mu}_t$ and $\hat{\sigma}_t^2$, and those from the Direct strategy by $\tilde{\mu}_t$ and $\tilde{\sigma}_t^2$. Let $f(\cdot; \mu, \sigma)$ be the Gaussian density function with mean μ and variance σ^2. The resultant predictive density from the hybrid strategy is then given by $f(\cdot; \hat{\mu}_t, \tilde{\sigma}_t^2)$.

One way to assess the quality of probabilistic forecasts is to use scoring rules that assign a numerical score to the predictive distribution given the event or value that materializes; see e.g. Gneiting and Raftery [2007] for an overview of scoring rules. One most commonly used scoring rule is the logarithmic score proposed by Good [1952], which is defined as

$$\text{Log}(F, y) = -\log(f(y)) \tag{3.12}$$

for a probabilistic forecast F of a random variable y. The logarithmic score has many attractive features and is particularly easy to implement. The lower the scores, the better the predicted distributions. We calculate the average logarithmic score of predicted distributions for our various models and report their results in Table 3.3. The hybrid forecasting strategy clearly outperforms the other two strategies for both the GP and linear regression models.

Table 3.3: Logarithmic scores for probabilistic forecasts.

Method	Log Scores
GP Model (DirRec)	7.16
GP Model (Direct)	7.47
GP Model (hybrid)	6.52
Regression Model (DirRec)	10.87
Regression Model (Direct)	17.15
Regression Model (hybrid)	7.25

The log scoring rule lends itself for model comparison. Note that we can re-write the log score

from the hybrid forecast as follows:

$$
\begin{aligned}
&-\sum_{t=1}^{120}\log(y_t|\hat{\mu}_t,\tilde{\sigma}_t^2)\\
=&-\sum_{t=1}^{120}\log(y_t|\hat{\mu}_t,\hat{\sigma}_t^2)-\sum_{t=1}^{120}\log\left(\frac{y_t|\hat{\mu}_t,\tilde{\sigma}_t^2}{y_t|\hat{\mu}_t,\hat{\sigma}_t^2}\right)=7.16-0.64=6.52,
\end{aligned}
$$

wherein 7.16 and 6.52 are the scores of the DirRec and hybrid strategy respectively. This calculation suggests that starting with the DirRec forecast, if we replace the predicted variances with their counterparts from the Direct strategy, the log score is improved by 0.64. Analogously, if we start with the Direct forecast and replace the predicted means with their counterparts from the DirRec strategy, the log score improves from 7.47 to 6.52 by 0.95.

We also note that the performance gap in terms of the log scores between the DirRec and Direct strategies is not as large as those based on their point estimates. Apparently the log score reflects the overall probabilistic forecasting performance of Gaussian Process models, which depend on both the point and variance estimates. As discussed above, the Direct strategy provides more reliable variance estimates. When it comes to the overall performance, its better variance estimation compensates its weakness in point estimates and reduces the performance gap.

Comparison of tail behavior

In addition to the overall forecasting performance, we shall pay particular attention to extreme events as they tend to have substantial impacts on the electricity market. For instance, an abrupt increase in electricity demand can be rather disruptive to the grid and may even bankrupt some retail providers. We therefore take a close look at the performance of our models under extreme situations.

It transpires that the log scoring rule can be tailored to evaluate predictions of tail events as well. To assess the tail behavior of forecast distributions, choosing extreme observations and then proceeding with the usual evaluation methods seems to be a reasonable choice. However, evaluation based on a small number of extreme observations may discredit even the most skillful forecast

available: if we evaluate forecast conditional on observed extreme outcomes, then always predicting the extremes becomes a winning strategy. To overcome this difficulty, Diks et al. [2011] and Lerch et al. [2017] proposed modified scoring rules that place particular emphasis on specific regions of the underlying distributions. We adopt the conditional likelihood (CL) score and censored likelihood (CSL) score proposed in these studies for our tail performance evaluation. In particular, the conditional likelihood and censored likelihood are defined as follows:

$$\text{CL}(F, y) = -w(y) \log \left(\frac{f(y)}{\int w(z) f(z) dz} \right) \tag{3.13}$$

$$\text{CSL}(F, y) = -w(y) \log f(y) - (1 - w(y)) \log \left(1 - \int w(z) f(z) dz \right) \tag{3.14}$$

where F is a predictive distribution and f its density for a random variable y and w is a non-negative weight function that specifies the region of interest. The main difference between these two scoring rules is that the CL scoring rule does not take into account the accuracy of the density forecast for the total probability of the region of interest but the CSL score does. These two scoring rules are both proper and can be tailored to specific regions. Similar to their unconditional counterpart, the lower the scores, the better the results.

In this investigation we focus on the upper tail of the distribution as unusually high electricity demand can be rather disruptive to the grid. We consider $w(z) = \mathbb{1}(z \geqslant r_\alpha)$, where r_α is the α-th percentile of F. We can vary α to focus on different regions of the underlying distribution. To examine the performance of our models under unexpected demand surge, we calculate the conditional and censored scores of the probabilistic forecasts at two high percentile levels with α being 80% and 90%. The results are reported in Table 3.4. The overall pattern is similar to that of log scores reported in Table 3.3. Under both percentile levels and according to both scoring rules, the GP model outperforms the regression model, and the hybrid strategy outperforms the other two forecasting strategies. These results suggest that the GP model and the hybrid strategy are well suited for predicting outrages of the electricity market.

Table 3.4: Conditional likelihood and censored likelihood scores for probabilistic forecasts

Method	Percentile (α)	CL	CSL
GP Model (DirRec)	80%	1.39	1.45
GP Model (Direct)	80%	1.56	1.75
GP Model (hybrid)	80%	1.26	1.32
Regression Model (DirRec)	80%	2.00	2.13
Regression Model (Direct)	80%	2.92	3.59
Regression Model (hybrid)	80%	1.35	1.43
GP Model (DirRec)	90%	0.82	0.89
GP Model (Direct)	90%	0.95	1.12
GP Model (hybrid)	90%	0.76	0.81
Regression Model (DirRec)	90%	1.29	1.39
Regression Model (Direct)	90%	1.68	2.22
Regression Model (hybrid)	90%	0.83	0.89

3.6 Economic applications

So far we have focused on the statistical performance of the proposed GP models for electricity demand forecasting. Ultimately our goal is to use these models to aid economic decision making and risk management. In this section, we illustrate the utility of the proposed models with some real world economic applications. We shall henceforth focus on the hybrid forecasting strategy, which is shown to outperform other strategies in the previous section.

3.6.1 Cost comparison under point forecasts

As we described above, a retail electric provider buys electricity from the generation companies and sells it to the consumers. In the ERCOT electricity market, the cost of a retail electric provider consists of two parts: day ahead cost and real time cost, which correspond to the day ahead market and real time market respectively. The day ahead cost is fixed once a company submits its hourly schedule in the day ahead market, while in the real time market, the retail provider needs to make instantaneous adjustments according to the actual demand. We focus on the more volatile real time cost in this investigation. Suppose first that firms use the point estimates from the GP forecasting models as their predicted demands. Denote by $Y_{forecast}$ the predicted demand from a statistical model and Y_{actual} the actual demand. Let P_{real} be the real time market price. For a given period i,

the real time cost for a firm can be represented as follows:

$$C_i = \begin{cases} (1+s_1) \times (Y_{actual,i} - Y_{forecast,i}) \times P_{real,i} & \text{if } Y_{forecast,i} \leq Y_{actual,i} \\ (1-s_2) \times (Y_{forecast,i} - Y_{actual,i}) \times P_{real,i} & \text{if } Y_{forecast,i} < Y_{actual,i} \end{cases} \tag{3.15}$$

Note here the cost function is asymmetric with positive 'friction' parameters s_1 and s_2 that reflect the stylized fact of extra cost associated with real time adjustments in the electricity market. When the actual demand exceeds the forecasted level, firms have to pay a premium beyond the going rate to procure extra electricity at a short notice. On the other hand when facing a lower than expected demand, firms can only unload the surplus electricity at a price lower than the going rate.

We now compare the economic performance of firms that use point forecasts from the GP and the conventional regression models. We use the average hourly real time cost as the evaluation criterion. For simplicity, we set $s_1 = s_2 = s$ in this application and calculate the real time cost for the 120 days used in our forecasting. The average results under different level of s are reported in the first and third columns of Table 3.5. It is seen that across multiple levels of s, the average real market cost based on the GP forecasts is 30% lower than that based on the regression method, implying the economic benefit of the proposed approach.

3.6.2 Cost optimization using probabilistic forecasts

Although point forecasts have been customarily used in decision making under uncertainty, we recognize that decision makers may further improve profitability by utilizing information contained in the probabilistic forecasts. In our second application, rather than setting the hourly supply schedule according to the point forecasts, we adopt a bidding strategy that aims to minimize the expected real time cost under the forecasted demand distributions.

Denote by x_i the quantity a firm submits as its bid for the i-th period and y_i the actual demand.

Its real time cost is given by

$$\hat{C}(y_i; x_i, \hat{P}_{real,i}) = \begin{cases} (1+s_1) \times (y_i - x_i) \times \hat{P}_{real,i} & \text{if } x_i \leq y_i \\ (1-s_2) \times (x_i - y_i) \times \hat{P}_{real,i} & \text{if } x_i < y_i \end{cases} \tag{3.16}$$

where $\hat{P}_{real,i}$ is the expected real time price. Denote by F_i the forecasted demand distribution for the i-th period. The optimal bid that minimizes the expected cost is then given by

$$\hat{x}_i = \arg\min_{x_i} \int \hat{C}(y_i; x_i, \hat{P}_{real,i}) dF_i(y_i) \tag{3.17}$$

The first order condition of this cost minimization is as follows

$$\int_{-\infty}^{\hat{x}_i} ((1-s_2)\hat{P}_{real,i}) dF_i(y_i) + \int_{\hat{x}_i}^{\infty} (-(1+s_1)\hat{P}_{real,i}) dF_i(y_i) = 0$$

yielding

$$F_i(\hat{x}_i) = \frac{1+s_1}{2+s_1-s_2}.$$

Thus this optimization problem has a simple analytical solution: the $(1+s_1)/(2+s_1-s_2)$ percentile of the forecasted distribution F_i. For simplicity, consider the case $s_1 = s_2 = s < 1$ such that the optimal bid is the $(1+s)/2$ percentile. Apparently in the presence of an asymmetric loss function wherein a heavier loss occurs with underbidding (when the actual demand exceeds the bid), it is optimal for firms to bid above the expected median outcome. The extent of overbidding should increase with s, which reflects the degree of asymmetry in the loss function. The optimal bid coincides with the point forecast only when $s = 0$, i.e. when the cost function is symmetric.

We undertake the optimization strategy based on the GP and the linear regression forecasts. For the latter, we assume that the error terms also follow a Gaussian distribution with mean zero and variance given by the estimated variance of the residuals. Although the expected real time cost depends on the expected real time price, the optimal solution given above does not depend on the real time price. For simplicity we use the observed real time prices in this experiment.

The resultant average costs from the optimal bidding strategy, under the same levels of s as in the previous example, are shown in the second and fourth columns of Table 3.5. Again, the GP model outperforms the linear regression model. As expected, the costs under the optimal bidding strategy are lower than when point forecasts are used as the bids, and the expected cost saving increases with the degree of asymmetry of the cost function.

Table 3.5: Expected real time market costs (unit: $/h)

s	GP	GP-Opt	Regression	Regression-Opt
0.1	3009.05	2969.66	4535.89	4494.91
0.2	3027.85	2949.91	4523.46	4376.01
0.3	3046.65	2894.60	4511.03	4192.70

3.7 Conclusions

We have proposed a novel hybrid Gaussian Process forecasting model that combines the strength of two different forecast strategies. The proposed GP model is shown to provide superior multi-step ahead point and probabilistic forecasts of hourly electricity demand. We expect the proposed hybrid forecasting strategy to find useful applications under a variety of different situations.

The proposed method has been applied to forecast the day ahead electricity demand in the southern region of Texas. We show that it can be used to reduce the expected cost of electricity supply relative to the conventional linear regression approach, and further economic benefit is obtained when the probabilistic forecasts are used to derive an optimal bidding strategy for the electricity suppliers. We conclude by noting that the above example is a mere illustration of how probabilistic forecasts can be fruitfully employed in a decision theoretic framework to optimize decision making. Generally speaking, the decision theory concerns decision making under uncertainty. Central to this framework is a loss function which specifies the loss of an action under a certain state of the world. Given uncertainty about the state, the decision maker chooses to minimize the expected loss, or risk. Interested readers are referred to Berger [1985] for a treatment

of decision theory. More complicated cost functions than the one considered in this study can be similarly entertained with little extra cost, albeit perhaps without simple analytical solutions. At the same time, we stress that reliable probabilistic forecasts that adequately reflect the underlying uncertainty is indispensable to the success of a decision theoretic procedure. Our investigation suggests that the proposed hybrid Gaussian Process forecasting approach can be a valuable tool for this purpose.

4. NON-STATIONARY MODELING OF CROP YIELD DISTRIBUTIONS WITH APPLICATIONS TO CROP INSURANCE

4.1 Introduction

The federal crop insurance program has been an important part of U.S. agricultural policy to stabilize farmers' income and protect against unpredictable risks for several decades. Since the 1990 Farm Bill, the crop insurance program has grown substantially. It covers more than 100 crops with a variety of yield-based, revenue-based and area-based policies.

An actuarially sound premium is critical to the effectiveness and robustness of crop yield insurances. Since the calculation of this parameter requires the knowledge of the future distribution of yields, one needs a reliable predictive yield distribution. A two stage approach has been customarily adopted for this purpose: in the first stage, a regression model is used to estimate the conditional mean of yield distribution in order to remove the influence of technological advancements and other factors; and in the second stage, de-meaned yields are used to estimate the yield distributions using some parametric or nonparametric methods. Despite its popularity, this two stage approach suffers from two potential limitations. First, if the conditional mean is not adequately modeled, the subsequent yield distribution estimation is compromised as it is based on the residuals from the first stage regression. Second, with a few exceptions a stationary yield distribution is often assumed in the second stage distribution estimation. Apparently if yield distributions evolve over time, this rigidity is overly restrictive.

In this study, we propose a new estimation approach for crop yields based on the method of Gaussian process (GP) regression. The GP regression is a powerful yet disciplined nonparametric method that has seen wide applications in statistical analysis and machine learning. This modeling approach is probabilistic in nature and yields not only point estimates but entire predictive distributions. This is particularly appealing to one of the primary focuses on the corp yield estimation, which is to obtain reliable predictive yield distribution. In particular, it offers two advantages. First

unlike the two stage estimation, it models the conditional mean and the entire yield distribution simultaneously. Second, the resultant yield distribution is free to vary over time and thus immune from the restriction of stationary distributions.[1]

In practice, crop yield estimations are often plagued by small sizes as typical studies on annual productions rely on yield histories no longer than 40-50 years. Furthermore, this problem is often exacerbated by the volatile yearly yield fluctuations. Fortunately when the estimations involve a large number of locations, one can resort to information pooling to alleviate this difficulty since yield distributions among geographically proximate locations tend to be similar. We therefore further construct for individual locations a predictive distribution that is based on mixture of all distributions in the analysis, wherein the weights of mixture is determined by some measure of similarity between distributions. These weighted estimates tend to be more reliable than individual estimates. In addition, the resultant predictive distributions constitute of mixture of Gaussian distributions are more flexible than the Gaussian predictions based on individual locations.

We evaluate the the proposed methods using simulations on the estimation of crop yield insurance polices based on historical corn yield data from Iowa. We consider a number of stationary and non-stationary data generating processes and different sample sizes. Our results suggest that the GP-based estimates compare favorably with conventional two stage estimations under all circumstances and excel when the underlying distributions are nonstationary. Furthermore, the weighted GP estimates considerably improve on those individual estimates.

Lastly we apply our estimators to an out-of-sample experiment of insurance policy selection, assuming the role of private insurance companies. In this experiment, insurance companies will choose policies they deem profitable and cede those deemed unprofitable. We are cognizant about the difficulty of decision making under uncertainty and that insurance companies might weigh the loss from taking an unprofitable policy differently than the forfeited benefit from ceding a profitable policy. Therefore we adopt a flexible decision-theoretic framework, wherein the opti-

[1]Even within the statistics and econometrics literature, the concept of 'stationarity' has multiple definitions. In this study, we use stationary distributions to refer to distributions that vary over time and non-stationary distributions for those vary over time.

mal action depends crucially on the decision makers' objective function. Previous studies tend to base their policy selection on the comparison between posted premiums by the government and insurance companies' own premium predictions. In contrast under the decision-theoretic framework, the recommendation is derived by minimizing the expected loss (or risk) with respect to a predictive distribution. Our results suggest that the proposed methods can be used to effectively identify profitable policies under both symmetric and asymmetric loss functions of the insurance companies.

The rest of this article is organized as follows. Section 4.2 gives a brief review of crop yield literatures. Section 4.3 introduces the Gaussian process models for crop yields and Section 4.4 presents a model averaging procedure. Simulations and an empirical illustration are presented in Sections 4.5 and 4.6. The last section concludes.

4.2 Literature

A two-stage estimation strategy is commonly used in studies of crop yield distributions. The first stage models the trend of yield distributions due to technological advancements and other reasons. A variety methods have been employed for this purpose, including a polynomial trend (Ramírez [1997] and Just and Weninger [1999]), the ARIMA process (Goodwin and Ker [1998]) ,the spline estimator (Harri et al. [2011] and Annan et al. [2013]), and the normal mixtures (Tolhurst and Ker [2014]). Some studies also examine possible heteroskedascitiy in the errors. For example, Just and Weninger [1999] and Harri et al. [2011] explored several forms of heteroskedasticity adjustment and its effects on crop insurance rate calculation.

In the second stage, the detrended residuals are used to model the yield distributions. Both parametric and nonparametric (including semiparametric) methods have been considered. Parametric methods assume a certain functional form of crop yield distributions. Commonly used specifications include the Normal (Botts and Boles [1958]), log-normal (Jung and Ramezani [1999]), Gamma (Gallagher [1987]) and their generalizations. The choice of parametric distributions is often based on the simplicity of estimation and inference or other practical considerations. In contrast, nonparametric methods provide a more flexible and data-driven way to model crop yield

distributions. Common nonparametric methods include inverse sin transformation (Moss and Shonkwiler [1993]), kernel density estimation (Goodwin and Ker [1998] and Ker and Goodwin [2000])) and normal mixtures (Goodwin et al. [2000]; Woodard and Sherrick [2011]; Tolhurst and Ker [2014] and Ker et al. [2015]). See also Ker and Coble [2003]; Stohs and LaFrance [2004]; Wu and Zhang [2012] and Tack et al. [2014] for application of alternative nonparametric methods to crop yield distributions.

One challenge in the modeling of crop yield distributions is that most of research are based on short time series data. For instance, systematic yield records at the county level in the U.S. started in the 1950s. Thus estimation of county level yield distributions typically rely on no more than 60 years of data. One possible way to mitigate this restriction is information pooling. Crop yields are heavily influenced by climate, weather and geographic factors and agricultural practice, all of which tend to be spatially correlated. As a result, the yield distribution of a certain region usually bears a high similarity to those of proximate regions. Researchers have devised information pooling estimation methods that take advantage of the similarity in crop yield distributions among geographically close regions. For example, Moss and Shonkwiler [1993] pooled information from neighboring counties to improve estimation efficacy. Ozaki and Silva [2009] incorporated temporal and spatial autocorrelation in hierarchical Bayesian models. Annan et al. [2013] employed formal distributional tests to determine whether or not to pool information from multiple counties. Ker et al. [2015] proposed a Bayesian model averaging approach to pool yield distributions from many regions. Zhang [2017] developed a density ratio estimator which features a common based line density for all regions and models individual distributions for each region as a deviation from the base line.

Despite its popularity, the two-stage estimation approach suffers a number of drawbacks. Firstly, the validity of this approach hinges critically on the premise that the underlying yield distributions are from the location-scale family. If this condition does not hold, the entire estimation is mis-specified. Secondly, within the location-scale family, even under the ideal condition that the conditional mean function is correctly specified and heteroskedasticity is properly accounted for,

the estimated residuals are subject to estimation variation due to the first stage estimation. In the presence of mis-specified conditional mean and/or variance function, the residuals do not consistently estimate the error term. The subsequent estimation of yield distributions, based on the first stage regression residuals, suffers from this inconsistency.

Lastly, an implicit assumption of many crop yield studies is that the residuals ε_t's are independently and identically distributed. Only a handful of studies have considered non-stationary yield distributions. In addition to the evolvement of the conditional mean and/or variance of the yield distribution, non-stationary models of yield distributions permit the evolvement of the entire distributions overtime. For instance, Zhu et al. [2011] proposed a time-varying yield distribution method to capture non-stationary nature of the yield distributions. They employed the Beta distribution to model observed crop yields and parametrized the coefficients of the Beta distribution as polynomials of time to accommodate time-varying yield distributions. Alternatively, Tolhurst and Ker [2014] employed mixture of normal distributions to model yield distributions, embedding time trend into the location parameters of the normal mixture components. These studies showed that their time-varying distributional model outperformed conventional stationary distributional models in many aspects.

4.3 Gaussian Process estimation

4.3.1 Preliminaries

We consider a new estimation approach that eschews the two stage estimation process and the i.i.d. assumption on the de-trended data in many conventional studies. Instead this approach directly estimates a time-evolving yield distribution using the approach of Gaussian process regression. Gaussian process is a powerful nonparametric machine learning tool for regression and classification. A Gaussian process is an infinite dimensional stochastic process that follows the Gaussian distribution. An important property that makes Gaussian process particularly useful in statistical analysis and machine learning is the so-called marginalization property: any subset of an (infinite-dimensional) Gaussian process retains its Gaussianity: it reduces to the familiar multi-

variate Gaussian distribution.

Gaussian processes provide a principled, practical, probabilistic approach for statistical learning; see the book by Rasmussen and Williams [2006] for an illuminating overview of this approach. The probabilistic nature of this approach distinguishes it from other machine learning techniques such as neural network. This approach produces not only point estimates but predictive densities that are essential towards the fulfillment of several goals of this study, including likelihood based model averaging, probabilistic yield forecasting, calculation of crop insurance premium, and decision-theoretic procedures for optimal decision making.

A GP regression model is formulated as follows. Consider a training set $\mathcal{D} = \{(X_i, Y_i), i = 1, \ldots, N\}$ of N pairs of input x_i and output y_i from an underlying relationship f. Here f is typically assumed to be a zero-mean Gaussian process with a covariance (kernel) function $k(\cdot, \cdot)$, and the observations y_i are given by

$$y_i = f(x_i) + \varepsilon_i, \qquad i = 1, \ldots, N, \tag{4.1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are white noises independent of $f(x_i)$.

Let $x = [x_1, x_2, \ldots, x_N]'$ and $y = [y_1, y_2, \ldots, y_N]'$. Denote the predictive distribution of outcome at a test location x_* by $f_* = f(x_*)$. The joint distribution of (f_*, y) is then given by

$$\begin{bmatrix} f_* \\ y \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} k_{**} & k'_{x_*} \\ k_{x_*} & \sigma^2 I + K_{xx} \end{bmatrix}\right), \tag{4.2}$$

where $k_{**} = k(x_*, x_*)$, $k_{x_*} = (k(x_1, x_*), \ldots, k(x_N, x_*))'$, K_{xx} is an $N \times N$ matrix with the $(i, j)^{th}$ entity $k(x_i, x_j)$ and I is a N-dimensional identity matrix. The predictive distribution of f_* given y is

$$f_*|y \sim \mathcal{N}(k'_{x_*}(\sigma^2 I + K_{xx})^{-1} y, k_{**} - k'_{x_*}(\sigma^2 I + K_{xx})^{-1} k_{x_*}). \tag{4.3}$$

The predictive mean $k'_{x_*}(\sigma^2 I + K_{xx})^{-1} y$ gives the point forecast of $f(x)$ at location x_*, whose uncertainty is measured by the predictive variance $k_{**} - k'_{x_*}(\sigma^2 I + K_{xx})^{-1} k_{x_*}$. Note that the point

forecast at location x_* depends on y and the various variance and covariance components, and is usually non-zero. The covariates influence the predictive outcome through the covariance. In this sense, the covariance is the determining factor of a GP predictor as it encodes our assumptions about the underlying relationship we wish to learn.

Given the observed output y, the dominant ingredient of a Gaussian process model is the covariance matrix, as it encodes our assumptions about the function which we wish to learn from the data. For instance, a popular choice of the covariance function is the Squared Exponential kernel:

$$k_{SE}(x_i, x_j) = \sigma_f^2 \exp\left[-\frac{(x_i - x_j)^2}{2l^2}\right], \tag{4.4}$$

where σ_f^2 reflects the maximum allowed covariance that usually increases with the variation of y. The so-called length scale l determines the relevancy of input x to the outcome y. To see this, note that the covariance between x_i and x_j vanishes under a sufficiently large length scale l, effectively removing it from the inference. A covariance function with this feature is called an Automatic Relevance Determination (ARD) covariance function. There exist a large selection of kernel functions that are suitable to model various functional relationship; see Chapter 4 of Rasmussen and Williams [2006] for details.

Roughly speaking the covariance function $k(x_i, x_j)$ captures the *similarity* between two inputs x_i and x_j and the GP regression can be viewed as a nonparametric smoother with an infinite number of basis functions, whose coefficients are modeled in a Bayesian fashion. The performance of GP models hinges on the configuration of the covariance function and its tuning parameters, which are referred to as the hyperparameters in the machine learning literature. Given model (4.1) and the SE kernel (4.4), the covariance function for the training set takes the form

$$k(x_i, x_j) = \sigma_f^2 \exp\left[-\frac{(x_i - x_j)^2}{2l^2}\right] + \sigma^2 \delta(x_i, x_j), \tag{4.5}$$

where $\delta(x_i, x_j)$ is the Kronecker delta function. The hyperparameter of this model then consists of $\theta = (\sigma_f^2, \sigma^2, l)$. One possibility of hyperparameter selection is via maximizing the marginal likeli-

hood $p(\theta|x, y)$ of the GP model given the observed data; this is also known as the type II Maximum Likelihood (ML-II) estimation, an empirical Bayesian approach that employs data-dependent priors. The log marginal likelihood is given by

$$\log p(y|x, \theta) = -\frac{1}{2} y' K_{xx}^{-1} y - \frac{1}{2} \log |K_{xx}| - \frac{N}{2} \log(2\pi), \tag{4.6}$$

where K_{xx} is an $N \times N$ matrix with the $(i, j)^{th}$ entity $k(x_i, x_j)$ given in (4.5). The first part of the likelihood function $-\frac{1}{2} y' K_{xx}^{-1} y$ reflects the goodness of fit. The second part $-\frac{1}{2} \log |K_{xx}|$ can be viewed as a complexity penalty that depends on the covariance function and the inputs. The third part is a normalization constant. What distinguishes the ML-II from the classical MLE is the presence of the complexity penalty, which effectively prevents overfitting. This estimator is similar in spirit to the penalized likelihood estimator whose objective function has an explicit (and sometimes ad hoc) complexity penalty. The trade-off between the log likelihood and the penalty is governed by a tuning parameter that determines the strength of the penalty. On the other hand, the ML-II is advantageous as its *complexity penalty* occurs inherently as part of the marginal likelihood and entails no additional tuning parameters.

4.3.2 GP model for crop yields

In this study we focus on the corn production in the state of Iowa, which is the largest corn producing state in United States. Our data consists of annual county corn yields per acre of 99 Iowa counties from year 1960 through 2010, obtained from the National Agricultural Statistics Service. Figure 4.1 shows the average annual yield for 99 counties of Iowa from 1960 to 2010, showing a clear increasing trend during the sample period.

We shall compare and contrast this method with conventional two-stage estimator throughout the rest of the text. To ease reference, we first present the two stage estimator used in this study. The two-stage approach first removes the time trend of crop yields and then model the distribution of the de-trended data under the assumption of a stationary distribution.

Following the benchmark model by the Risk Management Agency (RMA) of the USDA, we

use a two knots linear spline model[2]. Denote the yield in county i at time t as y_{it}. The first stage model is given by, for $i = 1, \ldots, N$ and $t = 1, \ldots, T$

$$y_{it} = \alpha_i + \beta_i t + \gamma_{1i}(t - k_1)_+ + \gamma_{2i}(t - k_2)_+ + \epsilon_{it}, \tag{4.7}$$

where k_1 and k_2 are the spline knots, $(x)_+ = \max(0, x)$, $(\alpha_i, \beta_i, \gamma_{1i}, \gamma_{2i})$ are unknown parameters to be estimated, and ϵ_{it} is random error with mean zero and finite variance. Denote the estimated residuals by $\hat{\varepsilon}_{it}$, which can be viewed as de-trended crop yields. In the second stage, we estimate the yield distributions based on these de-trended data (with proper adjustment for heteroskedasticity if needed) using some parametric or nonparametric method.

We next present a Gaussian Process model for crop yield. Owing to the flexibility of the GP models, a mean function, unless particularly desired, is usually not required. This convention is employed in this study. Instead our understandings and assumptions about the underlying relationship are encoded into the modeling process via careful configuration of the covariance. To capture the smooth rising trend we use a squared exponential (SE) covariance function given by

$$k_{SE}(t_i, t_j) = a_1^2 \exp(-\frac{(t_i - t_j)^2}{2a_2^2}). \tag{4.8}$$

The SE kernel is particularly suitable for smooth relationship. There are alternative kernels that can be used to model more rough relationships; for instance, the Matérn family of kernels. In our estimation, we use the commonly used Matérn $3/2$ kernel for this purpose. This kernel is given by

$$k_{3/2}(t_i, t_j) = a_3^2(1 + \frac{\sqrt{3}|t_i - t_j|}{a_4}) \exp(-\frac{\sqrt{3}|t_i - t_j|}{a_4}). \tag{4.9}$$

Lastly we capture the influence of idiosyncratic errors with the white noise kernel

$$k_{noise} = a_5^2 \delta(t_i, t_j), \tag{4.10}$$

[2]We configure the knots such that they divide the sample period equally into three sub-periods.

where δ is the Kronecker delta function.

The composite covariance function for our crop yield model is given by

$$k_{GP} = k_{SE} + k_{3/2} + k_{noise}. \tag{4.11}$$

We use squared exponential covariance function to capture the long run rising trend and Matérn $3/2$ covariance function to capture the the more volatile shorter run fluctuations. For each county-year, our Gaussian Process estimator can be written as

$$y_{it} = f^i_{GP}(t) \tag{4.12}$$

for $t = 1, 2, \cdots, T$ and $i = 1, 2, \cdots, N$. In particular, for each county, we fit the above GP model using its historical corn yields. Denote the hyperparameters by $\theta = (a_1, \ldots, a_5)$. These parameters are selected using the ML-II procedure in the previous section.

We now present an illustration of the proposed GP model for crop yields. The left panel of Figure 4.1 shows the historical average corn yields for Iowa from year 1960 to 2010. The data show an apparent increasing trend and also substantial year to year fluctuations. We plot in the same graph the predicted yields for the sample periods and forecasts for years 2011-2030. The fitted curve suggests a smoothly increasing trend of yields. The shaded area indicates plus and minus twice of the estimated standard deviation. As expected for the forecast period, the variation of the predictions increases steadily with the length of forecast horizon. The right panel shows the contribution of the two kernels used to model the time trend. The solid line is the prediction based on the squared exponential kernel, which captures the long term trend. The dash line is the contribution based on the Matérn kernel that reflects the shorter term fluctuations. The scale of the second component (the right hand side scale) is substantially smaller than that from the long run term. Also note that unlike the long run trend, the shorter term trend gradually dies off towards zero as the forecast horizon increases. This is because that in this example, the covariance based on the Matérn kernel decays rapidly with the distance in time.

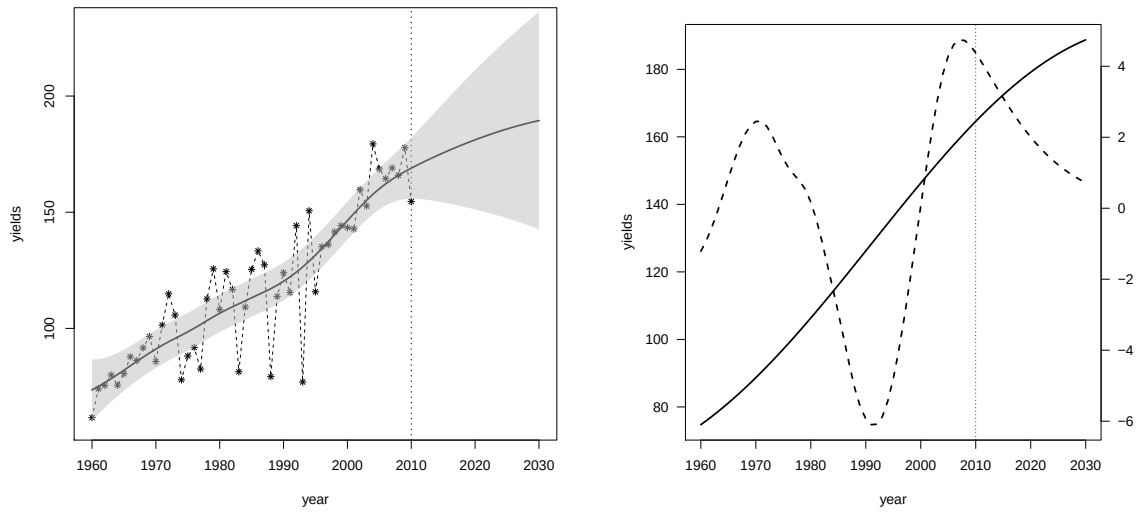


Figure 4.1: GP model for crop yields
Left: Historical yields (1960-2010, asterisks) and predicted yields (1960-2030, solid line); the shaded area indicates plus and minus twice of the standard deviation. Right: Long term trend predicted by the squared exponential kernel (solid, left hand scale) and shorter term trend predicted by the Matérn kernel (dash, right hand scale). In both plots, the vertical line indicates the end of the observation period and start of forecasting.

4.4 Performance weighted model averaging

Expansive U.S. crop production and yield data have been collected and made available by various sources. The most commonly used type of data are county level average yields. These are also the most relevant to the Group Area Insurance program, whose indemnity is determined by county level output. A critical restriction posed by most crop yield data is the relative short observation period. Typically they are available for at best 50-60 years. Agricultural economists have to rely on these rather short time series data to estimate crop yield distributions; the possibility that these distributions themselves are evolving over time makes this task even more difficult.

To improve the reliability of county level yield distribution estimation, we further develop a model averaging method to combine estimated Gaussian process estimators from individual counties. Model averaging exploits information from all candidate models and incorporates model uncertainty into the estimation. Like statistical estimation, model selection is subject to stochastic

errors due to sample variation. In contrast, combining the strength of multiple models/estimators can often lead to better performance in practice. A second motivation of the model averaging approach for the current study is that it improves the flexibility of the predictive distributions considerably. The GP regression is a versatile and yet principled nonparametric smoother that allows time varying distributions. However for each given point of time, the predicted distribution is Gaussian. A mixture of Gaussian distributions effectively solves this problem.

There are two major model averaging mechanisms: Bayesian model averaging (BMA) and frequentist model averaging (FMA). Bayesian model averaging provides a coherent mechanism for accounting for model uncertainties. Reviews of Bayesian literature can be found in the works of Hoeting et al. [1999] and Raftery et al. [1997]. For frequentist model averaging strategies, the most widely used methods are weighting strategies based on the AIC values proposed by Akaike [1974] or weighting framework proposed by Hjort and Claeskens [2003].

In this study, we will use the relative performance weights to construct mixture of Gaussian process models. A similar strategy is pursued by Ker et al. [2015] in their study of nonparametric crop yield estimation. The relative performance is evaluated by the marginal likelihood of individual Gaussian Processes estimators, with larger marginal likelihood value indicating a better fit. The general idea is rather straightforward. If an estimated model for county j provides a good fit for crop yields from county i, that indicates a high degree of similarity in yield distributions of these two counties. Accordingly in the construction of a mixture model for county i, a relatively heavy weight is assigned to county j. Alternatively, we can assign weights based on the geographical distances between the counties (for instance, the nearest neighbor smoothing or spatial autocorrelation). In contrast, the performance based approach adopted in this study is adaptive and avoids explicit assumptions on spatial relationships (for instance, exponential decays of spatial auto-correlations).

As is mentioned about, the marginal likelihood function has a built-in complexity penalty (similar to the explicit penalty on the number of coefficients in the AIC or BIC). For the proposed study, denote the estimated model by $\hat{f}_i$ for county i, and its corresponding marginal likelihood by L_i. To

assess how similar the yield distribution of county i is to that of county j, we evaluate the marginal likelihood of $\hat{f}_j$ using observed yields from county i. Denote this 'out-of-sample' marginal likelihood by L_j^i. Let the difference in likelihood scores $\Delta L_j^i = L_i - L_j^i$. We then construct the model mixing weights according to

$$\omega_j^i = \frac{\exp\{\frac{1}{2}\Delta L_j^i\}}{\sum_{j=1}^{N} \exp\{\frac{1}{2}\Delta L_j^i\}} \tag{4.13}$$

The mixed Gaussian process model for county i is then given by

$$\tilde{f}_i = \sum_{j=1}^{N} \omega_j^i \hat{f}_j \tag{4.14}$$

Since ΔL_j^i is a relative likelihood score based on out-of-sample evaluation, overfitting is prevented. The weight assigned to county j in the estimation of county i's distribution depends on the predictive power of county j's distribution on county i's observed yields.

A key benefit of this model averaging approach in the proposed study is the enhanced distributional flexibility afforded by the mixing of Gaussian process models. It is well known that mixture of Gaussian densities can approximate a distribution arbitrarily well. This desirable property extends naturally to the mixture of Gaussian process models. The ultimate purpose of the entire enterprise of yield distribution estimation is to inform and aid decision making in agricultural production and insurance programs. The mixture of Gaussian process models provides a flexible yet disciplined statistical tool to model time-varying distributions that are essential to these tasks.

4.5 Simulations

We use Monte Carlo simulations to evaluate the effectiveness of proposed non-stationary Gaussian Process based estimators for crop yields. Rather than generating arbitrary random samples, we based our data generating processes on distributions estimated from historical data. In particular, we use annual corn yields data of 99 Iowa counties from year 1960 through 2010. The entire yield distribution is rarely of direct interest but used instead to calculate quantities of economic interest. Thus we focus our assessment on a parameter of utmost importance: the premium rate of crop

insurance programs. The premium rate, associated with a certain yield guarantee y_G, is defined as the expected loss divided by the total liability:

$$r = \frac{1}{y_G} \int_0^{y_G} (y_G - y) f_Y(y|I) dy, \tag{4.15}$$

where I signifies the information set used in the prediction and $f_Y(y|I)$ is the predictive yield distribution based on I.

We consider different sample sizes of estimation data with $T = 20, 30$ and 40 and use the average of last five years' observations (2006-2010) for evaluation. We use an incremental forecasting scheme. For example with $T = 20$, we use observations from 1985 to 2005 to estimate the predictive distribution for 2006, and observations from 1985 to 2006 to estimate the predictive distribution for 2007, and so forth. As for the underlying distributions, we consider three possibilities: (i) stationary yield distributions; (ii) a non-stationary yield distributions; (iii) a combination of (i) and (ii), which is also non-stationary but to a lesser degree.

4.5.1 One-step ahead forecast

In our first scenario, we assume the underlying crop yield distribution is stationary. We first use the two knots spline model given in (4.7) to detrend the yield data for each county. We then estimate the distributions of the residuals $\{\hat{\varepsilon}_{it}, t = 1, \cdots, T\}$ using a three components normal mixture model[3]. Taking the estimated normal mixture densities as true yield densities, we repeatedly draw random samples of size T from these distributions to use in our simulations. Denote $\{\dot{\epsilon}_{it}, t = 1, \cdots, T\}$ an iid sample for the estimated distribution for county i, the simulated data are given by

$$\dot{y}_{it} = \hat{y}_{it} + \dot{\epsilon}_{it}, t = 1, \ldots, T, \tag{4.16}$$

where $\hat{y}_{it}$ is the estimated mean. By construction, the simulated samples, if properly de-trended, are distributed according to a stationary distribution that does not vary over time.

[3]A three components normal mixture model is sufficiently flexible for most smooth distributions. The fixed number of components also ensures that the DGP is not Gaussian so that the GP model does not enjoy an unfair advantage in this simulation.

For each simulated yield series, we estimate the GP model as given in Section 4.2. For comparison, we also estimate the conventional two-stage model. In particular, we use model (4.7) to de-trend the data and then estimate the distribution of the residuals using the flexible kernel density estimator (KDE). For both the GP and KDE estimates, we then construct the performance weighted estimates using the approach given in Section 4.4. We denote the resultant mixture densities by W-GP and W-KDE respectively. We then proceed to estimate the premium based on these estimated densities. Without loss of generality, we set the location of all predicted distributions to the actual and focus on the influence of the overall shape of the distribution on premium estimation. We set the guarantee yield to 90% of the location parameter. Note that the RMA uses the empirical distribution of the de-trended data in the calculation of premiums. For completeness, we also include thus-obtained estimates, denoted by EMP, in our comparisons.

We calculate the 'true' premium based on the distributions that are used to generate the random samples and use the mean squared errors (MSE) of the estimated premium rates to assess the performance of various estimators. We report in the top panel of Table 4.1 the average MSE across all counties and years (2006-2010) for simulations with $T = 20, 30$ and 40. As expected, the results generally improve with sample size. We also note the following: (i) both the EMP- and KDE-based estimates are consistent and the performance gaps between them decrease as sample size increases; (ii) the GP-based estimates mostly outperform the other estimates; (iii) model averaging via the mixture of individual densities substantially improves the estimates for both the KDE and GP models.

Overall, the W-GP model provides the best overall performance. The better performance of the GP models relative to the EMP and KDE models is remarkable as the latter use the true functional form of the conditional mean, correctly assume the stationarity of the error distribution, and estimate this distribution consistently. In contrast, the GP models do not assume any particular form for the conditional mean and allow the underlying distribution to vary overtime. Nonetheless, they are shown to be sufficiently flexible and adaptive to approximate the underlying DGP and yet disciplined enough to provide well-behaved estimates especially under small sample size.

We next consider the case where the underlying crop yield distribution is non-stationary. First we assume the underlying true distribution follows a Gaussian Process. In particular, we fit the same data used in the first experiment using the GP model proposed in section 4.2 and use the fitted models as the DGP's for our simulations.[4] As a result, the underlying distributions vary over time and thus are non-stationary.

We estimate the premium rates using the same models as considered in the first experiment. The average MSE's are reported in the middle panel of Table 4.1. The overall results are similar to those under the stationary distributions. We note that without exception, the GP and W-GP estimates outperform their counterparts based on the EMP and KDE estimates. In addition, the relative performance of the GP estimates to the KDE estimates improves under non-stationary distributions. The average MSE ratio of GP to KDE estimates is 88% under stationary distributions, and it improves to 85% under non-stationary distributions. A larger improvement, from 92% to 65%, is observed when we compare the W-GP to the W-KDE estimates.

The GP models are advantageous in the second experiment as they conform to the underlying DGP's. To explore the sensitivity of our results to this unrealistic situation, we consider a third experiment wherein we construct the underlying distribution as a combination of stationary and non-stationary distributions. In particular, we take the estimated three-component normal mixture distributions from the first experiment and the estimated GP distributions from the second experiment and use a 50-50% mixture of these two distributions as the DGP. The random samples draw from these distributions are then added to the estimated time trend from the first experiment. Under this DGP, both the two step estimator and the GP estimator have their respective advantages and disadvantages. The EMP- and KDE-based two stage estimators use the true functional form for the conditional mean, but incorrectly assume a stationary distribution. On the other hand, the GP models assume a non-stationary Gaussian distribution, while the true distribution is a combination of a stationary Gaussian mixture and a non-stationary Gaussian process.

The same estimation procedures as in the first two experiments are used in the third experiment.

[4]We still use the Gaussian Processes with zero mean and k_{GP} kernel function to model the crop yield series, hyperparameters are determined by maximizing the marginal likelihood.

Table 4.1: MSE (multiplied by 10^4) of estimated premium rates for one-step ahead forecast

DGP		EMP	KDE	W-KDE	GP	W-GP
Stationary	T = 20	0.9793	0.8899	0.5377	0.7762	0.4869
	T = 30	0.7937	0.7183	0.5758	0.6291	0.4443
	T = 40	0.4992	0.4813	0.2820	0.4275	0.3040
Non-Stationary	T = 20	1.1097	0.7397	0.5853	0.6158	0.3786
	T = 30	0.7863	0.5255	0.5377	0.4469	0.3823
	T = 40	0.4770	0.3164	0.2723	0.2738	0.1639
Mixed	T = 20	1.0190	0.8126	0.5459	0.7170	0.3969
	T = 30	0.7757	0.6089	0.5528	0.5250	0.3305
	T = 40	0.4600	0.3835	0.2642	0.3411	0.2143

The estimated average MSE's are reported in the bottom panel of Table 4.1. The same overall pattern is observed and the GP-based estimates again provide the best overall performance. Not surprisingly, the relative performance of the GP-based estimates to the KDE-based estimates are better than that from the first experiment (wherein the KDE models are favored) and worse than that from the second experiment (wherein the GP models are favored). For instance, the average MSE ratio of the W-GP estimates relative to the W-KDE estimates is 88%, falling between 92% from the first experiment and 65% from the second experiment.

4.5.2 Multi-step ahead forecast

Forward looking planning that spans multiple years is common in agricultural production and risk management. To explore how the proposed methods fare under longer forecast horizon, we consider multi-step forecast as well. In particular, we conduct simulations on 3- and 5-year ahead forecast. The same incremental forecasting scheme is used. For example in the case of 3-year ahead forecast, we use observations from years 1985-2003 to forecast the premium for year 2006, and years 1985-2004 to forecast the premium for year 2007, and so forth.

The same estimation procedures as those in the one-step ahead estimations are again used here. To save space, we only present the experiments with $T = 30$ and under the stationary and non-stationary distributions (as in the first and second experiments in the one-step ahead forecast simulations). Results for other cases are qualitatively similar and available from the authors upon

request.

The estimation results, reported in Table 4.2, exhibit a similar overall pattern as those in the one-step ahead estimations. Again, the W-GP estimates dominate other estimates across all scenarios. Unlike those in the one-step ahead simulations, the W-KDE estimates are worse than the KDE estimates under non-stationary distributions. Note that the KDE can not consistently estimate a time-varying distribution. We conjecture that the adverse consequence of this inconsistency worsens as the forecast horizon increases, and is further aggravated when multiple densities are combined (as is under the W-KDE estimation).

Table 4.2: MSE (multiplied by 10^4) of estimated premium rates for multi-step ahead forecast

	DGP	EMP	KDE	W-KDE	GP	W-GP
3-step forecast	stationary	0.8367	0.7394	0.6299	0.7706	0.5693
	non-stationary	1.1030	0.6342	0.7501	0.5155	0.4130
5-step forecast	stationary	0.9195	0.7923	0.7248	1.1226	0.8864
	non-stationary	1.4403	0.8246	0.9514	0.5810	0.4301

In sum, our simulations on crop insurance premium rates under various DGP's for one-step and multi-step simulations show that the proposed GP estimator compares favorably with the KDE-based two-stage estimator when the underlying distributions are stationary and excels when the underlying distributions are non-stationary. In addition, the performance weighted GP estimator considerably improves the GP estimator based in individual counties and provides the best overall performance.

4.6 Application to insurance policy rating

To illustrate the economic utility of the proposed estimator, we apply it to the rating of crop insurance policies. The U.S. federal crop insurance program is managed by the Risk Management Agency (RMA) of the USDA. RMA sets the premium rates of various crop insurance policies. An important feature of federal crop insurance program is these policies are sold through private

insurance companies to farmers. The insurance companies are allowed to select the policies they deem profitable and cede those they deem unprofitable. To improve their profitability, the insurance companies often develop their own premium estimates and based on which they assess the profitability of insurance policies. If the RMA premium for a given policy is higher than their own estimates, there is a good chance that this policy is overpriced. Accordingly, this policy is likely to be selected by insurance companies. In contrast, if the RMA rate is lower than their estimates, this policy is considered under-priced and shunned by insurance companies.

Unlike previous work, we adopt a flexible decision-theoretic approach for policy selection. In practice, people often need to decide how to act. Decision theory provides a framework to optimize decision making; see for example Berger [1985]. Central to this paradigm is a loss function that specifies the loss incurred by a certain action. For decision making under uncertainty, it is logical that one seeks to minimize the *expected loss*, or *risk*. Naturally the optimal decision is the one that minimizes the expected loss. For instance, the expected value is the optimal solution under a quadratic loss function, while the median is the optimal solution under an absolute value loss function.

In this investigation, we consider a general asymmetric loss function of policy selection. We assume that the loss due to selecting an unprofitable policy is heavier than the foregone benefit due to ceding a profitable policy. For simplicity, we consider the following loss function

$$L = \begin{cases} a, & \text{if retaining the policy and } \pi > \pi_p \\ b, & \text{if ceding the policy and } \pi < \pi_p \\ 0, & \text{if retaining the policy and } \pi < \pi_p \\ 0, & \text{if cedeing the policy and } \pi > \pi_p \end{cases} \tag{4.17}$$

where $0 < b < a$ and π_p denotes the premium set by the RMA. Accordingly the risk (expected

loss) of policy selection is given by

$$R = \begin{cases} a * \Pr(\pi > \pi_p), & \text{if retaining the policy} \\ b * \Pr(\pi < \pi_p), & \text{if ceding the policy} \end{cases} \tag{4.18}$$

It follows that the optimal solution under this loss function is to retain the policy if

$$a * \Pr(\pi > \pi_p) < b * \Pr(\pi < \pi_p),$$

or equivalently if

$$\Pr(\pi < \pi_p) > \frac{a}{a+b}. \tag{4.19}$$

Note that if $a = b$ (i.e., under a symmetric loss function), this rule suggests retaining the policy if the probability of overpricing is higher than 50%. However when $a > b$, this policy is retained only if the probability of overprice is higher than $a/(a+b)$. Clearly under the assumed asymmetric loss function wherein a heavier loss occurs when one retains an unprofitable policy, it is sensible to be more conservative in retained policies and the optimal degree of 'conservativeness' is governed by the relative severity of the losses.

Following Ker et al. [2015] and Zhang [2017], in this exploration we assume the role of a private insurance company and use the proposed estimator to select profitable policies. To avoid overfitting, we use out-of-sample performance for evaluation. In particular, we estimate the 2006 premium rates using the GP model based on yields from 1986-2005,[5] and then calculate the underwriting gains and losses using the actual 2006 yields. We repeat this process for years 2007, ..., 2010, each based on 20 years of historical yields. For each county-year, we consider a policy with a coverage level at 90% of predicted yield value given by model (4.7). The RMA rate is calculated based on the empirical distribution of the de-trended data with proper heteroskedascitity adjustment (see Harri et al. [2011] for details). Denote the estimated premium based on the GP model

[5]Ker et al. [2015] suggested using no more than 20 years of historical yield losses due to the evolution of yield distributions over time.

by $\hat{\pi}_{it}$. Let $\hat{\sigma}^2$ be the sample variance of $\{\hat{\pi}_{it}\}, i = 1, \ldots, 99, t = 2006, \ldots, 2010$. The predict distribution for the premium for county i and year t is then given by $\hat{f}_{it}(\pi) \sim N(\hat{\pi}_{it}, \hat{\sigma}^2)$.

For simplicity, we set $a = 1 + s$ and $b = 1 - s$ and experiment with $s = 0$ (symmetric loss) and $s = 0.1$ (asymmetric loss). We use policy loss ratio to assess the effectiveness of policy selection. Denote a set of insurance policies by Ω, its loss ratio is calculated as

$$\text{LR}_\Omega = \frac{\sum_{i \in \Omega} \max(0, y_G - y_i)}{\sum_{i \in \Omega} \hat{\pi}_{i,p}}, \tag{4.20}$$

where for each policy i, y_G is the yield guarantee, y_i is the actual yield and $\hat{\pi}_{i,p}$ is the RMA rate. For each experiment, we calculate the policy retained rates and the loss ratio of the corresponding retained and ceded policies. We use the bootstrap method as outlined in Ker et al. [2015] to test the hypothesis that the loss ratio of the retained policies is lower than that of the ceded policies.

Table 4.3: Out-of-sample rating game results

	s	Retain Rate (%)	LR(retained)	LR(ceded)	p-value
GP	0	12.12	0.8322	1.8006	0.0567
W-GP	0	23.23	0.8577	1.9982	0.0155
GP	0.1	6.26	0.3708	1.7678	0.0282
W-GP	0.1	12.92	0.2393	1.9538	0.0002

The out-of-sample evaluation results based on GP and W-GP estimates are reported in Table 4.3. The top panel shows the results under a symmetric loss function. Based the GP estimates, about 12% of policies with a loss ratio of 0.83; based on the W-GP estimates, about 24% of policies are retained with a loss ratio of 0.85 and 1.99. Policy selection based on the W-GP estimates is shown to effectively double the number of retained policies while maintains essentially the same loss ratio as that under the GP estimates. The loss ratios for the ceded policies are 1.80 and 1.99 based on the GP and W-GP estimates, and statistically different from those of the retained policies (with a p-value of 0.0577 and 0.0155 respectively).

The bottom panel of Table 4.3 shows the results under an asymmetric loss function with $s = 0.1$. Apparently an insurance company with such a loss function is loss averse and therefore more conservative in policy selection. This conjecture is confirmed in our experiments. Based the GP and W-GP estimates, the proportions of retained policies are reduced to about 6% and 13% respectively. Thanks to the more cautious policy selection, their loss ratios of the retained polices improve to 0.37 and 0.24. Compared with the case under the symmetric loss function, the proportion of retained policies is reduced by roughly 50% under either strategy while their loss ratio is reduced by 55% and 72% respectively. Not surprisingly, the difference in the loss ratios between the retained and ceded polices are more pronounced under this more selective policy election.

Our experiments suggest that the proposed methods are policy selection based on the proposed GP models are effective in selecting profitable policies under both a symmetric and an asymmetric loss function. In addition, the W-GP approach, which is heavily favored according to our simulations, is shown to be able to select a larger proportion of policies without compromising the loss ratio. This is certainly desirable for the insurance companies as it implies a higher total profit.

4.7 Concluding remarks

In this study, we propose a non-stationary modeling approach based on the method of Gaussian process regression for crop yields. This approach departs from the conventional two-step estimation procedure and allows the yield distributions to vary over time. We further develop a performance weighted model averaging method to construct densities as mixture of Gaussian processes. This method not only facilitates information pooling but greatly improves the flexibility of the resultant predictive density of crop yields. Our simulation results on crop insurance premium estimation show that the proposed method is comparable and often preferred to the conventional two-step estimation procedures regardless of whether the underlying distributions are stationary. When the underlying distributions are non-stationary, our method consistently outperforms its competitors.

We demonstrate the utility of the proposed method with an application to crop insurance policies selection from the insurance companies' point of view. We are cognizant about the difficulty

of decision making under uncertainty and its contingency upon decision maker's objective. Therefore we adopt the decision theoretic framework that tailors the decision process according to the loss function of the decision maker and derives a feasible solution by minimizing the expected loss (or risk) with respect to a distribution regarding the source of uncertainty. Our results suggest that the proposed method provides an effective tool in identifying profitable insurance policies. We illustrate the usefulness of this framework using a simple stylized asymmetric loss function for the insurance companies. Note that more complex loss functions can be similarly accommodated with little extra cost, and it can be used by all stakeholders (e.g. the insurance companies, the RMA and the farmers) alike. We conclude by stressing that a key ingredient to the successful implementation of this approach is reliable predictive distributions of crop yields. In addition to being a flexible yet principled estimator, the proposed Gaussian process approach, thanks to its probabilistic nature, lends itself to this task.

5. CONCLUSION

In the first essay, I propose an Averaged Normal Mixture model for density estimation based on normal mixture models. Instead of selecting the appropriate number of components in a normal mixture model, I first estimate a series of normal mixture models with different number of components, then I take these estimated normal mixture models as given and mix all these models. This new method is more stable and generally more accurate than the best selected normal mixture models. I propose two methods to find the appropriate weights in the Averaged Normal Mixture model, one is based on likelihood cross validation and the other one is based on BIC weights. I have established the theoretical properties of the proposed estimator and the simulation results demonstrate its good performance on different kind of densities. Finally, I illustrate that our proposed estimator behaves well on a real world data set.

The second essay focuses on forecasts of the day ahead electricity demand in the southern region of Texas. I have proposed a novel hybrid Gaussian Process forecasting model that combines the strength of two different forecast strategies. The proposed GP model is shown to provide superior multi-step ahead point and probabilistic forecasts of hourly electricity demand. Except for the statistical side, I show that the proposed method can be used to reduce the expected cost of electricity supply relative to the conventional linear regression approach, and further economic benefit is obtained when the probabilistic forecasts are used to derive an optimal bidding strategy for the electricity suppliers.

In the third essay, I propose a non-stationary modeling approach based on the method of Gaussian process regression for crop yields. This approach departs from the conventional two-step estimation procedure and allows the yield distributions to vary over time. I further develop a performance weighted model averaging method to construct densities as mixture of Gaussian processes. This method not only facilitates information pooling but greatly improves the flexibility of the resultant predictive density of crop yields. The simulation results on crop insurance premium estimation show that the proposed method is comparable and often preferred to the conventional two-step

estimation procedures regardless of whether the underlying distributions are stationary. When the underlying distributions are non-stationary, our method consistently outperforms its competitors. At last, I demonstrate the utility of the proposed method with an application to crop insurance policies selection from the insurance companies' point of view. I adopt the decision theoretic framework that tailors the decision process according to the loss function of the decision maker and derives a feasible solution by minimizing the expected loss (or risk) with respect to a distribution regarding the source of uncertainty. Results suggest that the proposed method provides an effective tool in identifying profitable insurance policies.

REFERENCES

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

Miltiadis Alamaniotis, Stylianos Chatzidakis, and Lefteri H Tsoukalas. Monthly load forecasting using kernel based gaussian process regression. In *9th Mediterranean Conference on Power Generation, Transmission, Distribution, and Energy Conversion, Athens, Greece, pp. 1-7, November 2014*, Athens, Greece, 2014. IET.

Francis Annan, Jesse Tack, Ardian Harri, and Keith Coble. Spatial pattern of yield distributions: implications for crop insurance. *American Journal of Agricultural Economics*, 96(1):253–268, 2013.

Jushan Bai and Serena Ng. Tests for skewness, kurtosis, and normality for time series data. *Journal of Business & Economic Statistics*, 23(1):49–60, 2005.

J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

Vincenzo Bianco, Oronzio Manca, and Sergio Nardini. Electricity consumption forecasting in italy using linear regression models. *Energy*, 34(9):1413–1421, 2009.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. An improvement of the nec criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20(3): 267–272, 1999.

Christophe Biernacki, Gilles Celeux, and Gérard Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725, 2000.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Manuel Blum and Martin Riedmiller. Electricity demand forecasting using gaussian processes. *Power*, 10:104, 2013.

Ralph R Botts and James N Boles. Use of normal-curve theory in crop insurance ratemaking. *Journal of Farm Economics*, 40(3):733–740, 1958.

Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

Steven T Buckland, Kenneth P Burnham, and Nicole H Augustin. Model selection: an integral part of inference. *Biometrics*, pages 603–618, 1997.

Kenneth P. Burnham and David R. (biologiste) Anderson. *Model selection and multimodel inference : a practical information-theoretic approach*. Springer, New York, 2002. ISBN 0-387-95364-7. URL `http://opac.inria.fr/record=b1100695`. Model selection and inference, cop. 1998.

B. Chen, M. Chang, and C. Lin. Load forecasting using support vector machines: A study on eunite competition 2001. *IEEE transactions on power systems*, 19(4):1821–1830, 2004a.

Hanfeng Chen, Jiahua Chen, and John D Kalbfleisch. Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1): 95–115, 2004b.

Jiahua Chen and J.D. Kalbfleisch. Penalized minimum-distance estimates in finite mixture models. *Canadian Journal of Statistics*, 24(2):167–175, 1996. ISSN 1708-945X. doi: 10.2307/3315623. URL `http://dx.doi.org/10.2307/3315623`.

Adrian Corduneanu and Christopher M Bishop. Variational bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, volume 2001, pages 27–34. Morgan Kaufmann Waltham, MA, 2001.

David R Cox. Prediction by exponentially weighted moving averages and related methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 414–422, 1961.

Didier Dacunha-Castelle, Elisabeth Gassiat, et al. Testing the order of a model using locally conic parametrization: population mixtures and stationary arma processes. *The Annals of Statistics*, 27(4):1178–1209, 1999.

Cees Diks, Valentyn Panchenko, and Dick Van Dijk. Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2):215–230, 2011.

Erkan Erdogdu. Electricity demand analysis using cointegration and arima modelling: A case

study of turkey. *Energy policy*, 35(2):1129–1146, 2007.

Shu Fan and Rob J Hyndman. Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems*, 27(1):134–141, 2012.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

Paul Gallagher. US soybean yields: estimation and forecasting with nonsymmetric disturbances. *American Journal of Agricultural Economics*, 69(4):796–803, 1987.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Irving John Good. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 107–114, 1952.

Barry K Goodwin and Alan P Ker. Nonparametric estimation of crop yield distributions: implications for rating group-risk crop insurance contracts. *American Journal of Agricultural Economics*, 80(1):139–153, 1998.

Barry K Goodwin, Matthew C Roberts, and Keith H Coble. Measurement of price risk in revenue insurance: implications of distributional assumptions. *Journal of Agricultural and Resource Economics*, 25:195–214, 2000.

Bruce E Hansen. Least squares model averaging. *Econometrica*, 75(4):1175–1189, 2007.

Bruce E Hansen and Jeffrey S Racine. Jackknife model averaging. *Journal of Econometrics*, 167 (1):38–46, 2012.

Ardian Harri, Keith H Coble, Alan P Ker, and Barry J Goodwin. Relaxing heteroscedasticity assumptions in area-yield crop insurance rating. *American Journal of Agricultural Economics*, 93(3):707–717, 2011.

Henrique Steinherz Hippert, Carlos Eduardo Pedreira, and Reinaldo Castro Souza. Neural networks for short-term load forecasting: A review and evaluation. *IEEE Transactions on power systems*, 16(1):44–55, 2001.

Nils Lid Hjort and Gerda Claeskens. Frequentist model average estimators. *Journal of the Ameri-*

can Statistical Association, 98(464):879–899, 2003.

Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

Tao Hong. *Short term electric load forecasting*. PhD dissertation, North Carolina State University, Sep 10th, 2010, 2010.

Tao Hong and Shu Fan. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, 32(3):914–938, 2016.

S. Huang and K. Shih. Short-term load forecasting via arma model identification including non-gaussian process considerations. *IEEE Transactions on power systems*, 18(2):673–679, 2003.

Lancelot F James, Carey E Priebe, and David J Marchette. Consistent estimation of mixture complexity. *Annals of Statistics*, pages 1281–1296, 2001.

AR Jung and CA Ramezani. Valuing risk management tools as complex derivatives: an application to revenue insurance. *Journal of Financial Engineering*, 8:99–120, 1999.

Richard E Just and Quinn Weninger. Are crop yields normally distributed? *American Journal of Agricultural Economics*, 81(2):287–304, 1999.

Hiroyuki Kasahara and Katsumi Shimotsu. Testing the number of components in normal mixture regression models. *Journal of the American Statistical Association*, 110(512):1632–1645, 2015.

Kadir Kavaklioglu. Modeling and prediction of turkeyâĂŹs electricity consumption using support vector regression. *Applied Energy*, 88(1):368–375, 2011.

Alan P Ker and Keith Coble. Modeling conditional yield densities. *American Journal of Agricultural Economics*, 85(2):291–304, 2003.

Alan P Ker and Barry K Goodwin. Nonparametric estimation of crop insurance rates revisited. *American Journal of Agricultural Economics*, 82(2):463–478, 2000.

Alan P Ker, Tor N Tolhurst, and Yong Liu. Bayesian estimation of possibly similar yield densities: implications for rating crop insurance contracts. *American Journal of Agricultural Economics*, 98(2):360–382, 2015.

Douglas J Leith, Martin Heidl, and John V Ringwood. Gaussian process prior models for elec-

trical load forecasting. In *Probabilistic Methods Applied to Power Systems, 2004 International Conference on*, pages 112–117. IEEE, 2004.

Sebastian Lerch, Thordis L Thorarinsdottir, Francesco Ravazzolo, and Tilmann Gneiting. Forecaster's dilemma: extreme events and forecast evaluation. *Statistical Science*, 32(1):106–127, 2017.

Brian G Leroux et al. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20 (3):1350–1360, 1992.

Gilbert Leung and Andrew R Barron. Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, 52(8):3396–3410, 2006.

Hua Liang, Guohua Zou, Alan TK Wan, and Xinyu Zhang. Optimal weight choice for frequentist model average estimators. *Journal of the American Statistical Association*, 2012.

Qingfeng Liu and Ryo Okui. Heteroscedasticity-robust cp model averaging. *The Econometrics Journal*, 16(3):463–472, 2013.

JM Lourenco and PJ Santos. Short-term load forecasting using a gaussian process model: The influence of a derivative term in the input regressor. *Intelligent Decision Technologies*, 6(4): 273–281, 2012.

J Steve Marron and Matt P Wand. Exact mean integrated squared error. *The Annals of Statistics*, pages 712–736, 1992.

Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Geoffrey J McLachlan. On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Applied statistics*, pages 318–324, 1987.

Geoffrey J McLachlan and Kaye E Basford. Mixture models. inference and applications to clustering. *Statistics: Textbooks and Monographs, New York: Dekker, 1988*, 1, 1988.

Hiroyuki Mori and Masatarou Ohmi. Probabilistic short-term load forecasting with gaussian processes. In *Proceedings of the 13th International Conference on Intelligent Systems Application to Power Systems, 2005*, pages 6–pp. IEEE, 2005.

Charles B Moss and J Scott Shonkwiler. Estimating yield distributions with a stochastic trend and

nonnormal errors. *American Journal of Agricultural Economics*, 75(4):1056–1062, 1993.

Dirk Ormoneit and Volker Tresp. Averaging, maximum penalized likelihood and bayesian estimation for improving gaussian mixture probability density estimates. *IEEE Transactions on Neural Networks*, 9(4):639–650, 1998.

Vitor A Ozaki and Ralph S Silva. Bayesian ratemaking procedure of crop insurance contracts with skewed distribution. *Journal of Applied Statistics*, 36(4):443–452, 2009.

Adrian E Raftery, David Madigan, and Jennifer A Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191, 1997.

Octavio A Ramírez. Estimation and use of a multivariate parametric model for simulating heteroskedastic, correlated, nonnormal random variables: the case of corn belt corn, soybean, and wheat yields. *American Journal of Agricultural Economics*, 79(1):191–205, 1997.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, 2006.

Kathryn Roeder and Larry Wasserman. Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902, 1997.

Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

Antti Sorjamaa and Amaury Lendasse. Time series prediction using dirrec strategy. *In M. Verleysen, editor, ESANN, European Symposium on Artificial Neural Networks, pages 143âĂŞ148. European Symposium on Artificial Neural Networks, April 26-28 2006*, 2006.

Stephen M Stohs and Jeffrey T LaFrance. A learning rule for inferring local distributions over space and time. In *Spatial and Spatiotemporal Econometrics*, pages 295–331. Emerald Group Publishing Limited, 2004.

Jesse Tack, Andrew Barkley, and Lawton Lanier Nalley. Heterogeneous effects of warming and drought on selected wheat variety yields. *Climatic change*, 125(3-4):489–500, 2014.

Souhaib Ben Taieb, Gianluca Bontempi, Amir F Atiya, and Antti Sorjamaa. A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting

competition. *Expert systems with applications*, 39(8):7067–7083, 2012.

James W Taylor. Short-term electricity demand forecasting using double seasonal exponential smoothing. *Journal of the Operational Research Society*, 54(8):799–805, 2003.

James W Taylor and Roberto Buizza. Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power Systems*, 17(3):626–632, 2002.

James W Taylor and Patrick E McSharry. Short-term load forecasting methods: An evaluation based on european data. *IEEE Transactions on Power Systems*, 22(4):2213–2219, 2007.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Tor N Tolhurst and Alan P Ker. On technological change in crop yields. *American Journal of Agricultural Economics*, 97(1):137–158, 2014.

Rafal Weron. *Modeling and forecasting electricity loads and prices: A statistical approach*. Wiley, Chichester., 2007.

Michael P Windham and Adele Cutler. Information ratios for validating mixture analyses. *Journal of the American Statistical Association*, 87(420):1188–1192, 1992.

Mi-Ja Woo and TN Sriram. Robust estimation of mixture complexity for count data. *Computational statistics & data analysis*, 51(9):4379–4392, 2007.

Joshua D Woodard and Bruce J Sherrick. Estimation of mixture models using cross-validation optimization: implications for crop yield distribution modeling. *American Journal of Agricultural Economics*, 93(4):968–982, 2011.

Ximing Wu and Yu Yvette Zhang. Nonparametric estimation of crop yield distributions: a panel data approach. In *AAEA 2012 Annual Meeting, Seattle, WA*, number 124630, 2012.

Tao Xiong, Yukun Bao, and Zhongyi Hu. Beyond one-step-ahead forecasting: evaluation of alternative multi-step-ahead forecasting models for crude oil prices. *Energy Economics*, 40:405–415, 2013.

Yuhong Yang. Mixing strategies for density estimation. *The Annals of Statistics*, 28(1):75–87,

2000.

Yuhong Yang. Adaptive regression by mixing. *Journal of the American Statistical Association*, 96 (454):574–588, 2001.

Yu Yvette Zhang. A density-ratio model of crop yield distributions. *American Journal of Agricultural Economics*, 99(5):1327–1343, 2017.

Ying Zhu, Barry K Goodwin, and Sujit K Ghosh. Modeling yield risk under technological change: dynamic yield distributions and the us crop insurance program. *Journal of Agricultural and Resource Economics*, 36:192–210, 2011.

Zoran Zivkovic and Ferdinand van der Heijden. Recursive unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):651–656, 2004.

APPENDIX A

NOTATION AND ASSUMPTIONS FOR THEOREM 1 IN SECTION 2

Let $(\mathcal{X}, \mathcal{F})$ be a measurable space and let λ be a σ-finite measure on $\mathcal{F}$. Whenever we mention below that a probability measure on $\mathcal{F}$ has a density we means it has a Radon-Nikodym derivative with respect to λ.

Consider a family of normal mixture normal distributions $\mathcal{H} = \{\phi_{\theta_i}(x) : \theta \in \Theta \subset \mathbb{R}^d\}$[1] over $\mathcal{X}$. The class of k component mixtures f_k is defined as

$$\mathcal{C}_k = conv_k(\mathcal{H}) = \{f : f(x) = \sum_{i=1}^{k} \lambda_i \phi_{\theta_i}(x), \sum_{i=1}^{k} \lambda_i = 1, \lambda_i \geq 0, \theta_i \in \Theta\}.$$

In a similar way we define the class of continuous convex combinations

$$\mathcal{C} = conv(\mathcal{H}) = \{f : f(x) = \int_{\Theta} \phi_{\theta_i}(x) P(d\theta), P \; is \; a \; probability \; measure \; on \; \Theta\}.$$

Assumptions for Theorem 1

Assumption A.1. Assume basis $0 < a \leq \phi_\theta(x) \leq b. \forall x \in \mathcal{X}, \phi_\theta(x) \in \mathcal{H}$.

Assumption A.2. Assume underlying function $0 < a \leq f(x) \leq b, \forall x \in \mathcal{X}$.

Assumption A.3. Given the data, we have the likelihood function $L(\Theta) = \sum_{i=1}^{n} \log p(x_i|\Theta)$, we assume the matrix of second derivatives of $L(\Theta)$ is defined and negative definite for all Θ,. Then there is a unique maximum-likelihood, and the estimators generated by EM algorithm will converge to this value (Redner and Walker, 1984).

[1]For ease of notation, we use $\phi_{\theta_i}(x)$ to denote $\phi(x; \mu_i, \Sigma_i)$.

APPENDIX B

PROOF OF THEOREM 1 IN SECTION 2

Before we prove for Theorem 1, we need to introduce several Lemmas, we will denote $f_i = f(x_i)$. Let $\varepsilon_1, \cdots, \varepsilon_n$ be i.i.d. Rademacher random variables, i.e. $P(i = -1) = P(i = +1) = 1/2$.

Lemma1 (Comparison Inequality for Rademacher Processes)

If $\phi_i : \mathbb{R} \to \mathbb{R} (i = 1, \cdots, n)$ are contractions ($\phi_i(0) = 0$ and $|\phi_i(s) - \phi_i(t)| \leq |s - t|$), then

$$E_\varepsilon \sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} \varepsilon_i \phi_i(f_i)| \leq 2 E_\varepsilon \sup_{f \in \mathcal{F}} |\sum_{i=1}^{n} \varepsilon_i f_i|$$

Lemma2 (Symmetrization)

Consider the following processes:

$$Z(x) = \sup_{f \in \mathcal{F}} |Ef - \frac{1}{n} \sum_{i=1}^{n} f_i|, R(x) = \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f_i|$$

Then

$$EZ(x) \leq 2ER(x)$$

Lemma3 (McDiarmid's Inequality)

Let $x_1, \cdots, x_n, x_1', \cdots, x_n' \in \Omega$ be i.i.d. random variables and let Z: $\Omega^n \to \mathbb{R}$ such that

$$\forall x_1, \cdots, x_n, x_1', \cdots, x_n', |Z(x_1, \cdots, x_n) - Z(x_1, \cdots, x_{i-1}, x_i', x_{i+1}, x_n)| \leq c_i,$$

then

$$P(Z - EZ > \varepsilon) \leq exp(-\frac{2\varepsilon^2}{\sum_{i=1}^{n} c_i^2})$$

We choose $\{f_1, f_2, \cdots, f_K\}$ as the underlying mixture basis. f is arbitrary fixed density, $h \in C$ and $x \in \mathcal{X}$. Our proof is based on the proof of Theorem 4.1 (Rakhlin, Panchenko and Mukherjee,

2005).

Proof

First, we apply Lemma 3 to the random variable $Z(x_1, \cdots, x_n) = \sup_{h \in \mathcal{C}} |\frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - E \log \frac{h}{f}|$. Let $t_i = \log \frac{h(x_i)}{f(x_i)}$ and $t_i' = \log \frac{h(x_i')}{f(x_i')}$. The bound on the martingale difference follows:

$$
\begin{aligned}
&|Z(x_1, \cdots, x_i', \cdots, x_n) - Z(x_1, \cdots, x_i, \cdots, x_n)| \\
&\qquad = |\sup_{h \in \mathcal{C}} |E \log \frac{h}{f} - \frac{1}{n}(t_1 + \cdots + t_i + \cdots + t_n)| \\
&\qquad - \sup_{h \in \mathcal{C}} |E \log \frac{h}{f} - \frac{1}{n}(t_1 + \cdots + t_i' + \cdots + t_n)|| \\
&\qquad \leq \sup_{h \in \mathcal{C}} \frac{1}{n} |\log \frac{h(x_i')}{f(x_i')} - \log \frac{h(x_i)}{f(x_i)}| \\
&\qquad \leq \frac{1}{n}(\log \frac{b}{a} - \log \frac{a}{b}) \\
&\qquad = \frac{1}{n} 2 \log \frac{a}{b} = c_i
\end{aligned}
$$

Applying McDiarmid's inequality(Lemma 3),

$$\forall \mu > 0, P(Z - EZ > \mu) \leq exp(-\frac{2\mu^2}{\sum c_i^2}) = exp(-\frac{n\mu^2}{(2\sqrt{2} \log \frac{b}{a})^2})$$

Therefore,

$$\sup_{h \in C} |\frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - E \log \frac{h}{f}| \leq E \sup_{h \in C} |\frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - E \log \frac{h}{f}| + 2\sqrt{2} \log \frac{b}{a} \sqrt{\frac{t}{n}}$$

with probability at least $1 - e^{-t}$, in which $t = \frac{n\mu^2}{(2\sqrt{2} \log \frac{b}{a})^2}$

Here we assume that $\mu \to 0, n\mu^2 \to \infty$. The reason we make this assumption is that when $n \to \infty$, we will have $\mu \to 0$ and $t \to \infty$.

By Lemma 2 we have

$$E\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\log\frac{h(x_i)}{f(x_i)}-E\log\frac{h}{f}|\leq 2E\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\log\frac{h(x_i)}{f(x_i)}|$$

Combining these two inequalities,

$$\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\log\frac{h(x_i)}{f(x_i)}-E\log\frac{h}{f}|\leq 2E\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\log\frac{h(x_i)}{f(x_i)}|+2\sqrt{2}\log\frac{b}{a}\sqrt{\frac{t}{n}}$$

with probability at least $1-e^{-t}$.

Now we need to bound the above expectation of the Rademacher average.

Let $p_i=\frac{h(x_i)}{f(x_i)}-1$ and note that $\frac{a}{b}-1\leq p_i\leq\frac{b}{a}-1$. Consider $\phi(p)=\log(1+p)$. The largest derivative of $\log(1+p)$ on the interval $p\in[\frac{a}{b}-1,\frac{b}{a}-1]$ is at $p=\frac{a}{b}-1$ and is equal to $\frac{b}{a}$. So, $\frac{a}{b}\log(p+1)$ is 1-Lipschitz. Also, $\phi(0)=0$. By Lemma 1,

$$\begin{aligned}
2E\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\log\frac{h(x_i)}{f(x_i)}| &= 2E\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\phi(p_i)|\\
&\leq 4\frac{b}{a}E\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{h(x_i)}{f(x_i)}-\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i|\\
&\leq 4\frac{b}{a}E\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{h(x_i)}{f(x_i)}|+4\frac{b}{a}E|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i|\\
&\leq 4\frac{b}{a}E\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{h(x_i)}{f(x_i)}|+4\frac{b}{a}\frac{1}{\sqrt{n}}.
\end{aligned}$$

Let $h_i=h(x_i), f_i=f(x_i)$. Assume $\phi_i(h_i)=a\frac{h_i}{f_i}$. Note that $|\phi_i(h_i)-\phi_i(g_i)|=\frac{a}{|f_i|}|h_i-g_i|\leq|h_i-g_i|$. Therefore,

$$4\frac{b}{a}E\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\frac{h(x_i)}{f(x_i)}|\leq 8\frac{b}{a^2}E\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i h(x_i)|$$

Combining these inequalities, with probability at least $1 - e^{-t}$

$$E \sup_{h \in C} |\frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - E \log \frac{h}{f}| \leq 8 \frac{b}{a^2} E \sup_{h \in C} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(x_i)| + 2\sqrt{2} \log \frac{b}{a} \sqrt{\frac{t}{n}} + 4 \frac{b}{a} \frac{1}{\sqrt{n}}$$

Here we use the fact that the Rademacher averages of a class are equal to those of the convex hull. Consider $\sup_{h \in C} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(x_i)|$ with $h(x) = \int \phi_\theta(x) P(d\theta)$, the above supremum is equal to $\sup_\theta |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi_\theta(x_i)|$, the corresponding supremum on the basis functions ϕ. Therefore, $E \sup_{h \in C} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(x_i)| = E \sup_\theta |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi_\theta(x_i)|$.

Then we use the classical result from van der Vaart and Wellner (1996),

$$E \sup_{\phi \in \mathcal{H}} |\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \phi_\theta(x_i)| \leq \frac{c}{\sqrt{n}} \int_0^b \log^{\frac{1}{2}} D(\mathcal{H}, \varepsilon, d_n) d\varepsilon$$

where $D(\mathcal{H}, \varepsilon, d_n)$ is the covering number of the famliy $\mathcal{H}$, d_n is the empirical distance with respect to the set S.

Combining all the results together, the following inequality holds with probability at least $1 - e^{-t}$,

$$\sup_{h \in C} |\frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - E \log \frac{h}{f}| \leq \frac{c}{\sqrt{n}} \int_0^b \log^{\frac{1}{2}} D(\mathcal{H}, \varepsilon, d_n) d\varepsilon] + 2\sqrt{2} \log \frac{b}{a} \sqrt{\frac{t}{n}} + 4 \frac{b}{a} \frac{1}{\sqrt{n}}$$

Note that we also assume when $n \to \infty$, $\mu \to 0$ and $n\mu^2 \to \infty$. So when $n \to \infty$, the last two terms $2\sqrt{2} \log \frac{b}{a} \sqrt{\frac{t}{n}}$ and $4 \frac{b}{a} \frac{1}{\sqrt{n}}$ will converge to 0 and $1 - e^{-t}$ will converge to 1.

So we can get the conclusion that when $n \to \infty$,

$$\sup_{h \in C} |\frac{1}{n} \sum_{i=1}^{n} \log \frac{h(x_i)}{f(x_i)} - E \log \frac{h}{f}| \leq \frac{c}{\sqrt{n}} \int_0^b \log^{\frac{1}{2}} D(\mathcal{H}, \varepsilon, d_n) d\varepsilon$$

with probability 1.

Then we take $\omega^* = argmin D(f||\sum_j \omega_j f_j)$

$$
\begin{aligned}
D(f||\sum_j \hat{\omega}_j \hat{f}_j) - D(f||\sum_j \omega_j^* f_j) &= E\log\frac{f}{\sum_j \hat{\omega}_j \hat{f}_j} - E\log\frac{f}{\sum_j \omega_j^* f_j} \\
&= (E\log\frac{f}{\sum_j \hat{\omega}_j \hat{f}_j} - \frac{1}{n}\sum_{i=1}^n \log\frac{f(x_i)}{\sum_j \hat{\omega}_j \hat{f}_j(x_i)}) \\
&+ (\frac{1}{n}\sum_{i=1}^n \log\frac{f(x_i)}{\sum_j \omega_j^* f_j(x_i)} - E\log\frac{f}{\sum_j \omega_j^* f_j}) \\
&+ (\frac{1}{n}\sum_{i=1}^n \log\frac{f(x_i)}{\sum_j \hat{\omega}_j \hat{f}_j(x_i)} - \frac{1}{n}\sum_{i=1}^n \log\frac{f(x_i)}{\sum_j \omega_j^* f_j(x_i)}) \\
&\leq 2\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^n \log\frac{h(x_i)}{f(x_i)} - E\log\frac{h}{f}| + \frac{1}{n}\sum_{i=1}^n \log\frac{\sum_j \omega_j^* f_j(x_i)}{\sum_j \hat{\omega}_j \hat{f}_j(x_i)}
\end{aligned}
$$

For the first term $\sup_{h\in C}|\frac{1}{n}\sum_{i=1}^n \log\frac{h(x_i)}{f(x_i)} - E\log\frac{h}{f}|$, the bound is given above, we then need to bound the second term $\frac{1}{n}\sum_{i=1}^n \log\frac{\sum_j \omega_j^* f_j(x_i)}{\sum_j \hat{\omega}_j \hat{f}_j(x_i)}$.

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \log\frac{\sum_j \omega_j^* f_j(x_i)}{\sum_j \hat{\omega}_j \hat{f}_j(x_i)} &= \frac{1}{n}\sum_{i=1}^n \log\frac{\sum_j \omega_j^* f_j(x_i)}{\sum_j \omega_j^* \hat{f}_j(x_i)}\frac{\sum_j \omega_j^* \hat{f}_j(x_i)}{\sum_j \hat{\omega}_j \hat{f}_j(x_i)} \\
&= \frac{1}{n}\sum_{i=1}^n \log\frac{\sum_j \omega_j^* f_j(x_i)}{\sum_j \omega_j^* \hat{f}_j(x_i)} + \frac{1}{n}\sum_{i=1}^n \log\frac{\sum_j \omega_j^* \hat{f}_j(x_i)}{\sum_j \hat{\omega}_j \hat{f}_j(x_i)}
\end{aligned}
$$

Note that we choose empirical $\hat{\omega}$ by cross validation.

$$\hat{\omega} = argmax[\sum_{i=1}^n \log\sum_j \omega_j \hat{f}_j^{(-i)}(x_i)]$$

so $\sum_{i=1}^n \log\sum_j \hat{\omega}_j \hat{f}_j^{(-i)}(x_i)$ has the largest likelihood, when n is large, we can regard

$\sum_{i=1}^{n} \log \sum_j \hat{\omega}_j \hat{f}_j(x_i)$ has the largest likelihood, so the second term $\frac{1}{n}\sum_{i=1}^{n} \log \frac{\sum_j \omega_j^* \hat{f}_j(x_i)}{\sum_j \hat{\omega}_j \hat{f}_j(x_i)} \leq 0$. For the first term, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \log \frac{\sum_j \omega_j^* f_j(x_i)}{\sum_j \omega_j^* \hat{f}_j(x_i)} &= \frac{1}{n}\sum_{i=1}^{n} \log \frac{\sum_j \omega_j^* f_j(x_i) - \sum_j \omega_j^* \hat{f}_j(x_i) + \sum_j \omega_j^* \hat{f}_j(x_i)}{\sum_j \omega_j^* \hat{f}_j(x_i)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \log(1 + \frac{\sum_j \omega_j^* f_j(x_i) - \sum_j \omega_j^* \hat{f}_j(x_i)}{\sum_j \omega_j^* \hat{f}_j(x_i)}) \\
&\approx \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_j \omega_j^* f_j(x_i) - \sum_j \omega_j^* \hat{f}_j(x_i)}{\sum_j \omega_j^* \hat{f}_j(x_i)} \to 0
\end{aligned}$$

the last two steps hold since when n is large, $f_j(x_i) \to \hat{f}_j(x_i)$, Therefore, when $n \to \infty$, with probability 1,

$$D(f||\sum_j \hat{\omega}_j \hat{f}_j) - D(f||\sum_j \omega_j^* f_j) \leq \frac{c_1}{\sqrt{n}} \int_0^b \log^{\frac{1}{2}} D(\mathcal{H}, \varepsilon, d_n) d\varepsilon \to 0$$

which means

$$\frac{D(f||\sum_j \hat{\omega}_j \hat{f}_j)}{\inf_\omega D(f||\sum_j \omega_j f_j)} \xrightarrow{p} 1$$

APPENDIX C

PROOF OF THEOREM 2 IN SECTION 2

In this section we prove that BIC weights are consistent in selecting the true model (if the true model is in the candidates set) or the quasi-true model (if the true model is not in the candidates set). This proof is mainly based on section 6.3 and 6.4 of Burnham and Anderson [2002].

Case 1

Assume that we have a sequence of models $g_1, \cdots, g_t, \cdots, g_K$ and that the true model, g_t, is in this sequence.Then by using BIC criterion, we can select the true model g_t with probability 1 as n gets large. Also we know the BIC weights are

$$P(g_i) = \frac{\exp\{-\frac{1}{2}\Delta BIC_i\}}{\sum_{j=1}^{M} \exp\{-\frac{1}{2}\Delta BIC_j\}},$$

since there is a true model, g_t, in the set then $P(g_t)$ goes to 1 as n goes to infinity; and of course $P(g_i)$ goes to 0 for all other models, we will consistently select the true model when using the BIC weights.

Case 2

Assume that we have a sequence of models $g_1, \cdots, g_K$ and that the true model is not in this sequence. As sample size $n \to \infty$, the model selected by BIC is consistent for the quasi-true model in the model set. Now we prove the BIC weights are consistent in selecting the quasi-true model. For a random sample we can write $D(f||g_i) = nD_1(f||g_i)$, where $D_1(f||g_i)$ being for $n = 1$ is a

constant as regards sample size. Hence, $D(f||g_i) - D(f||g_j) = n(D_1(f||g_i) - D_1(f||g_j))$.

$$BIC_i - BIC_j \approx 2n(D_1(f||g_i) - D_1(f||g_j)) + (\text{dim}(i) - \text{dim}(j)) \log(n).$$

In the case $t = K$,

$$2n(D_1(f||g_i) - D_1(f||g_K)) > 0,\ i < K$$

Hence, as $n \to \infty$ all these differences diverge to ∞, BIC criterion will select the quasi-true model g_K with certainty as $n \to \infty$. Also the BIC weights $P(g_K)$ goes to 1 as n goes to infinity.

In the case $t < K$, which model g_t nested in models $g_i, i > t$. The relevant differences are

$$BIC_i - BIC_t \approx 2n[D_1(f||g_i) - D_1(f||g_t)] + (\text{dim}(i) - \text{dim}(t)) \log(n), i < t,$$

$$BIC_i - BIC_t \approx -\chi_i^2 + (\text{dim}(i) - \text{dim}(t)) \log(n), i > t.$$

Here, χ_i^2 is a central chi-square random variable on $\text{dim}(i) - \text{dim}(t)$ degrees of freedom. For all $i < t$ the differenes $BIC_i - BIC_t$ become infinite as $n \to \infty$, with probability 1, hence model g_t is always selected over models g_1 to g_{t-1} and the BIC weights $P(g_t)$ goes to 1. For all $i > t$ the differences $BIC_i - BIC_t$ become infinite as $n \to \infty$, with probability 1, because as long as $\text{dim}(i) > \text{dim}(t)$ term $\log(n)$ will diverge to infinity, so model g_t is always selected and BIC weights $P(g_t)$ goes to 1.