SPARSITY AND SHRINKAGE IN PREDICTIVE DENSITY ESTIMATION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Gourab Mukherjee
May 29, 2013

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Iain Johnstone, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Persi Diaconis

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

David Donoho

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

ABSTRACT

We develop new perspectives on the roles of sparsity and shrinkage in predictive density estimation under Kullback-Leibler loss. Our results explain and extend some recently observed information theoretic connections between predictive density estimation and the well-studied normal mean estimation problem. We find new phenomena in sparse minimax prediction which contrast with point estimation theory results and are explained by the new notion of risk diversification. We generalize these new uncertainty sharing ideas to address the nature of optimal shrinkage over unconstrained parameter spaces. Our density estimates can be used to construct competitively optimal probability forecasts and our results give some theoretical support to log-optimality based forecasting techniques used in the fields of weather forecasting, financial investments and sports betting. Motivational stories and examples from the world of sports, stock markets and wind speed profiles are used to suggest the scope of the theory developed in this thesis.

ACKNOWLEDGMENTS

I am indebted to Professor Iain M. Johnstone for his encouragement, mentorship and undivided attention. I am grateful for his patience and counsel which greatly helped me to wade through the technical difficulties and to assimilate new ideas into statistical prediction analysis. This thesis also owes its debt to Professors David L. Donoho, Persi W. Diaconis and Thomas M. Cover for deeply influencing our research trajectory. The shrinkage results in Chapter 3 and many other frequentist perspectives in this thesis, were inspired by conversations with Professor Donoho. Professor Diaconis drew our attention to issues regarding the choice of loss function. It led us to Professor Cover and numerous enjoyable interactions with him yielded a better understanding of the entropy loss. This helped us in extending the ideas in the thesis to provide statistical support to some of the information-theory based repeated investment strategies. I would like to express my profound gratitude to Professors Trevor J. Hastie, Joseph P. Romano, Guenther Walther and Chiara Sabatti for their help and support during my Ph.D. life.

My stay at Stanford was very enjoyable because of the awesome company I had in and around the Statistics department. I greatly enjoyed hanging out with my friends from the Sequoia Hall, who in addition to being exceptionally talented, are also admirable individuals. I thank Biswajoy Roy Chaudhuri, Adrish Sen and Nandini Sen for their kindness and assistance in introducing me to some very exciting biology research problems. Collaborations and time spent with them in the medical school, played an important role in my scientific maturation.

Last, but not least, I am deeply appreciative of the unstinting support from my family which has made everything seem possible.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

We discuss the role of predictive distributions in forecasting problems. We describe the traditional set-ups and their information theoretic implications as we gradually narrow down to our high dimensional predictive framework. We provide examples and illustrations of our model and its extensions.

This thesis aims to increase our current understanding of high dimensional probability forecasting problems by incorporating their deep connections with high dimensionality issues seen in point estimation theory. We would like to state in the very beginning that our focus is on:

(a) prediction not estimation and

(b) predicting distributions (more specifically densities) not point-predictions.

As in statistical prediction analysis, here our objective is to choose a probability distribution which will be good in predicting the behavior of future samples (Aitchison & Dunsmore 1975). If the observed past data $\mathbf{X}$ and the unobserved future data $\mathbf{Y}$ are generated from a joint density $f(\mathbf{x}, \mathbf{y})$, the objective is to estimate the future conditional density of $f(\mathbf{y} \,|\mathbf{X} = \mathbf{x})$, also referred to as the predictive density (Geisser 1971). For practical purposes, we usually need to forecast functionals of the predictive density. Good predictive performances can be ensured by using functions of predictive density estimates which are optimally chosen based on appropriate goodness of fit measure.

1.1 Density Estimates in Prediction Analysis

A general framework for predictive density estimation outlined in the pioneering work of Aitchison & Dunsmore (1975) and Geisser (1993) has been subsequently adapted in the different fields of statistical decision theory, information theory, game theory, econometrics, machine learning and mathematical finance. Here, we consider a flexible, parametric predictive model which can accommodate most dependencies in the data. Suppose we observe $\mathbf{X}$ with $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_{m_1}$ independently generated from a parametric density $f_{(\boldsymbol{\theta},a_i)}(\cdot)$ indexed by unknown parameters $\boldsymbol{\theta}$ and known parameters $A = \{a_i : 1 \leq i \leq m_1\}$. The future data $\mathbf{Y} = \{\mathbf{Y}_1, \cdots, \mathbf{Y}_{m_2}\}$ are generated from successive independent parametric densities $\{f_{(\boldsymbol{\theta},b_i)}(\cdot) : 1 \leq i \leq m_2\}$ with time-invariant unknown parameters $\boldsymbol{\theta}$ and known parameters $B = \{b_i : 1 \leq i \leq m_2\}$. In such a predictive model the dependence between the past and the future is based on the time-invariant parameters $\boldsymbol{\theta}$.

Generic Parametric Predictive Model: M

$$\begin{array}{lll} \text{PAST} & \text{OBSERVATIONS:} & \mathbf{X}_i \overset{\text{indep.}}{\sim} f_{(\boldsymbol{\theta},\, a_i)}(\cdot)\,, \quad i = 1, \cdots, m_1 \\ \text{FUTURE} & \text{OBSERVATIONS:} & \mathbf{Y}_j \overset{\text{indep.}}{\sim} f_{(\boldsymbol{\theta},\, b_j)}(\cdot)\,, \quad j = 1, \cdots, m_2. \end{array}$$

If $\boldsymbol{\theta}$ is fixed the true predictive density of $\mathbf{Y}$ would be $f_{(\boldsymbol{\theta},B)}(\cdot) = \prod_{j=1}^{m_2} f_{(\boldsymbol{\theta},b_j)}(\cdot)$. We would like to estimate it by density estimates $\widehat{p}(.|\mathbf{X} = \mathbf{x})$. A perspective in prediction inference is to use the concept of predictive likelihoods (Hinkley 1979, Lauritzen 1974) and its variants (Bjørnstad 1990), to infer about the future $\mathbf{Y}$ based on $\mathbf{X}$, with $\boldsymbol{\theta}$ playing the role of a nuisance parameter. Most predictive likelihoods (Butler 1986) are functions of the future conditional density and can be effectively evaluated based on efficient predictive density estimates. In this context we would like to emphasize that the statistical techniques discussed here are different from the widely used Kernel based non-parametric density estimation techniques (Silverman 1986, Ferraty & Vieu 2006, Chapter5). However since our model includes parametric densities of any dimensions, we can incorporate non-parametric regression problems following the techniques in Nussbaum (1996), Grama & Nussbaum (2002) and Efromovich (1999).

One aspect of predictive modeling is that it explicitly avoids the problem of overfitting the data. In practice it is generally carried out through cross-validation (Stone 1974). In our results on predictive density estimation (Chapter 3) we would see similarities with the theory of cross-validation in regression models (Yang 2007). Also, there are both Frequentist and Bayesian approaches to predictive density estimation (Smith 1999) though the term predictive density may seem to be more associated with Bayesian perspective. We would consider both the approaches and would introspect the relation between their corresponding optimal estimators. To evaluate the performance of density estimates we need a goodness of fit measures between densities.

Goodness of fit & Predictive Risk

Given any discriminatory measure L between two densities the loss between the true predictive density $f_{(\boldsymbol{\theta},B)}$ and its estimate $\widehat{p}(\cdot|\mathbf{X}=\mathbf{x})$ can be evaluated at each parametric value $\boldsymbol{\theta}$

$$\mathrm{L}\big(\boldsymbol{\theta},\, \widehat{p}\big(\,\cdot\,\big|\mathbf{x}\big)\big) = \mathrm{L}\big(\, f_{(\boldsymbol{\theta},B)}(\cdot), \widehat{p}\big(\,\cdot\,\big|\mathbf{X}=\mathbf{x}\big)\big)$$

and averaging the loss over the past $\mathbf{X}$, we have the can evaluate the predictive risk of the strategy $\widehat{p}$ at $\boldsymbol{\theta}$ as

$$\rho_{\mathrm{L}}\big(\boldsymbol{\theta},\, \widehat{p}\big) = \int f_{(\boldsymbol{\theta},A)}(x) \,\times\, \mathrm{L}\big(\,\boldsymbol{\theta},\, \widehat{p}\big(\,\cdot\,\big|\mathbf{x}\big)\big)\, d\mathbf{x}.$$

There are different choice for a goodness of fit measure like Total Variation, Chi-squared, Hellinger, Kullback-Leibler, etc (Cha 2007). Any notion of loss on functionals of the predictive density will depend on the choice of the goodness of fit measure L. So, the choice of L will be based on the forecasting problem.

Some Examples of the Parametric Predictive Model M are presented next. Consider a simple forecasting set-up where in the absence of any covariates we observe m_1 i.i.d. past vectors $\{\mathbf{X}_i : 1 \leq i \leq m_1\}$. Each of these past vectors $\mathbf{X}_i$ has co-ordinates independently generated from the same parametric distribution but with different parameters. For example, we can consider $\mathbf{X}_i$ being i.i.d. from a d–dimensional product Poisson or Bernoulli distribution with the associated d-dimensional mean parameter being fixed but unknown. The objective is to predict

the distribution of m_2 future vectors $\{\mathbf{Y}_i : 1 \leq i \leq m_2\}$ which are generated from the same product distribution as the past. Discrete predictive models of these kinds are used for count or binary data. Next, we introduce the Gaussian Regression model. Often approximate normalization transformations of the data are available and inferences based on Gaussian predictive models can be extended to a wide range of forecasting problems.

Homoscedastic Gaussian Regression Model: M.1

$$\text{Past: } \mathbf{X} \sim N(A\,\boldsymbol{\theta},\, \sigma_p^2\, I) \quad \text{and Future } \mathbf{Y} \sim N(B\,\boldsymbol{\theta},\, \sigma_f^2\, I)$$

where σ_p and σ_f are known and represent the past and future standard deviations respectively. The past and the future vectors can be of different lengths. A and B are known design matrices with dimensions $m_1 \times p$ and $m_2 \times p$ respectively. $\mathbf{X}$ and $\mathbf{Y}$ are of lengths m_1 and m_2 and related through linear transformation of the p–dimensional latent parameter $\boldsymbol{\theta}$. If $\boldsymbol{\theta}$ were known, then $\mathbf{X}$ and $\mathbf{Y}$ would have been independent. Statistical analysis in **M.1** depends on proper understanding of the associated orthogonal model,

Orthogonal Gaussian Predictive Model: M.2

$$\mathbf{X} \sim N(\boldsymbol{\theta},\, \sigma_p^2\, I) \quad \text{and} \quad \mathbf{Y} \sim N(\boldsymbol{\theta},\, \sigma_f^2\, I)$$

the past $\mathbf{X}$ and the future vector $\mathbf{Y}$ are of equal length. $\mathbf{X}$ and $\mathbf{Y}$ can be n dimensional vectors related through the unknown n dimensional location parameter $\boldsymbol{\theta}$. Here, in both the models **M.1** and **M.2** we are concerned with only one past and one future observations. These one-sample models are theoretically tractable and phenomena seen here can be replicated in multi-sample models through simulations. A good attribute of **M.2** is that multi-sample models of that kind can be reduced to a one-sample model by using the corresponding sufficient statistics.

Predictive Difficulty of the problem: By r we will be denoting the ratio of the past to future volatilities. So, $r = \sigma_f^2/\sigma_p^2$ can lie between 0 to ∞ and is known. If

this ratio r decreases then predictive difficulty of the problem increases as we will be predicting a less volatile future density based on comparatively noisy past observations. For the non-orthogonal model, the role of r is played by the eigen values of the matrix $(A'A)^{-}B'B$.

Sequential and Multi-sample Models: Next we describe a multiple sample model where m_1 realizations each of n–dimensions are observed and we would like to predict m_2 future realizations which will be generated from the same distribution. The locations may be shifted by a linear transformation on latent factor Θ. The errors E^p and E^f are independent i.i.d ensembles. Usually we would also impose the additional assumption of Gaussianity. An interesting related case would be if the error matrices have i.i.d. rows from $N(0, \Sigma)$. If the data matrices S^p and S^f are known then the problem will be a *fixed design* problem. For *random design* problems, we generally assume that the rows of S^p and S^f are generated independently from a fixed, known n–dimensional distribution.

Multiple Regression Model: M.3

$$X_{m_1 \times n} = S^p_{m_1 \times p} \Theta_{p \times n} + E^p_{m_1 \times n}$$
$$Y_{m_2 \times n} = S^f_{m_2 \times p} \Theta_{p \times n} + E^f_{m_2 \times n}$$

Sequential models can be built based on multi-sample models. Here, the observations are sequentially ordered and at each step the objective is to estimate the predictive density of the next observation based on the observed past. For example, the objective at the i^{th} step is to estimate the predictive density $f_{\boldsymbol{\theta},a_i}$ for the i^{th} observation where a_i is known based on observing $\{X_j \stackrel{d}{=} f_{(\theta_j,a_j)} : j = 1, \cdots, i\}$ where $\{a_j : j \leq i\}$ are measurable with respect to the filtration $\mathcal{F}\{i-\}$. The performance of any predictive scheme is usually based on the cumulative error over a number of steps which reduces to finding the optimal density estimate separately at each step. So, it would involve repeated solving of a multi-sample problem. However, the information about Θ also increases in each step and it may happen that two predictive density estimates which differ a lot initially (when the sample size is small) have reasonably close cumulative loss over a large number of steps. So, optimal estimators in multi-sample

model does not necessarily produce efficient sequential strategies. Here, we mainly study **M.1, M.2** and probable extensions of their results to M.3 or to a sequential framework.

Evaluation of Predictive Schemes is another important issue. Theoretically we can calculate expressions of the predictive risk (at least for the predictive density estimate) at each parametric value $\boldsymbol{\theta}$. However, in practice $\boldsymbol{\theta}$ is unknown. Developing statistically consistent scoring rules for evaluating sequential predictive schemes is an interesting and active area of research (Lai, Gross & Shen 2011, Lichtendahl & Winkler 2007). As we mainly concentrate on non-sequential models, we will skip this issue entirely. We will use a sort of "oracle" estimate by waiting and using (a lot) of future data to construct a very consistent estimate $\widehat{\boldsymbol{\theta}}_f$ of the true $\boldsymbol{\theta}$. In our data illustrations, the predictive schemes will be evaluated at the parametric value $\widehat{\boldsymbol{\theta}}_f$. We will use $\widehat{\boldsymbol{\theta}}_f$ only for evaluation purpose.

Forecasting with High-dimensional Data: Traditionally, much of parametric statistics has been about fixed dimensional parametric models. In the context of our model **M**, this would mean that $\boldsymbol{\theta} \in \mathbb{R}^d$ where d is fixed. Almost all the results in predictive density estimation until the last decade are in this fixed dimensional paradigm. These fixed dimensional results are usually large sample and distribution independent (holds for a large class of parametric families which obey standard regularity conditions) sequential results (Aslan 2006).

However, due to the huge recent progress in data gathering and storing technologies, most modern data sets are quite massive and complex. These complicated data sets need to be modeled by a large number (which may be unbounded) of interacting parameters and the traditional fixed dimensional techniques are ineffective. Recent work of Xu (2007), Ghosh, Mergel & Datta (2008) and Maruyama & Strawderman (2012) have explored estimation of predictive density in high dimensional Gaussian probability spaces. This thesis builds on these ideas and will deal mainly with high-dimensional Gaussian models. In the models that were described before, the high-dimensional framework involves:

M.1 $m_1, m_2 \to \infty$ and $m_1/m_2 \to c \in (0, 1)$ and $p \to \infty$.

M.2 $n \to \infty$.

M.3 $m_1, m_2, n, p \to \infty$ and $m_1/m_2 \to c \in (0, 1)$.

In some cases we will have further prior constraints on the model which will put realistic and tractable restrictions on the parametric spaces. Afterwards, we describe detailed illustrations of these models.

1.2 Prediction with Relative Entropy Loss

Here, we use the information theoretic measure of Kullback & Leibler (1951) as the goodness of fit measure between the true and estimated distributions

$$\mathrm{L}\left(\boldsymbol{\theta}, \widehat{p}\left(\ \cdot\ \middle|\mathbf{x}\right)\right) = \int f_{(\boldsymbol{\theta},B)}(\mathbf{y})\ \log\left(\frac{f_{(\boldsymbol{\theta},B)}(\mathbf{y})}{\widehat{p}\big(\mathbf{y}\big|\mathbf{x}\big)}\right)\ d\mathbf{y}\ .$$

Averaging over the past observations $\mathbf{X}$, the predictive risk of the density estimate $\widehat{p}(\cdot\,|\mathbf{X} = \mathbf{x})$ at $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\rho}\left(\boldsymbol{\theta}, \widehat{p}\right) = \iint\ f_{(\boldsymbol{\theta},A)}(\mathbf{x})\ f_{(\boldsymbol{\theta},B)}(\mathbf{y}) \log\left(\frac{f_{(\boldsymbol{\theta},B)}(\mathbf{y})}{\widehat{p}\big(\mathbf{y}\big|\mathbf{x}\big)}\right)\ d\mathbf{y}\ d\mathbf{x}\ . \tag{1.1}$$

The relative entropy predictive risk $\boldsymbol{\rho}\left(\boldsymbol{\theta}, \widehat{p}\right)$ measures the exponential rate of divergence of the joint likelihood ratio over a large number of independent trials (Larimore 1983). In classical fixed-dimensional parametric analysis, the minimal predictive risk estimate would maximize the expected growth rate in repeated investment scenarios (Cover & Thomas 1991, Chapter 6 and 15). Competitive optimal predictive schemes (Bell & Cover 1980) for gambling, sports betting, portfolio selection, etc can be constructed from predictive density estimates with optimal Kullback-Leibler (KL) risk properties. In data compression set-up $\mathrm{L}\left(\boldsymbol{\theta}, \widehat{p}\left(\ \cdot\ \middle|\mathbf{x}\right)\right)$ reflects the excess average code length that we need if we use the conditional density estimate $\widehat{p}$ instead of the true density to construct a uniquely decodable code for the data $\mathbf{Y}$ given the past $\mathbf{x}$ (McMillan 1956). The notion can be extended to a sequential framework

where minimizing the predictive risk would result in the minimum description length (Rissanen 1984, Barron, Rissanen & Yu 1998) based estimate of the true parametric density (Liang & Barron 2005).

In statistical prediction analysis with this entropy loss, the interesting functionals of predictive density estimates are:
(a) simultaneous Probability estimates of several prespecified events,
(b) predicting behavior of functions of linear transformations of the future Y, in particular the moments of $\log(c'Y)$ where c is a fixed vector.
The modern data deluge has influenced a rapid evolution of these statistical methods towards simultaneous multi-parametric analyses (Donoho 2000) and traditional decision making techniques based on fixed dimensional predictive densities needs extension to high-dimensional parametric models in the following applications:

Data compression: Coding of high-dimensional data (Liang & Barron 2004, Candès 2006, Guo, Shamai & Verdu 2005) needs construction of a decodable code for $\mathbf{Y}$ given the value of $\mathbf{X}$. If the high-dimensional parameter $\boldsymbol{\theta}$ is known the optimal expected length of such a code would have been based on the true density $f_{(\boldsymbol{\theta},B)}$. In universal data compression (Rissanen 1984), without any prior knowledge of $\boldsymbol{\theta}$, a choice of predictive density $\widehat{p}\,(\mathbf{Y}|\mathbf{x})$ will be used instead of the true density to construct the code. The excess average code length in that case is given by:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\log_2(1/\widehat{p}(\mathbf{Y}|\mathbf{x})) - \log_2(1/f_{(\boldsymbol{\theta},B)}(\mathbf{Y}))\right] = \mathrm{L}\big(\boldsymbol{\theta}, \widehat{p}\,(.|\mathbf{x})\big)/\log 2 \text{ bits}. \tag{1.2}$$

We ignore the issue of discretization as the loss will be the limit redundancy based on infinitesimally fine choice of discretizations (Csiszár 1973, Vajda 2002). Now, if the parameter $\boldsymbol{\theta}$ was generated from the distribution π then the minimal excess average code length is given by the Bayes risk of the prior π on $\boldsymbol{\theta}$ (Liang 2002). The integrated Bayes risk of the estimator $\widehat{p}\big(\,\cdot\,\big|\mathbf{x}\big)$ with respect to the prior π is given by:

$$B(\pi, \widehat{p}) = \int \rho(\boldsymbol{\theta}, \widehat{p})\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

The marginal density of the past is given by $m_\pi(\mathbf{x}) = \int f_{(\boldsymbol{\theta},A)}(\mathbf{x})\ \pi(\boldsymbol{\theta})\ d\boldsymbol{\theta}$ and

for any $\mathbf{x}$ with $m_\pi(\mathbf{x}) < \infty$ the posterior density is given by The Bayes estimator with respect to the prior π is given by $\pi(\boldsymbol{\theta}|\mathbf{x}) = f_{(\boldsymbol{\theta},A)}(\mathbf{x})\ \pi(\boldsymbol{\theta})\{m_\pi(\mathbf{x})\}^{-1}$. The Bayes risk $B(\pi)$ of a prior π is $\min_{\widehat{p}} B(\pi, \widehat{p})$ and the minimum is attained at the Bayes predictive density. Under the KL loss and for any family $\{f_{(\boldsymbol{\theta},.)} : \boldsymbol{\theta} \in \mathbb{R}^d\}$ of uniformly bounded (above) densities, the Bayes predictive density is given by

$$\widehat{p}_\pi(y|\mathbf{x}) = \int f_{(\boldsymbol{\theta},B)}(y)\ \pi(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta} \quad \text{for all } \mathbf{x} \text{ such that } m_\pi(\mathbf{x}) < \infty.$$

The formal proof and associated assumptions are stated in Chapter 2. Now, if the prior π and the class of uniformly bounded densities f is such that $m_\pi(\mathbf{X}) < \infty$ almost surely, then the Bayes estimator also minimizes the mutual conditional information $I_\pi(\Theta; \mathbf{Y}|\mathbf{X})$ between the unknown parameter $\boldsymbol{\theta}$ and a future observation $\mathbf{Y}$ given the past $\mathbf{X} = \{\mathbf{X}_1, \cdots, \mathbf{X}_m\}$. By definition, we have

$$I_\pi(\Theta; \mathbf{Y}|\mathbf{X}) = \mathbb{E}\left[\log\left(\frac{p(\mathbf{y}, \boldsymbol{\theta}|\mathbf{x})}{p(\mathbf{y}|\mathbf{x})\, p(\boldsymbol{\theta}|\mathbf{x})}\right)\right]$$

where the expectation is over the true joint density

$$p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta})\, f_{(\boldsymbol{\theta},A)}(\mathbf{x})\, f_{(\boldsymbol{\theta},B)}(y) \text{ and } p(y, \boldsymbol{\theta}|\mathbf{x})\{p(\boldsymbol{\theta}|\mathbf{x})\}^{-1} = p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})\{p(\mathbf{x}, \boldsymbol{\theta})\}^{-1}$$

is the ratio of the true conditional distributions and $p(\mathbf{x}, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta})\, f_{(\boldsymbol{\theta},A)}(\mathbf{x})$. And so, the minimum of $I_\pi(\Theta; \mathbf{Y}|\mathbf{X})$ over the class of all possible conditional densities $\widehat{p}(\mathbf{y}|\mathbf{x})$ will be attained at $\widehat{p}_\pi(\mathbf{y}|\mathbf{x})$ and the minimum value is $B(\pi)$.

Thus, $\min_{\widehat{p}} B(\pi, \widehat{p})$ equals the mutual conditional information $I_\pi(\Theta; \mathbf{Y}|\mathbf{X})$ between the unknown parameter $\boldsymbol{\theta}$ and the future data $\mathbf{Y}$ given the past $\mathbf{X}$. Mutual information is associated with the notion of association between random variables and zero conditional mutual information denotes conditional independence. Again, based on the decomposition of

$$I_\pi(\Theta; \mathbf{Y}|\mathbf{X}) = I_\pi(\Theta; \mathbf{X}, \mathbf{Y}) - I_\pi(\Theta; \mathbf{X})$$

it can be seen that minimizing the Bayes predictive risk $B(\pi, \widehat{p})$ would signify extracting the maximum possible dependence of $\mathbf{Y}$ and Θ based on $\mathbf{X}$. The information capacity of the channel C is given by the maximal mutual conditional information $\max_{\pi \in \mathcal{M}} I_{\pi}(\Theta; \mathbf{Y}|\mathbf{X})$ over an appropriate class of priors. Here, we will evaluate C by explicitly calculating the maximal Bayes risk $\max_{\pi \in \mathcal{M}} B(\pi)$.

Sequential Investment with side information: Investment schemes based on high-frequency trading need predictive strategies on financial instruments governed by a large number of parameters (Fan, Lv & Qi 2011). The log-optimal predictive strategies of Barron & Cover (1988) will depend on the high-dimensional density estimates minimizing the predictive risk in Equation 1.1.

Sports Betting: Online betting portals have not only increased traffic but also caused a massive transformation of the fixed-odds sports betting market (Buchdahl 2003). Betfair [1], one of the leading betting exchanges in U.K. matches 15 times as many daily transactions as the London Stock Exchange. These online stochastic markets allow bets with the most lucrative odds to be placed on the joint occurrences of several events (multiple bets). As historical data can be accessed through portal-supplied application programming interface, statistical techniques are being increasingly used in designing betting strategies (Magee 2011) and multi-parametric models are required to estimate the multiple-bets probabilities.

1.3 Some examples of the roles of Sparsity and Shrinkage

The notion of shrinkage developed in Stein (1956) and the theory of sparse estimation (Donoho, Johnstone, Hoch & Stern 1992, Donoho & Johnstone 1994*b*, Mallat 2009, Elad 2010) are very helpful in constructing statistically optimal procedures in a wide range of data problems. In the following examples, we demonstrate how incorporating

[1] `http://sports.betfair.com/`

these two notions in probability forecasting problems can increase the efficiency of prediction schemes.

1.3.1 Sports Betting, Proportional Gambling & Log-Optimality

We consider a typical betting scenario (shown in Table 1.1) based on a game whose outcomes are completely determined by the individual performances $X_1, \cdots, X_k$ of the k players involved in the game. By $\mathbf{X}$ we represent the performance vector and a gambler has the option to bet on m outcomes $J_1(\mathbf{X}), J_2(\mathbf{X}), \cdots, J_m(\mathbf{X})$ which depends solely on $\mathbf{X}$ and have respective odds given by the vector $\mathbf{o} = \{o_1, \cdots, o_m\}$. For simplicity, we initially assume that the outcomes are mutually exclusive and exhaustive. The gambler distributes his wealth over the m events such that b_i fraction is invested in the outcome J_i and so we have the following constraints on the bets:

$$b_i \geq 0 \text{ and } \sum_{i=1}^{m} b_i = 1.$$

So, at the end of the game the gambler will have multiplied his wealth by $b_i o_i$ if $\mathbf{X}$ is such that the outcome J_i occurs. His wealth is given by

$$S(\mathbf{X}) = \sum_{i=1}^{m} b_i o_i \, I(J_i) \tag{1.3}$$

where $I(J_i) = 1$ if J_i occurs and is 0 otherwise.

Assuming that the performance vector $\mathbf{X}$ are generated from a multivariate distribution, each of the outcomes has a inherent unknown probability $p_i = \mathrm{P}(J_i \text{ occurs})$ for $i = 1, \ldots, m$. Now, if we consider a repeated gambling scenario where the gambler reinvests his wealth in succession on a sequence of n independent games. Also, if we assume that each of the games involve the same k players and the performance vectors $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ from the n games are not only independent but also identically distributed, then the gambler's wealth after n races is given by $S_n = \prod_{i=1}^{n} S(\mathbf{X}_i)$. It

Outcomes	Probability	Odds	Bets
$J_1(\mathbf{X})$	p_1	o_1	t_1
.	.	.	.
.	.	.	.
$J_m(\mathbf{X})$	p_m	o_m	t_m

Table 1.1: A sports betting scenario with the odds constrained to be $t_i \geq 0$ and $\sum_{i=1}^{m} t_i = 1$. The probabilities $\mathbf{p} = \{p_1, \cdots, p_m\}$ are unknown and $\sum_{i=1}^{m} p_i = 1$.

follows from Cover & Thomas (1991, Theorem 6.1.1) that as $n \to \infty$ we have

$$S_n = 2^{nW(b,p)}(1 + \mathrm{o}(1)) \text{ where } W(b,p) = \mathbb{E}(\log S(\mathbf{X})) = \sum_{i=1}^{m} p_i \log(b_i o_i)$$

is the doubling rate of the gambler's wealth. On maximizing this long term growth rate over the varied choices of betting strategies $\mathbf{b}$, it can be seen that the optimal strategy is to allocate the bets proportional to the unknown probability vector $\mathbf{p}$.

Lemma 1.3.1.

If the events are mutually exclusive then the optimal doubling rate of the gambler's wealth is $\boldsymbol{W}^{\star}(\boldsymbol{p}) = \sum_{i=1}^{m} p_i \log(o_i/p_i)$ and is achieved by the proportional gambling scheme $\boldsymbol{b}^{\star} = \boldsymbol{p}$.

Proof. See Theorem 6.1.2 in Cover & Thomas (1991). □

So, to optimize the long term growth rate of the gambler's wealth it is important to precisely estimate the unknown probabilities $\mathbf{p}$ of the betting events. In particular if the odds are fair with respect to some distribution i.e.

$$\sum_{i=1}^{n} o_i^{-1} = 1 \text{ and } q_i = o_i \text{ for } i = 1, \cdots, n$$

then the doubling rate is

$$W(b,p) = D(p||q) - D(p||b) \tag{1.4}$$

where D denotes the KL loss between discrete probability distributions. Thus, the doubling rate in this case is the difference between the distance of the bookie's estimate from the true distribution and the gambler's estimate from the true distribution. Also, note that as long as the performance vectors $\{\mathbf{X}_i : 1 \leq i \leq n\}$ are independent these results carry over. So, here on, we drop the unrealistic assumption of invariance of the players and their intrinsic capabilities and consider the situation where the gambler's bets repeatedly over independent games played by possibly different players. Next, we will be relaxing some of other assumptions we made in the beginning as we move to a predictive set-up. For ease of demonstration we describe the results through the example of horse racing.

Horse Race Example:

Suppose there are k horses running in a race. Usually, k is between 10 to 15. Some of the possible events to bet on the race are:

- Win: Predict the winner.
- Exacta: Predict the top 2 in correct order.
- Trifecta: Predict the top 3 in correct order.
- Superfecta: Predict the top 4 in correct order

These outcomes are completely determined by all their individual finishing times. We further assume that the finishing times follow a multivariate Gaussian distribution and consider the set up where we observe only one k–dimensional observation $\mathbf{X}$ which is generated from the normal density with mean at $\boldsymbol{\theta}$ and covariance Σ_p. Based on $\mathbf{X}$, we would like to forecast the probability of the outcomes for future sample $\mathbf{Y}$ which is also generated from the Gaussian distribution with same mean. Thus,

$$\mathbf{X} \sim N(\boldsymbol{\theta}, \Sigma_p) \text{ and } \mathbf{Y} \sim N(\boldsymbol{\theta}, \Sigma_f).$$

In practice Σ_p and Σ_f are unknown and need to be estimated. However, as multi-sample Gaussian models are reduced to a one-sample framework through the sufficient statistics, we generally have $\Sigma_f = r\Sigma_p$ where r is known. Here, we will also assume that Σ_p is known. Maximizing the long term growth rate would imply maximizing the individual growth rates for each of these games. Now, if we use the predictive density estimate $\widehat{p}$ to forecast the future probabilities then for the fair odds case by Equation 1.4 we have the growth rate as,

$$\mathbf{Gw}(\boldsymbol{\theta}, \widehat{p}) = \rho_D(\boldsymbol{\theta}, q) - \rho_D(\boldsymbol{\theta}, \widehat{p})$$

where ρ_D is the corresponding discrete predictive entropy risk. Thus, the maximum growth rate over the class of $\mathcal{C}$ predictive strategies is given by

$$\mathbf{Gw}(\boldsymbol{\theta}) = \rho_D(\boldsymbol{\theta}, q) - \min_{\widehat{p} \in \mathcal{C}} \rho_D(\boldsymbol{\theta}, \widehat{p}).$$

We are interested in finding lower bounds on Growth rate as the unknown parameter $\boldsymbol{\theta}$ varies in the k–dimensional parametric spaces. For example, the worst case rate

$$\min_{\boldsymbol{\theta} \in \Theta} \mathbf{Gw}(\boldsymbol{\theta}) \geq \min_{\boldsymbol{\theta} \in \Theta} \rho_D(\boldsymbol{\theta}, q) - \max_{\boldsymbol{\theta} \in \Theta} \min_{\widehat{p} \in \mathcal{C}} \rho_D\big(\boldsymbol{\theta}, \widehat{p}\big).$$

Similarly, guarantees on other attributes of the growth rate will be related with the maximin value of the predictive risk – the maximum being on the set of possible values that the unknown parameter can take and the minimum is over the class of predictive strategies that we use.

Overlapping Events: Here, we discuss the scenario where the events are not mutually exclusive. In case of horse betting there is a hierarchy among the Superfecta, Trifecta, Exacta and Win bets. Based on a correct Superfecta bet, the successful Trifecta, Exacta and Win bets are completely determined and similar properties hold as we move higher up in the tree based on those events. Now, the wealth of the gambler at the end of a horse race is still given by Equation 1.3. However, as the events now can be overlapping the long-term or the expected growth rate is no longer the same,

i.e.

$$W(b,p) = \mathbb{E}(\log S(\mathbf{X})) \neq \sum_{i=1}^{m} p_i \log(b_i o_i).$$

A lower bound on this growth rate can be derived based on the fact that $\log S(\mathbf{X}) \geq \sum_i \log(b_i o_i)\, I(J_i)$ and so

$$W(b,p) \geq \sum_{i=1}^{m} p_i \log(p_i o_i) - \sum_{i=1}^{m} p_i \log(p_i/b_i).$$

With the odds being prefixed, maximizing the R.H.S. above over the class of betting strategies with $\mathbf{b} \geq 0$ and $\mathbf{1'b} = 1$ will yield a lower bound on the maximum growth rate

$$\mathbf{Gw}(p_i, o_i) \geq \sum_{i=1}^{m} p_i \log(p_i o_i) - \min_{\mathbf{b}\geq 0,\, \mathbf{1'b}=1} \sum_{i=1}^{m} p_i \log(p_i/b_i)$$

If the bets of these overlapping events are set based on their corresponding probability with respect to some distribution Q then we have the following lower bound on the growth rate $\mathbf{Gw}(p_i, o_i)$:

$$\sum_{i=1}^{m} p_i \log\big(p_i o_i\big) - \sum_{i=1}^{m} P(J_i) \log\Big(P(J_i)/Q(J_i)\Big) - \Big(\sum_{i=1}^{m} P(J_i)\Big) \log\Big(\sum_{i=1}^{m} Q(J_i)\Big).$$

Now, if the betting events are such that the maximum number of overlaps in that set is upper bounded by c (say) and both the true probability distribution P and the one used for betting Q are lebesgue measurable with respective densities being p and q, then by the following Lemma 1.3.2 we have,

$$\mathbf{Gw}(p_i, o_i) \geq \sum_{i=1}^{m} p_i \log(p_i o_i) - c\, D(p||q) - \log(c) \sum_{i=1}^{m} P(J_i)$$

where $D(p||q) = \int p(x) \log\{p(x)/q(x)\}\, dx$ is the differential relative entropy between P and Q. Thus, in the set-up of prediction games, where based on side-information

we use predictive densities to construct our bets, a lower bound on the worst-case long-term growth rate is given by

$$\mathbf{Gw}(p_i, o_i) \geq \sum_{i=1}^{m} p_i \log(p_i o_i) - c \min_{\widehat{p}} \rho\big(p, \widehat{p}\big) - \log(c) \sum_{i=1}^{m} P(J_i)$$

and similarly as in the non-overlapping case, guarantees on the Growth rate will involve the maximin value of the predictive entropy risk.

Now, for the proof of Lemma 1.3.2 consider the set-up with a countable collection of measurable sets $\mathcal{A} = \{A_i : i = 1, \cdots, m\}$ with $m \leq \infty$ in $\mathbb{R}^k$. The collection is exhaustive if $\cup_{i=1}^{m} A_i = \mathbb{R}^k$. We can construct a mutually disjoint partition $\mathcal{B} = \{B_i : 1 \leq i \leq 2^m\}$ of the collection $\mathcal{A}$ where $B_i = \cap_{j=1}^{m} A_j^{w[i,j]}$ where $w[i,j]$ is the j^{th} term in the binary expansion of i and for any set $A^0 = A^c$ and $A^1 = A$. We do not track null B_i in $\mathcal{B}$ and would ignore them through out. Let $\kappa(B_i)$ denotes the number of repetitions of the subset B_i in the collection $\mathcal{A}$ i.e $\kappa(B_i) = \text{card}\{j : B_i \cap A_j \neq \phi \text{ and } j = 1, \cdots, m\}$. Note that $\kappa(B_i) \in [1, m]$ and under finite overlaps we can assume that $\sup_{i=1}^{2^m} \kappa(B_i) = c < \infty$ and we define a weight function on $\mathbb{R}^k$ as $w(x) = \sum_{i=1}^{2^m} c^{-1}\, \kappa(B_i)\, \mathbb{I}_{B_i}(x)$. Note that, $w(x) \in (0, 1]$ acts as a tilt function for the densities $p(x)$ and $q(x)$. Under this set-up, we have the following lemma which provides a lower bound on the growth rate.

Lemma 1.3.2.

If the probability measure P and Q have densities p and q with respect to Lebesgue measure, then for any countable collection of exhaustive measurable sets $\mathcal{A}$ we have,

$$\sum_{i=1}^{m} P(A_i) \log\big\{P(A_i)/Q(A_i)\big\} \leq c\, D(p||q)$$

where $D(p||q) = \int p(x) \log\{p(x)/q(x)\}\, dx$ is the differential relative entropy between P and Q.

Proof. If the collection consists of mutually disjoint sets then the proof follows from

the data processing inequalities associated with quantization idea in information theory. The function $t \log t$ is strictly convex if $t > 0$. So for any positive random variable T and any sigma-finite measure, by Jensen's inequality we have, $E_\mu(T \log T) \geq E_\mu(T) \log E_\mu(T)$. For any measurable set A, with $T(x) = p(x)/q(x)$ and measure $\mu(x) = q(x)/Q(A)\, dx$ we have $P(A) \log P(A)/Q(A) \leq \int_A p(x) \log\{p(x)/q(x)\}\, dx$ and so the proof extends to mutually exclusive cases.

If the events are not mutually disjoints then we can construct its mutually disjoint partition $\mathcal{B} = \{B_i : 1 \leq i \leq 2^m\}$ as above and using the Log-Sum inequality (Cover & Thomas 1991, Theorem 2.7.1) separately on each A_i we have,

$$\sum_{i=1}^{m} P(A_i) \log\{P(A_i)/S(A_i)\} \leq \sum_{i=1}^{2^m} \kappa(B_i)\, P(B_i) \log\{P(B_i)/Q(B_i)\}$$

and again using the above quantization argument we can show that the R.H.S. above is less than $c \int w(x)p(x) \log\{p(x)/q(x)\}\, dx$ which we denote as $c\, D(w.p||w.q)$. Now, observe that

$$\begin{aligned} D(w.p||w.q) - D(p||q) &= \int \big(1 - w(x)\big)\, p(x) \log\big\{q(x)/p(x)\big\}\, dx \\ &\leq \log\left[\int \big(1 - w(x)\big)\, q(x)\, dx\right] \end{aligned}$$

by Jensen's inequality and the result follows as $\int \big(1 - w(x)\big)\, q(x)\, dx \leq 1$. □

If the collection is not exhaustive we can restrict our densities to the corresponding subsets of $\mathbb{R}^k$. The above calculation can be easily generalized to incorporate over other kinds of bets like show and place bets.

Sub-fair odds: Generally, the betting portals and the organizers take a cut of all the bets. Under this circumstances we have $\sum_{i=1}^{m} o_i^{-1} > 1$ and proportional gambling is sub-optimal. Different modifications of the proportional gambling schemes are used in practice. The partial-Kelly strategies (Thorp 2000) bet only a fraction of the money and leave the rest as cash. The Constant Relative Risk Aversion Strategies

(Kadane 2011) are based on stable utility functions such as

$$U(f) = \frac{1 - f^{1-\beta}}{\beta - 1} \text{ for } \beta > 0.$$

Studying the efficiency of these strategies would involve generalizing the entropy loss function over the class of α-divergences and more generally over the class of f-divergences which is also known as Csiszar-divergences, Csiszar-Morimoto divergences or Ali-Silvey distances (Cichocki & Amari 2010). Let p and q be two lebesgue measurable densities then for a convex function f such that $f(1) = 0$, the f-divergence of q from p is:

$$D_f(p||q) = \int f\big(t(x)\big)q(x)\,dx \quad \text{where } t(x) = p(x)/q(x).$$

Many common divergences, such as KL-divergence, Hellinger distance, and total variation distance, are special cases of f-divergence, coinciding with a particular choice of f. Table 1.2 lists many of the common divergences between probability distributions and the f function to which they correspond. In the class of α-divergences with $\alpha = 0$ we have 2 times the Hellinger distance and our predictive risk corresponds to $\alpha = 1$. In this thesis we restrict ourselves to the entropy loss. However, the proof techniques used can be extended to other loss functions. Particularly, the results in Chapter 3 can be extended to the class of α–divergences and phenomena described in Chapter 4 seem to hold for any unbounded α–divergences.

1.3.2 Role of Shrinkage

Above, in the context of sports betting we saw that it is essential to minimize the predictive risk over the different choices of predictive density estimates. In the cases, where no prior information about the multivariate parameter $\boldsymbol{\theta}$ is available, we want our density estimate $\widehat{p}$ to control $\rho(\boldsymbol{\theta}, \widehat{p})$ better than other estimates for all $\boldsymbol{\theta} \in \mathbb{R}^k$. As such we want to use admissible density estimates $\widehat{p}$ for which there does not exist

Distribution		$f(t)$
χ^2 –Distance		$t^2 - 1$
Total variation Distance		$\big\lvert t-1\big\rvert$
α–Divergence	$\alpha \in (-1,1)$ $\alpha = -1$ $\alpha = 1$	$4(1-\alpha^2)^{-1}\big[1 - t^{(1+\alpha)/2}\big]$ $-\log t$ $t \log t$

Table 1.2: The class of f–divergences.

any other estimates $\widehat{q}$ such that

$$\rho(\boldsymbol{\theta}, \widehat{q}) \leq \rho(\boldsymbol{\theta}, \widehat{p}) \text{ for all } \boldsymbol{\theta} \in \mathbb{R}^k$$

and strict inequality for at least one parameter value. Such an estimator will also minimize the maximum predictive risk over the entire parametric space.

The notion of shrinkage is helpful in constructing such admissible decision rules. We explain it in the simple orthogonal Gaussian Regression model **M.2**. The best invariant density estimate here is the multivariate product Gaussian density centered around $\mathbf{X}$ and with variance $(\sigma_f^2 + \sigma_p^2) \cdot I$. As we will be seeing in Chapter 2, the best invariant density is the Bayes predictive density from the uniform prior and is denoted by $\widehat{p}_U$. It has constant predictive risk

$$\rho(\boldsymbol{\theta}, \widehat{p}_U) = \frac{k}{2} \log(1 + r^{-1}) \text{ for all } \boldsymbol{\theta} \in \mathbb{R}^k.$$

Density estimates dominating $\widehat{p}_U$ can be constructed even within the Gaussian family by moving either to a better location choice $\widehat{\boldsymbol{\theta}}$ or to a better estimate of scale $\widehat{s}$ or both. The resultant estimates are of the form $N(\widehat{\boldsymbol{\theta}}, \widehat{s} \cdot I)$. Shrinking the canonical

location estimate $\mathbf{X}$ towards the overall mean provides good choices of $\widehat{\boldsymbol{\theta}}$ and using proper scale estimates $\widehat{s}$ which are lower than $\sigma_f^2+\sigma_p^2$ produces data-adaptive flattened density estimates with better risk properties. These ideas can be extended to address admissibility of non-gaussian estimates and also to other parametric model.

1.3.3 Prediction along a curve

Suppose we observe a series of observations $\mathbf{X} = \{X(t_i) : 1 \leq i \leq m_1\}$ each of which is a noisy evaluation of an unknown function f (with domain being a subset of $\mathbb{R}$) at that point. The entire function and its domain remain invariant in the future and noisy realizations $\mathbf{Y}$ sampled at equispaced $\{s_j : 1 \leq j \leq m_2\}$ points will be observed in the future. A point to note that the future and past sampling intervals are both equispaced but can be different. Many objects of interest depend on prediction of the behavior of the future sample $\mathbf{Y}$ and can be addressed by estimating a simultaneous predictive density of the future vector $\mathbf{Y}$ at the time points $\{s_j : 1 \leq j \leq m_2\}$ based on the past vector $\mathbf{X}$.

It follows from discussions later in Section 1.3.5, that in following set-up

PAST OBSERVATION: $X(t_i) = f(t_i) + \sigma\,\epsilon_{1,i},\ i = 1, \ldots, m_1$

FUTURE OBSERVATION: $Y(s_j) = f(s_j) + \sigma\,\epsilon_{2,j},\ j = 1, \ldots, m_2.$

the simultaneous predictive entropy risk of the predictive density estimate $\widehat{g}(\mathbf{Y}|\mathbf{X})$

$$R_{(m_1,m_2)}(f, \widehat{g}) = \frac{1}{m_2} \sum_{l=1}^{m_2} \mathbb{E}_{Y,X|f} \left(\log \frac{p\big(Y(s_l)\big|f(s_l)\big)}{\widehat{g}\big(Y(s_l)|X, Y[s_l-]\big)} \right)$$

is related to maximizing the growth rate in along this stationary curve. Here, $Y[s_l-]$ denotes the vector of future observations before the point s_l. As both m_1 and m_2 tend to infinity in a way such that $m_1/m_2 = r \in (0, \infty)$, the predictive risk converges to the integrated likelihood ratio over the domain of the function

$$\lim_{(m_1,m_2)\to\infty} R_{(m_1,m_2)}(f, \widehat{g}) \to \int_0^1 \mathbb{E}_{\mathbf{X},\mathbf{Y}|f} \log \big\{ p(\mathbf{Y}|f(u))/\widehat{g}(\mathbf{Y}|\mathbf{X}) \big\}\, du.$$

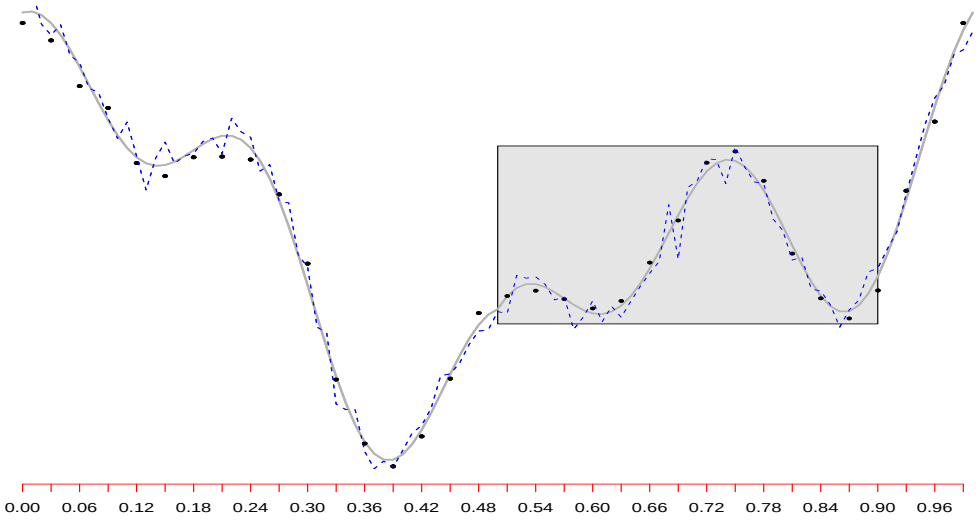

Figure 1.1: Schematic diagram showing the prediction problem along a unknown stationary curve which is represented by the smooth gray line. The black dots represent noisy observations sampled from the function at equispaced intervals and the dotted blue line denotes a future realization from the curve. An object of interest can be the event that the future blue line lies entire in the gray box. We usually need simultaneous prediction of several such events.

This set up, demonstrated pictorially in Figure 1.1, also corresponds to predictive density estimation problems in Non-parametric Regression models.

1.3.4 Role of Sparsity

Usually, the unknown function f of the curve prediction problem lies in a smooth class of functions. Appropriate orthonormal transformations can translate the smoothness of these function classes to sparsity of the basis coefficients. Suppose, for simplicity that m_2 is an integral multiple of m_1 and there exist suitable orthonormal transformation $\{\phi_i : 1 \leq i \leq m_2\}$ such that corresponding matrices Φ_A and Φ_B transform

the prediction problem into

$$\boldsymbol{U} = m_1^{-1}\Phi_A^t\mathbf{X} \text{ and } \boldsymbol{V} = m_2^{-1}\Phi_B^t\mathbf{Y}.$$

The smoothness of the function class can be characterized by sparsity assumptions on the transformed basis coefficients

$$\theta_i = \int_0^1 f(t)\,\phi_i(t)\,dt,\ i = 1, \ldots, m_2.$$

Under this transformation the curve prediction problem reduces to our orthogonal Gaussian regression model **M.2** where the transformed

$$\text{Past:}\quad U|\boldsymbol{\theta} \sim N\big(\tilde{\boldsymbol{\theta}}, m_1^{-1}\boldsymbol{I}\big) \quad \text{and the future} \quad V|\boldsymbol{\theta} \sim N\big(\boldsymbol{\theta}, m_2^{-1}\boldsymbol{I}\big) \text{ and } r = m_1/m_2$$

When $m_1 < m_2$, $\tilde{\boldsymbol{\theta}}$ is a sub-series of the m_2 dimensional parameter $\boldsymbol{\theta}$ indexed at the corresponding m_1 points. As m_2 becomes large, depending on the smoothness of the class of function, the high-dimensional parameter $\boldsymbol{\theta}$ is generally constrained to smaller subsets Θ of $\mathbb{R}^{m_2}$. We can incorporate these restrictions on the parameter space by using function spaces such as Besov and weak ℓ_p balls. Weak sparsity based on ℓ_p balls of varying moments (denoted by the shape parameter p) model the decay of the basis coefficients through power law by restricting the l^{th} largest coefficient to be lower than the radius C

$$|\theta|_{(l)} \le Cl^{-1/p},\ l = 1, 2, \ldots, m_2.$$

Studying the worst-case predictive risk over weak ℓ_p balls can be reduced to evaluating the minimax risk in Model **M.2** (also known as Gaussian sequence model) as the unknown parameter space Θ varies over a wide class of ℓ_p balls with shape parameter p and normalized mean radius $\eta_{m_2,p}$ both varying in $(0, \infty)$:

$$\ell_p \text{ balls: } \Theta_{m_2,p}(C_{m_2}) = \left\{ \sum_{i=1}^{m_2} |\theta_i|^p \le C_{m_2} \right\} \text{ with mean radius } \eta_{m_2,p} = \sigma^{-p}C_{m_2}/m_2.$$

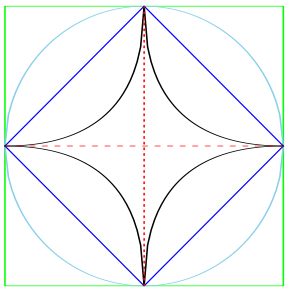

Figure 1.2: Schematic diagram representing the different shapes of ℓ_p sparsity. In green, skyblue, blue and black we have 2 dimensional ℓ_p spaces with moment $p = \infty, 2, 1$ and 0.5 respectively. The dotted red line shows exact (ℓ_0) sparsity.

We can also impose exact or ℓ_0 sparsity on our parameter space which upper bounds the number of non-zero co-ordinates of the parameter $\boldsymbol{\theta}$. As such,

$$\ell_0 \text{ sparsity: } \Theta_0(m_2, s) = \left\{ \boldsymbol{\theta} \in \mathbb{R}^{m_2} : \sum_{i=1}^{m_2} \mathbb{I}[\theta_i \neq 0] \leq s \right\}$$

can be used for sparse coding and for prediction in sparse networks. Figure 1.2 shows the different notions of ℓ_p sparsity. As such the worst-case or the minimax predictive risk for the smooth stationary curve prediction problem is equivalent to the minimax predictive density estimation in **M.2** under the varied ℓ_p sparsity restrictions on the transformed parameter space Θ.

Lemma 1.3.3.

For a smooth function class $\mathcal{F}$ with any $p \in (0, \infty)$ and fixed ratio $m1/m2 = r \in (0, 1)$ and with $\eta_{m_2,p} \to \eta$ as $m_2 \to \infty$ we have,

$$\min_{\widehat{g}} \max_{f \in \mathcal{F}} R_{(m_1,m_2)}(f, \widehat{g}) \sim \left(\min_{\widehat{p}} \max_{\boldsymbol{\theta} \in \Theta_{m_2,p}(\eta)} \rho(\boldsymbol{\theta}, \widehat{p}) \right) (1 + \mathrm{o}(1)) \text{ as } m_2 \to \infty.$$

Proof. The proof follows from minor modification of the proof of Theorem 2.1 in Xu & Liang (2010). □

1.3.5 Growth rate in stock investments

Consider a daily trading scenario where we trade at n–predetermined time points $0 \leq t_1 < t_2 \cdots < t_n$ in the day. The objective is to maximize the cumulative return at the end of each successive day. Let us consider the simplest case where we trade only on financial instruments associated with a single stock. Let the variable X_i denote the differential of the logarithm of the stock price S_i at time t_i. So, $X_i = \nabla \log S(t_i) = \log S_i/S_{i-1}$ for $i = 0 \cdots, n$. Note, that the profit from any concerned financial instrument based on that stock which expires or is exercised time t_i is a function of $f(y_i : 0 \leq i \leq n)$. We model these stock prices by linear combinations of past fluctuations of other related instruments. Let a_i be the fluctuations of the other related instruments in appropriate scale (log transformed in case of stock prices) and a_i is adaptive to the filtration $\mathcal{F}_i$ i.e. there are measurable functions with respect to the filtration produced by market knowledge before t_i. We model $x_i = a_i^T\,\theta + \epsilon_i$ where ϵ_i is white noise, θ reflects the market structure based on partial cross dependence and a_i may include lagged stock prices of related assets. Assuming that the market structure $\boldsymbol{\theta}$ is invariant in short term, the problem becomes that we observe $\mathbf{X} = A\boldsymbol{\theta} + \epsilon$ and the future price $\mathbf{Y} = B\boldsymbol{\theta} + \epsilon$ where B are known regressors.

Some examples of different forms of f are:

- Pay-off for a Call Option with strike price K bought at cost C and exercised after time t is $\max\{S_T - K, 0\}$ and hence the return is $C^{-1}e^{-rt}\max\{S_T - K, 0\}$. And, for a put option bought at price P with strike price K is $P^{-1}e^{-rt}\max\{K - S_T, 0\}$ where r is the risk-free rate of interest.

- There are different kind of exotic options which may depend on the entire path of the stock price like Bermuda options.

So, if l_i strategies executed at time i produces $f[i, l](\mathcal{F}_i)$ return then the return at the end of the day will be

$$\prod_{i=i}^{n}\prod_{l=1}^{l_i} f[i, l](e^{\mathbf{Y}}) = \exp\{W(\boldsymbol{\theta})\}(1 + o(1))$$

where $W(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\big[\sum_{i=i}^{n}\sum_{l=1}^{l_i}\log f[i,l](\mathbf{Y})\big]$ where the expectation is over the true density of $\mathbf{Y}$. Maximizing the growth rate would involve maximizing $W(\theta)$ which can be done by using Monte Carlo simulations. However, the true density of $\mathbf{Y}$ is unknown as the parameter $\boldsymbol{\theta}$ is unknown. So instead of the true density we can use predictive density estimate $\widehat{p}(\mathbf{Y}|\mathbf{X})$ to optimize the growth rate $W(\widehat{p})$ over the set of actions. The following lemma shows that difference between this pseudo growth rate $W(\widehat{p})$ and the true growth rate $W(\boldsymbol{\theta})$ is less than the predictive entropy risk. Thus, evaluation of the minimax predictive risk will provide guarantees on the most conservative growth rate achieved by actions based on predictive density estimates.

Lemma 1.3.4.

For any $\boldsymbol{\theta}$ and any predictive density estimate $\widehat{p}(\mathbf{Y}|\mathbf{X})$ we have

$$W(\boldsymbol{\theta}) - W(\widehat{p}) \leq \rho(\boldsymbol{\theta}, \widehat{p}).$$

Proof. The result follows by adapting the proof of Theorem 15.4.1 in Cover & Thomas (1991) for the predictive regime. □

Reduction: If A has full rank then the above model (based on **M.2**) can be further reduced to

$$\begin{array}{ll} \mathbf{X} &= A\boldsymbol{\theta} + \epsilon_1 \\ \mathbf{Y} &= B\boldsymbol{\theta} + \epsilon_2 \end{array} \quad \Leftrightarrow \quad \begin{array}{ll} C\mathbf{X} &= B\boldsymbol{\theta} + E_1 \\ \mathbf{Y} &= B\boldsymbol{\theta} + \epsilon_2 \end{array}$$

where $C = B(A'A)^{-1}A'$ and $E_1 = N(0, \Sigma)$ with $\Sigma = C'C$. Usually, $\boldsymbol{\theta}$ is high-dimensional and it will be beneficial to incorporate the notions of shrinkage in these models. Time variability can also be incorporated here through the location structure $\boldsymbol{\theta}$. Sequential probability models $\boldsymbol{\theta}_{t+1} \stackrel{d}{=} p\,\boldsymbol{\theta}_t + (1-p)\,F$ on the location structure $\boldsymbol{\theta}_t$ at time t with fixed but unknown probability $p \in (0,1)$ and known cdf F, correspond to rapidly trend changing environments. In such a model, augmenting change point analysis to the decision theoretic behavior of the relative entropy risk will yield sequential log-optimal investment strategies with side information.

1.4 Our contributions and layout of the thesis

Recently, decision theoretic parallels have been found between point estimation theory and the predictive density estimation problem. Table 1.3 summaries the connections across these two estimation theory regimes. Here, we build on these parallels and provide new insights on the roles of sparsity and shrinkage in the predictive regime. We develop the new notion of risk diversification to construct minimax optimal, sparse density estimates. This new concept of risk diversification is also related to the nature of optimal shrinkage in unrestricted parameter spaces.

Decision Theoretic Issues	Point Estimation	Predictive Density
Admissibility in Unrestricted space		
Shrinkage priors	Stein (1974) Strawderman (1971)	Komaki (2001) George, Liang & Xu (2006)
Complete Class & Bayes Rules	Brown & Hwang (1982)	Brown, George & Xu (2008)
Minimaxity over & Restricted space		
Ellipsoids	Pinsker (1980)	Xu & Liang (2010)
Sparsity Constraints	Donoho et al. (1992)	Evaluated here

Table 1.3: Parallels between Point Estimation of the Normal mean under quadratic loss and Predictive Density Estimation under KL loss

We then combine our information theoretic optimality results of the predictive risk with the rich theory around log-utility maximization (Kelly 1956, Breiman 1961) to produce log-optimal probability forecasting strategies for a wide range of prediction problems. In spite of some its drawbacks (Samuelson 1971, Samuelson 1979) the log-utility maximization philosophy is extremely popular and Kelly criterion based

methods have become a part of the mainstream investment theory (Poundstone 2006, MacLean, Thorp & Ziemba 2011). This thesis provides some theoretical support to many of these methods. Figure 1.3 displays the main ideas contained in the thesis. Next, in **Chapter 2** we introduce the different classes of predictive densities and elab-

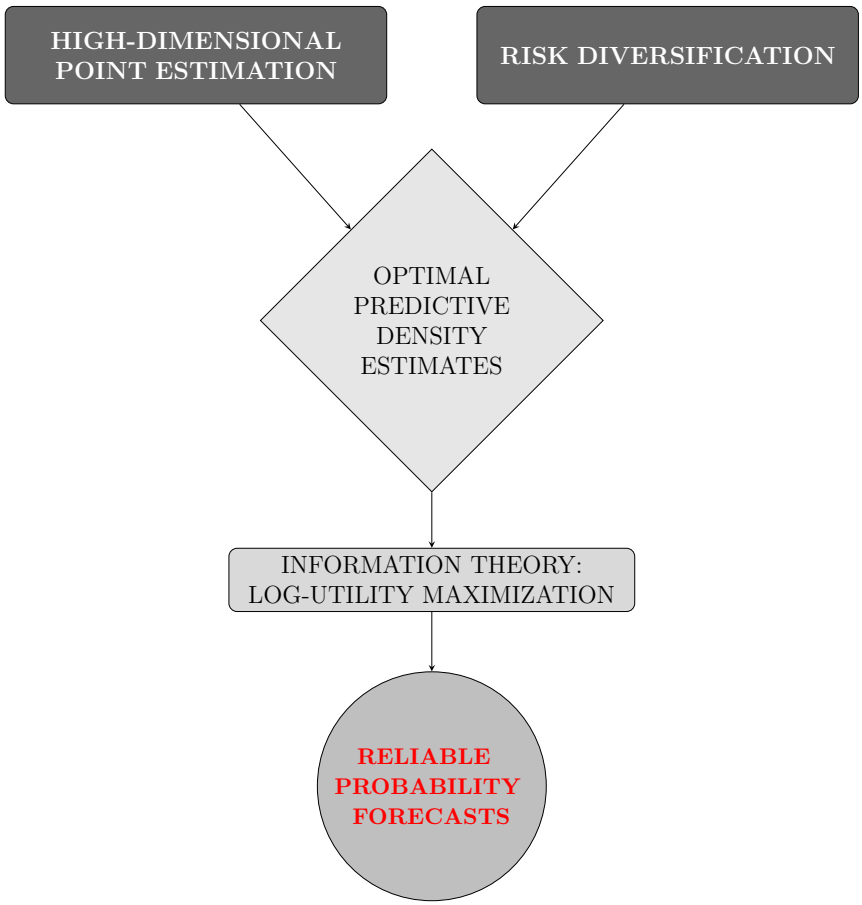

Figure 1.3: Diagrammatic representation of the main ideas contained in the thesis.

orate on some of the basic calculations with the predictive entropy loss. In **Chapter 3**, we describe the role of shrinkage in the class of all Gaussian density estimates. Here, we construct data-adaptive, linear predictive density estimates which are admissible in the unrestricted parametric space. **Chapter 4** is about prediction under sparsity restriction. We introduce the role of risk diversification to explain why we need to move outside the Gaussian family to construct minimax optimal estimates.

BIBLIOGRAPHY

Aitchison, J. & Dunsmore, I. R. (1975), *Statistical prediction analysis*, Cambridge University Press, Cambridge.

Aslan, M. (2006), 'Asymptotically minimax Bayes predictive densities', *Ann. Statist.* **34**(6), 2921–2938.

Barron, A. R. & Cover, T. M. (1988), 'A bound on the financial value of information', *IEEE Trans. Inform. Theory* **34**(5, part 1), 1097–1100.

Barron, A., Rissanen, J. & Yu, B. (1998), 'The minimum description length principle in coding and modeling', *IEEE Trans. Inform. Theory* **44**(6), 2743–2760. Information theory: 1948–1998.

Bell, R. M. & Cover, T. M. (1980), 'Competitive optimality of logarithmic investment', *Math. Oper. Res.* **5**(2), 161–166.

Bjørnstad, J. F. (1990), 'Predictive likelihood: a review', *Statist. Sci.* **5**(2), 242–265. With comments and a rejoinder by the author.

Breiman, L. (1961), Optimal Gambling systems for favourable games, *in* 'Fourth Berkeley Symposium on Mathematical Statistics and Probability', Vol. 1, Univ. Calif. Press, Berkeley, CA.

Brown, L. D., George, E. I. & Xu, X. (2008), 'Admissible predictive density estimation', *Ann. Statist.* **36**(3), 1156–1170.

Brown, L. D. & Hwang, J. T. (1982), A unified admissibility proof, *in* 'Statistical decision theory and related topics, III, Vol. 1 (West Lafayette, Ind., 1981)', Academic Press, New York, pp. 205–230.

Buchdahl, J. (2003), *Fixed Odds Sports Betting: Statistical Forecasting and Risk Management*, High Stakes Publishing, London.

Butler, R. W. (1986), 'Predictive likelihood inference with applications', *J. Roy. Statist. Soc. Ser. B* **48**(1), 1–38. With discussion and a reply by the author.

Candès, E. J. (2006), Compressive sampling, *in* 'International Congress of Mathematicians. Vol. III', Eur. Math. Soc., Zürich, pp. 1433–1452.

Cha, S.-H. (2007), ' Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions'.

Cichocki, A. & Amari, S.-i. (2010), 'Families of alpha- beta- and gamma-divergences: flexible and robust measures of similarities', *Entropy* **12**(6), 1532–1568.

Cover, T. M. & Thomas, J. A. (1991), *Elements of information theory*, Wiley-Interscience, New York, NY, USA.

Csiszár, I. (1973), Generalized entropy and quantization problems, *in* 'Transactions of the Sixth Prague Conference on Information Theory, Statistical Decision Functions, Random Processes (Technical Univ. Prague, Prague, 1971; dedicated to the memory of Antonín Špaček)', Academia, Prague, pp. 159–174.

Donoho, D. (2000), 'High-dimensional data analysis: The curses and blessings of dimensionality', *AMS Math Challenges Lecture* pp. 1–32.

Donoho, D. L. & Johnstone, I. M. (1994), 'Minimax risk over l_p-balls for l_q-error', *Probab. Theory Related Fields* **99**(2), 277–303.

Donoho, D. L., Johnstone, I. M., Hoch, J. C. & Stern, A. S. (1992), 'Maximum entropy and the nearly black object', *J. Roy. Statist. Soc. Ser. B* **54**(1), 41–81. With discussion and a reply by the authors.

Efromovich, S. (1999), *Nonparametric curve estimation*, Springer Series in Statistics, Springer-Verlag, New York. Methods, theory, and applications.

Elad, M. (2010), *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, 1st edn, Springer Publishing Company, Incorporated.

Fan, J., Lv, J. & Qi, L. (2011), 'Sparse high dimensional models in economics', *Annual review of economics* **3**, 291.

Ferraty, F. & Vieu, P. (2006), *Nonparametric functional data analysis*, Springer Series in Statistics, Springer, New York. Theory and practice.

Geisser, S. (1971), The inferential use of predictive distributions, *in* 'Foundations of statistical inference (Proc. Sympos., Univ. Waterloo, Ont., 1970)', Holt, Rinehart and Winston of Canada, Toronto, Ont., pp. 456–469. With comments by V. P. Godambe, I. J. Good, W. J. Hall and D. A. Sprott and a reply by the author.

Geisser, S. (1993), *Predictive inference*, Vol. 55 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, New York. An introduction.

George, E. I., Liang, F. & Xu, X. (2006), 'Improved minimax predictive densities under Kullback-Leibler loss', *Ann. Statist.* **34**(1), 78–91.

Ghosh, M., Mergel, V. & Datta, G. S. (2008), 'Estimation, prediction and the Stein phenomenon under divergence loss', *J. Multivariate Anal.* **99**(9), 1941–1961.

Grama, I. & Nussbaum, M. (2002), 'Asymptotic equivalence for nonparametric regression', *Math. Methods Statist.* **11**(1), 1–36.

Guo, D., Shamai, S. & Verdu, S. (2005), 'Mutual information and minimum mean-square error in gaussian channels', *IEEE Trans. Inform. Theory* **51**, 1261–1282.

Hinkley, D. (1979), 'Predictive likelihood', *Ann. Statist.* **7**(4), 718–728.

Kadane, J. B. (2011), 'Partial-kelly strategies and expected utility: Small-edge asymptotics', *Decision Analysis* **8**(1), 4–9.

Kelly, J. L. (1956), 'A new interpretation of information rate'.

Komaki, F. (2001), 'A shrinkage predictive distribution for multivariate normal observables', *Biometrika* **88**(3), 859–864.

Kullback, S. & Leibler, R. A. (1951), 'On information and sufficiency', *Ann. Math. Statistics* **22**, 79–86.

Lai, T. L., Gross, S. T. & Shen, D. B. (2011), 'Evaluating probability forecasts', *Ann. Statist.* **39**(5), 2356–2382.

Larimore, W. E. (1983), 'Predictive inference, sufficiency, entropy and an asymptotic likelihood principle', *Biometrika* **70**(1), 175–181.

Lauritzen, S. L. (1974), 'Sufficiency, prediction and extreme models', *Scand. J. Statist.* **1**(3), 128–134.

Liang, F. (2002), *Exact minimax procedures for predictive density estimation and data compression*, ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Yale University.

Liang, F. & Barron, A. (2004), 'Exact minimax strategies for predictive density estimation, data compression, and model selection', *IEEE Trans. Inform. Theory* **50**(11), 2708–2726.

Liang, F. & Barron, A. (2005), *Exact Minimax Predictive Density Estimation and MDL*, Advances in Minimum Description Length: Theory and Applications (P. Grunwald, I. Myung and M. Pitt eds), MIT Press.

Lichtendahl, K. C. & Winkler, R. L. (2007), 'Probability elicitation, scoring rules, and competition among forecasters', *Management Science* **53**(11), 1745–1755.

MacLean, L. C., Thorp, E. O. & Ziemba, W. T., eds (2011), *The Kelly Capital Growth Investment Criterion: Theory and Practice*, Vol. 3, World Scientific Publishing Co. Pte. Ltd. "http://EconPapers.repec.org/RePEc:wsi:wsbook:7598".

Magee, C. (2011), *Automatic Exchange Betting*, London. "http://www.betwise.co.uk".

Mallat, S. (2009), *A wavelet tour of signal processing*, third edn, Elsevier/Academic Press, Amsterdam. The sparse way, With contributions from Gabriel Peyré.

Maruyama, Y. & Strawderman, W. E. (2012), 'Bayesian predictive densities for linear regression models under -divergence loss: Some results and open problems'.

McMillan, B. (1956), 'Two inequalities implied by unique decipherability', *Information Theory, IRE Transactions on* **2**(4), 115 –116.

Nussbaum, M. (1996), 'Asymptotic equivalence of density estimation and Gaussian white noise', *Ann. Statist.* **24**(6), 2399–2430.

Pinsker, M. S. (1980), 'Optimal filtration of square-integrable signals in Gaussian noise', *Problems in Information Transmission* .

Poundstone, W., ed. (2006), *Fortune's Formula: The Untold Story of the Scientific Betting System That Beat the Casinos and Wall Street*, Hill and Wang; 1st edition.

Rissanen, J. (1984), 'Universal coding, information, prediction, and estimation', *IEEE Trans. Inform. Theory* **30**(4), 629–636.

Samuelson, P. A. (1971), 'The fallacy of maximizing the geometric mean in long sequences of investing or gambling', *Proc Natl Acad Sci USA* **68**(10), 2493–6.

Samuelson, P. A. (1979), 'Why we should not make mean log of wealth big though years to act are long', *Journal of Banking & Finance* **3**(4), 305–307.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis*, Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Smith, R. L. (1999), Bayesian and frequentist approaches to parametric predictive inference, *in* 'Bayesian statistics, 6 (Alcoceber, 1998)', Oxford Univ. Press, New York, pp. 589–612.

Stein, C. (1956), Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *in* 'Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I', University of California Press, Berkeley and Los Angeles, pp. 197–206.

Stein, C. (1974), Estimation of the mean of a multivariate normal distribution, *in* 'Proceedings of the Prague Symposium on Asymptotic Statistics (Charles Univ., Prague, 1973), Vol. II', Charles Univ., Prague, pp. 345–381.

Stone, M. (1974), 'Cross-validatory choice and assessment of statistical predictions', *J. Roy. Statist. Soc. Ser. B* **36**, 111–147. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

Strawderman, W. E. (1971), 'Proper Bayes minimax estimators of the multivariate normal mean', *Ann. Math. Statist.* **42**(1), 385–388.

Thorp, E. O. (2000), 'The kelly criterion in blackjack, sports betting, and the stock market', *Finding the Edge: Mathematical Analysis of Casino Games* pp. 163–213.

Vajda, I. (2002), 'On convergence of information contained in quantized observations', *IEEE Trans. Inform. Theory* **48**(8), 2163–2172.

Xu, J. (2007), A closed form for the harmonic-prior Bayes estimator with associated confidence sets for the mean of a multivariate normal distribution, PhD thesis. `"http://repository.upenn.edu/dissertations/AAI3271836"`.

Xu, X. & Liang, F. (2010), 'Asymptotic minimax risk of predictive density estimation for non-parametric regression', *Bernoulli* **16**(2), 543–560.

Yang, Y. (2007), 'Consistency of cross validation for comparing regression procedures', *Ann. Statist.* **35**(6), 2450–2473.

CHAPTER 2

TYPES OF PREDICTIVE DENSITY ESTIMATES

We build the decision-theoretic foundation for the predictive regime by discussing four natural sub-classes of predictive density estimates and their corresponding risk calculations. We describe how the KL risk of these predictive density estimates can be expressed in terms of the quadratic risk of their corresponding location estimates.

We consider the Gaussian sequence model **M.2** described in Chapter 1 where we observe only one n–dimensional past observation vector $\mathbf{X}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, I)$ and the future observation $\mathbf{Y}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, rI)$. If $\boldsymbol{\theta}$ were known and fixed then $\mathbf{X}$ and $\mathbf{Y}$ would have been independent. Our risk calculations will depend on the future to the past variability r. As r is known, a natural class of density estimates will be densities in the same parametric class as the future density. They are called plug-in density estimates as they are constructed by replacing the unknown location parameter $\boldsymbol{\theta}$ by estimates $\widehat{\theta}(\mathbf{X})$ in the parametric form $\phi(\mathbf{Y}|\,\boldsymbol{\theta}, r)$ of the true density. By $\phi(\cdot|\boldsymbol{\theta}, r)$ we denote multivariate Gaussian density with mean $\boldsymbol{\theta}$ and variance $r \cdot I$. In bold we represent vectors and the dimension of the multivariate Gaussian density is equal to the dimension of $\boldsymbol{\theta}$. Thus, $\phi(\cdot|\,\theta, r)$ will denote a univariate normal with center at θ and variance r.

2.1 The class of Plug-in densities

It is the class of all n–variate Gaussian densities with variance matrix rI and centered around location estimates $\widehat{\boldsymbol{\theta}}$,

$$\mathcal{P} = \Big\{ \phi\big(\ \cdot\ \big|\, \widehat{\boldsymbol{\theta}},\, r\big) \text{ where, } \widehat{\boldsymbol{\theta}} \text{ is any location estimate} \Big\}.$$

Noting that the true and estimated log-likelihoods are

$$\log \phi\big(\mathbf{Y}\big|\, \boldsymbol{\theta}, r\big) = -\frac{n}{2}\log(2\pi r) - \frac{\|\mathbf{Y}-\boldsymbol{\theta}\|^2}{2r}$$
$$\log \phi\big(\mathbf{Y}\big|\, \widehat{\boldsymbol{\theta}}, r\big) = -\frac{n}{2}\log(2\pi r) - \frac{\|\mathbf{Y}-\widehat{\boldsymbol{\theta}}\|^2}{2r}$$

we have the risk of the plug-in estimator $\phi\big(y\big|\, \widehat{\theta}(\mathbf{X}), r\big)$ is given by:

$$\rho\Big(\boldsymbol{\theta},\, \phi\big(\ \cdot\ \big|\, \widehat{\boldsymbol{\theta}}, r\big)\Big) = \mathbb{E}_{\boldsymbol{\theta}}\left[\log\left(\frac{\phi\big(\mathbf{Y}\big|\, \boldsymbol{\theta}, r\big)}{\phi\big(\mathbf{Y}\big|\, \widehat{\boldsymbol{\theta}}, r\big)}\right)\right] = \mathbb{E}_{\boldsymbol{\theta}}\left(-\frac{\|\mathbf{Y}-\boldsymbol{\theta}\|^2}{2r} + \frac{\|\mathbf{Y}-\widehat{\theta}(\mathbf{X})\|^2}{2r}\right)$$

where the expectation is over $\phi\big(\mathbf{X}\big|\, \widehat{\boldsymbol{\theta}}, r\big) \times \phi\big(\mathbf{Y}\big|\, \widehat{\boldsymbol{\theta}}, r\big)$ – the joint density of $(\mathbf{X}, \mathbf{Y})$ under the true location $\boldsymbol{\theta}$. Expanding the second term we get that the risk equals

$$\mathbb{E}_{\boldsymbol{\theta}}\left(-\frac{\|\mathbf{Y}-\boldsymbol{\theta}\|^2}{2r} + \frac{\|\mathbf{Y}-\boldsymbol{\theta}\|^2 + \|\widehat{\theta}\,(\mathbf{X})-\boldsymbol{\theta}\|^2 + 2\,(\mathbf{Y}-\boldsymbol{\theta})'\,(\,\widehat{\theta}\,(\mathbf{X})-\boldsymbol{\theta})}{2r}\right).$$

The cross-product term above vanishes due to time-independence and we have

$$\rho\Big(\boldsymbol{\theta},\, \phi\big(\ \cdot\ \big|\, \widehat{\boldsymbol{\theta}}, r\big)\Big) = \frac{\mathbb{E}_{\boldsymbol{\theta}}\|\widehat{\theta}(\mathbf{X})-\boldsymbol{\theta}\|^2}{2r}.$$

Thus, the plug-in KL predictive risk is the quadratic location risk discounted by the future variability, and so the risk properties of plug-in density estimates follow directly from point estimation theory. Plug-in densities are also called estimative densities and the plug-in predictive risk will sometimes be also denoted by $r_E(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}})$.

Next, we would like to extend the class of density estimates outside the parametric

class of the true future density. We consider the class of Linear predictive densities which are product Gaussian density estimates with varying scale parameters but their location and scale parameters are related.

2.2 The class of Linear predictive densities

This class consists of Bayes density estimates based on conjugate product normal priors $\prod_{i=1}^{n} \phi(\,.\,|0, l_i)$ where $l_i, i = 1, 2, \ldots, n$ are the non-negative prior variances. They are referred to as "Linear" predictive densities because by Diaconis & Ylvisaker (1979) they are analogous to linear diagonal point estimates in the quadratic error setting.

We exhibit the calculation for the univariate case. As the normal prior $N(0, l)$ is conjugate, the posterior is also normal. Setting $\alpha = (1 + l^{-1})^{-1}$, we have

$$\pi(\theta \,|\, x) \propto \phi(x - \theta \,|\, 0, 1) \,\times\, \phi(\theta \,|\, 0, l) \propto \phi(\theta \,|\, \alpha x, \alpha) \sim N(\alpha x,\, \alpha)$$

The corresponding predictive density is a convolution of Gaussians

$$\widehat{p}_l(y|x) = \int \phi(y - \theta \,|\, 0, r)\, \phi(\theta - \alpha x \,|\, 0, \alpha)\, d\theta$$

and so is also Gaussian: $\widehat{p}_l(y|x) \sim N(\alpha\, x,\, r + \alpha)$.
And given the past x, its loss is given by:

$$\begin{aligned}
&\mathrm{L}\big(\theta, \widehat{p}_l\,(\,.\,|\,X = x)\big) = \mathbb{E}_\theta \log \phi(Y - \theta \,|\, 0,\, r) - \mathbb{E}_\theta \log \phi(Y - \alpha x \,|\, 0,\, r + \alpha) \\
&= -\frac{1}{2}\left(\log(2\pi r) + \frac{\mathbb{E}_\theta (Y - \theta)^2}{r} \right) + \frac{1}{2}\left(\log\big(2\pi(r + \alpha)\big) + \frac{\mathbb{E}_\theta\,(Y - \alpha x)^2}{r + \alpha} \right) \\
&= \frac{1}{2} \log\left(1 + \frac{\alpha}{r}\right) + \frac{\mathbb{E}_\theta\,(Y - \alpha x)^2}{2\,(r + \alpha)} - \frac{1}{2}
\end{aligned}$$

and by Bias-Variance decomposition, we have, $\mathbb{E}_\theta\,(Y - \alpha x)^2 = (\theta - \alpha x)^2 + r$.
To evaluate the risk we take expectation over the past X and again by Bias-Variance

decomposition $\mathbb{E}_\theta (Y - \alpha X)^2 = (1-\alpha)^2 \theta^2 + \alpha^2 + r$. Hence the risk is

$$\begin{aligned}\rho\big(\theta, \widehat{p}_l\big) &= \frac{1}{2}\log\left(1+\frac{\alpha}{r}\right) + \frac{(1-\alpha)^2\theta^2 - \alpha(1-\alpha)}{2\,(r+\alpha)} \\ &= \frac{1}{2}\log\left(1+\frac{\alpha}{r}\right) + \frac{(1-\alpha)^2}{2\,(r+\alpha)}\,(\theta^2 - l).\end{aligned}$$

Figure 2.1 shows the predictive risk of different linear density estimates $\widehat{p}_l$ as the

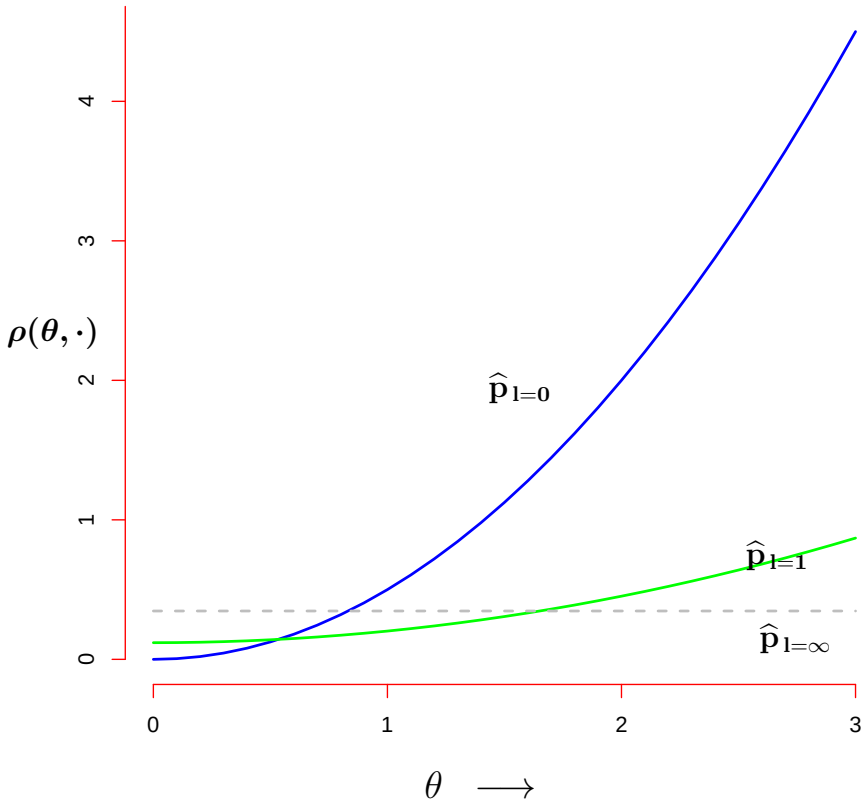

Figure 2.1: The plot depicts the quadratic nature of the risk of linear univariate predictive densities. Here, we have $r = 1$ and the risk $\theta^2/2$ of the zero estimator is plotted in blue. The dotted gray line at $\log 2$ shows the risk of $\widehat{p}_U$ and the green line portrays the risk of the linear estimator with unit prior variance.

parameter varies along the positive orthant. Also, as $l \to \infty, \quad \alpha \to 1$, we get the

uniform prior Bayes Predictive Density ($\widehat{p}_U$). It is the best invariant density as well as minimax (Liang & Barron 2004) in the unrestricted parametric space (will be referred as 'canonical minimax estimator'). Thus:

$$\widehat{p}_U\big(y \,|\, X = x\big) = N(x, 1+r) \quad \text{and} \quad \rho(\theta, \widehat{p}_U) = \frac{1}{2}\log\left(1 + r^{-1}\right).$$

Again, as $l \to 0, \quad \alpha \to 0$ and we get the zero density $\phi(\,\cdot\,|\,0, r)$ with $\theta^2/(2r)$ as its risk.

Returning to the multivariate case, the risk of the class of Linear predictive densities

$$\mathcal{L} = \left\{ \prod_{i=1}^{n} \phi\big(\,.\,|\,\alpha_i\, X[i],\, (\alpha_i + r)\big) : \alpha_i = (1 + l_i^{-1})^{-1},\, l_i \geq 0,\; i = 1, \ldots, n \right\}$$

is quadratic in the parameter $\boldsymbol{\theta}$ and is given by

$$\rho\big(\,\boldsymbol{\theta}, \widehat{p}_{\,L}\,\big) = \frac{1}{2}\sum_{i=1}^{n} \log\left(1 + \frac{\alpha_i}{r}\right) + \sum_{i=1}^{n} \frac{(1-\alpha_i)^2}{2\,(r+\alpha_i)}\,(\theta_i^2 - l_i).$$

Now, we further extend the class $\mathcal{P}$ and $\mathcal{L}$ to the class of all product Gaussian density estimates with equal co-ordinate wise variance.

2.3 The class of Gaussian Predictive densities

The class of of all n–dimensional Gaussian distributions with positive definite (p.d.) covariances is given by

$$\mathcal{G}_n = \Big\{ g : \mathbb{R}^n \to \mathbb{R}^+ \text{ such that } g = N(\mu, \Sigma) \text{ where } \mu \in \mathbb{R}^n \text{ and } \Sigma \text{ p.d.} \Big\}.$$

By $\mathcal{G}_n[p]$ we denote the sub-class of all n–dimensional product Gaussian densities

$$\mathcal{G}_n[p] = \Big\{ g[\mu_n, D_n] : \mu_n \in \mathbb{R}^n \;\&\; D_n \text{ is any } n \times n \text{ p.d. diagonal matrix} \Big\}$$

where $g[\mu_n, D_n]$ is a normal density with mean μ_n and diagonal covariance $\sigma_f^2 D_n$. We consider a further sub-family $\mathcal{G}_n[1]$ of $\mathcal{G}_n[p]$. $\mathcal{G}_n[1]$ contains Gaussian densities with only one data-adaptive scale estimate. So, it includes density estimates of the form $N(\widehat{\theta}\,(\mathbf{X}), \widehat{c}\,(\mathbf{X})\, rI)$ with data adaptive mean $\widehat{\theta}\,(\mathbf{X})$ and variance $r\,\widehat{c}\,(\mathbf{X})I$ which makes it a non-linear estimate. Thus

$$\mathcal{G}_n[1] = \left\{ g\big[\,\widehat{\boldsymbol{\theta}},\, \widehat{c}\,\big] = \phi\big(\,\cdot\,\big|\,\widehat{\boldsymbol{\theta}},\, \widehat{c}\, r\big) \text{ where } \widehat{\boldsymbol{\theta}} \;\&\; \widehat{c} \text{ are any location and scale estimate} \right\}.$$

Note that, $\widehat{\boldsymbol{\theta}}$ is a n–dimensional vector whereas $\widehat{c}$ is a scalar which is used in specifying n–variate product predictive densities. Conditioned on $\mathbf{X} = \mathbf{x}$ the loss is given by,

$$\begin{aligned} \mathrm{L}\Big(\boldsymbol{\theta}, g\big[\,\widehat{\boldsymbol{\theta}},\, \widehat{c}\,\big]\big(\,\cdot\,\big|\mathbf{X} = \mathbf{x}\big)\Big) &= \mathbb{E}_{\boldsymbol{\theta}} \log\left(\frac{\phi\big(\mathbf{Y}\big|\,\boldsymbol{\theta}, r\big)}{\phi\big(\mathbf{Y}\big|\,\widehat{\theta}\,(\mathbf{x}), \widehat{c}\,(\mathbf{x}) r\,\big)}\right) \\ &= \frac{n}{2}\log \widehat{c}\,(\mathbf{x}) - \frac{\mathbb{E}_{\boldsymbol{\theta}}\|\mathbf{Y} - \boldsymbol{\theta}\|^2}{2r} + \frac{\mathbb{E}_{\boldsymbol{\theta}}\|\mathbf{Y} - \widehat{\theta}\,(\mathbf{x})\|^2}{2\,\widehat{c}\,(\mathbf{x})\, r}. \end{aligned}$$

Now, $\mathbb{E}_{\boldsymbol{\theta}}\|\mathbf{Y} - \boldsymbol{\theta}\|^2 = nr$. And, similarly as in the case of linear density estimates, by Bias-variance decomposition $\mathbb{E}_{\boldsymbol{\theta}}\|\mathbf{Y} - \widehat{\theta}\,(\mathbf{x})\|^2 = \|\widehat{\theta}\,(\mathbf{x}) - \boldsymbol{\theta}\|^2 + nr$. So, we have

$$\mathrm{L}\Big(\boldsymbol{\theta}, g\big[\,\widehat{\boldsymbol{\theta}},\, \widehat{c}\,\big]\big(\,\cdot\,\big|\mathbf{X} = \mathbf{x}\big)\Big) = \frac{n}{2}\left[\log \widehat{c}\,(\mathbf{x}) + \frac{1}{\widehat{c}\,(\mathbf{x})} - 1\right] + \frac{\|\widehat{\theta}\,(\mathbf{x}) - \boldsymbol{\theta}\|^2}{2\,\widehat{c}\,(\mathbf{x})\, r}.$$

The KL predictive loss decomposes into two parts. One of them is independent of the parameter and only involves the scale estimate while the other contains the adjusted quadratic loss. For any fixed $\widehat{\theta}\,(\mathbf{x})$ and known $\boldsymbol{\theta}$, the KL loss can be minimized over $\widehat{c}\,(\mathbf{x})$. The minimum value is attained at $c_{\text{opt}} = 1 + (nr)^{-1}\,\|\widehat{\theta}\,(\mathbf{x}) - \boldsymbol{\theta}\|^2$ yielding a loss of $(n \log \widehat{c}_{\text{opt}})/2$. Note that, $c_{\text{opt}} \geq 1$. It signifies that among the Gaussian densities centered at $\widehat{\boldsymbol{\theta}}$ the one closest to $N(\boldsymbol{\theta}, rI)$ in KL loss has variance greater than r and is flattened proportional to the distance between $\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\theta}}$. Figure 2.2 provides a pictorial illustration of this phenomenon.

If $\boldsymbol{\theta}$ is unknown, an estimate $\widehat{c}_{\text{opt}}$ can be constructed based on the quadratic risk estimates of the location estimator $\widehat{\boldsymbol{\theta}}$. We will elaborately describe it in the next chapter.

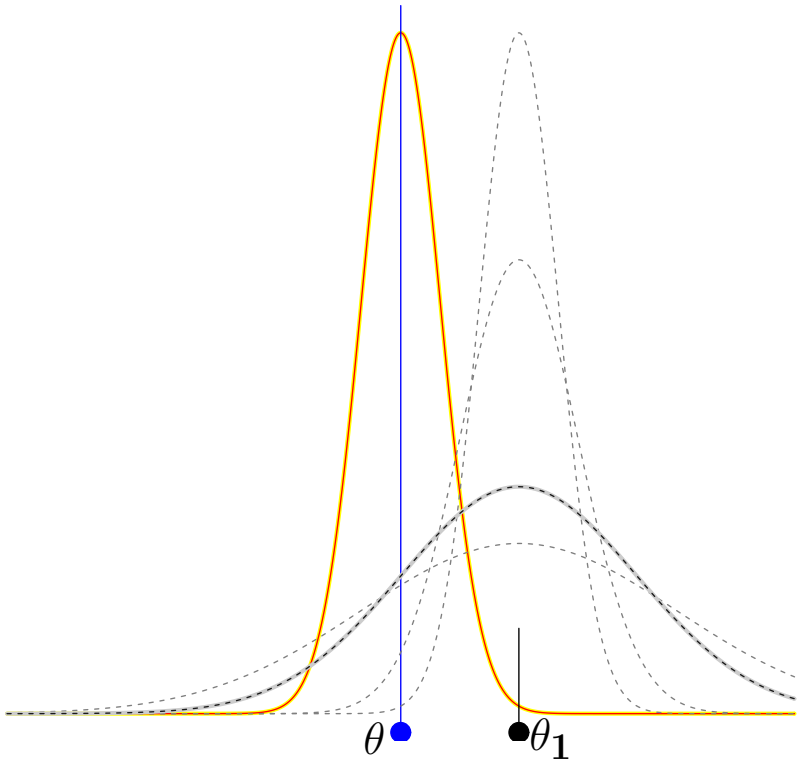

Figure 2.2: Pictorial depiction of the decomposition of the entropy loss. In yellow we represent the true univariate $N(\theta, 1)$ density. In dotted lines are representative densities from $\mathcal{G}$ around a fixed location θ_1. The one in gray is optimally flattened among all Gaussian densities centered at θ_1 and is closest to $N(\theta, 1)$ in terms of the KL loss. Figure drawn to scale with $\theta = 0$, $\theta_1 = 3$, $c_{\text{opt}} = 10$.

Integrating the loss over the past X, we find the risk $\rho\big(\boldsymbol{\theta}, g\big[\,\widehat{\boldsymbol{\theta}},\, \widehat{c}\,\big]\big)$ is given by,

$$\frac{n}{2}\left[\left\{\mathbb{E}_{\boldsymbol{\theta}}(\log \widehat{c}(\mathbf{X})) + \mathbb{E}_{\boldsymbol{\theta}}\left(\frac{1}{\widehat{c}\,(\mathbf{X})}\right) - 1\right\} + \frac{1}{nr}\mathbb{E}_{\boldsymbol{\theta}}\left(\frac{\|\widehat{\theta}\,(\mathbf{X}) - \boldsymbol{\theta}\|^2}{\widehat{c}\,(\mathbf{X})}\right)\right].$$

In chapter 3, we will see that in high dimensions this KL predictive risk can also be decomposed into convenient parts which can be subsequently optimized and we can exactly characterize the asymptotic predictive risk of estimates in $\mathcal{G}$.

2.4 The class of Bayes predictive densities

Next, we have the class of all Bayes predictive densities ($\mathcal{B}$). Given a prior $\pi(\boldsymbol{\theta})$ in $\mathbb{R}^n$ the Bayes predictive density is given by:

$$\widehat{p}_\pi\left(\mathbf{Y}|\mathbf{X}=\mathbf{x}\right) = \int \phi(\mathbf{Y}|\,\boldsymbol{\theta}, r)\,\pi(\boldsymbol{\theta}|\,\mathbf{x})\,d\boldsymbol{\theta} \;\text{ where }\; \pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{\phi(\mathbf{x}|\,\boldsymbol{\theta}, 1)\,\pi(\boldsymbol{\theta})}{m_\pi(\mathbf{x})}$$

is the posterior distribution and the marginal $m_\pi(\mathbf{x}) = \int \phi(\mathbf{x}|\,\boldsymbol{\theta}, 1)\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$.

The integrated Bayes risk of a density estimate $\widehat{p}\big(\mathbf{y}\big|\mathbf{x}\big)$ with respect to a prior π is $B(\pi, \widehat{p}) = \int \rho(\boldsymbol{\theta}, \widehat{p})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$. The Bayes risk $B(\pi)$ of a prior π is $\min_{\widehat{p}} B(\pi, \widehat{p})$ and the minimum is attained by the Bayes predictive density.

As the true density $\phi(\cdot|\boldsymbol{\theta}, r)$ in **M.2** is bounded above by $c_n = (2\pi r)^{-n/2}$, we can restrict the action set $\mathcal{A}_n$ comprising of all densities in $\mathbb{R}^n$ to the set of all c_n bounded densities

$$\mathcal{A}(n, c_n) = \left\{ p : \mathbb{R}^n \to \mathbb{R} \text{ such that } \int_{\mathbb{R}^n} p\,(\mathbf{y})\,d\mathbf{y} = 1 \text{ and } p \in [0, c_n] \right\}.$$

Lemma 2.4.1.

For any $p \in \mathcal{A}_n$ but not in $\mathcal{A}(n, c_n)$ there exists $p_b \in \mathcal{A}(n, c_n)$ that dominates p in the sense $\mathrm{L}(\boldsymbol{\theta}, p_b) \leq \mathrm{L}(\boldsymbol{\theta}, p)$ for all $\boldsymbol{\theta} \in \mathbb{R}^n$.

Details of the proof can be found in Brown et al. (2008, Lemma2). For any positive density $p \in \mathcal{A}_n$ but not in $\mathcal{A}(n, c_n)$ the general idea for constructing a better estimate $p_b \in \mathcal{A}(n, c_n)$ is to truncate p on the set $S_p = \{\mathbf{y} \in \mathbb{R}^n : p(\mathbf{y}) > c_n\}$ and to lift it on S_p^c, i.e.

$$p_b(\mathbf{y}) = \begin{cases} \{\int_{S_p^c} p(\mathbf{y})\,d\mathbf{y}\}^{-1} \cdot \{1 - c_n \mathrm{Vol}(S_p)\} \cdot p(\mathbf{y}) & \text{if } \mathbf{y} \in S_p^c \\ c_n & \text{if } \mathbf{y} \in S_p \end{cases}.$$

As the KL loss for estimators in $\mathcal{A}(n, c_n)$ is always defined (can be infinite though),

they are mathematically more convenient for risk calculations than estimator in $\mathcal{A}_n \setminus \mathcal{A}(n, c_n)$. In the light of Lemma 2.4.1, with out any loss of generality we restrict ourselves to estimators in $\mathcal{A}(n, c_n)$ only. Next we show that the Bayes predictive density defined in equation (4.2) actually minimizes the integrated Bayes risk for any prior π in collection $\mathcal{P}(\mathbb{R}^n)$ of all probability measures on $\mathbb{R}^n$.

Lemma 2.4.2.

For any prior $\pi \in \mathcal{P}(\mathbb{R}^n)$ if $B(\pi, \widehat{p}_\pi) < \infty$ then we have

$$B(\pi, \widehat{p}_\pi) \leq B(\pi, \widehat{p}) \text{ for any } \widehat{p} \in \mathcal{A}_n.$$

Proof. Note that, here we have $\sigma_f^2 = r$ and $\sigma_p^2 = 1$. Now, as the true density $\phi(\cdot|\boldsymbol{\theta}, \sigma_f^2)$ is bounded above, the marginal density $m_\pi(\mathbf{x}) < \infty$ almost surely for all $\mathbf{x} \in \mathbb{R}^n$. Thus, $\widehat{p}_\pi$ is defined almost everywhere and by Lemma 2.4.1, with out loss of generality we can assume that $\widehat{p} \in \mathcal{A}(n, c_n)$. The difference in the integrated Bayes risk between any estimator $\widehat{p}$ and the Bayes estimator is,

$$B(\pi, \widehat{p}) - B(\pi, \widehat{p}_\pi) = \iiint \phi(\mathbf{x}|\boldsymbol{\theta}, \sigma_p^2)\, \phi(\mathbf{y}|\boldsymbol{\theta}, \sigma_f^2)\, \pi(\boldsymbol{\theta}) \log \frac{\widehat{p}_\pi(\mathbf{y}|\mathbf{x})}{\widehat{p}(\mathbf{y}|\mathbf{x})}\, d\boldsymbol{\theta}\, d\mathbf{y}\, d\mathbf{x}$$

as we can interchange the order of integrals by Fubini's theorem. Also, $m_\pi(\mathbf{x})\, \widehat{p}_\pi(\mathbf{y}|\mathbf{x}) = \int \phi(\mathbf{x}|\boldsymbol{\theta}, \sigma_p^2)\, \phi(y|\boldsymbol{\theta}, \sigma_f^2)\, \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, so we have,

$$B(\pi, \widehat{p}) - B(\pi, \widehat{p}_\pi) = \iint m_\pi(\mathbf{x})\, \widehat{p}_\pi(\mathbf{y}|\mathbf{x}) \log \frac{\widehat{p}_\pi(\mathbf{y}|\mathbf{x})}{\widehat{p}(\mathbf{y}|\mathbf{x})}\, d\mathbf{y}\, d\mathbf{x}$$

which is the KL divergence between the densities $m_\pi(\mathbf{x}) \times \widehat{p}_\pi(\mathbf{y}|\mathbf{x})$ and $m_\pi(\mathbf{x}) \times \widehat{p}(\mathbf{y}|\mathbf{x})$ and so it always non-negative. This completes the proof. □

$\mathcal{B}$ is a complete class of procedures i.e. given any density estimate $\widehat{p}$ there exists a Bayes density estimate which is at least as good as $\widehat{p}$ (Brown et al. 2008). Also members in $\mathcal{B}$ which are based on priors with sufficient growth and asymptotic flatness conditions are admissible. $\mathcal{B}$ is also a very wide class of procedures. However, $\mathcal{P}$ is not contained in $\mathcal{B}$. Also, $\mathcal{L}$ consists of Bayes predictive densities with respect to

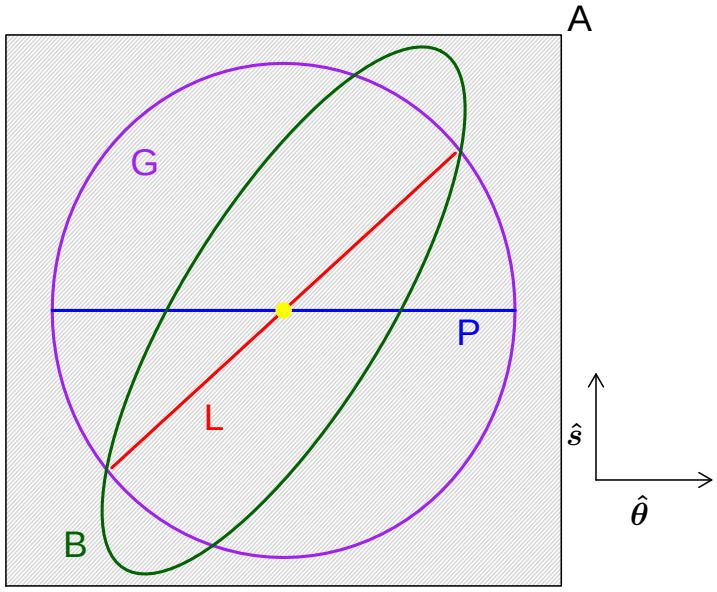

Figure 2.3: Schematic diagram of the different classes of predictive density estimates. The representation is for univariate predictive density estimates with their corresponding location estimates represented along the abscissa and scale estimate along the ordinate. The blue line represents the class of Plugin estimates ($\mathcal{P}$) which have fixed scale. The red lines represent linear estimates ($\mathcal{L}$) where the location and scale estimates are related in a linear fashion by a single parameter. $\mathcal{P}$ and $\mathcal{L}$ intersects at a point which is the zero density estimate. The violet circular body is the class $\mathcal{G}$ which contains both $\mathcal{L}$ and $\mathcal{P}$ and the green ellipsoidal body denotes the class $\mathcal{B}$ whcih contains $\mathcal{L}$ but not $\mathcal{P}$.

normal priors, so $\mathcal{L} \subset \mathcal{B}$ and $\mathcal{L} \cap \mathcal{P}$ is the singleton set $\{\phi(\,\cdot \mid 0, r)\}$. Figure 2.3 shows these classes in the action space $\mathcal{A}$ of all possible densities in $\mathbb{R}^n$. The true future density lies in $\mathcal{P}$ but its location parameter $\boldsymbol{\theta}$ is unknown. In absence of any prior knowledge about $\boldsymbol{\theta}$, we would like to characterize and compare the predictive risk of representative members (efficient in each class in the sense of asymptotic admissibility in that class) from each these four classes as the true density varies over $\mathcal{P}$.

As $\mathcal{B}$ is a very wide class of procedures it is extremely difficult to explicitly quantify the risk for each of its member. However, there exist parallels between the risk calculations in the predictive regime and point estimation theory (George et al. 2006). The risks of Bayes predictive densities will follow from the parallels by using known evaluation of the quadratic risk of the corresponding posterior mean. Next, we describe these connections through a path of experiments.

2.4.1 Relations with Point Estimation: Connecting Equations and path of experiments

The Kullback-Leibler (KL) predictive risk is connected to risk calculations in point estimation (PE) theory via the semi-futuristic random variable $W = v_w\,(X + r^{-1}Y)$ where $v_w = (1 + r^{-1})^{-1}$. W would have been the UMVUE for the unknown location parameter θ if the future Y were also known along with the past X. For simplicity, in this section we restrict ourselves to the univariate version of model **M.2** and the connections described here can be easily extended to multivariate orthogonal Gaussian models.

The connections between the predictive and point estimation theory center around the parallel to the Tweedie's formula (Efron 2011, Brown 1971, Robbins 1956) which gives a closed from expression of the Bayes estimate $\widehat{\theta}_\pi$ corresponding to prior π for the location θ estimation problem under quadratic loss

$$\widehat{\theta}_\pi\big(X\big) = X + \nabla \log m_\pi(X, 1) \tag{2.1}$$

where $m_\pi(Z, v) = \int \phi(Z \,|\, \theta, v)\, \pi(\theta)\, d\theta$ denote the marginal distribution of a Gaussian random variable Z with variance v and with prior distribution π on the location

parameter θ. Bayes predictive densities for KL loss in **M.2** is analogously related to the best invariant density estimate $\widehat{p}_{\mathrm{U}}$:

$$\widehat{p}_{\pi}\big(Y\big|\,X=x\big)=\big\{m_{\pi}\big(W_x,v_w\big)\,m_{\pi}^{-1}\big(x,1\big)\big\}\times\widehat{p}_{\mathrm{U}}\big(Y\big|\,X=x\big) \tag{2.2}$$

where $W_x = v_w\,(x + r^{-1}Y)$. As such, the Bayes risk in these two regimes are also related. By Brown et al. (2008, Theorem 1) the predictive risk of any prior π in the univariate model **M.2** with $\sigma_p^2 = 1$ is given by $m_\pi(z;1) < \infty$ for all $z \in \mathbb{R}$ we have,

$$\rho(\theta,\widehat{p}_{\pi})=\frac{1}{2}\int_{v_w}^{1}v^{-2}\,q(\theta,\widehat{\theta}_{\pi},v)\,dv \tag{2.3}$$

where $q(\theta,\widehat{\theta}_{\pi},v)$ denotes the quadratic risk $\mathbb{E}_{(\theta,v)}\|\,\widehat{\theta}_{\pi}-\theta\|^2$ based on the univariate model **M.2** with $\sigma_p^2 = v$ and $\sigma_f^2 = r\sigma_p^2$.

Through the connecting equations the Kullback-Leibler risk can be viewed as an weighted aggregation of the square error risk. As the variance of Z varies from 1 to $v_w = (1 + r^{-1})^{-1}$ it marks the gradual assimilation of the information in future Y to the existing information about θ in X through a *path of memoryless experiments* conducted separately for $\alpha \in [0, 1]$. It is pictorially demonstrated in Figure 2.4. At each stage α along the path we observe $(X, Y[\alpha]\,)$ where $Y[\alpha]$ is a Gaussian random variable around the true unknown location θ and with variability $r \cdot \alpha^{-2}$. Along the path, the information about the unknown location θ percolates through the sufficient statistics $\{Z[\alpha] : \alpha \in [0, 1]\}$ where $Z[\alpha]$ is the UMVUE, of θ based on observing $(X, Y[\alpha])$. As α increases in $[0, 1]$, v decreases from 1 to v_w and we have,

$$Z[\alpha]=\frac{X[\alpha]+{}^{\alpha^2}\!/_{r}\,Y[\alpha]}{1+{}^{\alpha^2}\!/_{r}}\quad\text{where }\alpha=r^{1/2}(v^{-1}-1)^{1/2}\in[0,1]\text{ and} \tag{2.4}$$

$$\widehat{\theta}_{\pi}^{v}=Z[\alpha]+v\,\nabla\log m_{\pi}(Z[\alpha],v) \tag{2.5}$$

As shown before in this chapter, the plug-in risk $\rho_E^v(\theta,\widehat{\theta})$ equals $q(\theta,\widehat{\theta}_{\pi},v)/(2v)$ and we see that $\rho(\theta,\widehat{p}_{\pi})$ is equal to $\int_{v_w}^{1}v^{-1}\rho_E^v(\theta,\widehat{\theta}_{\pi})\,dv$ which implies that the predictive KL risk is a linearly weighted (according to precision) accumulation of the corresponding plug-in risk. Using the connecting equation 2.3 most of the calculations involving risks

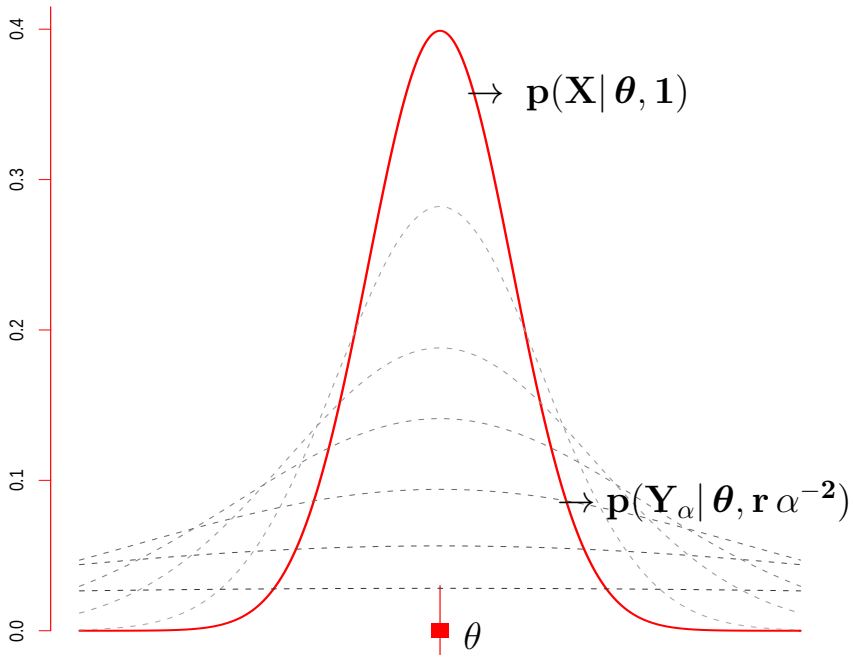

Figure 2.4: *Path of Experiments:-* In red we have the true density of the observed random variable X around the unknown location θ. In different shades of gray (dark to light) we have respectively the density of Y_α for α equal 0.10, 0.20, 0.33, 0.50, 0.67 and 1.00. Hence the corresponding v equal 1.00, 0.98, 0.95, 0.89, 0.82 and 0.67. When, $\alpha = 1$, Y_α corresponds to the true future density around θ with known future variability $r = 2$.

of different Bayes predictive densities would follow easily by using known evaluation of the quadratic risk of the corresponding posterior mean.

Sometimes calculations from definition will cater new knowledge about the predictive regime (eg. Chapter 4, Section 4.4). However these connecting equations along with the path of memoryless experiments are pivotal to our understanding in the predictive regime.

Later on, we will be seeing that some of the popular shrinkage priors are actually scale invariant in the sense that the prior π on the n–dimensional space has the

property:

$$\pi\big(c\cdot\boldsymbol{\theta}\big)\propto\pi(\boldsymbol{\theta})\text{ for all }\boldsymbol{\theta}\in\mathbb{R}^n\text{ and }c\in\mathbb{R}.$$

In point estimation the harmonic prior $\pi_H(\boldsymbol{\theta})\propto||\boldsymbol{\theta}||^{-(n-2)}$ in Stein (1981) is admissible and dominates the best invariant estimator. π_H is spherically symmetric and so possesses the above scale invariance property. By Komaki (2001, Theorem) the Bayes predictive density estimate based on the harmonic prior π_H dominates $\widehat{p}_U$ and is given by:

$$\widehat{p}_H(\mathbf{y}|\mathbf{x})=v_w^{-(n-2)/2}\Gamma_p\big(v_w^{-1/2}\mathbf{w}\big)\big\{\Gamma_p(||\mathbf{x}||)\big\}^{-1}\widehat{p}_U(\mathbf{y}|\mathbf{x})\text{ where }\mathbf{w}=v_w(\mathbf{x}+r^{-1}\mathbf{y}),$$
$$\Gamma_p(u)=u^{-(p-2)}\int_0^{u^2/2}v^{d/2-2}\,\exp(-v)\,dv,\ v_w=(1+r^{-1})^{-1}$$
$$\text{and }\widehat{p}_U(\mathbf{y}|\mathbf{x})=\{2\pi(1+r)\}^{-n/2}\exp\left(-\frac{||\mathbf{y}-\mathbf{x}||^2}{2(1+r)}\right).$$

$\widehat{p}_H$ is a non-linear, non-gaussian, improved minimax estimator. It is admissible (Brown et al. 2008) and so it is important to understand the direction of shrinkage produced by this non-gaussian density estimate. Just by reformatting Equation 2.3, a simplified expression of the predictive risk of a scale invariant prior is attained in the following lemma.

Lemma 2.4.3.

For any π satisfying

$$\pi\big(c\cdot\boldsymbol{\theta}\big)\propto\pi(\boldsymbol{\theta})\text{ for all }\boldsymbol{\theta}\in\mathbb{R}^n\text{ and }c\in\mathbb{R},$$

the predictive risk of the corresponding Bayes predictive density estimate $\widehat{p}_\pi$ in the orthogonal Gaussian predictive model **M.2** with $\sigma_p=1$ and $\sigma_f^2=r$ is given by

$$\rho(\boldsymbol{\theta},\widehat{p}_\pi)=\frac{1}{2}\int_{v_w}^1 v^{-1}\cdot q\big(\boldsymbol{\theta}/\sqrt{v},\,\widehat{\theta}_\pi,1\big)\,dv\text{ where }v_w=(1+r^{-1})^{-1}.$$

Lemma 2.4.3 will be used in Chapter 3 to compare the nature of shrinkage produced by $\widehat{p}_H$ with those produced by of data-adaptive linear predictive densities.

Akin to the connecting equations, the information-theoretic relation between mutual information and minimum mean square error can be explicitly expressed in independent Gaussian channels (Guo et al. 2005). In the succeeding chapters, through the connecting equations, inferences in the predictive regime will be inter weaved with well understood notions from Point Estimation theory.

BIBLIOGRAPHY

Brown, L. D. (1971), 'Admissible estimators, recurrent diffusions, and insoluble boundary value problems', *Ann. Math. Statist.* **42**, 855–903.

Brown, L. D., George, E. I. & Xu, X. (2008), 'Admissible predictive density estimation', *Ann. Statist.* **36**(3), 1156–1170.

Diaconis, P. & Ylvisaker, D. (1979), 'Conjugate priors for exponential families', *Ann. Statist.* **7**(2), 269–281.

Efron, B. (2011), 'Tweedies formula and selection bias', *Journal of the American Statistical Association* **106**(496), 1602–1614.

George, E. I., Liang, F. & Xu, X. (2006), 'Improved minimax predictive densities under Kullback-Leibler loss', *Ann. Statist.* **34**(1), 78–91.

Guo, D., Shamai, S. & Verdu, S. (2005), 'Mutual information and minimum mean-square error in gaussian channels', *IEEE Trans. Inform. Theory* **51**, 1261–1282.

Komaki, F. (2001), 'A shrinkage predictive distribution for multivariate normal observables', *Biometrika* **88**(3), 859–864.

Liang, F. & Barron, A. (2004), 'Exact minimax strategies for predictive density estimation, data compression, and model selection', *IEEE Trans. Inform. Theory* **50**(11), 2708–2726.

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

Stein, C. M. (1981), 'Estimation of the mean of a multivariate normal distribution', *Ann. Statist.* **9**(6), 1135–1151.

CHAPTER 3

WITHIN–FAMILY PREDICTIVE RISK: OPTIMAL FLATTENING & SHRINKAGE

The within-family prediction error is the minimal risk among estimates in the class $\mathcal{G}$ of all Gaussian densities. We produce asymptotically sharp upper and lower bounds on the within-family prediction errors for various subfamilies of $\mathcal{G}$. We exhibit instances where the within-family error can be attained by data-adaptive linear density estimators. Also, in some special cases, the optimal risk can be expressed in terms of the mean square error of the associated location estimator and the nature of shrinkage can be determined based on only point estimation theory results.

Here, for the Homoscedastic Gaussian Predictive Model

$$\textbf{M.1} \qquad \mathbf{X} \sim N(A\,\boldsymbol{\theta},\, \sigma_p^2\, I) \quad \text{and} \quad \mathbf{Y} \sim N(B\,\boldsymbol{\theta},\, \sigma_f^2\, I)$$

described in Chapter 1, we discuss efficient estimators in the class $\mathcal{G}_n$ of all n–dimensional Gaussian distributions with positive definite (p.d.) covariances as the

dimension increases, i.e.,

$$\mathcal{G} = \left\{ g : \mathbb{R}^n \to \mathbb{R}^+ \text{ such that } g = N(\mu, \Sigma) \text{ where } \mu \in \mathbb{R}^n \text{ and } \Sigma \text{ p.d.} \right\}.$$

For any class $\mathcal{C}$ we define the predictive risk of the class as

$$\boldsymbol{\rho}_{\mathcal{C}}(\boldsymbol{\theta}) = \inf_{\widehat{p} \in \mathcal{C}} \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}).$$

As the true parametric density is also Gaussian, $\boldsymbol{\rho}_{\mathcal{G}}(\boldsymbol{\theta})$ represents the within-family predictive risk. Under mild regularity conditions, in the sub-family where the covariance structure is represented by a single data dependent parameter $\widehat{\Sigma} = \widehat{d} \cdot I$, the Kullback-Leiber risk has a tractable decomposition which can be subsequently minimized to yield optimally flattened predictive density estimates. We also evaluate the predictive risk of the sub-family $\mathcal{G}[p]$ which contains all product Gaussian densities. In Chapter 4 we will also make inferences on $\mathcal{G}$ under sparsity restrictions in the parameter space. As in point estimation, risk calculations in **M.1** would intrinsically depend on risk calculations in the orthogonal model:

Orthogonal Gaussian Predictive Model

$$\textbf{M.2} \qquad \mathbf{X} \sim N(\boldsymbol{\theta},\, \sigma_p^2\, I) \quad \text{and} \quad \mathbf{Y} \sim N(\boldsymbol{\theta},\, \sigma_f^2\, I)$$

where $\mathbf{X}$ and $\mathbf{Y}$ are both n – dimensional vectors. Most of our calculations will be in high-dimensions (which means $n \to \infty$ in the orthogonal model) though dimension independent bound will also be provided. As $n \to \infty$, **M.2** represents the Gaussian sequence model (Nussbaum 1996) and has been widely studied in the function estimation framework (Johnstone 2012). Estimation in **M.1** can be linked with the decision theoretic results in **M.2** through the procedure outlined in Donoho, Johnstone & Montanari (2011).

Our Contributions

Efficacy of predictive density estimates has been a subject of considerable interest in predictive inference. (Aitchison 1975, Aslan 2006, Komaki 1996, Hartigan 1998) determined asymptotically optimal (admissible) Bayes predictive density estimates in fixed dimensional parametric family whereas minimax optimality in restricted parameter spaces has been discussed in (Fourdrinier, Marchand, Righi & Strawderman 2011) and (Kubokawa, ric Marchand, Strawderman & Turcotte 2013). Recently, (George et al. 2006, Brown et al. 2008, Ghosh et al. 2008) extended the admissibility results to high dimensional Gaussian models. However, the optimal estimates are not necessarily Gaussian and using them in high-dimensional problems would involve computationally intensive methods. Here, we find optimal predictive density estimates within the Gaussian family and also compute their predictive risk. It is computationally easier to construct predictive attributes based on our optimal Gaussian predictive density estimates and the optimal Gaussian predictive risk assures guaranteed performances of our strategies.

Minimizing the Gaussian predictive risk involves simultaneous estimation of the location and scale parameters. The issue of joint estimation of location and scale (and to a degree the shape) has not been addressed before in one sample Gaussian models. However, separate estimation of location (Tibshirani 2011) and covariance (Friedman, Hastie & Tibshirani 2008) are well-studied topics in constrained Gaussian estimation. Also, as reviewed in (George, Liang & Xu 2012) decision theoretic parallels exist between point estimation theory under quadratic loss and predictive density estimation under Kullback-leibler loss in high-dimensional Gaussian models. Here, our results demonstrate that some of these decision theoretic parallels (in the class $\mathcal{G}$) can be explained by second moment based concentration properties on the quadratic loss of location point estimators in high dimensions. The moment based approach used here for estimating the scale parameter bears resemblance to concepts seen elsewhere in prediction theory, particularly in the the theory of cross validation (Yang 2007) and covariance penalties for model selection (Efron 2004, Ye 1998) .

3.1 Description of the main results

We describe optimality characterized in terms of asymptotic admissibility and through oracle inequalities of density estimates in class $\mathcal{G}$ as the parameter space is unrestricted over $\mathbb{R}^n$. As dimension n increases the role of shrinkage becomes important. In order to describe the results, we need to introduce the following notations.

Notation and Preliminaries

As some of our results are dimension dependent, henceforth we refrain from using bold representation for vectors and denote the dimension in the subscript. Given any fixed sequence θ_∞ we represent the first n values by the n–dimensional vector θ_n whereas $\theta(n)$ denotes the n^{th} value, i.e $\theta_{n+1} = (\theta_n, \theta(n+1))$. By $\mathcal{G}_n[p]$ we denote the class of all n–dimensional product Gaussian densities

$$\mathcal{G}_n[p] = \left\{ g[\mu_n, D_n] : \mu_n \in \mathbb{R}^n \ \& \ D_n \text{ is any } n \times n \text{ p.d. diagonal matrix} \right\}$$

where $g[\mu_n, D_n]$ is a normal density with mean μ_n and diagonal covariance $\sigma_f^2 D_n$. We represent the minimal Gaussian predictive risk by $\rho_{\mathcal{G}}(\theta_n) := \inf_{\widehat{p} \in \mathcal{G}_n} \rho(\theta_n, \widehat{p})$.

Our shrinkage results will mostly refer to the sub-family $\mathcal{G}_n[1]$ of $\mathcal{G}_n[p]$. $\mathcal{G}_n[1]$ contains Gaussian densities with only one data-adaptive scale estimate

$$\mathcal{G}_n[1] = \left\{ g[\mu_n, c] : \mu_n \in \mathbb{R}^n \text{ and } c \in \mathbb{R}^+ \right\}.$$

where $g[\widehat{\theta}_n, c]$ denotes a normal density with mean μ_n and covariance $c\,\sigma_f^2\, I$. A typical density estimate in $\mathcal{G}_n[1]$ is represented as $g[\widehat{\theta}_n, \widehat{c}\,(n)]$ where $\widehat{\theta}_n$ is a location estimate and $\widehat{c}(n)$ is the scale estimate based on observing an n–dimensional past observation X_n. For any fixed location estimate $\widehat{\theta}_n$, the optimal risk of density estimates in $\mathcal{G}_n[1]$ centered around $\widehat{\theta}_n$ is given by

$$\rho_0(\theta_n, \widehat{\theta}_n) = \inf_{\widehat{c}(X_n) \in \mathbb{R}^+} \rho(\theta_n, g[\widehat{\theta}_n, \widehat{c}(X_n)]).$$

The quadratic risk of the location estimate is denoted by

$$q(\theta_n, \widehat{\theta}_n) = \mathbb{E}_{\theta_n} \| \widehat{\theta}\,(X_n) - \theta_n \|^2$$

where the expectation is over the observed past X_n. The notation used here for representing the quadratic risk is a bit different than that used in Chapter 2 and Chapter 4. Here we have dropped of the symbol v representing the variability of the point estimation problem in $q(\theta_n, \widehat{\theta}_n, v)$ –the notation for quadratic risk used in the other chapters. It should be noted that here $q(\theta_n, \widehat{\theta}_n)$ means that the noise variance is always σ_p^2.

Later, we show that if the value of $q(\theta_n, \widehat{\theta}_n)$ were known, then the optimal choice for scale is

$$\mathrm{IF}_{\theta_n}(\widehat{\theta}_n) = 1 + n^{-1} r^{-1} q(\theta_n, \widehat{\theta}_n)$$

which will be called as the Ideal Flattening coefficient for $\widehat{\theta}_n$ at θ_n. Here, given a location estimate $\widehat{\theta}_n$ we construct suitable estimates $\widehat{c}(n)$ of the scale such that asymptotically when $n \to \infty$ the density estimate $g[\widehat{\theta}_n, \widehat{c}(n)]$ is optimally flattened in the sense that $\rho(\theta_n, g[\widehat{\theta}_n, \widehat{c}(n)]) - \rho_0(\theta_n, \widehat{\theta}_n) \leq \mathrm{O}\,(1)$. However, for proving optimality of the flattening coefficient we need the following mild regularity conditions on the location estimate $\widehat{\theta}_n$:

$$q(\theta_n, \widehat{\theta}_n) \leq \mathrm{O}\,(n). \tag{3.1}$$

$$\mathbb{V}ar_{\theta_n}\left(\| \widehat{\theta}_n - \theta_n \|^2 \right) \leq \mathrm{O}\,(n) \tag{3.2}$$

and the existence of a suitable estimate $\mathrm{U}[\widehat{\theta}](X_n)$ for the quadratic risk of $\widehat{\theta}_n$ at θ_n with the following properties:

$$\left| \mathbb{E}_{\theta_n}\big(\widehat{U}_n\big) - q(\theta_n, \widehat{\theta}_n) \right| \leq \mathrm{O}\left(n^{1/2} \right). \tag{3.3}$$

$$\mathbb{V}ar_{\theta_n}\big(\widehat{U}_n\big) \leq \mathrm{O}\,(n). \tag{3.4}$$

$$\mathbb{V}ar_{\theta_n}\left\{ \left(1 + (nr)^{-1} \widehat{U}_n\right)^{-1} \right\} \leq \mathrm{O}\,(n^{-1}). \tag{3.5}$$

These properties are fairly mild and in Section 2 we show that most popular point estimators obey these above conditions. We call these conditions Reasonable Asymptotic Square Loss (RASL) properties and the set of point estimators in the sequence model (where the action set is $\mathbb{R}^\infty$) which satisfies these conditions is denoted by $\mathcal{A}$. Also, we denote the ratio of the future to past variances by $r := \sigma_f^2/\sigma_p^2$. Our results will depend on r. For sequences, the symbol $a_n \sim b_n$ means $a_n = b_n(1 + \mathrm{o}(1))$ and $a_n \approx b_n$ means $a_n/b_n \in (k_1, k_2)$ where k_1 and k_2 are constants.

Results

We present the results under $\sigma_p^2 = 1$ and $\sigma_f^2 = r$ which will assumed throughout the rest of the chapter. The results can be easily modified for general σ_p though. We show that in high dimensions, the minimum predictive entropy risk of Gaussian density estimates around reasonable location estimate $\widehat{\theta}_n$ can be expressed in terms of the corresponding quadratic risk of $\widehat{\theta}_n$. The minimum predictive risk can be attained by optimally flattening the normal density estimate around $\widehat{\theta}_n$. The choice of the optimal flattening coefficient is not unique. An asymptotically efficient choice based on a reasonable estimate $U[\widehat{\theta}_n](X_n)$ of the quadratic risk of $\widehat{\theta}_n$ can be made.

Theorem 3.1.1.

For any estimator $\widehat{\theta}$ in $\mathcal{A}$ we have

$$\left| \rho_0(\theta_n, \widehat{\theta}_n) - \frac{n}{2} \log\left(1 + (nr)^{-1} \cdot q(\theta_n, \widehat{\theta}_n)\right) \right| \leq \mathrm{O}\left(1\right) \text{ as } n \to \infty. \tag{3.6}$$

And if $\widehat{c}(X_n) = 1 + (nr)^{-1} U[\widehat{\theta}](X_n)$ is based on a suitable estimate $\widehat{U}_n$ of the quadratic risk as defined in Equations (3.3)–(3.5) then

$$\rho(\theta_n, g[\widehat{\theta}_n, \widehat{c}(n)]) - \rho_0(\theta_n, \widehat{\theta}_n) \leq \mathrm{O}\,(1). \tag{3.7}$$

By $g[\widehat{\theta}_n]$ we will represent density estimates of the form $g[\widehat{\theta}_n, \widehat{c}(n)]$ which are based on scale estimates $\widehat{c}(n) = 1 + (nr)^{-1}\widehat{U}_n$ involving suitable quadratic risk estimate $\widehat{U}$.

By the above theorem and under the aforementioned regularity conditions $g[\widehat{\theta}_n]$ will be an asymptotically optimal choice among all Gaussian density estimates centered around $\widehat{\theta}_n$. Based on the asymptotic relations between $\rho_0(\theta_n, \cdot)$ and the Mean Square Error (MSE) $q(\theta_n, \cdot)$, we can characterize the predictive risk of $g[\,\widehat{\theta}_n]$ easily by plugging in standard oracle inequalities from point estimation theory. We check that RASL conditions defined in Equations (3.1)–(3.5) hold for the James-Stein estimator (Stein 1981)

$$\widehat{\theta}_n^{JS} = X_n \left(1 - \frac{n-2}{\|X_n\|^2}\right)$$

and its positive part estimator $\widehat{\theta}^{JS+}$. For the James-Stein estimator we determine the deviations from the optimal risk in terms of dimension dependent bounds.

Theorem 3.1.2.

For any dimension $n \geq 10$ and for any $\theta_n \in \mathbb{R}^n$ we have,

$$\rho\big(\theta_n, g[\widehat{\theta}_n^{JS}]\big) - \rho_0\big(\theta_n, \widehat{\theta}_n^{JS}\big) \leq 2^{-1}\big(a_n^{1/2}\, b_n^{1/2}\, r^{-3/2} + (a_n + b_n + l_n)\, r^{-2} + a_n\, r^{-3}\big)$$

where the constants a_n, b_n, l_n are independent of the parameter but depend on the dimensions n and are given by

$$a_n = 3\big(1 - (n-2)^{-1}\big)^{-2},\ b_n = 4(2 + a_n + k_2(n)),$$
$$l_n = 3(1 - 2/n)^{-2},\ k_2(n) = \max\{e(n), f(n)\} \text{ with}$$
$$e_n = \sqrt{3}\prod_{i=1}^{4}(1 - (2i+1)/n)\}^{-1/2} \text{ and } f_n = (1 - (\log n/n)^{1/2})^{-2}.$$

Also, $\rho\big(\theta_n, g[\widehat{\theta}_n^{JS}]\big)$ can be approximated by using the following bound

$$\left|\rho\big(\theta_n, g[\widehat{\theta}_n^{JS}]\big) - \frac{n \log \mathrm{IF}_{\theta_n}(\widehat{\theta}_n)}{2}\right| \leq \frac{\big(a_n^{1/2}\, b_n^{1/2}\, r^{-3/2} + (a_n + b_n)\, r^{-2} + a_n\, r^{-3}\big)}{2}.$$

These bounds hold for any value of $r \in (0, \infty)$. As the ratio of the future to past variances r decreases, we need to estimate the future observations based on increasingly

noisy past observations and so, the difficulty of the density estimation problem also increases. So, as expected when r decreases the bounds also increases. These bounds can be made dimension independent.
In particular, for all dimension $n \geq 20$, for any $\theta_n \in \mathbb{R}^n$ and for any fixed value of $r \in (0, \infty)$, we have,

$$\rho\big(\theta_n, g[\widehat{\theta}_n^{JS}]\big) - \rho_0\big(\theta_n, \widehat{\theta}_n^{JS}\big) \leq 5.3\, r^{-3/2} + 19.6\, r^{-2} + 1.7 r^{-3}. \tag{3.8}$$

In point estimation theory, there exist sharp oracle bounds on the quadratic risk $q(\theta_n, \widehat{\theta}_n^{JS})$ of the James-Stein estimator $\widehat{\theta}^{JS}$(Johnstone 2012, Chapter 2) which along with Theorem 3.1.2 produce the following oracle bound on the predictive risk of shrinkage predictive density estimates. Assuming that the value $||\theta_n||^2$ is known, the risk of the ideal linear predictive density estimate is given by

$$\mathrm{IL}\,(\theta_n) = \frac{n}{2}\, \log\left(1 + r^{-1}\, \frac{a_n}{1 + a_n}\right) \text{ where } a_n = ||\theta_n||^2/n. \tag{3.9}$$

The difference in the risk of $g[\widehat{\theta}_n^{JS}]$ and the optimal oracle linear risk is

$$\rho\big(\theta_n, g[\widehat{\theta}_n^{JS}]\big) - \mathrm{IL}\,(\theta_n) \leq 0.1\, r^{-1} + 5.3\, r^{-3/2} + 18.1\, r^{-2} + 1.7 r^{-3}. \tag{3.10}$$

and the following

$$\mathrm{B}(||\theta_n||, n, r) = \frac{n}{2}\, \log\left\{1 + r^{-1}\, ||\theta_n||^2 (n + ||\theta_n||^2)^{-1}\right\} + 0.1\, r^{-1} + 5.3\, r^{-3/2} + 18.1\, r^{-2} + 1.7 r^{-3} \tag{3.11}$$

upperbounds the predictive risk $\rho\big(\theta_n, g[\widehat{\theta}_n^{JS}]\big)$. Figure 3.2 and Figure 3.1 show these bounds for different values of r and with dimension n being 20 and 1000 respectively. From the figures, it is evident that the upperbound $\mathrm{B}(||\theta_n||, n, r)$ is not very good for low dimensions but becomes better as n increases. Figure 3.3 shows that even in low dimensions the risk of $g[\widehat{\theta}_n^{JS}]$ is close to the ideal linear risk $\mathrm{IL}(\theta_n)$.

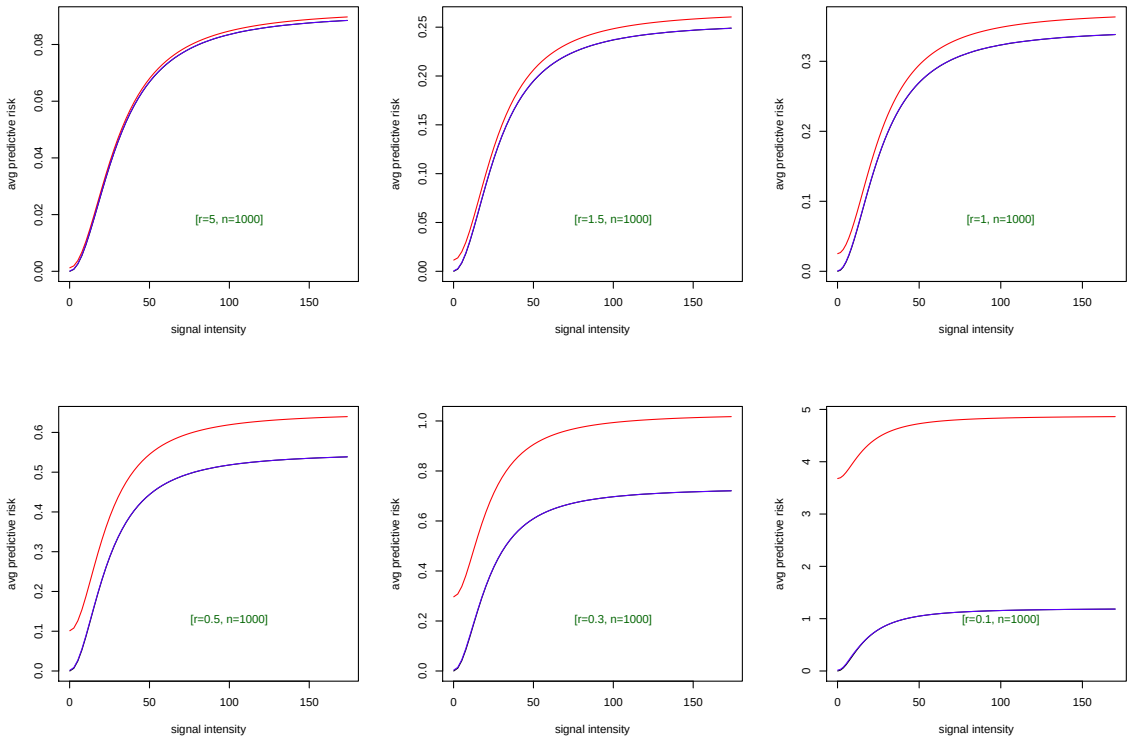

Figure 3.1: The plots show the average ideal linear oracle risk $\mathrm{IL}(\theta_n)/n$ (in black), the average predictive risk of $g[\widehat{\theta}_n^{\mathrm{JS}}]$ (numericaly evaluated and in blue) and the upper-bound $\mathrm{B}(||\theta_n||, n, r)$ (described in Equation 3.11) as the signal intensity $||\theta_n||$ varies along the abscissa. From the top-left, in anti-clockwise order, the plots correspond to $r = 5, 1.5, 1, 0.5, 0.3$ and 0.1. Here, dimension $n = 1000$.

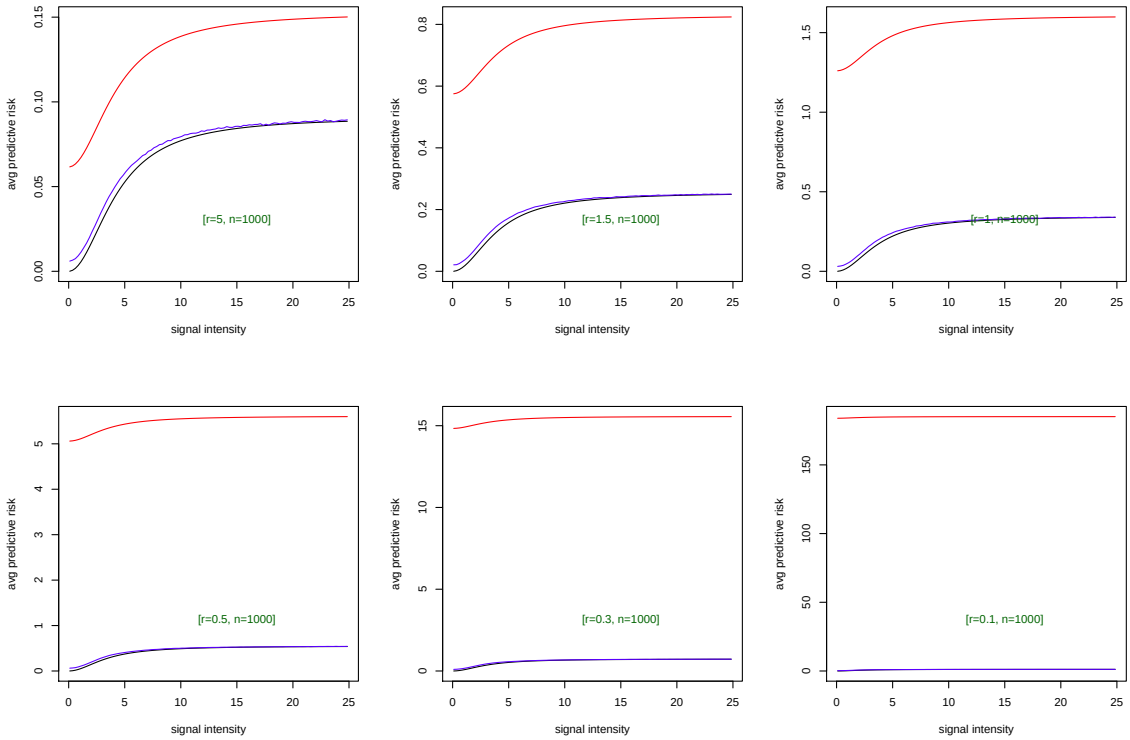

Figure 3.2: The plots show the average ideal linear oracle risk $\mathrm{IL}(\theta_n)/n$ (in black), the average predictive risk of $g[\widehat{\theta}_n^{\mathrm{JS}}]$ (numericaly evaluated and in blue) and the upper-bound $\mathrm{B}(||\theta_n||, n, r)$ (described in Equation 3.11) as the signal intensity $||\theta_n||$ varies along the abscissa. From the top-left, in anti-clockwise order, the plots correspond to $r = 5, 1.5, 1, 0.5, 0.3$ and 0.1. Here, dimension $n = 20$.

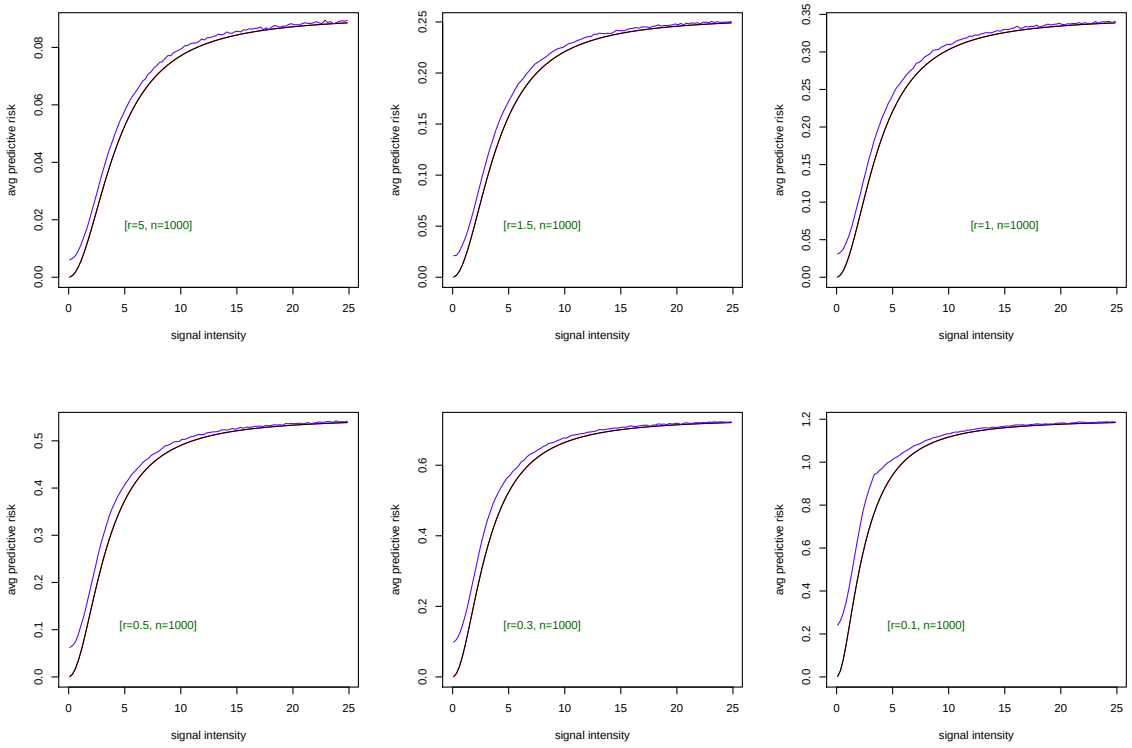

Figure 3.3: The plots show the difference between the average ideal linear oracle risk $\mathrm{IL}(\theta_n)/n$ (in black) and the average predictive risk of $g[\widehat{\theta}_n^{\mathrm{JS}}]$ (numericaly evaluated and in blue) as the signal intensity $||\theta_n||$ varies along the abscissa. From the top-left, in anti-clockwise order the plots correspond to $r = 5, 1.5, 1, 0.5, 0.3$ and 0.1. Here, dimension $n = 10$.

Comparing these bounds to the oracle bound of (Xu & Zhou 2011) which is derived based on an empirical Bayes perspective

$$\rho\big(\theta_n, g[\widehat{\theta}_n^{JS}]\big) - \mathrm{IL}\,(\theta_n) \leq 2\,r^{-1} + 5\,r^{-2} + 4r^{-3}, \tag{3.12}$$

the particular features of our moment based approach can be seen. As our oracle inequality is a by-product of the optimal Gaussian risk, for most values of r the bound in the Inequality 3.10 is coarser than that in Inequality 3.12. However, when $r = 0.1$, the RHS in the Inequality 3.10 is 3830 and is better than the bound (4520) in the latter. Thus, the moment based approach can be quite informative. The bounds derived on the predictive risk are sharp enough to derive decision-theoretic optimality. We can produce unrestricted improved minimax predictive densities which asymptotically behaves like ideally shrunk linear density estimates (as defined later). The following lemma shows the asymptotic improvement in the predictive risk over the best invariant predictive density $g[X_n, 1+r]$ (Liang & Barron 2004).

Lemma 3.1.3.

If $||\theta_n||^2 \to \infty$ as $n \to \infty$, then we have

[a] $\rho\big(\theta_n, g[X_n, 1+r]\big) = 2^{-1}\, n\, \log\big(1 + r^{-1}\big)\,.$

[b] $\rho\big(\theta_n, g[\widehat{\theta}_n^{JS}, r]\big) \sim (2r)^{-1}\, n\, a_n\, (1+a_n)^{-1}$ where $a_n = n^{-1}\, ||\theta_n||^2$.

[c] $\rho\big(\theta_n, g[\widehat{\theta}_n^{JS},\, 1+r]\,\big) \sim 2^{-1}\, n\, \big\{\log(1+r^{-1}) - (1+a_n)^{-1}(1+r)^{-1}\big\}\,.$

[d] $\rho\big(\theta_n, g[\widehat{\theta}_n^{JS}]\,\big) \sim 2^{-1}\, n\, \log\big\{1 + r^{-1}\, a_n\, (1+a_n)^{-1}\big\}\,.$

The improvement in predictive risk due to efficient choice of location is reflected by the risk of $g[\widehat{\theta}^{\,JS},\, 1+r]$ where as the effect of the optimal choice of scale after choosing an appropriate location estimate can be followed by evaluating the asymptotic predictive risk of $g[\widehat{\theta}^{\,JS}]$. The following lemma extends Theorem 3.1.1 to convex combinations of location estimates.

Lemma 3.1.4.

For any countable collection Λ of estimators $\widehat{\theta}[\lambda]$ in $\mathcal{A}$ and their convex collection $\widehat{\theta}^w = \sum_{\lambda\in\Lambda} w_\lambda\, \widehat{\theta}[\lambda]$ with $\sum_{\lambda\in\Lambda} w_\lambda = 1$, we have

$$\rho_0(\theta_n, \widehat{\theta}^w_n) - \frac{n}{2}\sum_{\lambda\in\Lambda} w_\lambda \log\Big(1 + (nr)^{-1}\cdot q\big(\theta_n, \widehat{\theta}_n[\lambda]\big)\Big) \leq \mathrm{O}\left(1\right) \text{ as } n\to\infty.$$

And, the predictive density estimate $\sum_{\lambda\in\Lambda} w_\lambda\, g\big[\widehat{\theta}_n[\lambda]\big]\big)$ is asymptotically optimal in the sense

$$\rho\left(\theta_n, \sum_{\lambda\in\Lambda} w_\lambda\, g\big[\widehat{\theta}_n[\lambda]\big]\right) - \rho_0\big(\theta_n, \widehat{\theta}^w_n\big) \leq \mathrm{O}\left(1\right) \text{ as } n\to\infty. \tag{3.13}$$

It should be noted that the RASL regularity conditions on the location point estimates do not necessarily extend to any convex collections. However, the predictive risk still concentrates and the optimal predictive risk ρ_0 can be determined based on the above lemma.

3.1.1 Organization of this chapter

Next we present some properties of the KL risk of density estimates in the class $\mathcal{G}[1]$. We report the asymptotic decomposition of the optimal KL risk under regularity conditions. After a formal summary of the RASL regularity conditions, we narrate the notion of optimal flattening, its relation with shrinkage and their decision theoretic implications. We end with a discussion on the applicability our approach in the wider classes $\mathcal{G}[1]$ and $\mathcal{G}$ of density estimates.

3.2 Optimal flattening and predictive risk in $\mathcal{G}[1]$

Hereon we will assume that $\sigma_p^2 = 1$ and $\sigma_f^2 = r$. The general predictive KL risk will not be affected by this restriction. However, the density estimates are usually based on statistics equivariant to the scale transformation and needs multiplication by σ_p.

Heuristic Idea: In the high dimensions the quadratic loss of a reasonable point estimator will concentrate around its risk And, so the KL risk of the corresponding Gaussian predictive density partitions into two parts involving (i) quadratic risk on the location parameter adjusted by the expected scale (ii) logarithm of the expected scale. As such, the risk $\rho\big(\theta_n, g[\widehat{\theta}_n, \widehat{c}_n]\big)$ of the normal predictive density estimate $g[\widehat{\theta}_n, \widehat{c}_n]$ is given by

$$\frac{n}{2}\left[\left\{\mathbb{E}_{\theta_n}(\log \widehat{c}(X_n)) + \mathbb{E}_{\theta_n}\left(\frac{1}{\widehat{c}\,(X_n)}\right) - 1\right\} + \frac{1}{nr}\mathbb{E}_{\theta_n}\left(\frac{\|\widehat{\theta}\,(X_n) - \theta_n\|^2}{\widehat{c}\,(X_n)}\right)\right].$$

In high dimensions, due to concentration of measure we expect

- $\mathbb{E}_{\theta_n} \log\big(\widehat{c}_n\big) \sim \log \mathbb{E}_{\theta_n} \widehat{c}_n$
- $\mathbb{E}_{\theta_n}\big(\widehat{c}_n^{-1}\big) \sim \big(\mathbb{E}_{\theta_n} \widehat{c}_n\big)^{-1}$
- $\mathbb{E}_{\theta_n}\Big(\|\widehat{\theta}\,(X_n) - \theta_n\|^2 \cdot \widehat{c}^{-1}\,(X_n)\Big) \sim \big(\mathbb{E}_{\theta_n} \widehat{c}^{-1}\,(X_n)\big)^{-1} q(\theta_n, \widehat{\theta}_n)$

which will lead to

$$\rho\big(\theta_n, g[\widehat{\theta}_n, \widehat{c}_n]\big) \sim \frac{n}{2}\left\{\log \mathbb{E}_{\theta_n} \widehat{c}_n + \frac{1 + (nr)^{-1} q(\theta_n, \widehat{\theta}_n)}{\mathbb{E}_{\theta_n} \widehat{c}_n} - 1\right\} + \mathrm{O}\,(1) \;\text{ as } n \to \infty.$$

This asymptotic decomposition of the predictive risk can be explicitly validated through the RASL properties. Because of this decomposition for any fixed point estimate $\widehat{\theta}_n$ at each parametric value θ_n we can minimize the above asymptotic value of $\rho(\theta_n, g[\widehat{\theta}_n, \,.\,])$ over the scalar quantity $\mathbb{E}_{\theta_n} \widehat{c}_n$. The minimum asymptotic value is given by

$$\rho\big(\theta_n, g[\widehat{\theta}_n]\big) \sim n/2 \cdot \log\big(1 + (nr)^{-1} q(\theta_n, \widehat{\theta}_n)\big)$$

and the optimal value is attained when

$$\mathbb{E}_{\theta_n}(\,\widehat{c}^{\,\mathrm{opt}}\,(X_n)\,) = 1 + (nr)^{-1}\, q(\theta_n, \widehat{\theta}_n) = \mathrm{IF}_{\theta_n}(\widehat{\theta}_n)$$

which is the ideal flattening coefficient. However, $\text{IF}_{\theta_n}(\widehat{\theta}_n)$ is unknown. But, it depends only on the parametric value θ_n. Thus a choice would be $\widehat{c}^{\,\text{opt}}(X_n) = 1 + (nr)^{-1}U[\widehat{\theta}_n](X_n)$, where $U[\widehat{\theta}_n](X_n)$ (to be abbreviated as $\widehat{U}_n$) is reasonable (i.e with reasonable bias and concentration properties) estimate of the quadratic risk of $\widehat{\theta}_n$. With very high probability such an optimal choice of $\widehat{c}_n$ will be greater than 1 reflecting a flattening of scale of the estimated predictive density (with respect to the true future variability). Intuitively, we are performing an appropriate flattening of the density based on empirical estimates of the quadratic loss. The optimal density $g[\widehat{\theta}_n]$ levels out with increasing inaccuracy in the location estimate $\widehat{\theta}_n$.

One of the popular Frequentist notion (which is better than plug-in density estimates) of constructing predictive densities in this parametric model is to use Gaussian density estimate around an efficient location $\widehat{\theta}_n$ and variance $r + \widehat{\mathbb{V}ar}(\widehat{\theta}_n)$. Estimates of these kind are natural extensions of confidence sets. The optimal density estimate $g[\widehat{\theta}]$ is quite similar except with a larger variance $r + \widehat{q}\,(\theta_n, \widehat{\theta}_n)$. And unless the bias of θ_n is negligible compared to its variance the above mention general notion produces sub-optimal density estimates. Next through the RASL conditions we will quantify some statistical regularities in the behavior of quadratic loss in high dimensions.

3.2.1 RASL Properties of a Location Point Estimate

In high dimensions, for any fixed location parameter θ_n and its estimate $\widehat{\theta}_n$ we expect the quadratic loss $\|\,\widehat{\theta}_n - \theta_n\|_2^2$ to be concentrated around its expected value (quadratic risk) $\mathbb{E}_{\theta_n}\|\,\widehat{\theta}_n - \theta_n\|_2^2$ and it would be reflected by its variance. We will also rule out very bad point estimators by neglecting those with too high risk as we do not want them for prediction purposes. Apart from these we also assume the existence of a statistic which estimates the quadratic risk within reasonable bias. These properties of point estimators are referred to as **R**easonable **A**symptotic **S**quare **L**oss properties and the corresponding location estimates as RASL estimates.

As dimension $n \to \infty$, for any fixed parametric value θ_n the location point estimate $\widehat{\theta}(X_n)$ is such that its quadratic loss has the following properties.

P1. Reasonable Risk:

$$\mathbb{E}_{\theta_{\mathbf{n}}} \| \widehat{\theta}_{\mathbf{n}} - \theta_{\mathbf{n}} \|^{\mathbf{2}} \leq \mathbf{O(n)}.$$

The canonical minimax point estimator X_n which is also the UMVUE (under square loss) in this case acts as the benchmark in weeding out the bad point estimators. For any parameter value θ_n, X_n has constant risk n. So, it is appropriate for our purpose to restrict ourselves to point estimators with risk of the O(n).

P2. Concentration property of Quadratic loss:

$$\mathbb{V}\mathbf{ar}_{\theta_{\mathbf{n}}} (\| \widehat{\theta}_{\mathbf{n}} - \theta_{\mathbf{n}} \|^{\mathbf{2}}) \leq \mathbf{O(n)}.$$

In high dimensions the estimator $\widehat{\theta}_n$ is such that its loss has variability less than $O(n)$. Again comparing with X_n, we see $\mathbb{V}ar_{\theta_n} (\| X_n - \theta_n \|^2) = 2n$ as $\| X_n - \theta_n \|^2$ is distributed as a central χ^2 random variable with n degrees of freedom.

P2 implies concentration of the loss function and would in turn also impose some concentration properties on well-behaved functions of the loss. As such, using Lemma **A.1**, we have

$$\textbf{P2.a} \qquad \mathbb{V}\mathbf{ar}_{\theta_{\mathbf{n}}} \left\{ \left(\mathbf{1} + \frac{\|\widehat{\theta}_{\mathbf{n}} - \theta_{\mathbf{n}}\|^{\mathbf{2}}}{\mathbf{nr}} \right)^{-\mathbf{1}} \right\} \leq \mathbf{O(n^{-1})}$$

following directly from **P2**. It is an important condition and will be used in our derivations.

P3. Reasonable Estimate of Quadratic Risk:

There exists an estimator $U[\widehat{\theta}_n](X_n)$ (will be abbreviated as $\widehat{U}_n$) of the quadratic

risk of $\widehat{\theta}_n$ satisfying the following:

P3.1. $$\left| \mathbb{E}_{\theta_\mathbf{n}}\big(\widehat{\mathbf{U}}_\mathbf{n}\big) - \mathbb{E}_{\theta_\mathbf{n}} \|\widehat{\theta}_\mathbf{n} - \theta_\mathbf{n}\|^\mathbf{2} \right| \leq \mathbf{O}(\,\mathbf{n^{1/2}}\,).$$

P3.2. $$\mathbb{V}\mathbf{ar}_{\theta_\mathbf{n}}\big(\widehat{\mathbf{U}}_\mathbf{n}\big) \leq \mathbf{O}(\mathbf{n}).$$

P3.3. $$\mathbb{V}\mathbf{ar}_{\theta_\mathbf{n}} \left\{ \left(\mathbf{1} + \frac{\widehat{\mathbf{U}}_\mathbf{n}}{\mathbf{nr}} \right)^{-\mathbf{1}} \right\} \leq \mathbf{O}(\mathbf{n^{-1}}).$$

P3.1 implies existence of a statistic which estimates the quadratic risk by not making significant bias. Bias exceeding $O(\sqrt{n})$ is considered significant here and the order is associated with the $O(n^{-1})$ asymptotic statements we would like to make. **P3.1** and **P3.2** are analogous to **P2** and **P2.a** respectively. They imply that the asymptotic concentration properties associated with the quadratic loss also holds for its estimator $\widehat{U}_n$. If $\widehat{U}_n$ is positive then **P3.2** follows directly from **P3.1** by Lemma **A.1**.

3.2.2 Validating the RASL properties

Given a location point estimator and its corresponding reasonable quadratic risk estimate the RASL conditions can be checked at least by simulations. However, existence of a 'reasonable' risk estimator (as defined in **P3**) is essential. For most widely used point estimates, we can construct risk estimates satisfying the three conditions in **P3** though the procedures can sometime get quite complicated.

If $\widehat{\theta}_n$ is the posterior mean – generalized Bayes estimate with respect to prior π, then by Tweedie's formula (Robbins 1956, Brown 1971) we have explicit expression

of an unbiased estimate of its risk as,

$$\widehat{U}_n^{\pi} = n - \left[\|\nabla \log m_\pi(X_n)\|^2 - 2\frac{\nabla^2 m_\pi(X_n)}{m_\pi(X_n)} \right] \text{ where } \nabla f \triangleq \sum_{i=1}^{n} D_i\, f \tag{3.14}$$

$$\nabla^2 m_\pi(X_n) = \sum_{i=1}^{n} D_i^2\, m_\pi(X_n) \text{ and } m_\pi(x_n) = \int \phi_n(x_n|\theta_n, 1)\pi(\theta_n)\, d\theta_n. \tag{3.15}$$

$\widehat{U}_n^{\pi}$ is a natural candidate for a 'reasonable estimate of the quadratic loss' though **P3.2** and **P3.3** are also to be checked separately. In particular, for **P3.3** to hold $\widehat{U}_n^{\pi}$ may need some modification by introducing some bias.
For spherically symmetric estimators, we can get candidates for 'reasonable' risk estimates by using Stein's unbiased (quadratic) risk estimates (SURE) or their modifications (like positive part, etc) (Stein 1974, Stein 1981). As mentioned before, here too we needed to introduce some bias to the the SURE estimate as the unbiased one does not has property P3.3.

These RASL conditions are quite mild and usually holds for reasonable point estimates and can be checked by Monte Carlo simulations for arbitrary point estimates. Next, we check these conditions analytically for the following popular point estimators:

$\widehat{\theta}^{JS}$: James Stein estimator

$\widehat{\theta}^{JS+}$: Positive-part James Stein estimator

$\widehat{\theta}^{H}$: Posterior mean of the harmonic prior $\pi_H(\theta_n) \propto ||\theta_n||^{-(n-2)}$.

All these 3 point estimators are linear estimates of the form $s(X_n)X_n$ where $s(X_n)$ is a data-dependent shrinkage term. They are better than the canonical minimax estimator X_n. While $\widehat{\theta}^{H}$ is admissible, $\widehat{\theta}^{JS}$ and $\widehat{\theta}^{JS+}$ are both inadmissible. As such both $\widehat{\theta}^{JS+}$ and $\widehat{\theta}^{H}$ dominates $\widehat{\theta}^{JS}$. However, in high dimensions, they behave similarly and have near ideal linear risk properties. We will construct reasonable risk estimates for each of these estimators. While verifying the RASL conditions for the JS estimator we would also compute the bound explicity for each n. It will be needed afterwards in Theorem 3.1.2. Since, the estimators are spherically symmetric it will be more

informative to derive bounds depending on $\|\theta_n\|^2$. Hence forth in this section, by a_n we denote $\|\theta_n\|^2/n$. A convenient fact about this spherically symmetric estimators is that the n-dimensional parameter θ_n can be substituted by $(||\theta_n||, 0, \ldots, 0)$ while checking the asymptotic behavior of square loss. As these estimators are not Lipchitz functions of the normal random variable X, we can not directly use well-established Gaussian concentration inequalities (Dembo & Zeitouni 1993, Ledoux 2001).

James Stein Location Estimator

The James-Stein estimator and its unbiased risk estimate is given by:

$$\widehat{\theta}_n^{JS} = X_n \left(1 - \frac{n-2}{\|X_n\|^2}\right), \quad \text{and} \quad U(\widehat{\theta}_n^{JS}) = \left(n - \frac{(n-2)^2}{\|X_n\|^2}\right).$$

RASL property **P1.** holds as the JS is better than the canonical estimator X_n. As such a good upper bound on its risk is also known

$$E_{\theta_n} \| \widehat{\theta}^{JS} - \theta_n \|^2 \leq 2 + \frac{(1-2/n)\, a_n}{(1-2/n) + a_n}.$$

Lemma 3.2.1.

$$\mathbb{V}ar_{\theta_n}\left(||\widehat{\theta}_n^{JS} - \theta_n||^2\right) \leq 4\Big[2n + (n-2)^4 n^{-3} k_1(n) + n k_2(n)\Big].$$

Proof. We decompose $||\widehat{\theta}_n^{JS} - \theta_n||^2$ into 3 parts as

$$||\widehat{\theta}_n^{JS} - \theta_n||^2 = \|X_n - \theta_n\|^2 + (n-2)^2 \|X_n\|^{-2} + 2(n-2) M_n$$

where $M_n = \langle X_n - \theta_n,\, X_n \|X_n\|^{-2} \rangle$. Then we use the naive inequality that for any three random variables $Z_i, i = 1, 2, 3$

$$\mathbb{V}ar\left(\sum_{i=1}^{3} Z_i\right) \leq \sum_{j=0}^{3} \mathbb{V}ar\left(\sum_{i=1}^{3} (-1)^{\mathbb{I}\{j=i\}} Z_i\right) = 4 \sum_{i=1}^{3} \mathbb{V}ar(Z_i)$$

to get the following bound on $\mathbb{V}ar_{\theta_n}\big(||\widehat{\theta}_n^{JS} - \theta_n||^2\big)$

$$\leq 4\Bigg\{\mathbb{V}ar_{\theta_n}\big(\|X_n - \theta_n\|^2\big) + \mathbb{V}ar_{\theta_n}\left(\frac{n-2}{\|X_n\|^2}\right) + 4\,(n-2)^4\mathbb{V}ar_{\theta_n}\big(M_n\big)\Bigg\}.$$

Now $\|X_n - \theta_n\|^2$ has a central chi-square distribution with n degrees of freedom and hence its variance is $2n$. The bounds on the other quantities follow from Lemma 3.2.2 and Lemma 3.5.1. □

Lemma 3.2.2.

For $n \geq 10$ we have $\mathbb{V}ar_{\theta_n}(M_n) \leq n^{-1}k_2(n)$ where

$$k_2(n) = \max\{h(n), k(n)\} \text{ where } e_n = \frac{\sqrt{3}}{\prod_{i=1}^4\{1-(2i+1)/n\}^{1/2}}$$
$$\text{and } f_n = \frac{1}{(1-(\log n/n)^{1/2})^2}.$$

Proof. The variance of M_n is same as the variance of $\langle\theta_n, X_n\rangle\|X_n\|^{-2}$ whose distribution is spherically symmetric in θ_n as it can be written as sum of two spherically symmetric terms $H_n = \langle\theta_n, X_n - \theta_n\rangle||X_n||^{-2}$ and $J_n = \|\theta_n\|^2||X_n||^{-2}$, . So, with out loss of generality we can assume that $\theta_n = (\theta, 0, \ldots, 0)$ where $\theta = \|\theta_n\|$. We also divide the proof into two cases depending on the magnitude of θ.

When $\theta \leq \sqrt{n}$ we have, $\mathbb{V}ar_{\theta_n}(M_n) \leq 2(\,\mathbb{V}ar_{\theta_n}(H_n) + \mathbb{V}ar_{\theta_n}(J_n))$ with the later being less than n^{-1} by Lemma 3.5.3. And, the former is bounded above by $E(H_n^2)$. Now, with $Z \stackrel{d}{=} N(0,1)$ and $W \stackrel{d}{=} \chi^2_{n-1}(0)$ and $V = (Z+\theta)^2 + W$ it can be rewritten as

$$E\big(\theta^2 Z^2 V^{-2}\big) \leq \sqrt{\theta^4 E Z^4\, E V^{-4}} \leq \sqrt{3\theta^4 E(W^{-4})} \leq \left\{\frac{3\theta^4}{\prod_{i=1}^4(n-2i-1)}\right\}^{1/2}$$

which is less than $n^{-1}\sqrt{3}\prod_{i=1}^4(1-(2i+1)/n)^{-1/2}$.

When $\theta > n$, we first recall that $M_n \stackrel{d}{=} (\theta + Z)/W$ and so

$$\begin{aligned} E(M_n^2) &\leq E\big\{V^{-2}\mathbb{I}_{\{|\theta+Z|\leq 1\}}\big\} + E\big\{(\theta+Z)^{-2}\mathbb{I}_{\{|\theta+Z|>1\}}\big\} \\ &\leq E\{V^{-2}\} + 2\int_1^\infty x^{-2}\phi(x-\theta)\,dx \\ &\leq [(n-3)(n-5)]^{-1} + \widetilde{\Phi}(\sqrt{\log n}) + \big\{\theta - \sqrt{\log n}\big\}^{-2} \\ &\leq [(n-3)(n-5)]^{-1} + n^{-1}(\log n)^{-1} + n^{-1}(1-(\log n/n)^{1/2})^{-2} \end{aligned}$$

Hence the result follows. □

Though it is very tempting but we can not use the unbiased risk estimate $U(\widehat{\theta}_n^{JS})$ as the estimate can be negative and violates **P3.3**.

Lemma 3.2.3.

For any fixed n and r, $E_0\big[\big\{1+(nr)^{-1}U(\widehat{\theta}_n^{JS})\big\}^{-1}\big]$ does not exist.

We will instead use $\widehat{U}_n^+$ the positive part of $U(\widehat{\theta}_n^{JS})$ and the scale estimate $\widehat{c}_n^{JS+} = 1+(nr)^{-1}\widehat{U}_n^+$. RASL condition **P3.1** can be easily checked as

$$\mathbb{V}ar_{\theta_n}(\widehat{U}_n^+) \leq \mathbb{V}ar_{\theta_n}(\,U(\,\widehat{\theta}_n^{JS}\,)\,) = (n-2)^4\mathbb{V}ar_{\theta_n}(\,\|X_n\|^{-2}) = O(n)$$

by Lemma 3.5.1 and **P3.2** follows from Lemma 3.5.3. As such, an exact dimension dependent bound can also be derived.

Lemma 3.2.4.

For any fixed $n \geq 3$ we have

$$|Bias_{\theta_n}(\widehat{c}_n^{\mathrm{JS}+})| \leq k_3(n)\,n^{-1/2} \text{ where } k_3(n) = \frac{\sqrt{2}+5n^{-1/2}}{1-2/n}.$$

Proof. Noting that $\text{Bias}_{\theta_n}(\widehat{c}_n^{\text{JS}+}) = n^{-1/2}\,\mathbb{E}_{\theta_n}(\widehat{U}_n^-)$ and

$$n\left|\mathbb{E}_{\theta_n}(\widehat{U}_n^-)\right| \leq \mathbb{E}_{\theta_n}\left[\left(\frac{n}{Y}-1\right)\cdot\mathbb{I}\{Y\leq n\}\right]$$

where Y follows Chi-square with degree n and non-centrality parameter $||\theta_n||^2$. We know that $Y \overset{d}{=} \chi^2_{n+2N}$ where $N \overset{d}{=} \text{Poisson}(||\theta_n||^2/2)$ and the above expectation can be written as

$$\mathbb{E}_{||\theta_n||^2}\left(\mathbb{E}\left[\left(\frac{n}{Y_{n+2N}}-1\right)\cdot\mathbb{I}\{Y_{n+2N}\leq n\}\middle| N\right]\right) \leq \mathbb{E}\left[\left(\frac{n}{Y_n}-1\right)\cdot\mathbb{I}\{Y_n\leq n\}\right]$$

where Y_{n+2N} is a central chi-square random variable with $(n+2N)$ degrees of freedom and the second inequality follows as for any $N \geq 0$, $(n/Y_{n+2N}-1)\cdot\mathbb{I}\{Y_{n+2N}\leq n\}$ is stochastically dominated by $N=0$. Now,

$$\mathbb{E}\left[\left(\frac{n}{Y}-1\right)\cdot\mathbb{I}\{Y\leq n\}N\right] = \int_0^n \frac{n-y}{y}\,\frac{y^{n/2-1}e^{-y/2}}{2^{n/2}\Gamma(n/2)}\,dy \leq \frac{E|W_n-n|}{n-2}$$

where $W_n \sim \text{Gamma}(n/2-1, 1/2)$ and so

$$E|W_n-n| \leq 1 + E(W_n-n)^2 \leq 1+4+Var(W_n) = 5+\sqrt{2n}$$

where the second inequality follows by Bias-Variance decomposition. Thus, we get our result. □

James-Stein Positive-Part Location Estimator

We consider the positive part of the JS estimator and a reasonable estimate of its loss as

$$\widehat{\theta}_n^{JS+} = X_n\left(1-\frac{n-2}{\|X_n\|^2}\right)_+ \quad \text{and} \quad U\big(\widehat{\theta}_n^{JS+}\big) = \left(n-\frac{(n-2)^2}{\|X_n\|^2}\right)_+.$$

There exists unbiased estimator of the quadratic risk of $\widehat{\theta}_n^{JS+}$ (Johnstone 2012, Exercise 2.13). We use a biased estimator here mainly to highlight the fact that even

biased estimators will work.

P1. follows from the fact that $\widehat{\theta}_n^{JS+}$ is better than $\widehat{\theta}^{JS_n}$(Johnstone 2012, Exercise 2.8).

For checking **P2**, define C_n to be the event $\{X_n : \widehat{\theta}^{JS+}(X_n) \neq 0\} = \{X_n : \widehat{\theta}_n^{JS+} = \widehat{\theta}_n^{JS}\}$. And the idea is to relate the variance of the loss in JS+ case with the case of JS estimator.

$$\begin{aligned}
&\mathbb{V}ar_{\theta_n}\big(\|\widehat{\theta}_n^{JS+} - \theta_n\|^2\big) = \mathbb{E}\|\widehat{\theta}_n^{JS+} - \theta_n\|^4 - \mathbb{E}^2\|\widehat{\theta}_n^{JS+} - \theta\|^2 \\
&= \mathbb{E}\big\{\|\widehat{\theta}_n^{JS} - \theta_n\|^4 I_{C_n}\big\} - \mathbb{E}^2\big\{\|\widehat{\theta}_n^{JS} - \theta_n\|^2 I_{C_n}\big\} + ||\theta_n||^4 P(C_n^c) - ||\theta_n||^4 P^2(C_n^c) \\
&= \mathbb{V}ar_{\theta_n}\big(\|\widehat{\theta}_n^{JS} - \theta_n\|^2 \big| C_n\big) \cdot P_{\theta_n}(C_n) + ||\theta_n||^4 P_{\theta_n}(C_n) P_{\theta_n}(C_n^c) \\
&\leq \mathbb{V}ar_{\theta_n}\big(\|\widehat{\theta}_n^{JS} - \theta_n\|^2\big) + ||\theta_n||^4 P_{\theta_n}(C_n^c) \\
&\text{as } \mathbb{V}ar_{\theta_n}\big(\|\widehat{\theta}_n^{JS} - \theta_n\|^2\big) \geq \mathbb{E}_{\theta_n}\Big(\mathbb{V}ar_{\theta_n}\big(\|\widehat{\theta}_n^{JS} - \theta_n\|^2 \big| C_n\big)\Big)
\end{aligned}$$

We know that $\mathbb{V}ar_{\theta_n}\big(\|\widehat{\theta}_n^{JS} - \theta_n\|^2\big)$ is $O(n)$ and lemma A.5 shows $||\theta_n||^4 P_{\theta_n}(C_n^c) \leq O(n)$. So, we have the desired bound.

Condition P.3.1 We will condition on the event C_n again and express **P.3.1** in terms of the James-Stein estimator

$$\mathbb{E}_{\theta_n} U(\widehat{\theta}_n^{JS+}) - q(\theta_n, \widehat{\theta}_n^{JS+}) = \mathbb{E}_{\theta_n}\left\{\left(U(\widehat{\theta}_n^{JS}) - \|\widehat{\theta}_n^{JS} - \theta_n\|^2\right) I_{C_n}\right\} - \|\theta_n\|^2 P\big(C_n^c\big).$$

When $\theta_n = 0$ then the R.H.S for large n reduces to,

$$I_n = E\left\{\left(n - \frac{n^2}{Y_n} - \left(1 - \frac{n}{Y_n}\right)^2 Y_n\right) I_{C_n}\right\}$$

where Y_n is an central chi-squared random variable with n degrees of freedom. Now, we decompose I_n into

$$I_n = I_n^1 + 2I_n^2 \text{ where}$$
$$I_n^1 = E\left\{(n - Y_n)\, I_{C_n}\right\} \text{ and } I_n^2 = E\left\{(n - n^2/Y_n)\, I_{C_n}\right\}$$

We standardize Y_n as $Z_n = (Y_n - n)/\sqrt{2n}$. We can use concentration inequalities on Z_n and have, [need to make rigorous]

$$I_n^1 \leq \sqrt{2n} \cdot EZ_+ \text{ as } n \to \infty$$
$$I_n^2 = \sqrt{n} \cdot E\left\{\left(\frac{Z_n}{Z_n/\sqrt{n}+1}\right) I_{C_n}\right\} \leq \sqrt{n}\, E|Z_n| \to \sqrt{n}E|Z|$$

as on C_n, $Z_n \geq 0$. Thus $I_n \leq O(\sqrt{n})$.

Condition P3.2 By Lemma A.8 we have

$$\mathbb{V}ar_{\theta_n}(\, U(\, \widehat{\theta}_n^{JS+}\,)\,) \leq \mathbb{V}ar_{\theta_n}(\, U(\, \widehat{\theta}_n^{JS}\,)\,) = O(n).$$

Condition P.3.3 Follows from Lemma A.1

Harmonic Prior Bayes Location Estimator

The mean of the posterior density in model **M.2** based on the harmonic prior π_H is given by

$$\widehat{\theta}_H(X_n) = X_n + \frac{\nabla m_H(X_n)}{m_H(X_n)} \text{ where } m_H(x_n) = \int ||\theta_n||^{2-p}\phi(x_n|\theta_n, 1)d\theta_n.$$

Based on Xu (2007, Lemma 3, Page 17) we have closed form expressions of the following quantities associated with the harmonic prior:

$$\textbf{If } n \textbf{ is even: } m_H(x_n) = ||x_n||^{2-n}\left(1 - e^{-||x_n||^2/2}\sum_{k=0}^{n/2-2}\frac{(||x_n||^2/2)^k}{k!}\right),$$

$$\nabla m_H(x_n) = x_n(2-n)||x_n||^{-n}\left[1 - e^{-||x_n||^2/2}\right],$$

$$\nabla^2 m_H(x_n) = -n(n-2)||x_n||^{-n}e^{-||x_n||^2/2}\frac{(||x_n||^2/2)^{n/2}}{(n/2)!}.$$

$$\textbf{If } n \textbf{ is odd: } m_H(x_n) = ||x_n||^{2-n}\left(2\Phi(||x_n||) - 1 - \sqrt{2/\pi}e^{-||x_n||^2/2}\sum_{k=0}^{(n-1)/2-2}\frac{||x_n||^{2k+1}}{(2k+1)!!}\right),$$

$$\nabla m_H(x_n) = x_n(2-n)||x_n||^{-n}\left[2\Phi(||x_n||) - 1 - \sqrt{\frac{2}{\pi}}\,e^{-||x_n||^2/2}||x_n||\right],$$

$$\nabla^2 m_H(x_n) = 2(2-n)||x_n||^{-(n-3)}\phi(||x_n||).$$

Using Equation 3.14, an unbiased estimate of the quadratic risk of $\widehat{\theta}_H$ is given by

$$\widehat{U}_H(x_n) = n - \frac{||\nabla m_H(x_n)||^2}{m_H^2(x_n)} + 2\frac{\nabla^2 m_H(x_n)}{m_H(x_n)}.$$

The RASL conditions conditions can be checked by using the positive part of $\widehat{U}_H$ as a suitable estimate of the quadratic risk of $\widehat{\theta}_H$.

3.2.3 Determining ρ_0 for RASL point estimators

In this section, we will show that in high dimension with very high precision we can express $\rho_0(\widehat{\theta}_n)$ – the minimum Predictive Entropy risk of the class of Gaussian density estimates around location $\widehat{\theta}_n$ in terms of the Mean Square estimation error of θ_n by $\widehat{\theta}_n$. We initially prove bounds on the error rates which holds for all dimensions but are dimension dependent. Then, we would show that in high dimensions those bounds are asymptotically sharp.

Lower Bound on $\rho_0(\theta_n, \widehat{\theta}_n)$:

Next, we produce a lower bound on the prediction error. The bound ultimately will be a function of θ_n though it depends on the form of $\widehat{\theta}_n$. It involves expectation of a quantity which usually is neither a parameter nor a statistic and hence can not be computed in closed form.

Lemma 3.2.5.

For any dimension n, any parameter value θ_n and any location point estimate $\widehat{\theta}\,(X_n)$, we have

$$\rho_0(\theta_n, \widehat{\theta}_n) \geq \frac{1}{2}\,\mathbb{E}_{\theta_n}\left\{\log\left(1 + r^{-1}\cdot n^{-1}\,\cdot\,\|\,\widehat{\theta}_n - \theta_n\|^2\right)\right\}.$$

Proof. For any fixed n, the risk of the predictive density q_n which is a n-dimensional product normal with with data adaptive mean $\widehat{\theta}\,(X_n)$ and data and parameter dependent variance $\widehat{c}\,(\theta_n, X_n)\,r$ (for all co-ordinates) is given by $2\rho(\theta_n, q_n)$

$$= n\left\{\mathbb{E}_{\theta_n}(\log\widehat{c}\,(\theta_n, X_n)) + \mathbb{E}_{\theta_n}\left(\frac{1}{\widehat{c}\,(\theta_n, X_n)}\right) - 1\right\} + \frac{1}{r}\mathbb{E}_{\theta_n}\left(\frac{||\widehat{\theta}\,(X_n) - \theta_n||^2}{\widehat{c}\,(\theta_n, X_n)}\right)$$

$$= n\,\mathbb{E}_{\theta_n}\left\{\log\widehat{c}\,(\theta_n, X_n) + \frac{1 + (nr)^{-1}||\,\widehat{\theta}\,(X_n) - \theta_n||^2}{\widehat{c}\,(\theta_n, X_n)} - 1\right\}.$$

For any fixed value of θ_n and for each x_n,

$$\log\widehat{c}\,(\theta_n, x_n) + \widehat{c}^{-1}(\theta_n, x_n)\{1 + (nr)^{-1}||\,\widehat{\theta}(x_n) - \theta_n||^2\} - 1$$

is minimized at $\widehat{c}^{\,\mathrm{opt}}\,(\theta_n, x_n) = 1 + (nr)^{-1}||\widehat{\theta}(x_n) - \theta_n||^2$ and the minimum value is given by $\log(1 + (nr)^{-1}\,\|\,\widehat{\theta}(x_n) - \theta_n\|^2)$. Hence, the result follows. □

Though $\widehat{c}^{\,\mathrm{opt}}\,(\theta_n, X_n)$ is the best possible flattening coefficient, it depends on the parameter and can not be used in practice. As such, $\widehat{c}^{\,\mathrm{opt}}\,(\theta_n, X_n)$ is the ideal flattening coefficient. In high dimensions due to statistical regularity we expect $\widehat{c}^{\,\mathrm{opt}}\,(\theta_n, X_n)$ to

be very close to its expected value

$$\mathbb{E}_{\theta_n}\{\widehat{c}^{\,\mathrm{opt}}(\theta_n, X_n)\} = 1 + (nr)^{-1}\mathbb{E}_{\theta_n}\|\widehat{\theta}_n - \theta_n\|^2$$

which can be viewed as the (near) **I**deal **F**lattening coefficient and is referred to as $\mathrm{IF}_{\theta_n}(\widehat{\theta}_n) = 1 + n^{-1}r^{-1}q(\theta_n, \widehat{\theta}_n)$. Here flattening coefficients are usually called scale and it should be noted that the corresponding variance needs to be multiplied by r.

From Lemma 3.2.5 we can derive a worse but more tractable bound

$$\rho_0(\theta_n, \widehat{\theta}_n) \geq 2^{-1}\,\mathbb{E}_{\theta_n}\left\{\log\left(\|\,\widehat{\theta}_n - \theta_n\|^2/(nr)\right)\right\}. \tag{3.16}$$

Upper Bound for $\rho_0(\theta_n, \widehat{\theta}_n)$:

We now produce an upper bound on the risk of any Gaussian density estimate. Henceforth, SD would mean Standard Deviation and by Bias of the scale estimate $\widehat{c}_n$ we would mean the expected deviation from the near ideal flattening coefficient $\mathrm{IF}_{\theta_n}(\widehat{\theta}_n)$. With scales estimators based on the statistic $U[\widehat{\theta}_n](X_n)$ and of the form $\widehat{c}\,(X_n) = (1 + n^{-1}U[\widehat{\theta}_n](X_n))$ we have $\mathrm{Bias}_{\theta_n}(c_n) = (nr)^{-1}\left[\mathbb{E}_{\theta_n}U[\widehat{\theta}_n](X_n) - q(\theta_n, \widehat{\theta}_n)\right]$.

Lemma 3.2.6.

For any fixed dimension n, parameter value θ_n, location point estimate $\widehat{\theta}(X_n)$ and any scale estimate $\widehat{c}\,(X_n) > 0$ almost surely and of the form $\widehat{c}\,(X_n) = 1 + (nr)^{-1}U[\widehat{\theta}_n](X_n)$, we have

$$\rho\big(\theta_n, g[\widehat{\theta}_n, \widehat{c}_n]\big) - \frac{n}{2}\cdot\log\left(\mathrm{IF}_{\theta_n}(\widehat{\theta}_n)\right) \leq \frac{n}{2}\cdot\left[A_{\theta_n}\big(\widehat{\theta}_n, \widehat{c}_n\big) + B_{\theta_n}\big(\widehat{\theta}_n, \widehat{c}_n\big)\right]$$

$$\text{where } A_{\theta_n}\big(\widehat{\theta}_n, \widehat{c}_n\big) = \mathrm{IF}_{\theta_n}\big(\widehat{\theta}_n\big)\left\{\mathbb{E}_{\theta_n}(\widehat{c}_n)\right\}^{-1}\mathrm{SD}_{\theta_n}(\widehat{c}_n)\mathrm{SD}_{\theta_n}(\widehat{c}_n^{-1})$$

$$+ r^{-1}SD_{\theta_n}\left(\frac{\|\widehat{\theta}_n - \theta_n\|^2}{n}\right)\mathrm{SD}_{\theta_n}\big(\widehat{c}_n^{-1}\big) \text{ and}$$

$$B_{\theta_n}\big(\widehat{\theta}_n, \widehat{c}_n\big) = \mathrm{Bias}^2_{\theta_n}\big(\widehat{c}_n\big)\left\{\mathrm{IF}_{\theta_n}(\widehat{\theta}_n)\right\}^{-1}\left\{\mathbb{E}_{\theta_n}(\widehat{c}_n)\right\}^{-1}.$$

Proof. The risk of the normal predictive density estimate $g[\widehat{\theta}_n\,,\,\widehat{c}_n\,]$ is given by $2\rho(\theta_n, \widehat{g}_n)$

$$= n\left\{\mathbb{E}_{\theta_n}(\log\widehat{c}(X_n)) + \mathbb{E}_{\theta_n}\left(\frac{1}{\widehat{c}(X_n)}\right) - 1\right\} + \mathbb{E}_{\theta_n}\left(\frac{||\,\widehat{\theta}\,(X_n) - \theta_n||^2}{r\,\widehat{c}(X_n)}\right).$$

Now, we replace $\mathbb{E}_{\theta_n}(\widehat{c}_n^{-1})$ by $\mathbb{E}_{\theta_n}^{-1}\widehat{c}_n$ and $\mathbb{E}_{\theta_n}\big(\,||\widehat{\theta}_n - \theta_n||^2 \times \widehat{c}_n^{-1}\,\big)$ by $\mathbb{E}_{\theta_n}||\widehat{\theta}_n - \theta_n||^2 \times \mathbb{E}_{\theta_n}^{-1}\widehat{c}_n$ in the above expression to get $\tilde{\rho}(\theta_n, \widehat{g}_n)$

$$\begin{aligned}
&= \frac{1}{2}\left[n\left\{\mathbb{E}_{\theta_n}(\,\log\widehat{c}\,(X_n)\,) + \frac{1}{\mathbb{E}_{\theta_n}(\,\widehat{c}\,(X_n)\,)} - 1\right\} + \frac{1}{r}\left(\frac{\mathbb{E}_{\theta_n}||\widehat{\theta}\,(X_n) - \theta_n||^2}{\mathbb{E}_{\theta_n}(\,\widehat{c}\,(X_n)\,)}\right)\right] \\
&= \frac{n}{2}\mathbb{E}_{\theta_n}(\,\log\widehat{c}\,(X_n)\,) + \frac{n}{2}\left\{\frac{1 + (nr)^{-1}\mathbb{E}_{\theta_n}\|\widehat{\theta}\,(X_n) - \theta_n\|^2}{\mathbb{E}_{\theta_n}(\,\widehat{c}\,(X_n)\,)} - 1\right\} \qquad (3.17) \\
&= \frac{n}{2}\mathbb{E}_{\theta_n}(\,\log\widehat{c}\,(X_n)\,) - \frac{n}{2}\frac{\mathrm{Bias}_{\theta_n}\big(\widehat{c}\,(X_n)\big)}{\mathbb{E}_{\theta_n}(\,\widehat{c}\,(X_n)\,)}
\end{aligned}$$

and the distortion caused thereby $(n/2)^{-1}(\rho(\theta_n, \widehat{g}_n) - \tilde{\rho}(\theta_n, \widehat{g}_n))$ equals

$$\mathbb{E}_{\theta_n}\left(\frac{1 + (nr)^{-1}||\widehat{\theta}\,(X_n) - \theta_n||^2}{\widehat{c}\,(X_n)}\right) - \frac{1 + (nr)^{-1}\mathbb{E}_{\theta}||\widehat{\theta}\,(X_n) - \theta_n||^2}{\mathbb{E}_{\theta_n}(\,\widehat{c}\,(X_n)\,)}. \qquad (3.18)$$

Next we will show that $(n/2)^{-1}|r(\theta_n, \widehat{g}_n) - \tilde{r}(\theta_n, \widehat{g}_n)| \leq A_{\theta_n}(\widehat{\theta}_n, \widehat{c}_n)$. Before that, note that if $\widehat{U}_n$ is unbiased then the second term in Equation 3.17 vanishes and we have the result stated in Corollary 3.2.7.
Now, note that $2n^{-1}\,\tilde{r}(\theta_n, \widehat{g}_n)$ equals

$$\log \mathrm{IF}_{\theta_n}(\widehat{\theta}_n) + \mathbb{E}_{\theta_n}\left[\log\left(1 + \frac{\widehat{c}\,(X_n) - \mathrm{IF}_{\theta_n}(\widehat{\theta}_n)}{\mathrm{IF}_{\theta_n}(\widehat{\theta}_n)}\right)\right] - \frac{\mathrm{Bias}_{\theta_n}\big(\widehat{c}\,(X_n)\big)}{\mathbb{E}_{\theta_n}(\,\widehat{c}\,(X_n)\,)}$$

and using the inequality $\log(1 + x) \leq x$ for all $x > -1$ on the second term on the

right hand side it follows that

$$2n^{-1}\,\tilde{r}(\theta_n,\widehat{g}_n) \le \log \mathrm{IF}_{\theta_n}(\widehat{\theta}_n) + \frac{\mathrm{Bias}_{\theta_n}\big(\widehat{c}\,(X_n)\big)}{\mathrm{IF}_{\theta_n}(\widehat{\theta}_n)} - \frac{\mathrm{Bias}_{\theta_n}\big(\widehat{c}\,(X_n)\big)}{\mathbb{E}_{\theta_n}(\,\widehat{c}\,(X_n)\,)}$$
$$= \log \mathrm{IF}_{\theta_n}(\widehat{\theta}_n) + \frac{\mathrm{Bias}^2_{\theta_n}\big(\widehat{c}\,(X_n)\big)}{\mathrm{IF}_{\theta_n}(\widehat{\theta}_n)\mathbb{E}_{\theta_n}(\,\widehat{c}\,(X_n)\,)} = B_n.$$

Now, we write $2n^{-1}\{r(\theta_n,\widehat{g}_n) - \tilde{r}(\theta_n,\widehat{g}_n)\} = H_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n) + J_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)$ where,

$$H_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n) = \mathrm{IF}_{\theta_n}\big(\widehat{\theta}_n\big)\left\{\mathbb{E}_{\theta_n}\left(\frac{1}{\widehat{c}_n}\right) - \frac{1}{\mathbb{E}_\theta \widehat{c}_n}\right\} \quad \text{and}$$
$$J_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n) = (nr)^{-1}\cdot\mathbb{E}_{\theta_n}\left[||\widehat{\theta}_n - \theta_n||^2\left\{\frac{1}{\widehat{c}_n} - \mathbb{E}_{\theta_n}\left(\frac{1}{\widehat{c}_n}\right)\right\}\right].$$

Note that the second term in $H_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)$ can be rewritten as,

$$\mathbb{E}_{\theta_n}\left(\frac{1}{\widehat{c}_n}\right) - \frac{1}{\mathbb{E}_{\theta_n}\widehat{c}_n} = \frac{-1}{\mathbb{E}_\theta \widehat{c}_n}\cdot\mathbb{E}_{\theta_n}\left[\left(\widehat{c}_n - \mathbb{E}_{\theta_n}\widehat{c}_n\right)\left(\frac{1}{\widehat{c}_n} - \mathbb{E}_{\theta_n}\left(\frac{1}{\widehat{c}_n}\right)\right)\right]$$

which by Cauchy-Schwartz (C-S) inequality has lower absolute value than

$$(\mathbb{E}_{\theta_n}\widehat{c}_n)^{-1}\left\{\mathbb{V}ar_{\theta_n}(\widehat{c}_n)\times\mathbb{V}ar_{\theta_n}\big(\widehat{c}_n^{-1}\big)\right\}^{1/2}.$$

Thus, $|H_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)| \le \mathrm{IF}_{\theta_n}\big(\widehat{\theta}_n\big)\mathbb{E}_{\theta_n}\widehat{c}_n)^{-1}\mathrm{SD}_{\theta_n}(\widehat{c}_n)\times\mathrm{SD}_{\theta_n}\big(\widehat{c}_n^{-1}\big)$.
Again, rewriting $J_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)$ as,

$$J_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n) = (nr)^{-1}\mathbb{E}_{\theta_n}\left[\left\{||\widehat{\theta}_n - \theta_n||^2 - \mathbb{E}_{\theta_n}||\widehat{\theta}_n - \theta_n||^2\right\}\left\{\widehat{c}_n^{-1} - \mathbb{E}_{\theta_n}\big(\widehat{c}_n^{-1}\big)\right\}\right]$$

and applying C-S inequality we get

$$|J_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)| \le (nr)^{-1}\mathrm{SD}_{\theta_n}\left(\|\widehat{\theta}_n - \theta_n\|^2\right)\cdot\mathrm{SD}_\theta\left(\widehat{c}_n^{-1}\right).$$

So $(n/2)^{-1}\,|r(\theta_n,\widehat{g}_n) - \tilde{r}(\theta_n,\widehat{g}_n)| \le |H_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)| + |J_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)| \le A_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)$ and we have our desired result. □

Corollary 3.2.7.

If $\widehat{U}_n$ is an unbiased estimate of the parameter $q(\theta_n, \widehat{\theta}_n)$ and $\widehat{c}_n = 1 + (nr)^{-1}\widehat{U}_n > 0$ almost surely, then we have,

$$\rho_0(\theta_n, \widehat{\theta}_n) \leq \rho(\theta_n, g[\widehat{\theta}_n\,,\, \widehat{c}_n]) \leq \frac{1}{2}\,\mathbb{E}_{\theta_n}\left\{\log\left(1 + (nr)^{-1}\widehat{U}_n\right)\right\} + A_{\theta_n}(\widehat{\theta}_n, \widehat{c}_n)/2.$$

The corollary follows from the above Lemma. The upper bound derived here involves expectation of a statistic along with a distortion term $A_{\theta_n}(\widehat{\theta}_n, \widehat{c}_n)$ which will be negligible under the RASL conditions. Ignoring it for the time being we can say that an upper bound is produced when $||\widehat{\theta}_n - \theta_n||^2$ in the lower bound of Lemma 3.2.6 can be replaced by a good statistic. Lemma 3.2.6 has an upper bound based on $\mathrm{IF}_{\theta_n}(\widehat{\theta}_n)$ and next we show that the lower bound and the upper bound are fairly close.

Lemma 3.2.8.

For any point estimate $\widehat{\theta}_n$ and location parameter $\theta_n \in \mathbb{R}^n$ we have,

$\rho_0(\theta_n,\, \widehat{\theta}_n)) \geq 2^{-1} \log \mathrm{IF}_{\theta_n}\big(\widehat{\theta}_n\big) - L_{\theta_n}(\widehat{\theta}_n)/2$ where,

$$L_{\theta_n}(\widehat{\theta}_n) = \big(nr\big)^{-1} \cdot SD_{\theta_n}\big(\|\widehat{\theta}_n - \theta_n\|^2\big) \cdot SD_{\theta_n}\left\{\left(1 + (nr)^{-1}\,\|\widehat{\theta}_n - \theta_n\|^2\right)^{-1}\right\}$$

Proof. From Lemma 3.2.5 we have

$$\begin{aligned}\log \mathrm{IF}_{\theta_n}\big(\widehat{\theta}_n\big) - 2\rho_0(\theta_n, \widehat{\theta}_n) &\leq \log \mathrm{IF}_{\theta_n}\big(\widehat{\theta}_n\big) - \mathbb{E}_{\theta_n}\big\{\,\log\big(1 + (nr)^{-1}\|\,\widehat{\theta}_n - \theta_n\|^2\,\big)\big\} \\ &= \mathbb{E}_{\theta_n}\left\{\log\left(1 - \frac{\bar{l}(\theta_n, \widehat{\theta}_n)}{nr + \|\,\widehat{\theta}_n - \theta_n\,\|^2}\right)\right\}\end{aligned}$$

where $\bar{l}(\theta_n, \widehat{\theta}_n) = \|\,\widehat{\theta}_n - \theta_n\,\|^2 - q(\theta_n, \widehat{\theta}_n)$and using Jensen's inequality and $\log(1+x) \leq x$

consecutively, the difference becomes

$$\leq -\mathbb{E}_{\theta_n}\left(\frac{\bar{l}(\theta_n,\widehat{\theta}_n)}{nr+\|\,\widehat{\theta}_n-\theta_n\,\|^2}\right)$$
$$= -\mathbb{E}_{\theta_n}\left[\bar{l}(\theta_n,\widehat{\theta}_n)\cdot\left\{\frac{1}{nr+\|\,\widehat{\theta}_n-\theta_n\,\|^2}-\mathbb{E}_{\theta_n}\left(\frac{1}{nr+\|\,\widehat{\theta}_n-\theta_n\,\|^2}\right)\right\}\right]$$

and by applying C-S inequality the magnitude of the said difference is

$$\leq \mathrm{SD}_{\theta_n}(\|\,\widehat{\theta}_n-\theta_n\,\|^2)\times\mathrm{SD}_{\theta_n}\left\{(nr+\|\,\widehat{\theta}_n-\theta_n\,\|^2)^{-1}\right\}=L_{\theta_n}(\widehat{\theta}_n).$$

This completes the proof. □

Corollary 3.2.9.

Under the conditions of Lemma 3.2.6 we have

$$[i.]\ 0\leq\rho\big(\theta_n,g[\widehat{\theta}_n,\widehat{c}_n\,]\big)-\rho_0(\theta_n,\widehat{\theta}_n)\leq 2^{-1}\left\{L_{\theta_n}(\widehat{\theta}_n)+[A+B]_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)\right\}$$
$$[ii.]\ \left|\rho_0(\theta_n,\widehat{\theta}_n)-2^{-1}\log\mathrm{IF}_{\theta_n}(\widehat{\theta}_n)\right|\leq 2^{-1}\max\left\{L_{\theta_n}(\widehat{\theta}_n),[A+B]_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)\right\}.$$

The corollary follows directly by combining the above lemma with Lemma 3.2.6. It bounds the deviation of the predictive risk from a continuous, increasing function of the MSE . The RASL conditions ensure the existence of at least one candidate for the statistic $\widehat{U}_n$ such that $c(X_n)>0$ almost surely (follows from RASL condition P3.3) and each of the associated terms $A_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n), B_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n)$ and $L_{\theta_n}(\widehat{\theta}_n)$ is of the order of $O(n^{-1})$. Hence, Theorem 3.1.1 follows.

Proof of Theorem 3.1.1. Note that under the RASL conditions we have $A_{\theta_n}(\widehat{\theta}_n,\widehat{c}_n), B_n$ and $L_{\theta_n}(\widehat{\theta}_n)$ to be of the order of $O(n^{-1})$. Also, note that the fact that $c>0$ almost surely is taken care in the the RASL property P3.3. □

3.2.4 Violation of RASL conditions

Based on the lower bound of 3.2.5 and concentrating around the expectation by using Chebyshev's inequality we have for any a in $(0,1)$

$$\rho_0(\theta_n, \widehat{\theta}_n) \geq \frac{1}{2}\log\left(1 + \frac{(1-a)\, q(\theta_n, \widehat{\theta}_n)}{nr}\right)\left\{1 - \frac{a^2\, \mathbb{V}ar_{\theta_n}(||\widehat{\theta}_n - \theta_n||^2)}{q^2(\theta_n, \widehat{\theta}_n)}\right\}.$$

So if **P1** of the RASL condition is violated i.e for some $\theta'_n \in \mathbb{R}^n$ we have $q(\theta'_n, \widehat{\theta}_n) > O(n)$ then if **P2** holds or we have $\mathbb{V}ar_{\theta'_n}(||\widehat{\theta}_n - \theta'_n||^2) < q(\theta'_n, \widehat{\theta}_n)$ then $\rho_0(\theta'_n, \widehat{\theta}_n) > 1/2\log(1 + r^{-1})$ which is the minimax risk of the best invariant density estimate and so the class of density estimates in $\mathcal{G}[1]$ centered around $\widehat{\theta}$ does not have any minimax estimator. Thus, we can exclude bad point estimators in most conditions (also see Equation 3.16).

Among the cases where RASL conditions does not hold the only exciting case is when P2 is violated but P1 holds. In those cases the asymptotic predictive entropy risk can not be characterized in closed form. A example of a point estimator of this kind is:

$$\delta_n(i) = \begin{cases} \delta_1(X_1) & \text{if } i = 1 \\ X_i & \text{if } i = 2, \cdots, n \end{cases}$$

where the univariate point estimator δ_1 is given by

$$\delta_1(x) = \begin{cases} n^{1/2}\,(2\log n)^{-1/2}\, x & \text{if } x < (2\log n)^{1/2} \\ x & \text{if } x \geq (2\log n)^{1/2} \end{cases}$$

3.3 Decision Theoretic implications of optimal flattening

The asymptotic relation between the predictive risk and the mean square risk will help us in deriving oracle inequalities on the predictive risk of g_n. The bounds will be sharp enough to discuss asymptotic optimality in the class $\mathcal{G}$. We would first relate

the class $\mathcal{G}$ with the other decision-theoretic classes of predictive densities. Then, we would compare the predictive risk of the respective classes in unrestricted parametric spaces.

In the above context, we consider the following 6 predictive estimates:

- $\widehat{\mathbf{p}}_{\mathbf{L}}$: As an representative of the class of all Linear predictive density estimates ($\mathcal{L}$) we choose the predictive density $g[X_n,\, 1+r]$. It is the Bayes predictive density with respect to the uniform prior, has constant risk and is inadmissible in $\mathcal{L}$. It is the best invariant predictive strategy and is also minimax among all procedures(Liang & Barron 2004).

- $\widehat{\mathbf{p}}_{\mathbf{E}}$: We choose the James-Stein positive part plug-in predictive density estimate $g[\widehat{\theta}_n^{JS+},\, r]$ as a representative of $\mathcal{P}$. Though the positive part James-Stein estimator is inadmissible as a point estimate, it is difficult to find estimators that have significant improvements over it. And, for all practical purposes the JS+ estimator can be considered as a 'nearly' admissible point estimate. In that respect we can consider

$$\widehat{\mathbf{p}}_{\mathbf{E}} = \mathbf{g}[\widehat{\theta}^{\,\mathbf{JS+}},\, \mathbf{r}] \;\text{ where }\; \widehat{\theta}_n^{JS+} = X_n \left(1 - \frac{(n-2)}{\|X_n\|^2}\right)_+$$

  as an efficient representative from the class of Plug-in predictive densities ($\mathcal{P}$). The subscript stands for the class of estimative (plug-in) distributions.

- $\widehat{\mathbf{p}}_{\mathbf{H}}$: We consider the Bayes predictive density estimate from the harmonic prior π_H as a representative of the class of all Bayes predictive density estimates ($\mathcal{B}$). It is an admissible rule. As such, it also dominates $\widehat{p}_L$(Komaki 2001, Ghosh et al. 2008).

- Next, we consider 3 member of $\mathcal{G}$ which we will use to compare the risk of the predictive densities from the above 3 classes.

- $\mathbf{g}[\widehat{\theta}^{\mathbf{JS+}}, \mathbf{1}+\mathbf{r}]$: A non-linear, fixed variance predictive density estimator around the JS+ estimator. It is uniformly better than $\widehat{p}_L$. It is also denoted by g_M.
- $\mathbf{g}[\widehat{\theta}^{\mathbf{JS+}}]$: The optimal member in $\mathcal{G}(\widehat{\theta}^{JS+})$ which we will use to compare with $\widehat{p}_E$ and $\widehat{p}_L$.
- $\mathbf{g}[\widehat{\theta}^{\mathbf{H}}]$: The optimal member in $\mathcal{G}(\widehat{\theta}^{H})$. We would like to compare its performance with $\widehat{p}_H$. Also, $g[\widehat{\theta}^{H}]$ is asymptotically inadmissible among the procedures in $\mathcal{G}$.

In Table 3.1 we evaluate the predictive performance of each of these density estimates on a data set.

Oracle inequalities and Implications

Lemma 3.1.3 describes the predictive risk of density estimates center around $\widehat{\theta}^{JS}$.

Proof of Lemma 3.1.3. The results follows from Theorem 3.1.1 and by using Proposition 2.6 and Exercise 2.8 of (Johnstone 2012) □

The lemma will not be useful in very very low signal-to-noise ratio. It can be used effectively when $a_n > O(n^{-1})$. Note that, we can partition the improvement in the asymptotic prediction error over $\widehat{p}_L$ in two parts.

- We first shrink the location estimate while keeping the scale unperturbed and move to a better estimate $g[\widehat{\theta}^{JS+}, 1+r]$. The improvement is denoted by d_n^1.
- Next, we optimize the scale keeping the location fixed and arrive at $g[\widehat{\theta}^{JS+}]$. The improvement at this stage is denoted by d_n^2.

And, based on the above lemma we have,

$$d_n^1 \sim \frac{1}{2}\,\alpha_n \text{ and } d_n^2 \sim \frac{1}{2}\,\log(1-\alpha_n)^{-1} \text{ where } \quad \alpha_n = \{(1+a_n)(1+r)\}^{-1}.$$

As $\alpha_n < 1$, d_n^1, d_n^2 as well as $d_n^2 - d_n^1$ are all positive and increasing in α_n. It means we are actually making more improvement by adapting the scale than that we got by shifting location and their difference is also decreasing in both a_n and r.

Prediction error for shrinkage estimators

By shrinkage point estimators we define estimators of the form $s(X_n)\, X_n$ where $s(X_n)$ is an almost everywhere differentiable function. If $||\theta_n||^2$ were known, then spherically symmetric shrinkage estimators of the form $s(a_n)\, X_n$ where $a_n = ||\theta_n||^2/n$ and $s(a_n) \leq 1$ would be efficient. Let $\mathcal{S}$ denotes the class of normal predictive densities based on ideal point location estimators. Such an estimate satisfies the RASL condition P2 and so Lemma 3.2.8 can be used to calculate an optimal lower bound on the predictive risk of the family of density estimators based on S – the class of all shrinkage point estimators conditioned on a_n.

Note that by Bias-Variance decomposition the quadratic risk of the ideal point estimator $s(a_n)\, X_n$ is given by

$$\mathbb{E}_{\theta_n}\big(\|s_n\, X_n - \theta_n\|^2\big) = s_n^2\, n + \bar{s}_n^2 ||\theta_n||^2 \text{ where } \bar{s}_n = 1 - s_n \text{ and } s_n = s(a_n).$$

Based on Lemma 3.5.3 we have $L_{\theta_n}(\widehat{\theta}_n) \leq (nr)^{-2}\mathbb{V}ar_{\theta_n}(\|\widehat{\theta}_n - \theta_n\|^2)$ and for an estimator in S we have,

$$\begin{aligned}
&\mathbb{V}ar_{\theta_n}\big(\|s_n\, X_n - \theta_n\|^2\big) = \mathbb{V}ar_{\theta_n}(s_n^2||X_n - \theta_n||^2 + 2\, s_n\, \bar{s}_n\langle X_n - \theta_n, \theta_n\rangle) \\
&\leq 2\big[s_n^4\, \mathbb{V}ar_{\theta_n}\big(\|X_n - \theta_n\|^2\big) + 4\, s_n^2\, \bar{s}_n^2\, \mathbb{V}ar_{\theta_n}(\langle X_n - \theta_n, \theta_n\rangle)\big] \\
&= 2\, n\, s_n^2\, [\, s_n^2 + 4\, \bar{s}_n^2\, a_n\,]
\end{aligned}$$

which is obviously less than $O(n)$ if $\bar{s}_n^2\, a_n = O(1)$. Otherwise,

$$\begin{aligned}
\rho_0(\theta_n, s(a_n)X_n) &\geq 2^{-1}\mathbb{E}_{\theta_n} \log(||s_n X_n - \theta_n||^2/(nr)) \\
&\geq \mathbb{E}_{\theta_n} \log |\bar{s}_n||\theta_n|| - s_n\chi_n|/(\sqrt{nr}) \to \infty \text{ as } n \to \infty
\end{aligned}$$

and thus the optimal error in $\mathcal{S}$ is attained at $\mathrm{IL}(\theta_n)$ as defined in Equation (3.9).

Dimension independent bounds

For any location point estimator $\widehat{\theta}_n$ which obeys the RASL conditions, the difference between the predictive risk of its associated optimally-flattened gaussian density estimate $g[\widehat{\theta}_n]$ and the ideal linear predictive risk can be upper bounded by dimension-free quantities. These dimension-less bounds can be constructed by substituting explicit bounds (independent of n, θ_n and $\widehat{\theta}_n$) of the following 3 quantities in Corollary 3.2.9: $L_{\theta_n}(\widehat{\theta}_n)$, $A_{\theta_n}(\widehat{\theta}_n, \widehat{c}_n)$ and $B_{\theta_n}(\widehat{\theta}_n, \widehat{c}_n)$.
Now note that by construction we have $\mathbb{E}(\widehat{c}_n^{-1}) \leq 1$ and the ideal flattening coefficient $\mathrm{IF}_{\theta_n}(\widehat{\theta}_n^{JS}) \leq (1 + r^{-1})$. Also, from Lemma 3.5.3 it follows that $\mathrm{SD}_{\theta_n}(c_n^{-1}) \leq (nr)^{-1}\,\mathrm{SD}_{\theta_n}(\widehat{U})$. So, we get

$$\begin{aligned} nA_{\theta_n}(\widehat{\theta}_n, \widehat{c}_n)\big\} \leq\ & r^{-2}\, n^{-1}\mathrm{IF}_{\theta_n}(\widehat{\theta}_n^{JS})\, \mathbb{V}ar_{\theta_n}(\widehat{U}) \\ & + r^{-3/2}\, n^{-1}\, \mathrm{SD}_{\theta_n}(||\widehat{\theta}_n - \theta_n||^2/n)\, \mathrm{SD}_{\theta_n}(\widehat{U}) \\ nB_{\theta_n}(\widehat{\theta}_n, \widehat{c}_n)\big\} \leq\ & r^{-2}\, n^{-1}\, \mathrm{Bias}^2(\widehat{U}_n), \\ nL_{\theta_n}(\widehat{\theta}_n) \leq\ & r^{-2}\, n^{-1}\, \mathbb{V}ar_{\theta_n}(||\widehat{\theta}_n - \theta_n||^2). \end{aligned}$$

In the previous section for the JS estimator, we established upper bounds in terms of n and a_n for each of the terms in the RHS of the above bounds. With those, we from crude upper bounds which depends on n only and then maximizing them over dimension $n \geq 20$ produced the upper bound displayed in Equation 3.10.

Nature of shrinkage

The plug-in estimate $\widehat{p}_{\mathbb{E}}$ performs better than $\widehat{p}_L$ and $g[\widehat{\theta}^{\,JS+},\, 1 + r]$ when a_n is close to 0 but gets dominated with increasing values of a_n. And, $g[\widehat{\theta}^{\,JS+}]$ is asymptotically better than $\widehat{p}_{\mathbb{E}}$ throughout. The relationship between $g[\widehat{\theta}^{\,H}]$ and $\widehat{p}_H$ can not be expressed explicitly. By Lemma 2.4.3 we can express the risk of $\widehat{p}_H$ as:

$$\rho(\theta_n, \widehat{p}_H) = \frac{1}{2} \int_{(1+r^{-1})^{-1}}^{1} v^{-1} q(\, \theta_n/v,\, \widehat{\theta}^{\,H}\,)\, dv \tag{3.19}$$

where $\widehat{\theta}^H$ denotes the posterior mean of the Harmonic prior. Equation 3.19 can be used to numerically evaluate the risk of $\widehat{p}_H$ as the risk of $\widehat{\theta}_H$ has closed form. The fact that these estimators are spherically symmetric will also help. We also get the following crude bound

$$C \inf_{\beta_n \in A(\theta_n)} q(\, \theta_n/v, \, \widehat{\theta}^H \,) \leq \rho(\theta_n, \widehat{p}_H) \leq C \sup_{\beta_n \in A(\theta_n)} q(\, \theta_n/v, \, \widehat{\theta}^H \,)$$

where $A(\theta_n) = \left\{ \, \beta_n = k \, \theta_n : 1 \leq k \leq \sqrt{1 + r^{-1}} \, \right\}$ and $C = \log(1 + r^{-1})/2$.

Minimaxity over Unrestricted Spaces

For any dimension n, $\widehat{p}_L$ is a minimax estimator. However, in dimensions greater than 2, $\widehat{p}_L$ is inadmissible and so there exists improved minimax estimators. $\widehat{p}_H$ is an improved minimax estimator than $\widehat{p}_L$ for $n \geq 3$. $g[\widehat{\theta}^{JS+}]$ is also an asymptotic minimax estimator and with huge improvements over $\widehat{p}_L$ which can also be explicitly quantified. Using Theorem 3.1.1, asymptotically minimax predictive density estimates can be constructed around minimax location estimates.

3.3.1 An illustration with a Dataset

We consider the Baseball data that was used to show the advantage of shrinking location estimates in Efron & Morris (1977). The data set consists of 18 players (so, $n = 18$ which is not so high dimensions) with exactly 45 at-bats on a particular date during the 1970 season. The objective is to predict the performance of the players on the remainder of the season .

The number of hits (H) and the number of at-bats (N) over two portions of the season were

$$H_{ji} \overset{ind.}{\sim} \text{Binomial}(N_{ji}, p_i), \quad j = 1, 2; \quad i = 1, \ldots, n.$$

Where $j = 1$ denotes past data and $j = 2$ represents the unknown future. As the variance of the Binomial model depends of the mean parameter p_i, a variance stabilization transformation (Brown 2008) is conducted (which goes through as N_{ij}

are quite large). The transformation

$$X_{ji} = \arcsin\left(\frac{H_{ji}+1/4}{N_{ji}+1/2}\right)^{1/2} \tag{3.20}$$

reduces the binomial model to the normal model

$$X_{ji} \sim N(\theta_i, \sigma_{ji}^2) \text{ where } \theta_i = \arcsin\sqrt{p_i}\,,\ \sigma_{ij}^2 = \big(4N_{ji}\big)^{-1} \tag{3.21}$$

and $X_{\cdot i}$ independent for $1 \le i \le n$. With the past $P = X_{1\cdot}$ and the future $F = X_{2\cdot}$ we have the following predictive set-up :

$$F|\theta_n \sim N\big(\theta_n, v_y I_n\big)\ ; \quad P|\theta_n \sim N(\theta_n, v_x I_n). \tag{3.22}$$

We want joint predictive densities of the future performances of players in this standardized model. We use a very naive evaluation strategy by considering the entire season's batting average as the true parametric value. In the entire season the players ended up playing around 400 games on the average. So, evaluating the predictive densities at $\theta_i^0 = \arcsin\big(p_i^{\text{full}}\big)^{1/2}$ where p_i^{full} are the batting averages from the entire season will not be terrible. Evaluation procedures with guarantees may be developed in a sequential set-up (Lai et al. 2011). While using shrinkage on the location estimators we shrink towards the grand average. We evaluate the 6 different predictive strategies of Section 3.3 for different values of the future to past variability. The value of r will be close to 0.1 when we consider prediction on the entire remaining half of the season.

We find that for most choices of r, $\widehat{p}_H$ is the best one among the 6 estimators considered. However, the losses for $g[\widehat{\theta}^{\,H}]$ and $g[\widehat{\theta}^{\,JS+}]$ are similar to that of $\widehat{p}_H$ and always very close to the minimal loss. Also, $d_n^2 - d_n^1$ (as discussed in Section 3.3) is decreasing in r. The JS+ plug-in estimator $\widehat{p}_E$ behaves well when r is large and horribly for small values.

r	$\widehat{p}_P$	$\widehat{p}_L$	$g[\widehat{\theta}^{JS+}]$	$g[\widehat{\theta}^{H}]$	$\widehat{p}_H$
0.1	22.963	19.451	11.435	11.465	11.456
0.2	11.482	14.174	7.418	7.432	7.395
0.5	4.593	8.326	3.717	3.758	3.622
1	2.296	5.067	2.047	2.086	1.957
2	1.148	2.868	1.081	1.092	1.023
5	0.459	1.250	0.448	0.454	0.421
10	0.23	0.645	0.227	0.231	0.244

Table 3.1: Predictive loss for the different predictive density estimates as r varies.

3.4 Predictive risk of density estimates in $\mathcal{G}[p]$

A typical member in the class $\mathcal{G}[p]$ of all product Gaussian predictive densities is represented by $g[\widehat{\theta}_n, \widehat{D}_n] = \prod_{i=1}^n N(\widehat{\theta}(i), \widehat{d}(i)\ \sigma_f^2)$. Generalizing the argument in Lemma 3.2.5 we see that a lower bound on the minimum predictive risk $\rho_p(\theta_n, \widehat{\theta}_n)$ of all density estimates in $\mathcal{G}_n[p]$ that have mean $\widehat{\theta}_n$, is given by

$$\rho_p(\theta_n, \widehat{\theta}_n) \geq \frac{1}{2} \sum_{i=1}^n \mathbb{E}_{\theta(i)} \big\{ \log\big(1 + r^{-1}(\widehat{\theta}(i) - \theta(i))^2\big) \big\}. \tag{3.23}$$

The predictive risk of the estimate $g[\widehat{\theta}_n, \widehat{D}_n]$ is given by

$$2\rho(\theta_n, g[\widehat{\theta}_n, \widehat{D}_n]) = \sum_{i=1}^n \mathbb{E}_{\theta_n} \log(\widehat{d}(i)) + \mathbb{E}_{\theta_n} \sum_{i=1}^n \left\{ \frac{1 + (\widehat{\theta}(i) - \theta(i))^2) - \widehat{d}(i)}{\widehat{d}(i)} \right\}. \tag{3.24}$$

It is not necessarily true that

$$\rho_p(\theta_n, \widehat{\theta}_n) = \min_{\widehat{D}_n \in \mathbb{R}_+^n} \rho(\theta_n, g[\widehat{\theta}_n, \widehat{D}_n])$$

asymptotically equals the lower bound given in Equation (3.23). In the previous section we saw that under sufficient regularity conditions these bounds matches. The ideas there can be extended to block-wise estimators and to non-orthogonal models by using the concept of Mallow's unbiased risk estimates. In the ℓ_0 sparse predictive space as the degree of sparsity tends to zero, i.e., $s/n \to 0$ as $n \to \infty$, the lower bound given in Equation (3.23) is significantly greater than the minimax predictive risk of the class $\mathcal{G}[p]$. And so, procedure used in the previous section can not be used for finding the asymptotic minimax predictive Gaussian risk over sparse parameter spaces. In the following chapter in Section 4.7.2 we calculate the minimax predictive risk over $\mathcal{G}[p]$.

3.5 Appendix

Lemma 3.5.1.

Y_n is sequence of random variables such that $Y_n \stackrel{d}{=} \chi^2_n(\lambda_n)$ for a non-negative and increasing sequence $\{\lambda_n : n \geq 1\}$ then for $n \geq 5$ we have

$$\mathbb{V}ar\big(Y_n^{-1}\big) \leq k_1(n) \cdot n^{-3} \text{ where } k_1(n) = 3\,(1-2/n)^{-2}(1-4/n)^{-1}.$$

Proof. We observe that Y_n being a non-central chi-square random variable can be written as convolution of central Chi-square and Poisson random variables

$$Y_n \stackrel{d}{=} \chi^2_{n+2N} \text{ where } N_n \stackrel{d}{=} \text{Poisson}(\lambda_n/2).$$

Decomposing the variance by conditioning on the Poisson random variable we have,

$$\begin{aligned}\mathbb{V}ar\left(Y_n^{-1}\right) &= \mathbb{V}ar_{\lambda_n}\left(E\left(Y_n^{-1}|N_n\right)\right) + E_{\lambda_n}\left(\mathbb{V}ar\left(Y_n^{-1}|N_n\right)\right) \\ &= \mathbb{V}ar_{\lambda_n}\left(\frac{1}{n+2N_n-2}\right) + E_{\lambda_n}\left(\frac{2}{(n+2N_n-2)^2(n+2N_n-4)}\right)\end{aligned}$$

which follows from moments of central chi-square (gamma) distribution and as $N_n \geq 0$

the second term on the R.H.S is $\leq 2(n-2)^{-2}(n-4)^{-1}$ and by Lemma 3.5.3 we have

$$\begin{aligned}(n-2)^2\mathbb{V}ar_{\lambda_n}\left(\frac{1}{n+2N_n-2}\right) &= \mathbb{V}ar_{\lambda_n}\left(\frac{1}{1+2N_n/(n-2)}\right)\\ &\leq \left\{1+2\,E(N_n)/(n-2)\right\}^{-4}\mathbb{V}ar\left(\frac{2\,N_n}{n-2}\right)\\ &= \frac{4\,\lambda_n\,(n-2)^2}{(n-2+2\lambda_n)^4} \leq \frac{1}{2(n-2)}.\end{aligned}$$

Thus, $\mathbb{V}ar\,(Y_n^{-1}) \leq 3\,(n-2)^{-2}(n-4)^{-1}$. □

Lemma 3.5.2.

If $Y_n \stackrel{d}{=} \chi^2_n(\lambda_n)$ and λ_n is an increasing sequence then

$$\lambda_n^2\,P(Y_n \leq n-2) \leq O(n)$$

Proof. Holds trivially for $\lambda_n \leq O(\sqrt{n})$. So we will prove for all other sequences i.e sequence where $\lambda_n/\sqrt{n}$ is not bounded. Note that $P(Y_n \leq n-2) \leq P(Y_n \leq n)$. And as Y_n is a non-central chi-square we have

$$Y_n \stackrel{d}{=} V_{n+2N} \text{ where } N \stackrel{d}{=} \text{ Poisson}(\lambda_n) \text{ and } V_n \stackrel{d}{=} \chi^2_n(0)$$

Now, for any fixed n and N we have,

$$P(V_{n+2N} \leq n) \leq 2P(V_{m+2N} \leq m) \text{ for all } m \geq n \text{ such that } m-n \text{ is large.}$$

Because $P(V_{m+2N} \leq m | V_{n+2N} \leq n) \leq P(\chi^2_{m-n}(0) \leq m-n) \leq 1/2$. So,

$$\begin{aligned}\lim_{n\to\infty} P(Y_n \leq n) &= \lim_{n\to\infty} E_{\lambda_n}\left\{P\left(V_{n+2N} \leq n \middle| N\right)\right\}\\ &\leq 2\lim_{n\to\infty} E_{\lambda_n}\left\{\lim_{n\to\infty} P\left(V_{n+2N} \leq n \middle| N\right)\right\}\\ &= 2\lim_{n\to\infty} E_{\lambda_n}\left\{\Phi\left(\frac{-2N}{\sqrt{2n+4N}}\right)\right\}\end{aligned}$$

as we can do a normal approximation to the sequence of central chi-square random variables. Next, we interchange the integrals (by Fubini's as integrand is positive) and then use bounded convergence theorem to have,

$$\begin{aligned}\lim_{n\to\infty} P(Y_n \le n) &\le 2 \lim_{n\to\infty} \int \phi(z) P_{\lambda_n}\left(\frac{2N}{\sqrt{2n+4N}} \le z\right) dz \\ &= 2 \int \phi(z) \lim_{n\to\infty} P_{\lambda_n}\left(\frac{2N}{\sqrt{2n+4N}} \le z\right) dz\end{aligned}$$

Now for all large n, λ_n is large (as λ_n increasing and $\lambda_n/\sqrt{n}$ is not a bounded sequence). So each large n, we can separately do a normal approximation to the Poisson random variable N.

Consider the case first when $\lambda_n > O(n)$. In this case the following naive bound will work:

$$P_{\lambda_n}\left(\frac{2N}{\sqrt{2n+4N}} \le z\right) \le P_{\lambda_n}\left(\frac{\sqrt{N}}{\sqrt{n}} \le z\right) \sim \widetilde{\Phi}\left(\frac{\lambda_n - nz^2}{\sqrt{\lambda_n}}\right).$$

We will use this bound for all z such that $z^2 \le t_n$ where t_n equals $n^{-1}(\lambda_n - \sqrt{\lambda_n}\sqrt{4\log\lambda_n + 2\log n})$. Also note that,

$$\lambda_n^2 \widetilde{\Phi}\left(\frac{\lambda_n - nz^2}{\sqrt{\lambda_n}}\right) \le O(n) \text{ for all } z^2 \le t_n \text{ and } \widetilde{\Phi}(t_n) = O(n\lambda_n^{-2}).$$

And so, it follows that $\lambda_n^2 \lim_{n\to\infty} P(Y_n \le n) \le O(n)$. □

Lemma 3.5.3.

For any non-negative random variable Y

$$\mathbb{V}ar\big\{(1+Y)^{-1}\big\} \le \{1+E(Y)\}^{-4}\, \mathbb{V}ar\big(Y\big).$$

Proof. As Y is non-negative we have

$$\left(\frac{1}{1+Y} - \frac{1}{1+E(Y)}\right)^2 = \frac{(Y-E(Y))^2}{(1+Y)^2(1+EY)^2} \le \frac{(Y-E(Y))^2}{(1+EY)^2}.$$

Now, taking expectation on both sides and using Bias-Variance decomposition we get

$$\mathbb{V}ar\left(\frac{1}{1+Y}\right) + \left(E\left(\frac{1}{1+Y}\right) - \frac{1}{1+E(Y)}\right)^2 \leq \{1+E(Y)\}^{-4}\,\mathbb{V}ar(Y).$$

This completes the proof. □

Lemma 3.5.4.

For any random variable X we have $\mathbb{V}ar(X_+) \leq \mathbb{V}ar(X)$.

Proof. With the decomposition of $X = X_+ - X_-$ we have

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{E}(X^2) - \mathrm{E}^2(X) \\
&= \mathrm{E}(X_+^2) + \mathrm{E}(X_-^2) - \mathrm{E}^2(X_+) - \mathrm{E}^2(X_-) + 2\mathrm{E}(X_+)\,\mathrm{E}(X_-) \\
&= \mathrm{Var}(X_+) + \mathrm{Var}(X_-) + 2\mathrm{E}(X_+)\mathrm{E}(X_-)
\end{aligned}$$

and we get the stated result as all the terms in R.H.S. are non-negative. □

BIBLIOGRAPHY

Aitchison, J. (1975), 'Goodness of prediction fit', *Biometrika* **62**(3), 547–554.

Aslan, M. (2006), 'Asymptotically minimax Bayes predictive densities', *Ann. Statist.* **34**(6), 2921–2938.

Brown, L. D. (1971), 'Admissible estimators, recurrent diffusions, and insoluble boundary value problems', *Ann. Math. Statist.* **42**, 855–903.

Brown, L. D. (2008), 'In-season prediction of batting averages: a field test of empirical Bayes and Bayes methodologies', *Ann. Appl. Stat.* **2**(1), 113–152.

Brown, L. D., George, E. I. & Xu, X. (2008), 'Admissible predictive density estimation', *Ann. Statist.* **36**(3), 1156–1170.

Dembo, A. & Zeitouni, O. (1993), *Large deviations techniques and applications*, Jones and Bartlett Publishers, Boston, MA.

Donoho, D., Johnstone, I. & Montanari, A. (2011), 'Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising'. arXiv:1111.1041.

Efron, B. (2004), 'The estimation of prediction error: covariance penalties and cross-validation', *J. Amer. Statist. Assoc.* **99**(467), 619–642. With comments and a rejoinder by the author.

Efron, B. & Morris, C. (1977), 'Stein's Paradox in Statistics', *Scientific American* **236**, 119–127.

Fourdrinier, D., Marchand, É., Righi, A. & Strawderman, W. E. (2011), 'On improved predictive density estimation with parametric constraints', *Electron. J. Stat.* **5**, 172–191.

Friedman, J., Hastie, T. & Tibshirani, R. (2008), 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics* **9**(3), 432–441.

George, E. I., Liang, F. & Xu, X. (2006), 'Improved minimax predictive densities under Kullback-Leibler loss', *Ann. Statist.* **34**(1), 78–91.

George, E. I., Liang, F. & Xu, X. (2012), 'From minimax shrinkage estimation to minimax shrinkage prediction', *Statist. Sci.* **27**(1), 82–94.

Ghosh, M., Mergel, V. & Datta, G. S. (2008), 'Estimation, prediction and the Stein phenomenon under divergence loss', *J. Multivariate Anal.* **99**(9), 1941–1961.

Hartigan, J. A. (1998), 'The maximum likelihood prior', *Ann. Statist.* **26**(6), 2083–2103.

Johnstone, I. M. (2012), Gaussian estimation: Sequence and wavelet models. Available at: `"http://www-stat.stanford.edu/~imj"`.

Komaki, F. (1996), 'On asymptotic properties of predictive distributions', *Biometrika* **83**(2), 299–313.

Komaki, F. (2001), 'A shrinkage predictive distribution for multivariate normal observables', *Biometrika* **88**(3), 859–864.

Kubokawa, T., ric Marchand, Strawderman, W. E. & Turcotte, J.-P. (2013), 'Minimaxity in predictive density estimation with parametric constraints', *Journal of Multivariate Analysis* **116**(0), 382 – 397.

Lai, T. L., Gross, S. T. & Shen, D. B. (2011), 'Evaluating probability forecasts', *Ann. Statist.* **39**(5), 2356–2382.

Ledoux, M. (2001), *The concentration of measure phenomenon*, Vol. 89 of *Mathematical Surveys and Monographs*, American Mathematical Society, Providence, RI.

Liang, F. & Barron, A. (2004), 'Exact minimax strategies for predictive density estimation, data compression, and model selection', *IEEE Trans. Inform. Theory* **50**(11), 2708–2726.

Nussbaum, M. (1996), 'Asymptotic equivalence of density estimation and Gaussian white noise', *Ann. Statist.* **24**(6), 2399–2430.

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

Stein, C. (1974), Estimation of the mean of a multivariate normal distribution, *in* 'Proceedings of the Prague Symposium on Asymptotic Statistics (Charles Univ., Prague, 1973), Vol. II', Charles Univ., Prague, pp. 345–381.

Stein, C. M. (1981), 'Estimation of the mean of a multivariate normal distribution', *Ann. Statist.* **9**(6), 1135–1151.

Tibshirani, R. (2011), 'Regression shrinkage and selection via the lasso: a retrospective', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**(3), 273–282.

Xu, J. (2007), A closed form for the harmonic-prior Bayes estimator with associated confidence sets for the mean of a multivariate normal distribution, PhD thesis. `"http://repository.upenn.edu/dissertations/AAI3271836"`.

Xu, X. & Zhou, D. (2011), 'Empirical bayes predictive densities for high-dimensional normal models', *J. Multivariate Analysis* **102**(10), 1417–1428.

Yang, Y. (2007), 'Consistency of cross validation for comparing regression procedures', *Ann. Statist.* **35**(6), 2450–2473.

Ye, J. (1998), 'On measuring and correcting the effects of data mining and model selection', *J. Amer. Statist. Assoc.* **93**(441), 120–131.

CHAPTER 4

PREDICTION UNDER EXACT SPARSITY & RISK DIVERSIFICATION

We construct minimax optimal predictive densities over ℓ_0 sparsity constrained high-dimensional parametric spaces. We find that minimax optimal strategies lie outside the Gaussian family but can be constructed with threshold predictive density estimates. Under high sparsity, explicit expressions of the first order minimax risk along with its exact constant, asymptotically least favorable priors and optimal predictive density estimates are derived.

Sometimes, high-dimensional problems are aided with additional information about the parameter. It helps in effective estimation of the true parameter in an otherwise huge and intractable space. However, the statistical estimates need to adapted in accordance to these prior constraints. Here, we consider sparsity restriction on the parametric space. The notion of sparsity is innate to modeling problems involving a highly interactive system (usually represented by large number of interacting parameters) which is dominated by only few significant effects. Sparse modeling has been successfully employed in scientific, economic as well as engineering applications to an extent that it is one of the most popular choices for modeling high dimensional data sets. The homoscedastic Gaussian Model

High Dimensional Gaussian Predictive Model

$$\textbf{M.1} \qquad \mathbf{X} \sim N(A\,\boldsymbol{\theta},\, \sigma_p^2\, I) \quad \text{and} \quad \mathbf{Y} \sim N(B\,\boldsymbol{\theta},\, \sigma_f^2\, I)$$

along with the restriction that $\boldsymbol{\theta}$ is sparse with few non-zero coefficients is a widely used. In biological modeling sparse models are used to decode significant gene networks from high dimensional gene-expressions data (Tibshirani, Hastie, Narasimhan & Chu 2002, Tibshirani 2011). In signal processing, sparse signals and codes arise from a wide range of applications (eg. image reconstruction, speech segmentation, etc) giving rise to the sub-field of Compressed Sensing (Donoho 2006, Candès 2006, Candès, Romberg & Tao 2006, Lustig, Donoho & Pauly 2007), where as, while modeling financial transactions sparsity impositions dictate few active positions and controls transaction costs (Brodie, Daubechies, De Mol, Giannone & Loris 2009, Fan et al. 2011). We impose an $\boldsymbol{\ell_0}$ constraint on the parameter space:

$$\Theta(n, s) = \left\{ \boldsymbol{\theta} \in \mathbb{R}^n : \sum_{i=1}^{n} \mathbb{I}[\theta_i \neq 0] \leq s \right\}. \tag{4.1}$$

The predictive model **M.1** with $\boldsymbol{\ell_0}$ constraint on the location structure can be used for sparse coding and for prediction in sparse networks. Here, the minimax risk calculations will be based on the orthogonal Gaussian model which is the basic building block of complex sparse models. Like sparse point estimation (Zhang 2010, Raskutti, Wainwright & Yu 2011), constrained predictive density estimation in **M.1** would intrinsically depend on risk calculations in the orthogonal model:

Predictive Gaussian Sequence Model

$$\textbf{M.2} \qquad \mathbf{X} \sim N(\boldsymbol{\theta},\, \sigma_p^2\, I) \quad \text{and} \quad \mathbf{Y} \sim N(\boldsymbol{\theta},\, \sigma_f^2\, I)$$

where $\mathbf{X}$ and $\mathbf{Y}$ are both n–dimensional vectors. If $\boldsymbol{\theta}$ were known, then $\mathbf{X}$ and $\mathbf{Y}$ would have been independent. **M.2** is known as the homoscedastic Gaussian sequence model (Nussbaum 1996) and has been widely studied in the function estimation framework (Johnstone 2012). Optimal estimation in **M.1** can be linked with the

minimax decision theoretic results in **M.2** through the procedure outlined in Donoho, Johnstone & Montanari (2011).

Sparse Gaussian predictive density estimation has the attributes of a sparse prediction problem adapted to the peculiarities of the entropy loss function. Point prediction analyses of dense signals in **M.1** (Leeb 2009, Dicker 2012, Huber & Leeb 2012) relate the worst-case performance with the spectral distribution of the predictors. Here, we concentrate on the orthogonal model. We address some of the unresolved issues associated with the role of sparsity in prediction theory. Next, we describe in more detail the minimax predictive density estimation problem in the sparse orthogonal model **M.2**.

4.1 Introduction and main result

In order to help the reader to better understand the context and nature of the predictive problem, we provide a brief review of the literature around the predictive density estimation problem. Aitchison (1975), Murray (1977) and Ng (1980) showed that in most parametric models there exist Bayes predictive density estimates which are decision theoretically better than the maximum likelihood plug-in estimate. As the name suggests, a plug-in or estimative density estimate $f_{(\widehat{\boldsymbol{\theta}},B)}$ belongs to the same parametric family of the true density and has the point estimate $\widehat{\boldsymbol{\theta}}$ plugged in the place of the unknown parameter. Given a prior π over $\mathbb{R}^n$ the Bayes predictive density in $\mathbf{M}$ (along with some mild conditions) minimizes the integrated Bayes risk and is given by

$$\widehat{p}_{\,\pi}\left(\mathbf{y}|\mathbf{X}=\mathbf{x}\right)=\int f_{(\boldsymbol{\theta},B)}(\mathbf{y})\,\pi(\boldsymbol{\theta}|\,\mathbf{x})\,d\boldsymbol{\theta} \tag{4.2}$$

where the posterior distribution

$$\pi(\boldsymbol{\theta}\,|\mathbf{x})=\{m_\pi(\mathbf{x})\}^{-1}\,f_{(\boldsymbol{\theta},A)}(\mathbf{x})\,\pi(\boldsymbol{\theta}) \text{ and } \; m_\pi(\mathbf{x})=\int f_{(\boldsymbol{\theta},A)}(\mathbf{x})\,\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} \tag{4.3}$$

is the marginal distribution. An important issue in predictive inference has always been to compare the performance of the class $\mathcal{E}$ of point estimation (PE) based plug-in density estimates (Barndorff-Nielsen & Cox 1996) with that of the optimal predictive density estimate. In fixed dimensional parametric spaces, large sample attributes of the predictive risk of efficient plug-in and Bayes density estimates have been studied by Komaki (1996) and Hartigan (1998). These results are independent of specific distributional attributes of f (Aslan 2006) and reflect the predictive nature of the problem through the relative inefficiency of the maximum likelihood plug-in density estimates.

Recently, the predictive density estimation problem has been studied in high dimensional parametric spaces (Komaki 2004, George et al. 2006, Ghosh et al. 2008, Xu & Zhou 2011). Decision theoretic parallels between predictive density estimation under Kullback-Leibler loss and point estimation under quadratic loss have been explored in **M.2** (George et al. 2012). Fundamental techniques and results in unconstrained Gaussian point estimation theory (Stein 1974, Strawderman 1971, Brown 1971, Brown & Hwang 1982) can be extended to produce optimal predictive density estimates (Komaki 2001, Brown et al. 2008, Fourdrinier et al. 2011). The Bayes predictive density from the uniform prior $\widehat{p}_U$ is the best invariant as well as a minimax density estimate in the unconstrained parametric space. Its risk properties are similar to those of the canonical minimax point estimate $\mathbf{X}$. Both regimes exhibit inadmissibility of the best invariant estimates in their respective domains and improved minimax estimators are constructed.

Another important subclass of density estimates are Linear estimates ($\mathcal{L}$) which are Bayes rules based on the conjugate product normal priors. The resultant density estimates $\widehat{p}_L[\,\boldsymbol{\alpha}\,] = \prod_{i=1}^{n} N(\alpha_i X_i, \alpha_i + \sigma_f^2)$, with $\alpha_i \in [0,1]$, are still Gaussian but has larger variance than the future density $f_{\boldsymbol{\theta},\sigma_f^2}(\mathbf{y})$. We choose the name 'linear' because the conjugate prior implies linearity of the posterior mean in X. It should be noted that for linear estimates, shrinkage of the location estimate $\mathbf{X}$ is related to flattening of the variance and $\mathcal{E} \cap \mathcal{L}$ consists only the zero density estimate.

Xu & Liang (2010) showed that the class $\mathcal{L}$ is minimax optimal if the parameter space is restricted to ellipsoids with certain growth conditions. Here, we evaluate the

minimax risk over the $\boldsymbol{\ell}_0$ sparse parameter space $\Theta(n, s)$ in the asymptotic framework $\{n \to \infty \text{ and } s/n \to 0\}$. Sparse point estimation has been extensively studied in this asymptotic set-up by (Donoho & Johnstone 1994*a*, Donoho & Johnstone 1994*b*, Donoho et al. 1992, Foster & George 1994) and the results are the building blocks for popular sparse estimation methods (Zhang 2005, Candes & Tao 2007, Donoho, Maleki & Montanari 2011). It is natural to look for parallels in the predictive density regime.

4.1.1 Our contributions:

Instead of parallels, we found contrasting results for sparse estimation in the two regimes. The asymptotic minimax predictive risk reflects the nature of the predictive density estimation problem through the ratio r of the future to the past volatilities $r = \sigma_f^2/\sigma_p^2$. As r decreases, we need to estimate the future observations based on increasingly noisy past observations and so, the difficulty of the density estimation problem also increases. Unlike point estimation, sharp decision theoretic rates in the predictive density problem should depend on r. This dependence was not emphasized in the admissibility results in the unrestricted space.

In our ℓ_0 sparse prediction framework, as the proportion of non-zero signals goes to zero, we find that the order of the minimax rate does not depend on r. So, exact determination of the constants of the minimax risk is important here. Optimal minimax estimators can be constructed by incorporating the predictive nature of the problem through the notion of *diversification* of the future risk. Under sparsity constraints efficiency of the prediction schemes depend on careful coupling of the sparsity adjustment and the risk diversification mechanisms. The risk diversification notion can also be extended (though not done here) to dense unrestricted parametric spaces where future uncertainty can be effectively shared by optimally flattening probability densities based on the quadratic risk estimate of their corresponding location point estimator.

Here we also evaluate the minimax risk over the wide class $\mathcal{G}$ of all product *Gaussian density estimates* $\widehat{p}_G[\,\widehat{\boldsymbol{\theta}},\, \widehat{\mathbf{d}}\,] = \prod_{i=1}^n N(\widehat{\theta}_i, \widehat{d}_i)$. $\mathcal{G}$ contains both $\mathcal{L}$ and $\mathcal{E}$ and would

represent the infamily error rate of the sparse Gaussian predictive density estimation problem. We prove that the sub-class $\mathcal{G}$ is sub-optimal rand provide asymptotic miximax strategies as well as the sub-optimality rates of the sub-classes $\mathcal{E}, \mathcal{L}$ and $\mathcal{G}$.

4.1.2 Description of the main results

Notations and Preliminaries

To proceed further we need some notation. The action space $\mathcal{A}_n$ contains all possible densities in $\mathbb{R}^n$. The n-dimensional minimax risk of the prediction problem is given by

$$R(n, s, r) = \min_{\widehat{p} \in \mathcal{A}_n} \max_{\boldsymbol{\theta} \in \Theta(n,s)} \rho\left(\boldsymbol{\theta}, \widehat{p}\right).$$

We compute the limiting behavior of $R(n, s, r)$ in the asymptotic framework $\mathcal{F} = \{n \to \infty, s/n \to 0\}$. The minimax risk over the sub-class of Plugin ($\mathcal{E}$) density estimates is represented by

$$R(n, s, r, \mathcal{E}) = \min_{\widehat{\boldsymbol{\theta}}} \max_{\boldsymbol{\theta} \in \Theta(n,s)} \rho\left(\boldsymbol{\theta}, \widehat{p}_E[\widehat{\boldsymbol{\theta}}]\right) \text{ where } \widehat{p}_E[\widehat{\boldsymbol{\theta}}] = N(\widehat{\boldsymbol{\theta}}, \sigma_f^2 \mathrm{I}_n).$$

Similarly, the minimax risk over the sub-classes of Linear ($\mathcal{L}$) and Gaussian density estimates ($\mathcal{G}$) will be denoted by $R(n, s, r, \mathcal{L})$ and $R(n, s, r, \mathcal{G})$ respectively. The maximum Bayes risk over the class of priors $\mathcal{M}(n)$ on $\mathbb{R}^n$ is denoted by

$$B(r, \mathcal{M}(n)) = \max_{\pi \in \mathcal{M}(n)} \min_{\widehat{p}} B(\pi, \widehat{p}).$$

As defined in Chapter 1 this maximin value is also the information capacity. A prior maximizing this Bayes risk is said to be a least favorable prior for the prediction problem. We evaluate the supremum Bayes risk of the following class of priors

$$\mathcal{M}(n, s) = \left\{ \pi : \sum_{i=1}^{n} P_\pi(\theta_i \neq 0) \leq s \right\}.$$

Univariate Prediction Problem

In high dimensions, due to concentration of measure, the decision theoretic results in our multivariate set up $\mathcal{F}$ will intrinsically depend on the properties of the coordinate-wise univariate risk. The least favorable priors as well as the minimax density estimates will be product densities. So, computing the multivariate risk in the n–dimensional, s–sparse orthogonal Gaussian Model $\mathbf{M.2}(n, s, \sigma_p, \sigma_f)$ would involve studying the corresponding univariate model $\mathbf{M.2}(1, \eta, \sigma_p, \sigma_f)$ in which we relax the sparsity constraint to restriction on the univariate prior space

$$\mathfrak{m}(\eta) = \{\pi \in \mathcal{P}(\mathbb{R}) : \pi(\theta \neq 0) \leq \eta\}$$

where $\mathcal{P}(\mathbb{R})$ is the collection of all probability measures in $\mathbb{R}$. The maximin value $\sup_{\pi\in\mathfrak{m}(\eta)} \inf_{\widehat{p}} B(\pi, \widehat{p})$ of this univariate prediction game is given by the maximal Bayes risk $\beta(\eta, r) := \sup_{\pi\in\mathfrak{m}(\eta)} B(\pi)$. The minimax risk for this univariate prediction problem is given by

$$\rho_M(\eta, r) := \inf_{\widehat{p}} \sup_{\pi\in\mathfrak{m}(\eta)} B(\pi, \widehat{p}). \tag{4.4}$$

The minimax risk and the maximal Bayes risk over univariate sub-collection $\mathfrak{m}$ of priors of $\mathfrak{m}(\eta)$ are respectively denoted by $\rho_M(\eta, r, \mathfrak{m})$ and $\beta(\eta, r, \mathfrak{m})$. When the maximal Bayes risk (maximin) equals to the minimax risk, it is referred as the Bayes-Minimax risk for the prediction problem.

As in our asymptotic framework $\mathcal{F}$ the proportion of non-zero signals s/n goes to zero, the univariate risk calculation will be in the asymptotic regime $\eta \to 0$. The difference between the multivariate and univariate cases is notationally demonstrated through the bold representation of multivariate vectors. The other non-standard notations used are $\phi(|\boldsymbol{\theta}, r)$ for the multivariate normal density with center $\boldsymbol{\theta}$ and covariance $r\mathrm{I}$ while $\widetilde{\Phi} = 1 - \Phi$ with Φ being the standard normal distribution. For sequences, the symbol $a_n \sim b_n$ means $a_n = b_n(1 + \mathrm{o}(1))$ and $a_n \asymp b_n$ means $a_n/b_n \in (c_1, c_2)$ where c_1 and c_2 are constants.

Results

Consider the following symmetric univariate prior 3–point prior

$$\pi[\eta, r, 3] = (1-\eta) \cdot \delta_0 + \frac{1}{2}\eta \cdot \delta_{\nu_\eta} + \frac{1}{2}\eta \cdot \delta_{-\nu_\eta}$$

where ν_η is the positive root of the quadratic equation

$$v_w^{-1}\,\nu^2 + 2\,v_w^{-1/2}\,\nu\,a = \lambda_e^2$$

with $v_w = (1+r^{-1})^{-1}$, $a = \max(\sqrt{\log v_w \lambda_e^2}, 1)$ and $\lambda_e^2 = 2\sigma_p^2 \log\{(1-\eta)\,\eta^{-1}\}$ is close to the universal threshold (when $\eta = n^{-1}$) seen previously in PE (Donoho & Johnstone 1994*a*). As $\eta \to 0$, the solution $\nu_\eta \to \lambda_f = v_w^{-1/2}\lambda_e$.

Also, consider the discrete cluster prior $\pi[\eta, r, \mathrm{CL}]$ with $1-\eta$ probability at 0 and sharing the remaining mass among a cluster of support points. The non-zero support points start from $\pm\,\nu_\eta$ and span out symmetrically on either side in a geometric progression with common ratio $(1+2r)$ up to the universal threshold:

$$\pi[\eta, r, \mathrm{CL}] = (1-\eta)\cdot\delta_0 + \frac{\eta}{2K_\eta}\sum_{i=1}^{K_\eta}\left\{\,\delta_{\mu_i} + \delta_{-\mu_i}\,\right\} \quad \text{where} \tag{4.5}$$

$$K_\eta = \max\left\{i : (1+2r)^{\,i-1}\,\nu_\eta \le \lambda_e + a\right\}, \tag{4.6}$$

$$\mu_i = (1+2\,r)^{i-1}\,\nu_\eta,\ i = 1, 2, \ldots, K_\eta. \tag{4.7}$$

For any fixed $r \in (0, \infty)$ as $\eta \to 0$ we have

$$K(r) = \lim_{\eta\to 0} K_\eta = \left\lfloor \frac{\log(1+r^{-1})}{2\log(1+2r)} \right\rfloor.$$

We will use the Bayes predictive density $\widehat{p}\,(\cdot|x; \pi[\eta, r, \mathrm{CL}])$ derived from the cluster prior $\pi[\eta, r, \mathrm{CL}]$ to construct threshold estimates. Consider the following univariate threshold estimate which uses the best invariant density estimate $\widehat{p}\,(\cdot|x; \pi_U)$ from the

uniform prior above the threshold λ_e and $\widehat{p}\,(\cdot|x;\pi[\eta,r,\mathrm{CL}])$ below the threshold

$$\widehat{p}[\eta,\mathrm{T},\mathrm{CL},\mathrm{U}](y|x)=\begin{cases}\widehat{p}\,(y|x;\pi[\eta,r,\mathrm{CL}]) & \text{if } |x|\le\lambda_e\\ \widehat{p}\,(y|x;\pi_{\mathrm{U}}) & \text{if } |x|\ge\lambda_e\end{cases}. \tag{4.8}$$

The estimator $\widehat{p}[\eta,\mathrm{T},\mathrm{CL},\mathrm{U}]$ attains the univariate minimax risk as $\eta\to 0$.

Theorem 4.1.1.

For any fixed $r\in(0,\infty)$ as $\eta\to 0$ in the univariate prediction problem we have

$$\rho_M(\eta,r)=\beta(\eta,r)=\frac{\eta\,v_w\,\lambda_e^2}{2r}\Big(1+\mathrm{o}\,(1)\Big).$$

Also, $\pi[\eta,r,3]$ is an asymptotically least favorable prior and $\widehat{p}[\eta,\mathrm{T},\mathrm{CL},\mathrm{U}]$ is an asymptotically minimax optimal estimate.

Based on the univariate version, we can construct a multivariate, co-ordinate wise rule

$$\widehat{p}[n,s,\mathrm{T},\mathrm{CL},\mathrm{U}](\mathbf{y}|\mathbf{x})=\prod_{i=1}^{n}\widehat{p}\,[s/n,\mathrm{T},\mathrm{CL},\mathrm{U}](y_i|x_i)$$

which will be asymptotically minimax optimal in the high dimensional regime $\mathcal{F}$. Also the product discrete distribution

$$\pi[n,s,r,3](\boldsymbol{\theta})=\prod_{i=1}^{n}\pi[s/n,r,3](\theta_i)$$

based on the 3–point prior will be asymptotically least favorable. The following theorem, which is *our main result*, describes the minimaxity results for the predictive density estimation problem with ℓ_0 sparsity constraints in Model **M.2**.

Theorem 4.1.2.

As $n \to \infty$, if $s \to \infty$ but $s/n \to 0$ then for any fixed $r \in (0, \infty]$ we have:

a. The minimax risk $R(n, s, r) \sim (1 + r)^{-1} \, s \, \log(n/s)$.

b. $\pi[n, s, r, 3]$ is an asymptotically least favorable prior distribution, i.e.

$$B(\pi[n, s, r, 3]) \;\sim \sup_{\pi \, \in \, \mathcal{P}(\Theta(n,s))} \; \inf_{\widehat{p} \in \mathcal{A}_n} \; B(\pi, \widehat{p})$$

where $\mathcal{P}\big(\Theta(n, s)\big)$ is the collection of all probability measures over $\Theta(n, s)$.

c. The predictive density estimate $\widehat{p}[n, s, \mathrm{T}, \mathrm{CL}, \mathrm{U}]$ is minimax optimal, i.e.

$$\max_{\boldsymbol{\theta} \, \in \, \Theta(n,s)} \boldsymbol{\rho}\left(\boldsymbol{\theta}, \widehat{p}\,[n, s, \mathrm{T}, \mathrm{CL}, \mathrm{U}\,]\right) \sim R\,(n, s, r).$$

We compute the multivariate minimax risk over the different sub-classes of predictive density estimates. As an immediate corollary of the above theorem it follows that the class of plug-in estimators $\mathcal{E}$ is sub-optimal. The plug-in sup-optimality ratio $R(n, s, r, \mathcal{E})/R(n, s, r)$ asymptotically equals $1 + r^{-1}$ (see Lemma 4.7.1). As in point estimation, the class of linear estimates $\mathcal{L}$ performs very poorly.

Lemma 4.1.3.

For any fixed $r \in (0, \infty)$ and for all sequences s_n such that $s_n \to \infty$ and $s_n/n \to 0$ as $n \to \infty$, we have

$$\liminf_{n \to \infty} \quad \frac{R(n, s_n, r, \mathcal{L})}{R(n, s_n, r)} = \infty\,.$$

We also find that the performance of the wider class of all Gaussian density estimates is no better than that of plug-in estimates.

Lemma 4.1.4.

Under the condition of Theorem 4.1.2 we have

$$R\big(n, s, r, \mathcal{G}\big) \sim R\big(n, s, r, \mathcal{E}\big).$$

If the parametric space $\Theta(n, s)$ does not have any sparser representation with respect to the group of orthogonal transformations then the sub-optimality of the class $\widetilde{\mathcal{G}}$ comprising of all Gaussian densities $N(\boldsymbol{\theta}, \widehat{\Sigma})$ including non-diagonal covariances is $1 + r^{-1}$.

Now, $\widehat{p}\,[n, s, \mathrm{T}, \mathrm{CL}, \mathrm{U}]$ is not derived from a prior and we would like to construct a prior for which asymptotic minimaxity in Theorem 4.1.2 holds. Consider another symmetric univariate prior $\pi[\eta, r, \mathrm{INF}]$ whose support consists of the origin and infinite number of equidistant clusters each containing $2K_\eta$ points (defined before in equation [4.5]) in the same spatial alignment as for $\pi[\eta, r, \mathrm{CL}]$. As $\eta \to 0$, the clusters centers are separated by λ_e and, as they move away from zero they have geometrically decaying probability with η being the common ratio. However, within clusters all support points are not equally likely any more. They have geometrically decaying probability with common ratio $\log \eta^{-1}$.

$$\pi[\eta, r, \mathrm{INF}] = (1-\eta)\cdot\delta_0 + \frac{1-\eta}{2}\sum_{j=0}^{\infty}\eta^{j+1}\sum_{i=1}^{K_\eta} q_i\left[\delta_{\mu_{ij}} + \delta_{-\mu_{ij}}\right] \quad \text{where,}$$

$$\mu_{ij} = j\,\lambda_e + (1+2r)^{i-1}\,\nu_\eta\,,\ i = 1, \ldots, K_\eta \text{ and } j = 1, \ldots, \infty;$$

$$q_i = (\log \eta^{-1})^{-i} \text{ for } i = 2, \ldots, K_\eta \text{ and } q_1 = 1 - \frac{1 - (\log \eta^{-1})^{-K_\eta}}{\log \eta^{-1} - 1}.$$

Based on $\pi[\eta, r, \mathrm{INF}]$ we can construct a multivariate prior

$$\pi[n, s, r, \mathrm{INF}](\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi[s/n, r, \mathrm{INF}](\theta_i)$$

in $\mathbb{R}^n$ which will not only be least favorable but also yield a minimax optimal density estimate. As $\pi[n, s, r, \text{INF}]$ is a proper prior it is admissible. Though its support is not confined to $\Theta(n, s)$ it concentrates on it asymptotically. It represents an equilibrium solution for the sparse minimax prediction problem.

Theorem 4.1.5.

Under the conditions of Theorem 4.1.2 for any fixed $r \in (0, \infty]$, the proper prior distribution $\pi[n, s, r, \text{INF}]$ is an asymptotically least favorable and its corresponding Bayes predictive density is asymptotically minimax optimal.

These results reflect the predictive nature of the problem. The cluster prior $\pi[\eta, r, \text{CL}]$ in the minimax estimate $\widehat{p}[\eta, \text{T}, \text{CL}, \text{U}](y|x)$ diversifies the predictive risk over the constrained parametric space. The risk diversification notion is essential to construction of optimal estimates and can be extended to other different forms of asymptotically minimax predictive density estimates. This mechanism of uncertainty sharing in presence of sparsity has not been previously described in minimax decision theory. To rigorously interpret the results, we need the risk equations in George et al. (2006) which connect the Bayes predictive risk and with the square error of the posterior mean (see Section 2.4.1). Next, we provide an heuristic explanation of the implications of the results by an asymptotic (as $\eta \to 0$) risk analysis of univariate threshold estimators.

New Phenomena in Estimation Theory

To adjust for high sparsity we use threshold based non-linear estimates $\widehat{t}[\lambda, S]$ with the threshold cut off at λ, the best invariant estimate $\widehat{p}(\cdot|x; \pi_U)$ above the threshold and estimate/scheme S below the threshold. We found that for such an estimate the threshold choice is dictated by the level of sparsity η and can not be lower than λ_e.

Lemma 4.1.6.

For any fixed $r \in (0, \infty)$, scheme S and $u \in [0, 1)$, we have

$$\lim_{\eta \to 0} \frac{\sup_{\pi \in \mathfrak{m}(\eta)} B\big(\pi, \widehat{t}\,[u\,\lambda_e,\, S]\,\big)}{\rho_M(\eta, r)} = \infty.$$

However, unlike PE, here the non-zero support point of the least favorable prior is not at λ_e but at λ_f. So, the univariate asymptotic maximal Bayes risk $\beta(\eta, r) \sim (2\,r)^{-1}\eta\,\lambda_f^2$ is lower than the corresponding maximal quadratic Bayes risk $\beta_q(\eta, r) = \eta\,\lambda_e^2$ after adjustment by $2r$. Because of the threshold choice, the univariate threshold risk $\rho(\theta, \widehat{t}[\lambda_e, \cdot])$ is bounded when $|\theta| \geq \lambda_e$. So, we need to restrict the predictive univariate risk in the region $\{\theta \in (-\lambda_e, \lambda_e)\}$ below the minimax risk. For that purpose, unlike PE we can not just use the zero estimator $\phi(\cdot|0, \sigma_f^2)$ below the threshold because then $\rho(\theta, \widehat{t}[\lambda_e, 0])$ exceeds $\eta^{-1}\,\rho_M(\eta, r)$ (given by equation [4.4]) in the region $V = \{\theta : |\theta| \in [\lambda_f, \lambda_e]\}$. It leads to inefficiency of the optimal plug-in density estimates.

Instead using the Bayes density estimate from $\pi[\eta, r, 3]$ the risk can be controlled in the neighborhood around λ_f but exceeds $\beta(\eta, r)$ as θ moves further away. The univariate threshold estimate $\widehat{t}[\lambda_e, \pi[\eta, r, 3]]$ represents an unshared threshold prediction scheme. To control the risk through out we need to share the predictive risk for θ between λ_f and λ_e. The cluster prior serves the purpose by using a prior with probability $1 - \eta$ at 0 (which controls the risk at 0) and distributing the remaining mass η equally among a finite chain of points covering V. We also find that discreteness of the sharing scheme is important and continuous uniform sharing scheme will not work here. Also, the number of support points in the sharing scheme is proportional to r^{-1} reflecting the increasing difficulty of the prediction problem. Table 4.1 shows the number of support points in the cluster prior as r varies. In Figure [4.1], we have a schematic description of the asymptotic ($\eta \to 0$) behavior of $\rho(\theta, \widehat{t}[\lambda_e, S])$ for different type of schemes in S.

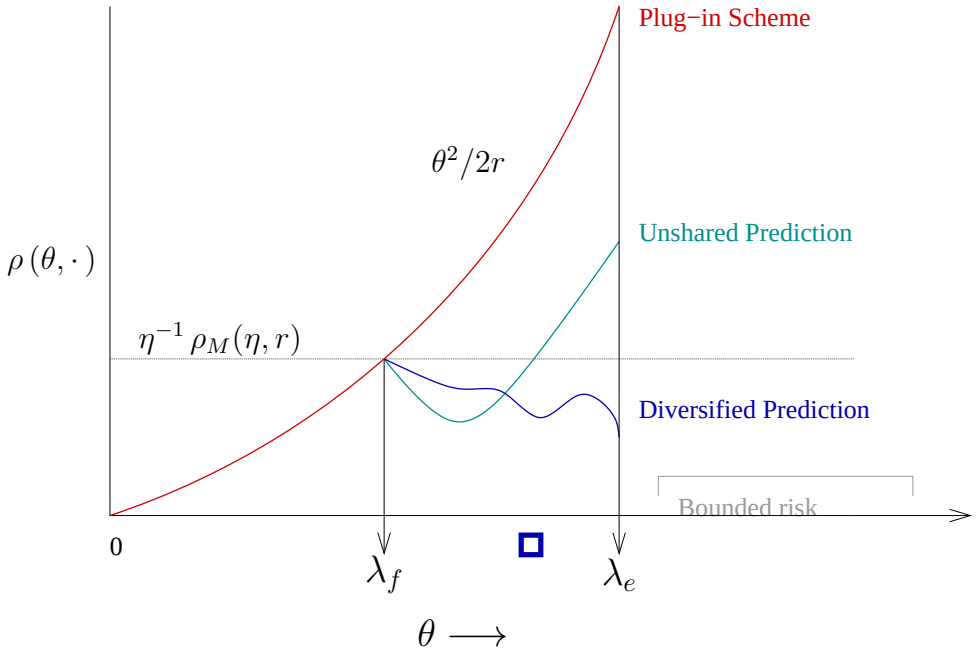

Figure 4.1: Schematic diagram of KL-risk functions for different Predictive Schemes. As the true parameter θ varies, the univariate asymptotic predictive risk $\lim_{\eta\to 0}\,\rho(\theta,\widehat{t}[\lambda_e,S])$ is represented on the ordinate. The blue box between λ_f and λ_e represents a support point of the cluster prior (representative of shared predictive schemes) which is not in the support of $\pi[\eta,r,3]$ or other unshared predictive schemes.

4.1.3 Organization of this Chapter

The proof of the results along with their implications are developed first in an overview fashion in Section 4.2 and Section 4.3 which may suffice for a first reading. Along with the general proof strategies it contains the proof of Theorem 4.1.1. Part of the proof of Theorem 4.1.2 extends over to Section 4.6. Technical proofs of all the statements in Section 4.3 are presented in Section 4.4 and Section 4.5 with some of the lemmas pushed to the Appendix to improve flow and readability. Theorem 4.1.5,

r	0.1073	0.1235	0.1465	0.1826	0.2485	0.4196	> 0.4196
K_η	7	6	5	4	3	2	1

Table 4.1: Number (K_η) of positive support points in the cluster prior $\pi[\eta,r,\mathrm{CL}]$ as r varies.

Lemma 4.1.3, Lemma 4.1.4 and Lemma 4.1.6 are proved in Section 4.7. The proofs involving direct risk calculations with the KL loss may be of independent interest and use in information theory.

4.2 Proof Overview

Hereon we will assume that $\sigma_p^2 = 1$ and $\sigma_f^2 = r$. The general predictive KL risk as well as the ℓ_0 constraints on the parametric space will not be affected by this restriction. However, the density estimates are usually based on statistics equivariant to the scale transformation and needs multiplication by σ_p. The proofs as well as the interpretation which will be mostly done on the reduced univariate model are presented for the case $\sigma_p^2 = 1$ and $\sigma_f^2 = r$. While extending the results to the multivariate set-up we will appropriately modify the estimators for general σ_p and σ_f. Proper interpretation of the predictive results will involve comparison with quadratic risk of point estimation in model $\mathbf{M.2}(1, \eta, r)$. Next, we describe the connections for the univariate version.

4.2.1 Bayes-Minimax Method

We will explicitly solve for the equilibrium of the univariate minimax problem in $\mathbf{M.2}(1, \eta, r)$. Using the minimax theorem here, we see that over the class $\mathfrak{m}(\eta)$ the maximal univariate Bayes risk $\beta(\eta, r) = \sup\{B(\pi) : \pi \in \mathfrak{m}(\eta)\}$ is always less than the minimax risk $\rho_M(\eta, r)$. So, if we can produce:

1. a lower bound on $\beta(\eta, r)$ by considering the Bayes risk of a particular prior π_0 (say);

2. an upper bound on the minimax risk $\rho_M(\eta, r)$ by considering the $\max_\theta \rho(\theta, \widehat{p}_0)$ for a particular estimator $\widehat{p}_0$;

3. such that the lower bound and upper bound matches asymptotically as $\eta \to 0$;

we can conclude that $\beta(\pi_0)$ is the supremum Bayes risk as well as the minimax risk and π_0 is asymptotically least favorable and $\widehat{p}_0$ is a minimax strategy for the univariate

predictive density estimation problem.

Once we have found the equilibrium for the univariate game, we extend the solution to the multivariate regime by following the general strategy outlined in Johnstone (2012, Section 4.11). Considering the class of exchangeable priors $\mathfrak{m}^e$ we can reduce the n–dimensional multivariate problem as repeated (n times) independent playing of the univariate minimax game and using the minimax theorem 4.8.3 we can show that $R(n, s, r)$ would be less than $n\beta(s/n, r)$ (see Lemma 4.6.1). As detailed in Section 4.6, we can actually show that $R(n, s, r) \sim n\beta(s/n, r)$ would follow from concentration properties of the multivariate least favorable prior $\pi[n, s, r, 3]$. Following this scheme, we now calculate the asymptotic univariate minimax risk.

The rest of this chapter is divided into 3 parts. We discuss univariate minimax optimality in the first and then extend those univariate rules to co-ordinate wise multivariate minimax strategies in the second. The last sub-chapter contains discussions on the extensions of these exact sparsity (ℓ_0) results to approximate (weak) sparsity (ℓ_p balls).

PART A: UNIVARIATE MINIMAX RISK UNDER HIGH ℓ_0 SPARSITY

4.3 The univariate asymptotic set-up

Here onward we would further restrict our univariate parametric space to the non-negative orthant. The corresponding prior space would be $\mathfrak{m}^+(\eta) = \{\pi(\theta) : \pi(0) \geq 1 - \eta\}$. It would simplify exposition and the results easily generalizes over $\mathfrak{m}(\eta)$ by symmetrization.

To produce a lower bound on the maximal Bayes risk, we consider the class of all 2–point priors in $\mathfrak{m}^+(\eta)$. We will see that the 2–point version of the prior $\pi[\eta, r, 3]$ will attain the maximal Bayes risk where

$$\pi[\eta, r, 2] = \begin{cases} 0 & \text{with prob } 1 - \eta \\ \nu_\eta & \text{with prob } \eta \end{cases}$$

and ν_η is the positive root of the quadratic equation

$$\frac{1}{2}\, v_w^{-1}\, \nu^2 + v_w^{-1/2}\, \nu\, a = \log\left\{(1-\eta)\,\eta^{-1}\right\} \tag{4.9}$$

where $v_w = (1 + r^{-1})^{-1}$ is the variance of the semi-futuristic random variable W and

$a = (2 \log \lambda_f)^{1/2}$ and $\lambda_f^2 = 2\, v_w\, \log\{(1-\eta)\, \eta^{-1}\}$.

To provide a detailed description of the univariate asymptotic regime we describe some fundamental quantities (functions of the sparsity level η) associated with our asymptotic calculations:

Universal Threshold: $\lambda_e = \left(2 \log\{(1-\eta)\, \eta^{-1}\}\right)^{1/2}]$ is the universal threshold for point estimation of θ based on the past X only Donoho & Johnstone (1994*a*). The subscript 'e' emphasizes its estimative purpose. Later on, we will see that λ_e is also the optimal threshold in the predictive regime.

Ideal Predictive Threshold: $\lambda_f = (2 v_w \log(1-\eta)\, \eta^{-1})^{1/2}$] is the universal threshold needed to devise minimax optimal threshold point estimate of θ based on observing the random variable W i.e both the past X and the complete future Y (which is equivalent to observing Y_α with $\alpha = 1$). The subscript f reflects its dependence on the future data.

Vantage Point: ν_η – the positive root of Equation [4.9] is the non-zero support point of the asymptotically least favorable 2–point prior in point estimation of θ under quadratic loss and noise variability v_w (compared with Johnstone (2012, Equation (8.36))) which again corresponds to location point estimation based on W. ν_η will be pivotal to our calculations. ν_η marks the beginning of the *Vulnerable Zone* which spans from $[\nu_\eta, \lambda_f]$. The calculation of the predictive risk on either side on ν_η displays the sparsity adjustments and uncertainty sharing dynamics.

Resolution Parameter, $a = (2 \log \lambda_f)^{1/2}$*:* As $\eta \to 0$, we need to compute the asymptotic predictive risk as the true parameter θ moves along the non-negative axis. In the asymptotic regime, we can exactly quantify the risk except at a few transition points. However, the discontinuity of our analysis will only be limited to $\mathrm{O}(a)$–neighborhood around the transition points. In our calculations, a will generally arise an overshoot/undershoot parameter, e.g. Equation [4.9]. As in PE, a is of the order of $(\log \log \eta^{-1})^{1/2}$ and the risk can be accurately approximated in a resolution coarser than a.

Now, note that Equation [4.9] reduces to $(v_w^{-1/2}\, \nu + a)^2 = \lambda_e^2 + a^2$ and so

$$\nu_\eta = \left(\lambda_f^2 + v_w\, a^2\right)^{1/2} - v_w^{1/2} a \geq \lambda_f - a v_w^{1/2}.$$

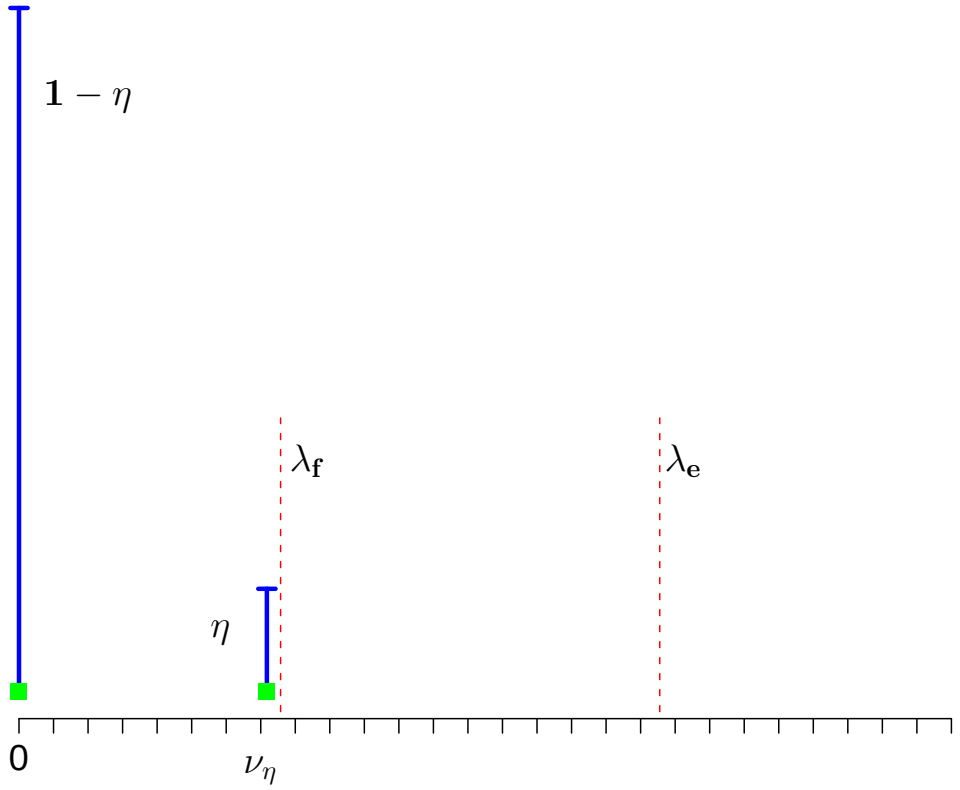

Figure 4.2: The figure shows the support and probability allocation of the sparse 2–point prior $\pi[\eta, r, 2]$ along with the universal threshold λ_e and the ideal predictive threshold λ_f. The abscissa is graduated in a units and is drawn according to the scale with $\eta = e^{-1000}$ and $r = 0.2$.

Thus $\nu_\eta \in (\lambda_f - a, \lambda_f)$. So as $\eta \to 0$, λ_f is quite close to the vantage point ν_η (in a–coarser resolution). Also, note that the ratio $\lambda_f : \lambda_e$ equals $v_w^{1/2} : 1$. Thus, as the future variability r decreases, the distance between λ_f and λ_e increases. Also as $\eta \to 0$, the threshold behaves as $\lambda_e \sim (2 \log \eta^{-1})^{1/2}$. Attributes in the asymptotic regime are pictorially represented in Figure 4.2.

$\pi[\eta, r, 2]$ is a *sparse prior* in the sense that repeated sampling from the prior would yield with a sparse signal as $\eta \to 0$. We will see that $\pi[\eta, r, 2]$ will be an asymptotically least favorable distribution for θ. To get a fair understanding of our strategies in the predictive regime we formulate the predictive density estimation problem as a two-person game between the Nature and a statistician.

4.3.1 Predictive Two-Player Game and Equilibrium Strategies

In this predictive game, Nature chooses a probability distribution $\pi(\theta)$ from $\mathfrak{m}^+(\eta)$ for the location parameter θ. Then, a particular sample point θ_0 is generated from $\pi(\theta)$ and based on the signal θ_0 realizations X and Y contaminated with white noise would be produced: $X = \theta_0 + \epsilon_1$ and $Y = \theta_0 + r^{1/2}\epsilon_2$ where ϵ_1 and ϵ_2 are independent. The statistician sees only X and he knows about the sparsity restrictions and the data generation scheme. He has to come up with a density estimate for Y. It is to be noted that under sufficient concentration properties the complicated sparse high dimensional minimax prediction problem is equivalent to repeated playing of this simple 2–player game with fixed strategies (from both) over independent trials.

As $\eta \to 0$, a minimax strategy of this predictive game is given by the positive version $\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}]$ of $\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}, \mathrm{U}]$ where

$$\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}](y|x) = \begin{cases} \widehat{p}\,(y|x; \pi[\eta, r, \mathrm{CL}^+]) & \text{if } X \le \lambda_e \\ \widehat{p}\,(y|x; \pi_U) & \text{if } X > \lambda_e \end{cases}$$

and the $\pi[\eta, r, \mathrm{CL}^+]$ is a sparse discrete prior (cardinality of the support set equals $(K_\eta + 2)$ with K_η defined in Equation [4.10]) with $(1 - \eta)$ probability at 0 and sharing the remaining mass on a cluster of $(K_\eta + 1)$ support points. The non-zero support

points approximately lies between λ_f and λ_e. The $(K_\eta + 1)$ non-zero support points start from $\mu_0 = \nu_\eta$ with ν_η given by the Equation [4.9] and span out in a geometric progression

$$\mu_i = (1+2r)^i \mu_0,\ i = 1, 2, \ldots, K_\eta \quad \text{and} \quad K_\eta = \max\left\{i : (1+2r)^i \mu_0 \leq \lambda_e + a\right\}.$$

As $\eta \to 0$, a first order approximation to the cardinality would be

$$K_\eta \sim \left\lfloor \frac{\log(\lambda_e/\lambda_f)}{\log(1+2r)} \right\rfloor = \left\lfloor \frac{\log(1+r^{-1})}{2\log(1+2r)} \right\rfloor. \tag{4.10}$$

The subscript in K_η would be dropped for simplicity. The non-zero support points are equally probable and in that way the cluster prior

$$\pi[\eta, r, \mathrm{CL}^+] = (1-\eta) \cdot \delta_0 + \frac{\eta}{K+1} \sum_{i=0}^{K} \delta_{\mu_i}(\theta)$$

is a probability distribution in $\mathfrak{m}^+(\eta)$ which is midway between the least favorable prior in PE based on X and $W = \mathrm{UMVUE}(X, Y)$ respectively. The intermediation is marked by equal sharing of probability among the finite support points laid between λ_f and λ_e. A schematic representation of this mass allocation is presented in Figure [4.3]. The alignment of (spacing between) the support points is also intrinsic to the nature of the predictive problem and will be discussed later (in Section 4.5). Note that as $\eta \to 0$, $\pi[\eta, r, \mathrm{CL}^+]$ is a sparse prior and $\pi[\eta, r, \mathrm{CL}^+] = \pi[\eta, r, 2]$ if $r > 0.42$.

Theorem 4.3.1.

As $\eta \to 0$, $\pi[\eta, r, 2]$ is an asymptotically least favorable prior for the univariate predictive game and $\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}+, \mathrm{U}](y|x)$ is a minimax estimator with the optimal asymptotic risk of $(2r)^{-1}\, \eta\, \lambda_f^2$.

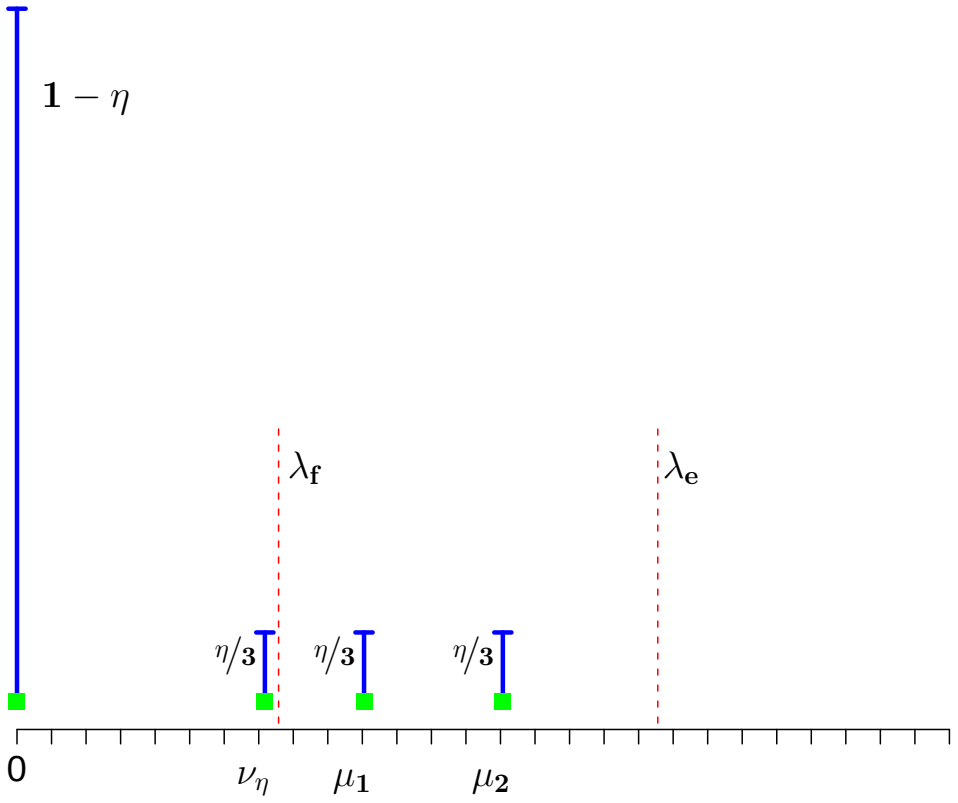

Figure 4.3: The figure shows the support and probability allocation of the Cluster prior $\pi[\eta, r, |Cl^+]$ along with the universal threshold λ_e and the ideal predictive threshold λ_f. Here with $r = 2$, we have 3 equally likely non-zero support points at $\mu_0 = \nu_\eta$, μ_1 and μ_2 which constitute a geometric progression with common ratio 1.4. The abscissa is graduated in a units and is drawn to the scale of $\eta = e^{-1000}$.

4.3.2 Proof of Theorems 4.1.1 and 4.3.1

As our set-up is symmetric it is enough to prove Theorem 4.3.1. We will use the the Bayes-Minimax strategy described before. So, we would calculate a lower bound on the Bayes risk of $\pi[\eta, r, 2]$ (see Lemma 4.3.2) and will produce a matching upper bound on the maximal risk of $\widehat{p}_T[\eta, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}]$ (see Theorem 4.3.4). To interpret the result in terms of predictive game note that the statistician's choice for a density estimate of Y will involve information about θ_0 stored in X as well as the Gaussianity of the noise distribution. And, as the parametric form of the true future density is known an effective density estimate will depend on efficient estimation of θ_0 which in turn depends only on the sufficient statistics. In particular if X and Y were both observed an optimal point estimate of θ_0 based on the sufficient statistics W would produce an optimal density estimate (we will choose the plug-in version). So, for Nature, who aims to set the most difficult predictive set-up, the goal is essentially setting up the most difficult point estimation case for θ_0 based on the fact that X and Y are both observed. She apprehends that the statistician may produce a near accurate point prediction $\widehat{Y}$ of Y based on X. So, the worst possible prior distribution involve point estimation of θ_0 based on observing $(X, Y_\alpha)\big|_{\alpha=1}$. Let us denote a typical 2–point prior in $\mathfrak{m}^+(\eta)$ with its only non-zero support point at ν by

$$\pi_{2\mathrm{pt}}[\eta, \nu](\theta) = \begin{cases} 0 & \text{with prob } 1-\eta \\ \nu & \text{with prob } \eta \end{cases}$$

Note that the sparse two point prior $\pi[\eta, r, 2] = \pi_{2\mathrm{pt}}[\eta, \nu_\eta]$.

Lemma 4.3.2.

$$B(\pi[\eta, r, 2]) \geq \eta \frac{\lambda_f^2}{2r}\,(1 + o(1)) \quad \text{as } \eta \to 0$$

Here, we provide an intuitive (and a bit non-rigorous) proof of the Lemma by using the connections with point estimation (PE) theory. We avoid the intricacies of overshoot term and present asymptotic arguments in the resolution higher than

the $\mathrm{O}(a)$. In Section 4.4 we have rigorous technical proofs of the exact asymptotic behavior of the predictive risk of any 2–point priors in $\mathfrak{m}^{+}$. For those detailed calculation, the connecting equations can not help much as we still need to dig into the asymptotic subtleties of the PE regime. Lemma 4.3.2 and the other 2–point priors results will follow directly from those calculations.

Proof. To compute the predictive Bayes risk of $\pi[\eta, r, 2]$, we will use the Connecting equation [2.3] and known properties of the quadratic risk of $\widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta, \nu]\big]$ – the posterior mean of the 2–point prior $\pi_{2\mathrm{pt}}[\eta, \nu]$.
From PE theory Johnstone (2012, Section 8.4) as $\eta \to 0$, the asymptotic quadratic risk $q\big(\theta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta, \nu]\big], 1\big)$ of the Bayes estimate of $\pi_{2\mathrm{pt}}[\eta, \nu]$ in an estimative set-up with unit noise variance, has the following properties:

Property 1: Because of the very high mass at 0, the risk at 0 will be insignificant (lower than the order of $\eta \log \eta^{-1}$) and the dominant proportion of the Bayes risk will be from the non-zero support point ν.

Property 2: As $\eta \to 0$, the quadratic risk at the non-zero point ν will be of the order of ν^2 as long as $\nu^2 \leq \lambda_e^2 - c\, a\, \lambda_e$ with $c > 0$. Once ν exceeds λ_e the quadratic risk at ν becomes negligible compared to its peak value. Thus, the maximal first order asymptotic quadratic risk is attained when $\nu = \lambda_e$ and

$$q\big(\theta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta, \nu]\big], 1\big) \sim \begin{cases} \nu^2 & \text{if } \nu^2 < \lambda_e^2 - c\, a\, \lambda_e \\ \mathrm{O}(1) & \text{if } \nu^2 \geq \lambda_e^2 + 2a\, \lambda_e \end{cases}.$$

Again, we know that the Gaussian estimative set up with noise variability v can be reduced to an unit variance problem by suitably scaling the observations as well as the location parameter by $v^{1/2}$. Posterior probabilities remain invariant to the transformation leading Bayes point estimates to be similarly scaled. And so, the quadratic risk of Bayes estimates under the variance stabilizing transformation is scaled by variability v, i.e.

$$q\big(\theta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta, \nu]\big], v\big) = v \cdot q\big(v^{-1/2}\theta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta, v^{-1/2}\nu]\big], 1\big).$$

Thus, while computing the predictive risk of the Bayes density estimate of $\pi[\eta, r, 2]$ at ν_η by the Equation [2.3] we have:

$$\begin{aligned}\rho\Big(\nu_\eta, \widehat{p}\big[\pi[\eta, r, 2]\big]\Big) &= \int_{v_w}^{1} \frac{1}{2v^2} q\big(\nu_\eta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta, \nu_\eta]\big], v\,\big)\, dv \\ &= \int_{v_w}^{1} \frac{1}{2v} q\big(v^{-1/2}\nu_\eta, \widehat{\theta}\big[\pi_{2\mathrm{pt}}[\eta, v^{-1/2}\nu_\eta]\big], 1\,\big)\, dv.\end{aligned}$$

Also, for all $v \in [v_w, 1]$ we have $(v^{-1/2}\nu_\eta)^2 \leq \lambda_e^2 - a\,\lambda_e$. So using the aforementioned Property 2, as $\eta \to 0$ we get

$$\rho\Big(\nu_\eta, \widehat{p}\big[\pi[\eta, r]\big]\Big) \geq \int_{v_w}^{1} \frac{1}{2v} \times \frac{\nu_\eta^2}{v}\, dv = \nu_\eta^2 \int_{v_w}^{1} \frac{1}{2v^2}\, dv = \frac{\nu_\eta^2}{2r} = \frac{\lambda_f^2}{2r}\,(1 + \mathrm{o}\,(1)) \text{ as } \eta \to 0$$

and the corresponding predictive Bayes risk satisfies

$$B(\pi[\eta, r, 2]) \geq \eta \times r\Big(\nu_\eta, \widehat{p}\big[\pi[\eta, r]\big]\Big) \geq \frac{\lambda_f^2}{2r}\,(1 + \mathrm{o}\,(1))\,.$$

This completes the proof. □

Actually we can infer more about the prior $\pi[\eta, r]$ and like PE, here too we can show that the prior $\pi[\eta, r]$ is asymptotically least favorable among all 2–points priors.

Lemma 4.3.3.

As $\eta \to 0$, $\pi[\eta, r, 2]$ maximizes the asymptotic Bayes risk in the class of all 2–point priors in $\mathfrak{m}^+$.

Proof. We know (by Property 1 described in Lemma 4.3.2) that the risk at the origin will have insignificant contribution (lower than the order of $\eta\lambda_e^2$) to the Bayes risk. Also, based on Property 2, for maximizing the risk at the non-zero support point the choice of ν is reduced to the set $\{\nu_k = k\lambda_e : k \in [0, 1]\}$. The predictive risk of

$\pi_{2\text{pt}}[\eta, \nu_k]$ at the non-zero support point will be given by,

$$\begin{aligned} B\big(\pi_{2\text{pt}}[\eta,\nu_k]\big) &\sim \eta \cdot \rho\Big(\nu_\eta, \widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu_k]\big]\Big) \\ &= \eta \int_{v_w}^{1} \frac{1}{2v^2} q\big(\nu_\eta, \widehat{\theta}\big[\pi_{2\text{pt}}[\eta,\nu_k]\big], v\,\big)\, dv \\ &= \eta \int_{v_w}^{1} \frac{1}{2v} q\big(v^{-1/2}\nu_\eta, \widehat{\theta}\big[\pi_{2\text{pt}}[\eta,v^{-1/2}\nu_k]\big], 1\,\big)\, dv. \end{aligned}$$

$$\text{Now, as } \eta \to 0,\ q\big(v^{-1/2}\nu_\eta, \widehat{\theta}\big[\pi_{2\text{pt}}[\eta,v^{-1/2}\nu_k]\big], 1\,\big) \sim \begin{cases} v^{-1}\, k^2\, \lambda_e^2 & \text{if } v > k^2 \\ \mathrm{O}\,(1) & \text{if } v \le k^2 \end{cases}$$

so, the asymptotic predictive Bayes risk is

$$\begin{aligned} B\big(\pi_{2\text{pt}}[\eta,\nu_k]\big) &= \eta\Big\{ 2^{-1}\, k^2\, \lambda_e^2 \int_{\max(k^2, v_w)}^{1} v^{-2}\, dv \Big\} \big(1 + \mathrm{o}\,(1)\big) \\ &= k^2 \left\{1 - (\max(k^2, v_w))^{-1}\right\} \eta\lambda_e^2/2 \,\big(1 + \mathrm{o}\,(1)\big) \end{aligned}$$

which is maximized at $k = v_w^{1/2}$. Thus the Bayes risk is maximized for the 2–point prior whose non-zero support point is at λ_f. □

In this context, note that the optimal asymptotic predictive risk is always lower than the square error of point estimation of θ_0 based on X. This is because in the predictive set up (for Nature) the future Y could disclose additional information about θ_0. As such, we will see afterward that the ratio of the optimal predictive to estimated risk is v_w. For the two extreme cases as r approaches 0 and ∞ the ratio tends to 0 and 1 respectively. It validates our intuition about this predictive set-up. As with $r \to \infty$, even knowing Y will not provide any additional information about X. So, for Nature the predictive problem will be as easy to set as the estimation one where only one sample X is explored. Similarly, as $r \to 0$, Y would disclose infinite amount of more information than X. Comparing the optimal risk in point estimation and the predictive regime we can say that predictive density estimation based on KL loss is an easier task than Point Estimation under quadratic loss. However, it does not say that prediction is easier than estimation.

On the contrary, constructing optimal predictive densities (for the statistician) is far more complicated than designing point estimates. The issues that need addressing are

Sparsity Regularization: The prior information of having at least $(1-\eta)$ mass at the origin has to incorporated.

Risk Diversification: The statistician does not see Y but can balance his ignorance about Y by sharing his future uncertainty.

As such, the interplay between sparsity-regularization needs and the dynamics of risk diversification (because of predictive purposes), requires to be explicitly tracked for calibrating optimal predictive schemes. We will see that the optimal strategies will lie outside the given parametric (Gaussian) family. However, there exists at least an asymptotic solution in the class of Gaussian mixture (finite) densities which is flexible enough to optimally balance the regularization vs diversification trade-off.

Theorem 4.3.4.

As $\eta \to 0$, for any $r \in (0, \infty)$ we have

$$\sup_{\pi \in \mathfrak{m}^+(\eta)} B(\pi, \widehat{p}[\eta, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}]) \leq \eta \lambda_f^2/(2r)(1 + \mathrm{o}\,(1)).$$

This theorem is proved in Section 4.5. For controlling the sparsity effect, the statistician can use a threshold density estimate. Threshold rules are a particular class of non-linear estimates which may be successfully employed to devise sparse minimax optimal estimates particularly in location estimation. The idea behind threshold rules is to use an unbiased estimate (generally unbiased or controlled bias) when the observed data is above the threshold and an adjusted one if the observation is below the threshold. Here we see that an optimal choice of threshold can depend entirely on the degree of sparsity η. As only X is observed the statistician is forced to use λ_e as the threshold. He can use the best invariant predictive density $\widehat{p}_U$ if the past observation X crosses λ_e. Below the threshold, his estimate has to account for both

sparsity effect and risk sharing. The rationale for this choice rests on the fact that, because of the severe sparsity constraints a near zero density estimate is required to control the risk at the origin. If nature places the entire remaining mass η between the parametric values $(0, \lambda_f]$, the statistician can perform under the optimal Bayes risk for this problem by using the zero density estimate $\phi(y|0, r)$. However, as the supremum Bayes risk is $\lambda_f^2/(2r)$ the zero estimator is unusable when the parametric value is greater than λ_f. Thresholding ensures that the predictive risk above λ_e is bounded. So, the need is to control the risk in $(\lambda_f, \lambda_e]$ by moving away from the zero estimator. But, this deformation should not be large to affect the risk at the origin and an approach would be to use the Bayes predictive density based on a prior with $(1 - \eta)$ probability at the origin (this leaves the sparsity restrictions untouched) and with η probability distributed approximately in $(\lambda_f, \lambda_e]$.

2–**Player Game and Equilibrium strategies**

Nature: Will choose a distribution on θ which will make point estimation of θ, under quadratic loss and based on observing both the past X and future Y, the most difficult.

Statistician: Will use threshold estimators. He is forced to use λ_e as threshold as he only observes X. The idea is to use 0 estimator when $\theta < \lambda_f$ and share his risk for θ between λ_f and λ_e. A predictive strategy can be constructed by using a prior with probability $1 - \eta$ at 0 and share equally the remaining mass η among a finite chain of points covering λ_f and λ_e.

Decision Theoretic Evaluation Game

Through the above sharing policy we control the transition of the corresponding Bayes predictive density (and subsequently the threshold version) under $\theta \in (0, \lambda_e]$. The Bayes predictive density will be sufficiently close to $\phi(y|0, r)$ till $\theta < \lambda_f$ and thereafter it gradually shifts rightwards in way that the risk at any $\theta \in (\lambda_f, \lambda_e]$ is under the desired limit. The interval $(\lambda_f, \lambda_e]$ increases with decrease of r and the statistician

is completely non-informative in that zone. As such $(\lambda_f, \lambda_e]$ can be regarded as his most **vulnerable zone**. A way to share his predictive risk across that zone would be to divide the probability η equally in a finite chain of point covering the interval. The non-informativeness in $(\lambda_f, \lambda_e]$ is reflected in uniform sharing of the future uncertainty. Distributing the vulnerability in finite locations across the interval is pivotal. As soon as the parametric value θ crosses λ_f we need sharp transitions from the zero-estimator for which there is need for non-zero mass around λ_f. So, continuous sharing policies which are independent of the degree of sparsity η will not work.

In Section 4.7, we will see several efficient alignments of the chain. $\widehat{p}_T(.|x)$ is discontinuous at $x = \lambda_e$ and the number of Gaussian mixtures in $\widehat{p}_T$ and their weights are based on the degree of sparsity η, the future volatility r and the past observation X. However, neither the Bayes density density corresponding to $\pi[\eta, r]$ attain minimax risk nor the cluster prior (which is the basis of $\widehat{p}_T$) is least favorable. But, based on the calculation we can trace an infinite support prior $\pi[\eta, r, \text{INF}]$ on θ which attains supremum risk and produces minimax Bayes density estimate.

4.4 Maximal Bayes risk of 2–point priors

Here, we will work directly with the predictive loss. Calculation will involve deriving closed forms for the Bayes predictive densities. In the process we will show that properties similar to those stated in Lemma 4.3.2 for the quadratic loss also exist for predictive densities.

Posterior probabilities based on $\pi_{2\text{pt}}[\eta, \nu]$ is given by

$$\pi_{2\text{pt}}[\eta, \nu](\theta = 0|\, x) = \frac{(1-\eta)\,\phi(x)}{\eta\,\phi(x-\nu) + (1-\eta)\,\phi(x)} \quad \text{and}$$
$$\pi_{2\text{pt}}[\eta, \nu](\theta = \nu|\, x) = 1 - \pi(\theta = 0|\, x).$$

And so, the corresponding Bayes predictive density is

$$\begin{aligned}\widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\big(y|x\big) =& \frac{(1-\eta)\cdot\phi(x)\cdot\frac{1}{\sqrt{r}}\phi\big(\frac{y}{\sqrt{r}}\big)+\eta\cdot\phi(x-\nu)\cdot\frac{1}{\sqrt{r}}\phi\big(\frac{y-\nu}{\sqrt{r}}\big)}{(1-\eta)\cdot\phi(x)+\eta\cdot\phi(x-\nu)}\\ =& \frac{1}{\sqrt{r}}\cdot\phi\left(\frac{y}{\sqrt{r}}\right)\frac{(1-\eta)+\eta\cdot e^{\nu(x+y/r)-\frac{1+r}{2r}\cdot\nu^2}}{(1-\eta)+\eta e^{\nu x-\frac{1}{2}\nu^2}}\\ =& \phi(y|0,r)\times h_\nu(x,y).\end{aligned}$$

For doing calculations under strong sparsity, we found it most convenient to represent the Bayes predictive densities as tiltings of the zero-density. The tilt function $h_\nu(X,Y)$ is given by

$$h_\nu(x,y)=\big\{1+\eta\,(1-\eta)^{-1}e^{\nu x-\frac{1}{2}\nu^2}\big\}^{-1}\big\{1+\eta\,(1-\eta)^{-1}e^{\nu(x+y/r)-\frac{1+r}{2r}\cdot\nu^2}\big\} \tag{4.11}$$

with both the numerator and denominator being greater than unity which implies that their logarithms are always positive.

Now, from definition we have predictive risk at 0 as,

$$\rho\bigg(0,\widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\bigg)=E_0\left(\log\left(\frac{\phi(Y|0,r)}{\widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\big(Y|X\big)}\right)\right)=-E_0\left\{\log h_\nu(X,Y)\right\}$$

where the expectation is over both X and Y which are independent Gaussian with common mean (denoted in subscript) and known variances 1 and r respectively. Also, note that though there is a negative sign the risk is always positive (as we have showed before that KL divergences are always positive).

The risk at ν is given by

$$\begin{aligned}\rho\Big(\nu, \widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\Big) &= E_\nu\left(\log\left(\frac{\phi(Y|\nu,r)}{\widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big](Y|X)}\right)\right)\\ &= E_\nu\left(\log\left(\frac{\phi(Y|\nu,r)}{\phi(Y|0,r)}\right)\right) - E_\nu\{\log h_\nu(X,Y)\}\\ &= \frac{1}{2r}\cdot E_\nu(2\nu Y-\nu^2) - E_\nu\{\log h_\nu(X,Y)\}\\ &= \frac{\nu^2}{2r} - E_\nu\left\{\log h_\nu(X,Y)\right\}.\end{aligned}$$

As $\eta \to 0$, for any 2–point prior with the very high probability $(1-\eta)$ at 0, the risk at the origin is always small irrespective of where the non-zero support ν is placed. We show in Lemma 4.4.1 that it remains bounded by η. At ν though, the asymptotic risk can be unbounded as $\eta \to 0$. We will chose an optimal ν maximizing this risk and the asymptotic maximal Bayes risk will be solely governed by risk at the non-zero support point in-spite of its low prior probability η.

Now as we move ν away from the origin, the Bayes density estimate at ν initially behaves like $\phi(.|0,r)$ giving rise to the first order asymptotic risk $\nu^2/2r$. In these cases the tilt function h_ν fails to sway the predictive density away from the origin when the true parametric value is ν. However, as we move ν further away from 0, $h_\nu(X,Y)$ will be successful in tilting the predictive density away from $\phi(.|0,r)$ and towards $\phi(y|\nu,r)$. Subsequently, the risk at ν will drop due to appreciable contribution from $E_\nu\left\{\log h_\nu(X,Y)\right\}$.

If the non-zero support point is ν_η (the positive root of the quadratic equation [4.9]) the tilt function is still unable to cause any significant reduction in the first order asymptotic risk at ν_η. The proof follows directly from Lemma 4.4.2. So, with $\eta \to 0$, the first order asymptotic risk of the Sparse 2–point prior $\rho(\nu_\eta, \widehat{p}_{\pi[\eta,r,2]}) \geq \nu_\eta^2/(2r)\,(1+o(1))$ which in turn reproves Lemma 4.3.2 as

$$\begin{aligned}B\big(\pi[\eta,r,2]\big) &= (1-\eta)\times r\big(0, \widehat{p}\big[\pi[\eta,r]\big]\big) + \eta \times r\big(\nu_\eta, \widehat{p}\big[\pi[\eta,r]\big]\big)\\ &\geq \eta\frac{\nu_\eta^2}{2r}\left(1+o(1)\right).\end{aligned}$$

Lemma 4.4.1.

For any $\eta \in [0,1)$ and $\nu \in [0,\infty)$ we have,

$$\rho\left(0, \widehat{p}\big[\pi_{\text{2pt}}[\eta,\nu]\big]\right) \leq \eta(1-\eta)^{-1}.$$

Proof. We use the representation of $h_\nu(X,Y)$ given by Equation [4.11] and so can assume that the logarithm of the numerator there is non-negative. Hence,

$$\begin{aligned}\rho\left(\nu, \widehat{p}\big[\pi_{\text{2pt}}[\eta,\nu]\big]\right) &\leq E_0 \log\left\{1+\eta\,(1-\eta)^{-1} e^{\nu X - \frac{1}{2}\nu^2}\right\} \\ &\leq \eta\,(1-\eta)^{-1} e^{-\nu^2/2}\, E_0\, e^{\nu X}\end{aligned}$$

by using the inequality $\log(1+x) \leq x$ which holds for all non-negative x. Again, as X is standard normal we have $E_0\left(\exp(\nu X)\right) = \exp(\nu^2/2)$ and we have the required result. □

Lemma 4.4.2.

For any $\eta \in [0,1]$ such that $a > 0$ and there exists a positive solution ν_η of Equation [4.9], we have

$$\log(1-\eta) \leq E_\nu\left\{\log h_\nu(X,Y)\right\} \leq 1 + v_w^{-1/2} a^{-2} \text{ for all } \nu \in (0, \nu_\eta].$$

Proof. We first show the upper bound. For that purpose we will use the representation of $h_\nu(X,Y)$ given by Equation [4.11] and so the logarithm of the denominator there is always positive and can be ignored while calculating the upper bound. We can rewrite the numerator in terms of the futuristic random variable $W = (1+r^{-1})^{-1}(X + Y/r)$ and its variance $v_w = (1+r^{-1})^{-1}$ as:

$$E_\nu \log\left\{1+\eta\,(1-\eta)^{-1} \exp\left(v_w^{-1}\,\nu\,W - \frac{1}{2}\,v_w^{-1}\nu^2\right)\right\}$$

where $W \stackrel{d}{=} N(\nu, v_w)$. We change the measure to standard normal $Z = v_w^{-1/2}(W - \nu)$ and it results in

$$E_0 \log \left\{ 1 + \eta \, (1-\eta)^{-1} \exp\left(v_w^{-1/2} \, \nu \, Z + \frac{1}{2} \, v_w^{-1} \nu^2 \right) \right\}$$

Now, as $\nu \leq \nu_\eta$ by Equation [4.9] we have,

$$\eta \, (1-\eta)^{-1} \exp\left(v_w^{-1/2} \, \nu \, Z + \frac{1}{2} \, v_w^{-1} \nu^2 \right) = c_\nu \, \exp\left(v_w^{-1/2} \, \nu \, (Z - a) \right)$$

where c_ν is a constant and $c_\nu \in [0, 1]$. Choosing $c_\nu = 1$ we get the upper bound,

$$E_\nu \left\{ \log h_\nu(X, Y) \right\} \leq E_0 \log \left(1 + \exp\left(v_w^{-1/2} \, \nu \, (Z - a) \right) \right).$$

Now, we decompose the expectation of the random variable on R.H.S. conditioned on the event $\{Z > a\}$. When $Z \leq a$, the random variable (whose expectation is considered) is bounded by $\log 2$.
When $Z > a$, we use the naive bound: '$\log(1 + x) \leq 1 + \log x$ if $x > 1$' for bounding the said random variable. Aggregating the two parts the ultimate bound would be

$$\begin{aligned} E_\nu \left\{ \log h_\nu(X, Y) \right\} &\leq \log 2 \cdot P(Z \leq a) + \left[P(Z > a) + v_w^{-1/2} \, \nu E(Z - a)_+ \right] \\ &\leq 1 + v_w^{-1/2} \, \nu E(Z - a)_+ \end{aligned}$$

and the truncated Gaussian expectation can be exactly computed as

$$E(Z - a)_+ = \int_a^\infty z \phi(z) \, dz - a \, \widetilde{\Phi}(a) = \phi(a) - a \, \widetilde{\Phi}(a) \leq a^{-2} \phi(a) \leq a^{-2} \lambda_f^{-1}$$

where the first inequality uses the result (4.22) about Mill's Ratio. As $\nu_\eta \leq \lambda_f$, for any $\nu \leq \nu_\eta$ we have $\nu E(Z - a)_+ \leq a^{-2}$. So, $E_\nu \left\{ \log h_\nu(X, Y) \right\}$ is also bounded by $1 + v_w^{-1/2} a^{-2}$.

For lower bound we can similarly neglect the numerator of $h_\nu(X,Y)$ and so,

$$\begin{aligned} E_\nu\left\{\log h_\nu(X,Y)\right\} &\geq -E_\nu \log\left(1+\eta\,(1-\eta)^{-1}e^{\nu X-\frac{1}{2}\nu^2}\right) \\ &\geq -\log\left(1+\eta\,(1-\eta)^{-1}e^{-\frac{1}{2}\nu^2}E_\nu e^{\nu X}\right) \end{aligned}$$

which follows by Jensen's inequality. Noting, that $X\sim N(\nu,1)$ the bound simplifies to $-\log\left(1+\eta\,(1-\eta)^{-1}\right)=\log(1-\eta)$. □

By Lemma 4.4.1 and Lemma 4.4.2 the asymptotic behavior of $B\big(\pi_{2\text{pt}}[\eta,\nu]\big)$ can be characterized when ν varies in $[0,\nu_\eta]$. Next, we track $\rho\left(\nu,\widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\right)$ (and hence $B\big(\pi_{2\text{pt}}[\eta,\nu]\big)$ by using Lemma 4.4.1) when $\nu>\lambda_f$.
We will prove that the risk remains bounded as ν crosses the threshold λ_e (Lemma 4.4.3). In between λ_f and λ_e we can show that the risk is decreasing (Lemma 4.4.4). However, the descent is gradual and there is no abrupt transition in the first order risk before λ_e. Thus, the maximal Bayes risk for 2–point prior is attained around $\nu=\lambda_f$ and we have effectively characterized the first order behavior of the risk with ν varying along the positive axis in resolution of a units. As such, as $\eta\to 0$

$$\rho\left(\nu,\widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\right)\sim\begin{cases} {}^{\nu^2}\!/_{2r} & \text{if } \nu\leq\nu_\eta \\ \text{decreasing} & \text{if } \lambda_f+(2\,v_w^{-1})^{1/2}a\leq\nu\leq\lambda_e-2a \\ \text{O}\,(1) & \text{if } \nu>\lambda_e+(2v_w^{-1})^{1/2}a \end{cases}.$$

We do not quantify the rate of descent of $\rho\left(\nu,\widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\right)$ though it can be approximated from our proof of Lemma 4.4.4. In Figure [4.4], we trace the asymptotic risk by Monte Carlo simulation. It depicts the gradual descent which compared to PE regime is a contrast.

Lemma 4.4.3.

For any $\eta\in(0,1)$ such that $a>0$ we have,

$$\rho\left(\nu,\widehat{p}\big[\pi_{2\text{pt}}[\eta,\nu]\big]\right)\leq(4\sqrt{\pi}\,a\,r)^{-1}+\log 2 \text{ for all } \nu>\lambda_e+(2v_w^{-1})^{1/2}a.$$

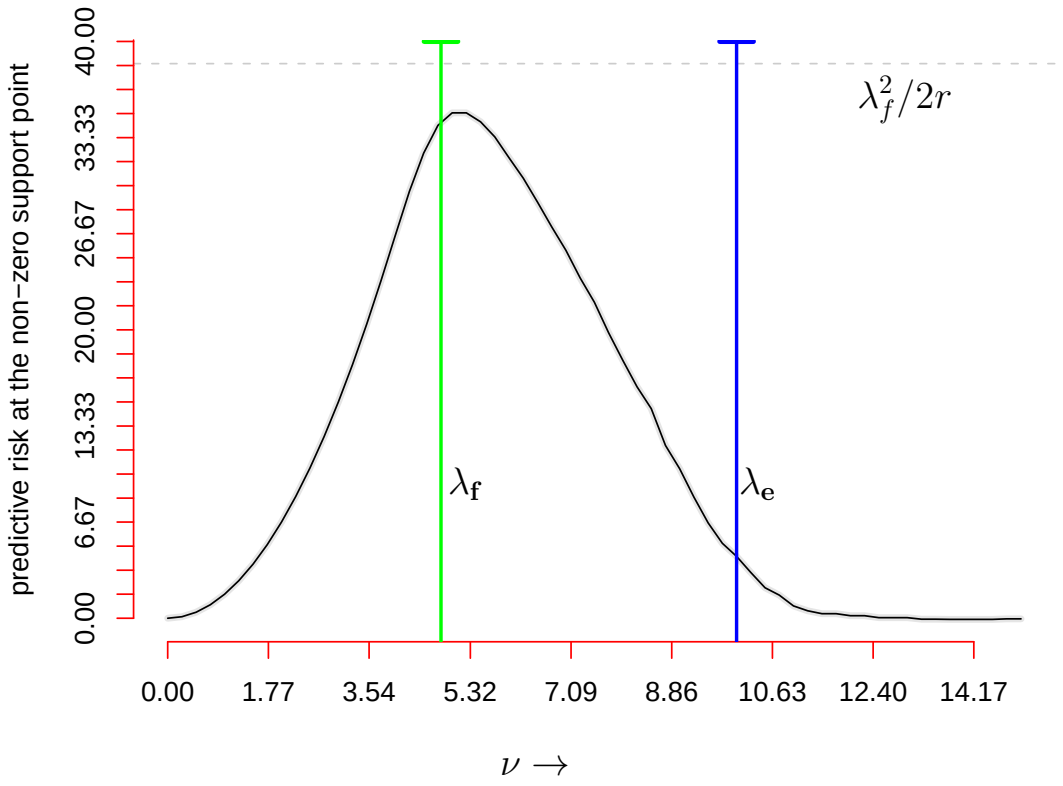

Figure 4.4: The plot shows the risk $\rho\left(\nu, \widehat{p}\left[\pi_{2\text{pt}}[\eta, \nu]\right]\right)$ of the Bayes predictive density from the two point prior $\pi_{2\text{pt}}[\eta, \nu]$ at the non-zero support point ν as ν is moved along the positive axis. λ_f and λ_e are marked by vertical lines and the horizontal gray line represents the first order maximal risk of $\lambda_f^2/2r$. Reflecting the resolution of our asymptotic calculations the abscissa is ticked at multiples of a while the ordinate is marked in multiple of $1/(2r)$ units to represent change in order of quadratic loss. The figure is actually drawn according to scale with $\eta = e^{-50}$, $r = 0.3$ producing $\lambda_f = 4.8$, $\lambda_e = 10$, $a = 1.77$.

Proof. We know that $\rho\left(\nu, \widehat{p}\big[\pi_{\text{2pt}}[\eta,\nu]\big]\right) = \nu^2/2r - E_\nu\{\log h_\nu(X,Y)\}$. As before we can standardize the measure to standard Gaussian. Because of the logarithm the numerator and denominator of h_ν separates and can be standardized simultaneously. After standardization, we have:

$$\rho\left(\nu, \widehat{p}\big[\pi_{\text{2pt}}[\eta,\nu]\big]\right) = \frac{\nu^2}{2r} - E_0 \log\left\{\frac{1+\eta\,(1-\eta)^{-1}\exp\left(v_w^{-1/2}\,\nu\,Z + \frac{1}{2}\,v_w^{-1}\nu^2\right)}{1+\eta\,(1-\eta)^{-1}\exp\left(\nu\,Z + \frac{1}{2}\nu^2\right)}\right\} \quad (4.12)$$

$$= \frac{\nu^2}{2r} - E_0 \log\left\{\frac{1+\exp\left(v_w^{-1/2}\,\nu\,Z + \frac{1}{2}\,v_w^{-1}\nu^2 - \frac{1}{2}\lambda_e^2\right)}{1+\exp\left(\nu\,Z + \frac{1}{2}\nu^2 - \frac{1}{2}\lambda_e^2\right)}\right\} \quad (4.13)$$

by substituting $\eta(1-\eta)^{-1}$ by $\exp(-\lambda_e^2/2)$. Note that $\nu > \lambda_e + (2v_w^{-1})^{1/2}a$ implies $2\Delta = \nu^2 - 2\nu b - \lambda_e^2 > 0$ where $b = (4\,v_w^{-1}\log\nu)^{1/2}$ and hence the expectation above can be rewritten as

$$E_0 A(Z) \text{ where } A(Z) = \log\left\{\frac{1+\exp\left(v_w^{-1/2}\,\nu\,(Z+bv_w^{1/2})\right)\cdot\exp(\Delta)\cdot\exp(\nu^2/2r)}{1+\exp\left(\nu\,(Z+b)\right)\cdot\exp(\Delta)}\right\}$$

where Z is standard normal distribution. Noting the following properties about $A(Z)$:

(i) $A(Z) \geq 0$ if $Z \geq -\nu$

(ii) $A(Z) \geq -\log 2$ if $Z < -\nu$

(iii) $A(Z) \geq \nu^2/(2r) + (v_w^{-1/2} - 1)\,\nu\,Z - \log 2$ if $Z > -bv_w^{1/2}$

it follows that $E_0\{A(Z)\} \geq \nu^2\,\Phi(\sqrt{4\log\nu})\,/(2r) - \log 2$, and eventually we have

$$\rho\left(\nu, \widehat{p}\big[\pi_{\text{2pt}}[\eta,\nu]\big]\right) \leq \frac{\nu^2}{2r}\cdot\widetilde{\Phi}\big((4\log\nu)^{1/2}\big) + \log 2 \leq (4ar\sqrt{\pi})^{-1} + \log 2$$

where the second inequality follows from Equation 4.22. Thus, we get the result. □

Lemma 4.4.4.

As $\eta \to 0$ and $\nu^2 \in \left(\left\{\lambda_f + (2\,v_w^{-1})^{1/2}a\right\}^2,\ \lambda_e^2 - 2\,a\,\lambda_e\right)$ the risk $\rho\left(\nu, \widehat{p}\big[\pi_{\text{2pt}}[\eta,\nu]\big]\right)$ is dominated by a decreasing function of ν which is bounded above by $\lambda_f^2/(2r)$.

Proof. The risk is given by Equation 4.12. Similarly as before we can show that as long as $\nu^2 < \lambda_e^2 - 2a\lambda_e$ the contribution from the denominator is insignificant and

$$E_0 \log\left\{1 + \exp\left(\nu\, Z + \frac{1}{2}\nu^2 - \frac{1}{2}\lambda_e^2\right)\right\} \leq 1 + v_w^{-1/2}$$

and whenever $\nu \geq \lambda_f + (2\,v_w^{-1})^{1/2}a$ we have,

$$E_0 \log\left\{1 + \exp\left(v_w^{-1/2}\,\nu\, Z + \frac{1}{2}\,v_w^{-1}\nu^2 - \frac{1}{2}\lambda_e^2\right)\right\} \geq \frac{1}{2}\left(v_w^{-1}\nu^2 - \lambda_e^2\right).$$

So, ultimately we will have

$$\begin{aligned}\rho\left(\nu, \widehat{p}\big[\pi_{\text{2pt}}[\eta,\nu]\big]\right) &\leq \frac{\nu^2}{2r} - \frac{1}{2}\left(v_w^{-1}\nu^2 - \lambda_e^2\right) + 1 + v_w^{-1/2} \\ &= \frac{1}{2}\left(\lambda_e^2 - \nu^2\right) + 1 + v_w^{-1/2}\end{aligned}$$

which is decreasing in ν and bounded above by $(\lambda_e^2 - \lambda_f^2)/2 = \lambda_f^2/(2r)$. □

It will be seen that to prove Theorem we only need Lemma 4.3.2 for the lower bound. In that regard, the extensive calculations in this section may seem to be a digression from the objective. However, these results not only provide more intuition about the predictive regime but will also help to follow the risk calculations in the next section where the risk of the density estimate from a multi (K) point prior compounded with thresholding complications is evaluated.

4.5 Minimax upper bound

To simplify notations the univariate strategy $\widehat{p}[\eta, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}]$ and the cluster prior $\widehat{p}[\eta, r, \mathrm{CL}^+]$ will be abbreviated as $\widehat{p}_T$ and $\pi[\eta, r, K]$ through out this section. K is the number of support point in the cluster prior and is a function of η and r. As mentioned before, the $\widehat{p}_T$ is governed by 2 major effects: thresholding and risk diversification. The risk diversification procedure involves (a) probability allocation (b) alignment of non-zero support points across the vulnerable zone, which in our $\widehat{p}_T$ is characterized by $\pi[\eta, r, K]$. In this section, the risk calculations of $\widehat{p}_T$ are carried out in a manner such that they can be easily generalized for other reasonable discrete probability sharing schemes (with $1-\eta$ probability at the origin). In particular, we incorporate the peculiar alignment of the non-zero support points of $\pi[\eta, r, K]$ only toward the end of the section and the results before Lemma 4.5.4 will be reused in Section 4.7 to display other feasible sharing schemes.

Proof of Theorem 4.3.4

We characterize the Bayes predictive density for the Cluster prior. Using Bayes formula, the posterior distribution of $\pi[\eta, r, K]$ is given by:

$$\pi[\eta, r, K](0\,|x) = \frac{(1-\eta)\,\phi\,(x)}{(1-\eta)\,\phi\,(x) + \eta/(K+1)\sum_{i=0}^{K}\phi\,(x-\mu_i)}$$
$$\pi[\eta, r, K](\mu_j\,|x) = \frac{\eta/(K+1)\,\phi\,(x-\mu_j)}{(1-\eta)\,\phi\,(x) + \eta/(K+1)\,\sum_{i=0}^{K}\phi\,(x-\mu_i)}, \quad j = 0, \ldots, K$$

and the predictive density $\widehat{p}\big[\,\pi[\eta, r, K]\,\big]\big(y\big|x\big)$ is given by

$$\pi[\eta, r, K](0\,|x)\,\cdot\,\phi(y|0, r) + \sum_{i=0}^{K}\pi[\eta, r, K](\mu_i\,|x)\,\cdot\,\phi(y|\mu_i, r)$$
$$= \frac{(1-\eta)\,\phi(x)\,\phi(y|0, r) + \eta/(K+1)\sum_{i=0}^{K}\phi(x-\mu_i)\,\phi(y|\mu_i, r)}{(1-\eta)\,\phi(x) + \eta/(K+1)\sum_{i=0}^{K}\phi(x-\mu_i)}$$

which like the 2-point prior case can be rewritten as a tilt function acting on the zero density $\widehat{p}\big[\pi[\eta, r, K]\big] = \phi(y|0, r) \times h\big[\,\pi[\eta, r, K]\,\big](x, y)$ where

$$h\big[\,\pi[\eta, r, K]\,\big](x, y) = \frac{1 + \frac{\eta(1-\eta)^{-1}}{K+1} \sum_{i=0}^{K} \exp\left\{\mu_i(x + \frac{y}{r}) - \frac{1}{2}(1 + r^{-1})\mu_i^2\right\}}{1 + \frac{\eta(1-\eta)^{-1}}{K+1} \sum_{i=0}^{K} \exp\left\{\mu_i x - \frac{1}{2}\mu_i^2\right\}} \,. \quad (4.14)$$

Observing $X = x$, its predictive loss at the point θ is expressed as :

$$\begin{aligned} \mathrm{L}\Big(\theta, \widehat{p}\,\big[\,\pi[\eta, r, K]\,\big]\big(.\,|X = x\big)\Big) &= \mathbb{E}_\theta\left[\log\left(\frac{\phi(Y|\theta, r)}{\widehat{p}[\,\pi[\eta, r, K]\,](Y|x)}\right)\right] \\ &= \frac{\theta^2}{2r} - \mathbb{E}_\theta\Big\{\,\log\big(h\big[\,\pi[\eta, r, K]\,\big](x, Y)\big)\Big\} \end{aligned}$$

where the discounted location distance $\theta^2/2r$ appears due to the predictive loss between $\phi(Y\,|\,\theta, r)$ and $\phi(Y\,|\,0, r)$.

We would need to study the behavior of the conditional expectation as well as the unconditional $\mathbb{E}_\theta\big\{\,\log\big(h\big[\,\pi[\eta, r, K]\,\big](\,X,\,Y\,)\big)\big\}$ (the expectation is over both X and Y) in details. For that purpose it will be helpful to change the measure to central Gaussian with X and Y having variances 1 and r respectively

$$\mathbb{E}_\theta\Big\{\,\log\big(h\big[\,\pi[\eta, r, K]\,\big](X, Y)\big)\Big\} = \mathbb{E}_0\big(N_\theta(X, Y) - D_\theta(X)\big)$$

where N_θ and D_θ are the logarithms of the numerator and denominator of $h\big[\,\pi[\eta, r, K]\,\big]$ and

$$\begin{aligned} N_\theta(x, y) &= \log\left[1 + \frac{\eta(1-\eta)^{-1}}{K+1} \sum_{i=0}^{K} \exp\left\{\mu_i\left(x + \frac{y}{r}\right) - \frac{1}{2}v_w^{-1}\mu_i^2 + v_w^{-1}\mu_i\theta\right\}\right] \\ &= \log\left[1 + \frac{\eta(1-\eta)^{-1}}{K+1} \sum_{i=0}^{K} \exp\left\{v_w^{-1}\left(\mu_i w - \frac{1}{2}\mu_i^2 + \mu_i\theta\right)\right\}\right] \end{aligned}$$

where W is the semi-futuristic random variable and $W \sim N(0, 1)$. Now, using the fact that $\eta^{-1}(1-\eta)$ actually equals $\exp(v_w^{-1}\lambda_f^2/2)$ (by Equation 4.9) we have,

$$N_\theta(x,y) = \log\left[1 + \frac{1}{K+1}\sum_{i=0}^{K}\exp\left\{v_w^{-1}\left(\mu_i w - \frac{1}{2}\mu_i^2 + \mu_i\theta - \frac{1}{2}\lambda_f^2\right)\right\}\right] \tag{4.15a}$$

$$= \log\left[1 + (K+1)^{-1}\sum_{i=0}^{K}\exp\left\{v_w^{-1}\Gamma_i(\theta)\right\} \times \exp\left\{v_w^{-1}\mu_i(w+a)\right\}\right] \tag{4.15b}$$

$$\text{where } \Gamma_i(\theta) = \mu_i\theta - \frac{1}{2}\mu_i^2 - a\mu_i - \frac{1}{2}\lambda_f^2. \tag{4.15c}$$

Thus when $v_w^{-1}\mu_i(W+a) \geq 0$ for all $i \in \{0, 1, \ldots, K\}$ a very naive lower bound is

$$N_\theta(x,y) \geq v_w^{-1}\Gamma(\theta) - \log(K+1) \text{ where } \Gamma(\theta) = \max_{i=0}^{K}\Gamma_i(\theta). \tag{4.15d}$$

Similarly, for the denominator we have

$$D_\theta(x) = \log\left[1 + \frac{\eta(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K}\exp\left\{\mu_i x - \frac{1}{2}\mu_i^2 + \mu_i\theta\right\}\right]. \tag{4.15e}$$

Next we calculate the risk of our threshold estimate

$$\widehat{p}_T(y|x) = \begin{cases} \widehat{p}\big(y|x;\pi[\eta,r,K]\big) & \text{if } X \leq \lambda_e \\ \widehat{p}\big(y|x;\pi_U\big) & \text{if } X > \lambda_e \end{cases}$$

We will later see that the threshold estimator $\widehat{p}_T(.|x)$ is discontinuous at $x = \lambda_e$. However, modifications of $\widehat{p}_U$ can be used to incorporate continuity correction. We are interested in finding the maximum risk of $\widehat{p}_T$. However, due to high prior probability concentration at 0 we have to treat the risk at the origin separately. Depending on X the loss of the threshold estimate is given by:

$$\mathrm{L}\Big(\theta, \widehat{p}\big[\pi[\eta,r,K]\big](.|x)\Big) = \frac{\theta^2}{2r} - \mathbb{E}_0 N_\theta(x-\theta, Y) + D_\theta(x-\theta) \qquad \text{if } x \leq \lambda_e$$

$$\mathrm{L}\big(\theta, \widehat{p}\big[\pi_U\big](.|x)\Big) = \frac{1}{2}\big(\log(1+r^{-1}) - (1+r)^{-1}\big) + \frac{(x-\theta)^2}{2(1+r)} \qquad \text{if } x > \lambda_e$$

where the loss of $\widehat{p}_{\pi_U}$ follows from the risk calculations of linear estimators (see Appendix). Now, averaging over the observed data X, the risk of $\widehat{p}_T$ will be given by:

$$\rho(\theta, \widehat{p}_T) = \rho_B(\theta) + \rho_A(\theta)$$

where $\rho_A(\theta)$ is the risk when X crosses the threshold

$$\rho_A(\theta) = \frac{1}{2}\big(\log(1 + r^{-1}) - (1 + r)^{-1}\big) P_\theta(X \geq \lambda_e) + \frac{\mathbb{E}_\theta[(X - \theta)^2 I_{\{X \geq \lambda_e\}}]}{2(1 + r)}$$

and the component of risk from below the threshold is

$$\begin{aligned}
\rho_B(\theta) &= \frac{1}{2r}\big[\, \rho_{B,1}(\theta) - \rho_{B,2}(\theta) + \rho_{B,3}(\theta)\,\big] \quad \text{where,} \\
&\rho_{B,1}(\theta) = \theta^2\, \Phi(\lambda_e - \theta) \\
&\rho_{B,2}(\theta) = 2r\mathbb{E}_0\left[N_\theta(X, Y) I_{\{X \leq \lambda_e - \theta\}}\right] \\
&\rho_{B,3}(\theta) = 2r\mathbb{E}_0\left[D_{_\theta}(X) I_{\{X \leq \lambda_e - \theta\}}\right].
\end{aligned}$$

As we have mentioned before, due to the very high probability of the parameter to concentrate at the origin we need to bound both $\rho_A(0)$ and $\rho_B(0)$ with high precision. Note that, by definition $N_\theta(X, Y) \geq 0$ and so $\rho_B(0) \leq 2r\mathbb{E}_0\big(D_0(X)\, I_{\{X \leq \lambda_e\}}\big)$ which again equals $\eta + o(\eta)$ as $\eta \to 0$ by Lemma 4.5.1. Though $\rho_A(0)$ will be significantly larger than $\rho_B(0)$, it will not be enough to carry the risk at 0 above the maximum value and

$$\begin{aligned}
\rho_A(0) \leq& \frac{1}{2} \log\left(1 + r^{-1}\right) \tilde{\Phi}(\lambda_e) + \frac{1}{2(1 + r)} \mathbb{E}_0\big(X^2 I_{\{X \geq \lambda_e\}}\big) && (4.16) \\
\leq& \frac{1}{2}\left(\log \frac{r + 1}{r} \cdot \tilde{\Phi}(\lambda_e) + \frac{1}{1 + r} \lambda\phi(\lambda_e) \right) && (4.17) \\
=& O(\eta\lambda) \quad [\text{as } \phi(\lambda_e) = \eta/\sqrt{2\pi}]. && (4.18)
\end{aligned}$$

The second inequality follows from calculation involving the risk of hard threshold point estimators where we know that $\mathbb{E}_0\big(X^2 I_{\{X \geq \lambda_e\}}\big) = \lambda_e \phi(\lambda_e)$ Johnstone (2012, Equation 8.15) and the third one uses the result in equation 4.22. Thus, the risk at

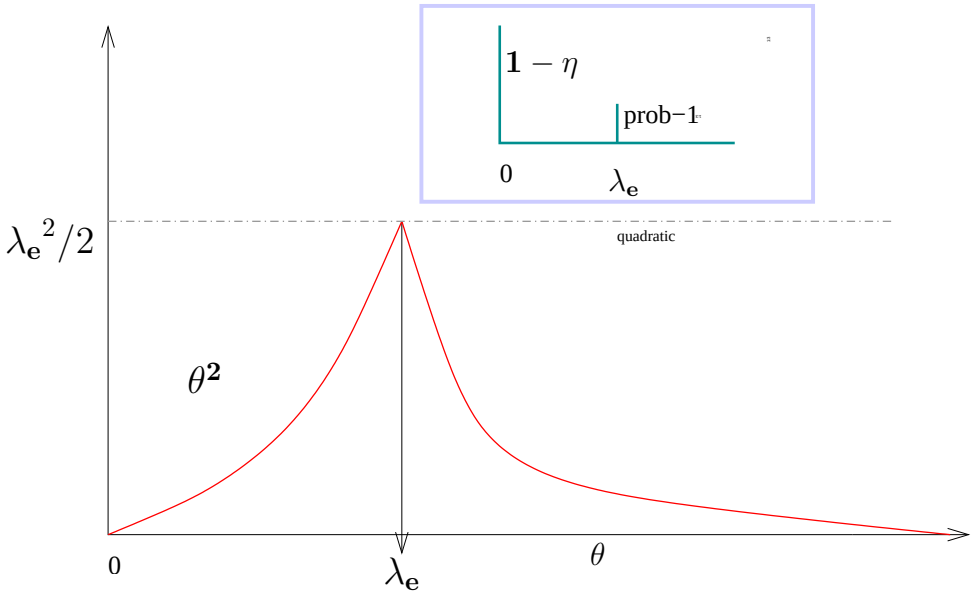

Figure 4.5: Schematic diagram of the behavour of the quadratic Risk under sparse minimax point estimation.

0 stays well below the maximal value.

Next we need to produce an upper bound on the maximum risk at any non-zero parametric point. While working with $\rho(0, \widehat{p}_T)$ we saw that the significant contribution came from $\rho_A(0)$. Whereas, over $\mathbb{R}^+$ the maximal predictive risk is governed solely by $\rho_B(\theta)$ which can be unbounded as $\eta \to 0$ while $\rho_A(\theta)$ remains bounded by $2^{-1}\{\log(1 + r^{-1}) + (1 + r)^{-1}\}$.

Now we trace the behavior of $\rho_B(\theta)$ as θ varies. It will vividly demonstrate how the dynamics of sharing future risk can be co-ordinated with sparsity prior information.

$\rho_{B,1}(\theta)$ is the dominant portion of quadratic risk of the Hard threshold point estimator of θ. From point estimation theory we know that it behaves as θ^2 until the threshold λ_e and then shrinks to 0 with a steep decent (see Figure [4.5]). $\rho_{B,2} - \rho_{B,3}$ is the diversification or aggregation effect. $\rho_{B,3}$ being based on X entirely will be insignificant before λ_e due to sparsity and negligible thereafter due to thresholding effect. It is technically proved in Lemma 4.5.1. So, $\rho_{B,2}$ portrays the diversification effect. It is dormant before λ_f. In between $\lambda_f + a$ and λ_e, $\rho_{B,2}$ produces significant contribution and is effective in bringing the predictive risk $\rho(\theta, \widehat{p}_T)$ below $\lambda_f^2/2r$. The

technical details of the following first order behavior of ρ_B is carried out in a serially in the Lemma 4.5.2, Lemma 4.5.3 and Lemma 4.5.4:

$$\rho_{B,1}(\theta) \sim \begin{cases} \theta^2 & \text{if } \theta < \lambda_e \\ \mathrm{O}\left(\lambda_f\right) & \text{if } \theta \geq \lambda_e + a \end{cases}$$

$$\rho_{B,2}(\theta) \sim \begin{cases} 2g(\theta) & \text{if } \theta \in [\mu_0 + a, \lambda_e + a] \\ 0 & \text{otherwise} \end{cases}$$

$$\rho_{B,3}(\theta) \sim 0 \text{ for all } \theta$$

$$\text{And, } \rho_{B,1}(\theta) - 2g(\theta) \leq \mu_0^2 \text{ for } \theta \in [\mu_0 + a, \lambda_e + a].$$

where $g(\theta) = 2(1+r)\,\Gamma(\theta)\,q(\theta) - 2r\log(K+1)$ with $\Gamma(\theta)$ and $q(\theta)$ defined in Lemma 4.5.3. So we have,

$\sup_{\theta>0}\ \rho(\theta, \widehat{p}_T) \leq \lambda_f^2/(2r)(1+o(1))$ and the minimax predictive risk of $\widehat{p}_T$ equals

$$\sup_{\pi \in \mathfrak{m}^+(\eta)} \int \rho(\theta, \widehat{p}_T)\,\pi(\theta)\,d\theta \leq (1-\eta)\,\rho(0, \widehat{p}_T) + \eta \sup_{\theta>0}\ \rho(\theta, \widehat{p}_T) \ \leq \eta \frac{\lambda_f^2}{2r}(1 + o\,(1)).$$

Figure 4.5 and Figure 4.1 show schematic diagrams of the univariate risk plot of threshold rules in the PE and density estimation framework. The sole difference between the two regimes is the reduction $\rho_{B,2}(\theta)$ in the predictive risk in the vulnerable zone $[\lambda_f, \lambda_e]$. Plug-in density estimates fail to attain this risk reduction and have risk properties like optimal threshold estimates in PE. To obtain the risk reduction we need to have diversified predictive schemes. Lemma 4.5.3 provides a crude lower bound on the decrement and Lemma 4.5.4 shows that it is sufficient enough to attain first order optimality. Figure 4.6 contains Monte-Carlo simulation of the different predictive risk curves.

Lemma 4.5.1.

For any $\eta \in (0,1)$ such that λ_e is well defined and greater than 1, we have

$$\textbf{(a)} \quad \mathbb{E}_0\left\{D_0(X)\,\mathbb{I}[X \le \lambda_e]\right\} \le -\log(1-\eta)$$
$$\textbf{(b)} \quad \mathbb{E}_0\left\{D_\theta(X)\,\mathbb{I}[X \le \lambda_e - \theta]\right\} \le \log 2$$

Proof. The first inequality follows directly from Jensen's inequality,

$$\begin{aligned}\mathbb{E}_0\left\{D_0(X)\,\mathbb{I}[X \le \lambda_e]\right\} &\le \log\left[\mathbb{E}_0\left\{\left(1 + \frac{\eta\,(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K} e^{\mu_i X - \frac{1}{2}\mu_i^2}\right)\mathbb{I}[X \le \lambda_e]\right\}\right] \\ &\le \log\left[1 + \frac{\eta\,(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K}\mathbb{E}_0\left(e^{\mu_i X - \frac{1}{2}\mu_i^2}\right)\right] \\ &= \log\left(1 + \eta\,(1-\eta)^{-1}\right) = -\log(1-\eta)\end{aligned}$$

For the second inequality, as $D_\theta(X)$ is an increasing function of X we have

$$\begin{aligned}\mathbb{E}_0\left\{D_\theta(X)\,\mathbb{I}[X \le \lambda_e - \theta]\right\} &\le D_\theta(\lambda_e - \theta)\,\Phi(\lambda - \theta) \quad \text{where,} \\ D_\theta(\lambda - \theta) &= \log\left(1 + \frac{\eta(1-\eta)^{-1}}{K+1}\sum_{i=0}^{K} e^{\mu_i\lambda_e - \frac{1}{2}\mu_i^2}\right).\end{aligned}$$

Also, as for each $i \in \{0,1,\ldots,K\}$ we have $0 \le \mu_i \le \lambda_e + a$, and so the maximum value of $\mu_i\lambda_e - \mu_i^2/2$ is at most $\lambda^2/2$ for all $i \in \{0,1,\ldots,K\}$ which would imply that $D_\theta(\lambda-\theta) \le \log\left(1 + \eta(1-\eta)^{-1}\exp(\lambda_e^2/2)\right) \le \log 2$. This completes the proof. □

Lemma 4.5.2.

For any $\theta \ge \lambda_e + a$, $\rho_{B,1}(\theta) \le O(\lambda_f)$ as $\eta \to 0$.

Proof. This Lemma follows from the risk calculations of threshold point estimates.

Taking derivative, we see

$$\rho'_{B,1}(\theta) = 2\theta\,\Phi(\lambda_e - \theta) - \theta^2\phi(\lambda_e - \theta) \text{ and so for } t \geq 0$$
$$\rho'_{B,1}(\lambda_e + t) = 2(\lambda_e + t)\Phi(t) - (\lambda_e + t)^2\phi(t) \leq \left(2t^{-1} - (\lambda_e + t)\right)(\lambda_e + t)\,\phi(t)$$

where the inequality follows from Equation [4.22]. Hence, for all $t \geq 2\,\lambda_e^{-1}$, $\rho'_{B,1}(\lambda_e + t)$ is negative. As, $\eta \to 0$, we have $a \geq 2\,\lambda_e^{-1}$ which implies that $\rho_1(\theta)$ is a decreasing function of θ as $\theta > \lambda_e + a$. So, as $\eta \to 0$,

$$\sup_{\theta \geq \lambda_e + a} \rho_{B,1}(\theta) \leq (\lambda_e + a)^2\tilde{\Phi}(a) \leq (\lambda + a)^2\,a^{-1}\phi(a) = (\lambda + a)^2 a^{-1}\lambda_e^{-1} = O(\lambda_f).$$

□

Lemma 4.5.3.

As $\eta \to 0$ and for $\theta \in [\nu_\eta + a, \lambda_e + a)$,

$$\begin{aligned}\rho_{B,2}(\theta) \;&\geq 2(1 + r)\,\Gamma(\theta)\,q(\theta) - 2r\log(K + 1) \quad \text{where,}\\ &\Gamma(\theta) = \max_{i=0}^{K}\left(\mu_i\theta - \frac{1}{2}\mu_i^2 - a\mu_i - \frac{1}{2}\lambda_f^2\right) \text{ and}\\ &q(\theta) = \Phi(\lambda_e - \theta) - \widetilde{\Phi}(a) - \widetilde{\Phi}(ar^{-1/2}).\end{aligned}$$

In particular when $\theta \in (\lambda, \lambda + a)$, the bound shown in the Lemma can be negative which certainly proves that its crudeness as we already know that $\rho_{B,2}$ is always non-negative. However, the bound is intentionally kept crude as it helped to increase clarity in some of the later proofs.

Proof. Using Inequality 4.15d and the fact that $N_\theta(X, Y)$ is non-negative, we get the following lower bound:

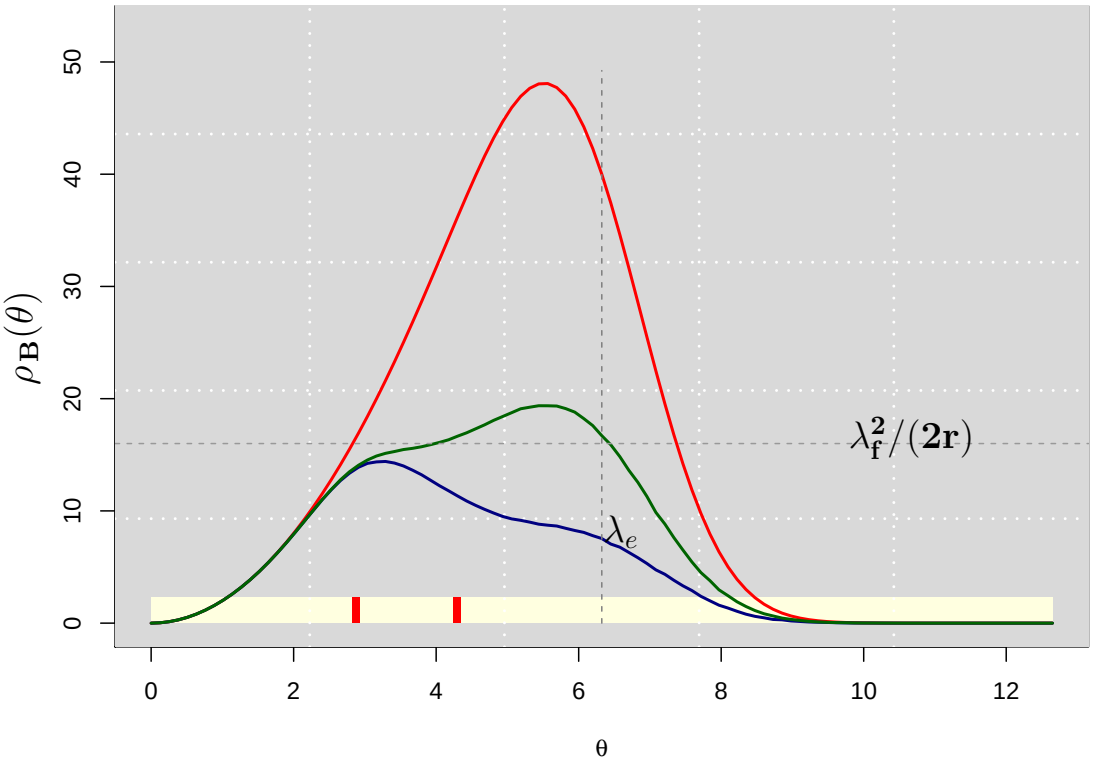

Figure 4.6: Plot of the dominant portion of the predictive risk $\rho_B(\theta)$ as θ varies over the positive axis. In red, green and blue are respectively the risks of the optimal hard threshold plug-in estimator, unshared prediction scheme $\widehat{p}[r, \mathrm{T}, \pi[\eta, r, 2], \mathrm{U}]$ and the minimax optimal density estimate $\widehat{p}[r, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}]$. Here, $r = 0.25$, $\eta = e^{-20}$, $\lambda_f = 2.83$ and $\lambda_e = 6.32$. The red boxes at 2.83 and 4.24 in the yellow bar zone show the non-zero support point of the cluster prior $\pi[\eta, r, \mathrm{CL}^+]$.

$$\begin{aligned}
&\mathbb{E}_0\Big\{ N_\theta(X,Y)\, \mathbb{I}[X \le \lambda_e - \theta] \Big\} + \log(K+1) \\
&\ge v_w^{-1}\, \Gamma(\theta) \;\times\; P_0\big(X \le \lambda_e - \theta \text{ and } v_w^{-1}\, \mu_i(W+a) \ge 0 \text{ for } i = 0, 1, \ldots, K\big) \\
&\ge v_w^{-1}\, \Gamma(\theta) \;\times\; P_0\big(X \le \lambda_e - \theta \text{ and } W \ge -a\big)
\end{aligned}$$

as each of the μ_i is positive and we the bound the probability by

$$\begin{aligned}
&P_0\big(X \le \lambda_e - \theta,\; W \ge -a\big) \\
&\ge P_0\big(-a \le X \le \lambda_e - \theta \text{ and } X + Y r^{-1} \ge -a(1 + r^{-1})\big) \\
&\ge P_0\big(-a \le X \le \lambda_e - \theta \text{ and } Y \ge -a\big) \\
&= P_0\big(-a \le X \le \lambda_e - \theta\big)\, P_0\big(Y \ge -a\big) \\
&= \big(\Phi(\lambda_e - \theta) - \widetilde{\Phi}(a)\big) \cdot \big(1 - \widetilde{\Phi}(a r^{-1/2})\big) \\
&\ge \Phi(\lambda_e - \theta) - \widetilde{\Phi}(a) - \widetilde{\Phi}(a r^{-1/2}).
\end{aligned}$$

Now, noting that $\rho_{B,2}(\theta) = 2r\mathbb{E}_0\Big\{ N_\theta(X,Y)\, \mathbb{I}[X \le \lambda_e - \theta] \Big\}$ the result follows. □

Lemma 4.5.4.

For $\theta \in [\lambda_f + a, \lambda_e + a)$, $\quad \rho_{B,1}(\theta) - \rho_{B,2}(\theta) \le \lambda_f^2\,(1 + o\,(1))$.

Proof. From Lemma 4.5.3, $\rho_{B,1}(\theta) - \rho_{B,2}(\theta)$ equals

$$\theta^2 \left\{ \widetilde{\Phi}(a) + \widetilde{\Phi}(a r^{-1/2}) \right\} \;+\; \left\{ \theta^2 - 2(1+r)\, \Gamma(\theta) \right\} q(\theta) + 2r \log(K+1).$$

In the similar way as in Lemma 4.5.2, we can show that as $\eta \to 0$, for all $\theta < \lambda_e + a$, $\theta^2 \widetilde{\Phi}(a) = O(\lambda_f)$ and $\theta^2 \widetilde{\Phi}(a r^{-1/2}) = o(\lambda_f^2)$.
And we show that the second sum involving $\theta^2 - 2(1+r)\, \Gamma(\theta)$ is bounded by $\lambda_f^2(1+o(1))$

when $\theta \in [\lambda_f + a, \lambda_e + a)$. For this purpose, note that

$$\begin{aligned}\theta^2 - 2(1+r)\Gamma(\theta) - \lambda_f^2 &= \min_{i=0}^{K} \left\{ f_i(\theta) + 2(1+r)a\mu_i \right\} \\ &\leq 2(1+r)\, a\, (\lambda_e + a) + \min_{i=0}^{K} f_i(\theta) \\ &\text{where } f_i(\theta) = \theta^2 - 2(1+r)\mu_i\theta + (1+r)\mu_i^2 + r\lambda_f^2.\end{aligned}$$

By construction of the cluster prior $\pi[\eta, K, r]$ the points μ_i were geometrically starting from $\mu_0 = \nu_\eta$ and with $\mu_{i+1} = (1+2r)\mu_i$ for all $i \in \{0, \ldots, K-1\}$. Also, the points end before $\lambda_e + a$. We have not used the properties of aligning rule anywhere before in our proof. A discrete, equi-probable prior distribute was all we utilized in the proofs before this stage.

Now, we will use the properties of μ_i. Note that for each $i \in \{0, \ldots, K\}$:

- f_i is convex in θ.

- $f_i(\mu_i) = \mu_i^2 - 2(1+r)\mu_i^2 + (1+r)\mu_i^2 + r\lambda_f^2 = -r(\mu_i^2 - \lambda_f^2) \leq -r(\nu_\eta^2 - \lambda_f^2)$ as μ_i are increasing.

- Define $\mu_{K+1} = (1+2r)\mu_K$. Then by choice of K, $\mu_{K+1} > \lambda_e + a$. Also,

$$\begin{aligned}f_i(\mu_{i+1}) &= \mu_{i+1}^2 - 2(1+r)\mu_i\mu_{i+1} + (1+r)\mu_i^2 + r\lambda_f^2 \\ &= (\mu_{i+1} - (1+r)\mu_i)^2 - (1+r)r\mu_i^2 + r\lambda_f^2\end{aligned}$$

  and using the common ration of the geometric progression, we have

$$f_i(\mu_{i+1}) = r^2\mu_i^2 - (1+r)r\mu_i^2 + r\lambda_f^2 = -r(\mu_i^2 - \lambda_f^2) \leq -r(\nu_\eta^2 - \lambda_f^2).$$

Convexity of f_i implies that if $\mu_i \leq \theta \leq \mu_{i+1}$ for some $i \in \{1, \ldots, K\}$, then $f_i(\theta) \leq O(\lambda_f^2 - \nu_\eta^2) = o(\lambda_f^2)$ as by Equation [4.9], $\nu_\eta^2 - \lambda_f^2 \leq 2v_w^{1/2} a\lambda_f$. Hence, for all $\theta \in [\lambda_f + a, \lambda_e + a]$ we have $\min_{j=0}^{K} f_j(\theta) = o(\lambda_f^2)$. This complete the proof. □

PART B: HIGH-DIMENSIONAL MINIMAX PREDICTIVE DENSITIES

4.6 Multivariate predictive risk

In this section, we would need to construct sequence of priors as (n, s) varies. For notational convenience we assume s as a function of n here. It can easily be generalized. We consider a tractable convex collection of probability measures in the n–dimensional space

$$\mathcal{M}(n, s_n) = \left\{ \pi(\boldsymbol{\theta}) : \sum_{i=1}^{n} P_\pi(\theta_i \neq 0) \leq s_n \right\}.$$

However, $\mathcal{M}(n_n, s)$ contains prior whose support is not confined to $\Theta(n, s_n)$. We consider the sub-class $\mathcal{M}_p(n, s_n)$ of all product priors in $\mathcal{M}(n, s_n)$. The least favorable prior in $\mathcal{M}_p(n, s_n)$ concentrates on $\Theta(n, s_n)$ when $s_n \to \infty$ and $s_n/n \to 0$ as $n \to \infty$.

Lemma 4.6.1.

For any n, s and r we have $R(n, s, r) \leq n\beta(s/n, r)$.

Proof. The set $\mathcal{M}(n,s)$ contains all Dirac priors $\delta_\theta \,\forall\, \theta \in \Theta(n,s)$ and is convex and weakly compact . So we can apply the Minimax Theorem 4.2.1 to have ,

$$R(n,s,r) \leq \sup\{B(\pi) : \pi \in \mathcal{M}(n,s)\} := B(\mathcal{M}(n,s),r)$$

and the result follows by Lemma 4.8.1. □

Based on the univariate least favorable 3–point prior, for each $\epsilon < 1$, we construct a sequence (in n) of prior $\pi[n,s_n,\epsilon,r]$ in $\mathcal{M}^p(\epsilon s_n, r)$ as

$$\pi[n,s_n,\epsilon,r](\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi[\epsilon s/n, r, 3](\theta_i).$$

The extension into the multivariate Bayes-Minimax set up can be conducted through the following lemma which is proved in the Appendix.

Lemma 4.6.2.

As $n \to \infty$, $s \to \infty$ then if for each $\epsilon < 1$ there exists an exchangeable product prior $\pi_{n,\epsilon}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_1[s/n,r](\theta_i)$ in $\mathcal{M}(n,s_n)$ satisfying the following conditions:

a. $B(\pi_{n,\epsilon}) \geq \epsilon B(r, \mathcal{M}(n, \epsilon\, s_n))$,

b. $\pi_{n,\epsilon}(\Theta(n,s)) \to 1$,

c. The Bayes predictive density based on the prior $t_{n,\epsilon} = \pi_{n,\epsilon}(\cdot|\Theta(n,s))$ is such that

$$\int_{\Theta^c(n,s)} \pi_{n,\epsilon}(\boldsymbol{\theta})\, \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{t_{n,\epsilon}})\, d\boldsymbol{\theta} = \mathrm{o}\left(B(r, \mathcal{M}(n,s_n))\right),$$

then we have

$$R(n,s,r) \sim B(r, \mathcal{M}(n,s_n)).$$

Proof of Theorem 4.1.2. We check the conditions of the lemma for our least favorable

prior. Consider the random variable N_n which is the number of non-zero coordinates in a random sample from $\pi[n, s_n, \epsilon, r]$. So, $N_n \sim \text{Binomial}(n, s_n/n)$.

As $s/n \to 0$ the 3-point prior is least favorable in $\mathfrak{m}(\epsilon s/n)$ and hence we have property (a) and Lemma 4.8.1 implies $B(\pi[n, s_n, \epsilon, r]) \geq \epsilon\beta(\epsilon s_n, r)$. Property (b) also holds as

$$\pi[n, s_n, \epsilon, r]\big(\Theta^c(n, \epsilon s_n)\big) = P\left(N_n \geq \epsilon s_n\right) \leq \frac{\mathbb{V}ar(N_n)}{(1-\epsilon)^2\mathbb{E}^2(N_n)}$$

which by Chebyshev's inequality goes to 0 as $s_n \to \infty$.

Now, note that the support of $\pi[n, s_n, \epsilon, r]$ is given by

$$S_{n,\epsilon} = \big\{\zeta : \zeta_i = 0 \text{ or } \pm\nu_\eta \text{ and } N_n(\zeta) \leq s_n\big\}$$

where ν_η is given by Equation [4.9] with $\eta = n^{-1}\epsilon s_n$. And, the univariate plug-in risk $\rho(\theta, \widehat{p}_E[0]) = \theta^2/(2r)$ and $\rho(\theta, \widehat{p}_E[\pm\nu_\eta]) = (\theta \pm \nu_\eta)^2/(2r)$.

So, by convexity of the relative entropy loss function we have,

$$\begin{aligned}
\boldsymbol{\rho}\left(\boldsymbol{\theta},\ \widehat{p}\big[\pi[n, s_n, \epsilon, r]\,\big]\right) &\leq \sup_{\zeta \in S_{n,\epsilon}} \boldsymbol{\rho}\left(\boldsymbol{\theta},\ \widehat{p}_E[\zeta]\right) \\
&\leq \frac{1}{2r}\left[\sum_{i:\zeta_i=0} \theta_i^2 + \sum_{i:\zeta_i \neq 0} (\theta_i \pm \nu_\eta)^2\right] \\
&\leq r^{-1}\big\{||\boldsymbol{\theta}||_2^2 + N_n\,\nu_\eta^2\big\} = 2r^{-1}N_n\nu_\eta^2.
\end{aligned}$$

Now integrating over the prior π, we have

$$\int_{\boldsymbol{\theta} \in \Theta^c(n, s_n)} \pi[n, s_n, \epsilon, r](\boldsymbol{\theta})\ \boldsymbol{\rho}\left(\boldsymbol{\theta}, \widehat{p}\big[\pi[n, s_n, \epsilon, r]\,\big]\right) d\boldsymbol{\theta} = 2r^{-1}\nu_\eta^2 E\{N_n;\, \Theta_n^c\}.$$

Extending the univariate minimax problem it follows that $B(r, \mathcal{M}(n, s_n)) \sim (2r)^{-1} s_n \nu_\eta^2$ and so the ratio

$$\big\{B(r, \mathcal{M}(n, s_n))\big\}^{-1} \left\{ \int_{\boldsymbol{\theta} \in \Theta^c(n, s_n)} \pi[n, s_n, \epsilon, r](\boldsymbol{\theta})\, \boldsymbol{\rho}\big(\boldsymbol{\theta}, \widehat{p}[\,\pi[n, s_n, \epsilon, r]\,]\,\big)\, d\boldsymbol{\theta} \right\}$$

is asymptotically equal to $\mathbb{E}[N_n \mathbb{I}\{\Theta^c(n, \epsilon s_n)\}]/\mathbb{E}(N_n)$, which converges to 0 as $n \to \infty$. The convergence is a consequence of the concentration of N_n as $N_n(\theta) = \mathbb{E}N_n\ (1+o(1))$ which follows directly from Chebyshev's inequality.

Thus, property (c) of the Lemma is also satisfied and $\prod_{i=1}^n \pi[n, s_n/n, r, 3](\theta_i)$ is the asymptotically least favorable prior and the theorem follows. □

4.7 Further Insights into Minimax strategies

4.7.1 The choice of Threshold

For having an optimal threshold density estimate we need to have a minimum threshold of λ_e. The working principle of threshold rule is that after the threshold zone they use an estimator with bounded bias and as a side-effect it has considerable loss at the origin. So, it needs ideal calibration of the threshold as the higher thresholds decreases the risk contribution from above the threshold ρ_A at the origin.

Based on the calculations in Section 4.5 it follows that we need to restrict $\rho_A(0)$ below the minimax risk $\eta\lambda_f^2/2r$. If the threshold is $t = q\lambda_e$, where $0 < q < 1$, then from the calculations in Equation [4.16] we have

$$\rho_A(0) = O\big(t\,\phi(t)\big) = O\big(te^{-t^2/2}\big) = O\big(\eta^{q^2}\big) >> O(\eta\lambda_f^2).$$

So, λ_e is the minimal threshold that is to be used. This proves Lemma 4.1.6. Note that the fact that the threshold is controlled entirely by the degree of sparsity resonates with general philosophy of this sparse predictive regime where the order of the optimal risk depends on sparsity and is scaled by future uncertainty.

4.7.2 Sub-optimality of $\mathcal{L}, \mathcal{E}$ and $\mathcal{G}$

Based on the minimax risk of sparse estimation of the normal mean and the calculations in Chapter 2, we see that Plug-in density estimates and in general the class of Gaussian predictive densities is minimax sub-optimal with the sub-optimality ratio being independent of η.

Lemma 4.7.1.

For any fixed $r \in (0, \infty)$, under the condition of Theorem 4.1.2, we have

$$R(n, s, r, \mathcal{E}) \sim (1 + r^{-1})\, R(n, s, r).$$

Also, minimax optimal density estimates lie outside $\mathcal{G}$. Lemma 4.1.4 follows by using Theorem 4.1.2 with the following lemma.

Lemma 4.7.2.

In model M.2 as $n \to \infty$ and $s/n \to 0$ we have

$$\liminf_{n\to\infty} \min_{\widehat{p}\in\mathcal{G}_n} \max_{\boldsymbol{\theta}\in\Theta_n(s)} \rho\,(\boldsymbol{\theta}, \widehat{p}) \;\geq\; (1 + \boldsymbol{r})^{-1}\, s\, \log(n/s)\big(1 + \mathrm{o}\,(1)\big).$$

Proof Outline. Following the previously described Bayes-Minimax procedure, the multivariate minimax problem can be reduced to univariate minimax problem with moment prior constraints

$$\mathfrak{m}(\eta) = \{\pi \in \mathcal{P}(\mathbb{R}) : \pi(0) \geq 1 - \eta\}$$

where $\mathcal{P}(\mathbb{R})$ is the collection of all probability measures on $\mathbb{R}$. By Theorem 4.1.1 we have the univariate minimax risk

$$\min_{\widehat{p}} \max_{\pi\in\mathfrak{m}(\eta)} \int \rho(\theta, \widehat{p})\, \pi(\theta)\, d\theta \sim (1 + r)^{-1}\, \eta\, \log \eta^{-1} \text{ as } \eta \to 0.$$

When restricted to the Gaussian family the minimax risk will be

$$\min_{\widehat{p}\in\mathcal{G}} \max_{\pi\in\mathfrak{m}(\eta)} \int \rho(\theta,\widehat{p})\,\pi(\theta)\,d\theta \sim f(\eta) \text{ as } \eta\to 0$$

where $f(\eta) = r^{-1}\,\eta\,\log\eta^{-1}$. In this univariate asymptotic set-up the lower bound in Equation (3.23) is much lower than the asymptotic rate $\eta\log\eta^{-1}$ and hence unusable. We get an upper bound on the minimax Gaussian risk as from point estimation theory (Donoho & Johnstone 1994*b*)it follows that the minimax plug-in risk in this asymptotic set-up is $f(\eta)$. For a lower bound consider the predictive risk of the normal density estimate $g[\widehat{\theta},\widehat{d}]$

$$\rho\big(\,\theta, g[\widehat{\theta},\widehat{d}]\,\big) = \mathbb{E}_\theta\big(\,\log\widehat{d}\,\big) + \mathbb{E}_\theta\big\{\,\widehat{d}^{-1}\cdot(1+(\widehat{\theta}-\theta)^2)-1\big\}. \tag{4.19}$$

And the idea is to establish the necessity of threshold zone as done in Johnstone & Silverman (2004). For $\rho\big(0, g[\widehat{\theta},\widehat{d}]\,\big)$ – the predictive risk of $g[\widehat{\theta},\widehat{d}]$ at the origin, to be lower than the order of η we need a threshold size of at least $\lambda(\eta) = \sqrt{2\log\eta^{-1}}$. And for density estimators of the form

$$\widehat{p}\,[\,\lambda(\eta)\,]\,(\cdot|X) = \begin{cases} N\big(\,0,\ \sigma_f^2\,\big) & \text{if } |X|\le\lambda(\eta) \\ N\big(\,\widehat{\theta}(X),\ \widehat{d}(X)\,\sigma_f^2\,\big) & \text{if } |X|>\lambda(\eta) \end{cases} \tag{4.20}$$

the supremum predictive risk at the non-zero support points is $f(\eta)$, i.e.,

$$\sup_{\theta\neq 0}\ \rho\big(\,\theta,\ \widehat{p}\,[\,\lambda(\eta)\,]\,\big) \sim f(\eta) \text{ as } \eta\to 0. \tag{4.21}$$

Thus, it follows that sup-optimality of the class $\mathcal{G}[p]$ is $1+r^{-1}$. □

The sub-optimality of Linear density estimates as described in Lemma 4.1.3 follows from the risk calculations in Chapter 2.

4.7.3 Risk sharing schemes and efficient alignments of support points

While constructing the minimax threshold estimator $\widehat{p}_T$ we have used the cluster prior $\pi[\eta, r, K]$ below the threshold. This choice is not unique and we can use other proper estimates in its place. By the proof in the Section 4.5, it follows that we can not use zero-threshold estimator as then ρ_B will only require $\rho_{B,1}$ and for optimality we also need the sharing effect from $\rho_{B,2}$. The proof structure outline before Lemma 4.5.4 goes through for any finite prior with $(1-\eta)$ probability at the origin and the remaining probability η being equally allocated across finite points between λ_f and λ_e.
So a prior with $(1-\eta)$ mass at the orgin and sharing the remaining mass η across the non-zero support points $\mu_0, \ldots, \mu_{K_1}$ will produce an optimal allocation if the alignment of these points (and the cardinality of the support) is such that Lemma 4.5.4 still holds.

For example, instead of the cluster prior $\pi[\eta, r, K]$ we choose a (K_1+2)-points prior whose non-zero support points are equispaced (unlike in geometric progression) and equiprobable. Let the spacing between the points be s. So, $\mu_0 = \nu_\eta$ and $\mu_i = \mu_0 + i\, s$ for $i \in \{0, \ldots, K_1\}$ where

$$K_1 = \max\{i : \mu_0 + i\, s \leq \lambda_e + a\} \text{ and so as } \eta \to 0,\ K_1 \sim \left\lfloor \frac{\lambda_e - \lambda_f}{s} \right\rfloor.$$

Again if we would like to equate $f_i(\mu_{i+1}) \leq -r(\mu_i^2 - \lambda_f^2)$ as in Lemma 4.5.4 we have,

$$\begin{aligned} f_i(\mu_{i+1}) &= (\mu_{i+1} - (1+r)\mu_i)^2 - (1+r)r\mu_i^2 + r\lambda_f^2 \\ &= (s - r\mu_i)^2 - (1+r)r\mu_i^2 + r\lambda_f^2 \\ &= -r\mu_i^2 + r\lambda_f^2 + s^2 - 2sr\mu_i. \end{aligned}$$

Now, we solve for the condition $s\mu_0(s\mu_0 - 2r\mu_i) \leq 0$. A solution is given by $s = 2r\lambda_f$ which produces a choice of $K_1 = \lfloor\, (2r)^{-1}\{(1+r^{-1})^{1/2} - 1\}\, \rfloor$.

We construct the following univariate prior with the non-zero support , equiprobable and equidistant support points lying in between μ_0 and $\lambda + a$

$$\pi[\eta, r, E]\theta) = (1-\eta)\cdot\delta_0(\theta) + \frac{\eta}{K_1+1}\sum_{i=0}^{K_1}\delta_{\mu_i}(\theta)$$
$$\text{where } \mu_i = (1+2ri)\mu_0,\ i = 1,\ldots,K_1 \ \text{ and } \ K_1 = \left\lfloor \frac{(1+r^{-1})^{1/2}-1}{2r} \right\rfloor .$$

As $\eta \to 0$, it will produce a first order minimax optimal predictive density estimate for the univariate restricted Bayes-Minimax problem over the constrained prior space $\mathfrak{m}^{+}(\eta)$.

In Figure 4.7 we have the empirical evaluations of optimal predictive schemes in the asymptotic regime. Figure 4.8 contains the risk plots under moderate sparsity.

4.7.4 Other Minimax Estimators

We consider the following non-negative analogue of $\pi[\eta, r, \text{INF}]$

$$\pi[\eta, r, \text{INF}^{+}](\theta) = (1-\eta)\cdot\delta_0(\theta) + (1-\eta)\sum_{j=0}^{\infty}\eta^{j+1}\sum_{i=0}^{K} s_i\,\delta_{\mu_{ij}}(\theta) \quad \text{where,}$$
$$\mu_{ij} = j\,\lambda_e + (1+2r)^i\,\nu_\eta\,;\ i = 0,\ldots,K \text{ and } j = 1,\ldots,\infty$$
$$s_i = (\log\eta^{-1})^{-i} \text{ for } i = 1,\ldots,K \text{ and}$$
$$s_0 = 1 - \frac{(1-(\log\eta^{-1})^{-K})}{(\log\eta^{-1}-1)} \sim 1-(\log\eta^{-1})^{-1} \text{ as } \eta\to 0.$$

The between cluster spacing and probability distribution on the clusters is motivated by the construction of second order minimax optimal point estimates of the normal mean in Johnstone (1994*b*). The with-in cluster mass distribution is more interesting. First note that for the univariate predictive density estimation problem $\pi[\eta, r, 2]$ or its corresponding infinite support geometric version is not the Bayes optimal strategy for the statistician because it is a prediction problem and he has to share his future risk. Also, neither $\pi[\eta, r, \text{CL}^{+}]$ nor its corresponding infinite support geometric version is least favorable as after sharing the statistician incurs the maximum risk at λ_f

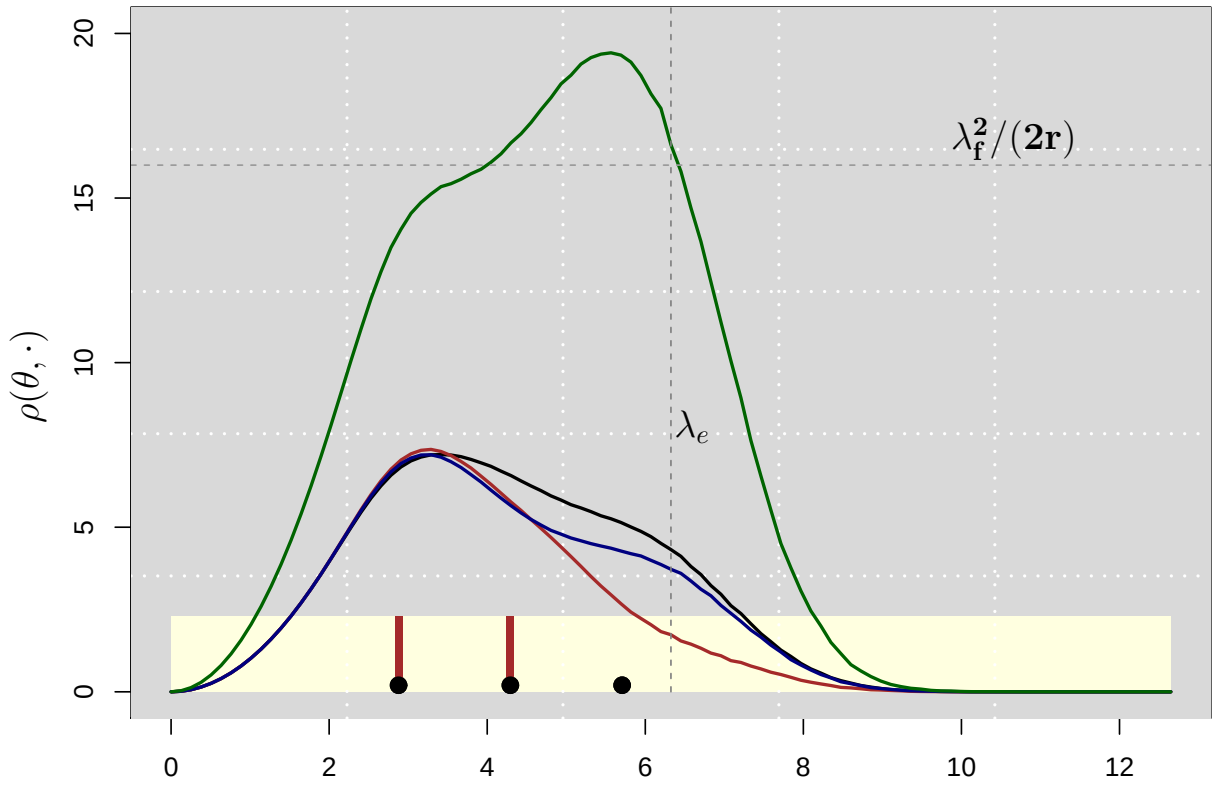

Figure 4.7: Plot of the predictive entropy risk $\rho(\theta, \cdot)$ for the different univariate predictive schemes as the parameter θ varies over $\mathbb{R}^+$. In green, blue, brown and black are respectively the risks of $\widehat{p}[\,r, \mathrm{T}, \pi[\eta, r, 2], \mathrm{U}\,]$, $\widehat{p}[\,r, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}\,]$, $\widehat{p}[\,r, \mathrm{T}, \pi[\eta, r, E], \mathrm{U}\,]$ and that of the Bayes predictive density estimate based on the infinite support prior $\pi[\eta, r, \mathrm{INF}]$. Here, $r = 0.25$, $\eta = e^{-20}$, $\lambda_f = 2.83$ and $\lambda_e = 6.32$. The brown boxes at 2.83 and 4.24 in the yellow zone show the non-zero support point of the cluster prior $\pi[\eta, r, \mathrm{CL}^+]$ and the black circles denote the non-zero support points of $\pi[\eta, r, E]$.

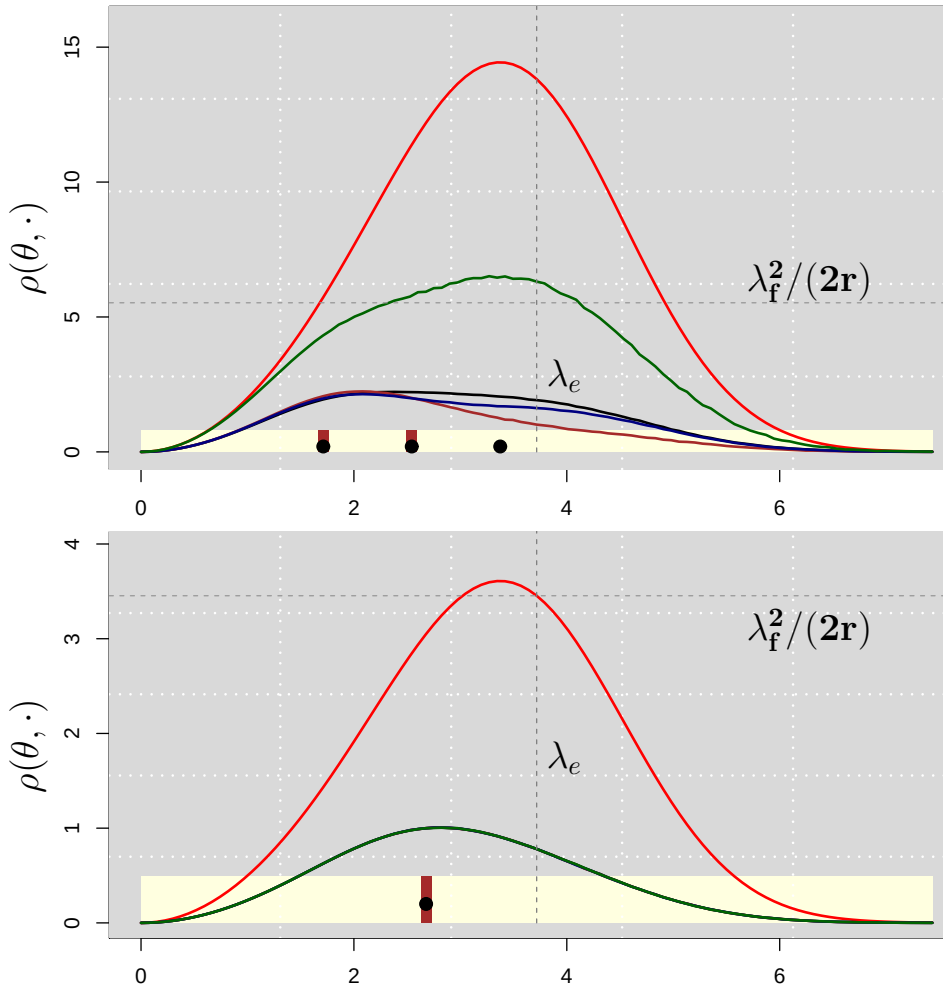

Figure 4.8: The figure shows the risk plots for the different univariate predictive schemes under moderate degree of sparsity ($\eta = 0.001$) for the two different values of the future to past variances: $r = 0.25$ (top) and $r = 1$. In red, green, blue, brown and black are respectively the risks of the optimal hard-threshold plug-in scheme, $\widehat{p}[\, r, \mathrm{T}, \pi[\eta, r, 2], \mathrm{U}\,]$, $\widehat{p}[\, r, \mathrm{T}, \mathrm{CL}^{+}, \mathrm{U}\,]$, $\widehat{p}[\, r, \mathrm{T}, \pi[\eta, r, E], \mathrm{U}\,]$ and that of the Bayes predictive density estimate based on the infinite support prior $\pi[\eta, r, \mathrm{INF}]$.

(which is expected) and the risks at the other non-zero support points is appreciably lower even in first order calculations. However, as $\eta \to 0$ a geometrically decreasing discrete probability sharing scheme with common ratio $\log \eta^{-1}$ solves this problem because $\log \eta^{-1} \to \infty$ when $\eta \to 0$ and hence dampens the first-order terms in the asymptotic limit.

Proof of Theorem 4.1.5. We prove the result in the corresponding univariate model $\mathbf{M.2}(1, \eta, r)$ with $\sigma_p = 1$ and the parameter space restricted to the non-negative axis. The risk of $\widehat{p}(y|x, \pi[\eta, r, \mathrm{INF}^+])$ at the point θ is given by:

$$\rho(\theta) = \frac{1}{2r}\left(\theta^2 - 2rE_0(N'_\theta(X,Y)) + 2rE_0(D'_\theta(X))\right) \quad \text{where,}$$

$$N'_\theta = \log\left[1 + \sum_{j=0}^{\infty}\frac{\eta^{j+1}}{K+1}\sum_{i=0}^{K}\exp\left\{\mu_{ij}\left(x+\frac{y}{r}\right) - \frac{1+r}{2r}\mu_{ij}^2 + \frac{1+r}{r}\mu_{ij}\theta\right\}\right]$$
$$D'_\theta = \log\left[1 + \sum_{j=0}^{\infty}\frac{\eta^{j+1}}{K+1}\sum_{i=0}^{K}\exp\{\mu_{ij}x - \frac{1}{2}\mu_{ij}^2 + \mu_{ij}\theta\}\right].$$

Now, $\rho(0) = \eta + o(\eta)$ as $N'_\theta(X,Y) \geq 0$ and

$$E_0(D'_0(X)) \leq \log\left[1 + \sum_{j=0}^{\infty}\sum_{i=1}^{K}\frac{\eta^{j+1}}{K+1}\right] = -\log(1-\eta) = \eta + O(\eta^2).$$

Next we note that, with probability 1 we will have,

$$E_0(N'_\theta(X,Y)) \geq \max_{\substack{i=1,\ldots,K\\ j=0,\ldots,\infty}}\left(\mu_{ij}\theta - \frac{1}{2}\mu_{ij}^2 - a\mu_{ij} - \frac{1}{2}(j+1)\mu_0^2\right)\left(1+\frac{1}{r}\right)$$
$$- \log(K+1) + \mathrm{O}(1) \quad \text{and}$$
$$E_0(D'_\theta(X)) \leq \max_{\substack{i=1,\ldots,K\\ j=0,\ldots,\infty}}\left(\mu_{ij}\theta - \frac{1}{2}\mu_{ij}^2 + a\mu_{ij} - \frac{1}{2}(j+1)\lambda^2\right)_+ + \mathrm{O}(1).$$

Also, the optimum value of j and i for both the numerator and denominator is same.

And so it follows

$$\begin{aligned}\rho(\theta) &\leq \frac{1}{2r} \max_{\substack{i=1,\ldots,K \\ j=0,\ldots,\infty}} (\theta^2 - 2\mu_{ij}\theta + \mu_{ij}^2 + 4a\mu_{ij}) + \mathrm{O}\,(1) \\ &\leq \frac{1}{2r} \max_{\substack{i=1,\ldots,K \\ j=0,\ldots,\infty}} (\theta^2 - 2\mu_{ij})^2 + o(\mu_0^2) \leq \mu_0/2r(1 + \mathrm{o}\,(1)).\end{aligned}$$

Risk calculations similar to Section 4.5 will show that the $B(\pi[\eta, r, \mathrm{INF}^+])$ attains the above lower bound. This will complete the proof. □

4.7.5 Quality of our results under moderate sparsity

The results we have discussed in this chapter are based in a highly sparse asymptotic regime where the degree of ℓ_0 sparsity $\eta_n \to 0$. In the univariate and non-negative parameter space version of model $\mathbf{M.2}(1, \eta, r)$ it corresponds to having the non-origin prior probability $\eta \to 0$. Here, we perform numerical experiments to study the effectiveness of our asymptotic results under different levels of sparsity. An object of special interest is to record the performance of the previously discussed asymptotically optimal predictive schemes under moderate sparsity. Depending on the degree of sparsity, we consider the following 3 different regimes:

- Moderate Sparsity: $\eta = 0.1$.
- High Sparsity: $\eta = 0.001$.
- Very High Sparsity: $\eta = 10^{-10}$.

Figure 4.11, Figure 4.10 and Figure 4.9 respectively show the univariate risk plots in these 3 different sparsity regimes (moderate to very high) for the following 3 predictive density estimates:

- hard threshold Plug-in density $\widehat{p}[\, r, \mathrm{T}, 0, \mathrm{U}\,]$
- Unshared predictive density $\widehat{p}[\, r, \mathrm{T}, \pi[\eta, r, 2], \mathrm{U}\,]$
- Cluster prior based diversified density $\widehat{p}[\, r, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}\,]$.

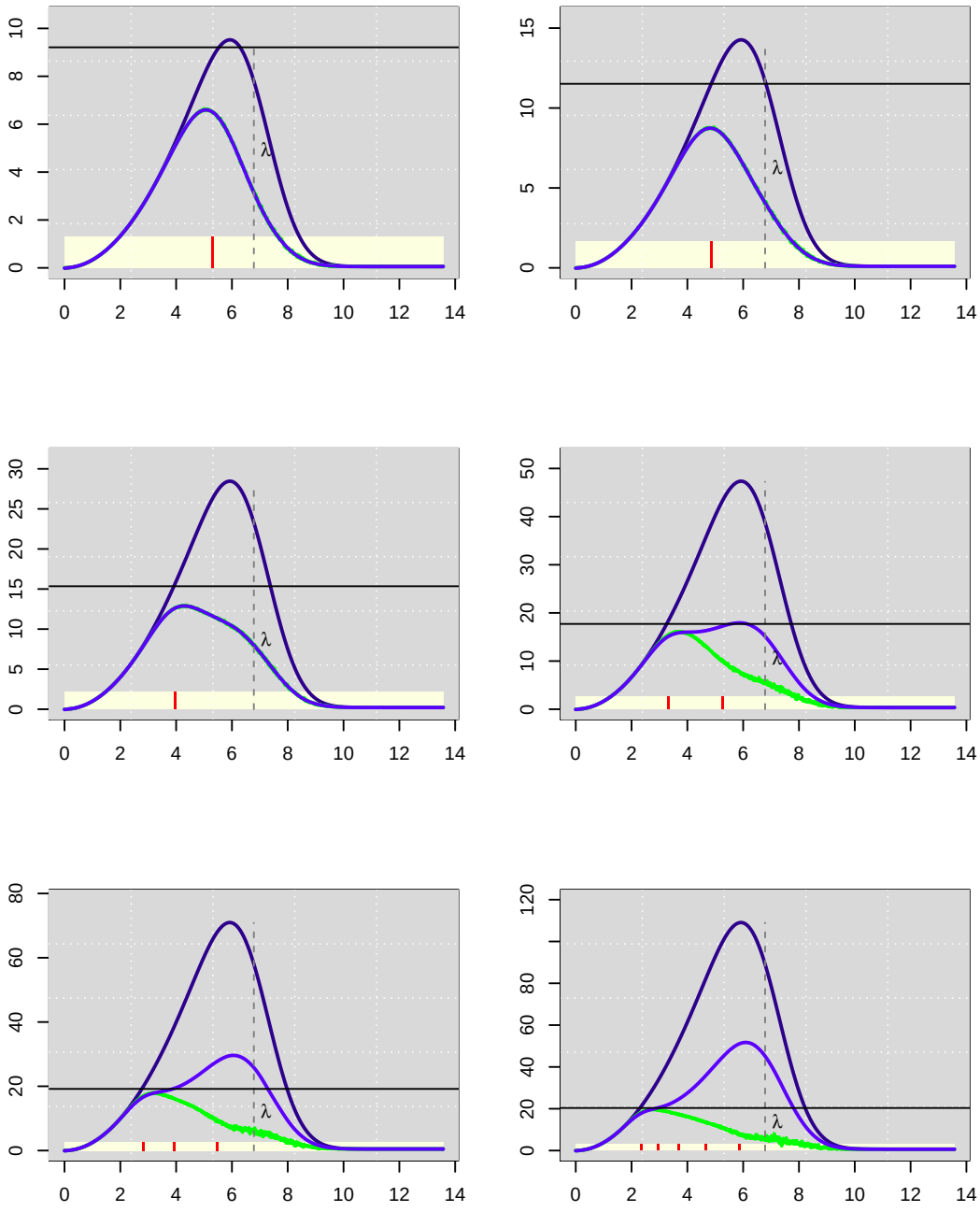

Figure 4.9: *Risk plots under very high sparsity,* $\eta = 10^{-10}$: As the parameter θ varies over $\mathbb{R}^+$ these plots show the risk $\rho(\theta, \cdot)$ for the 3 different univariate predictive densities (i) hard thereshold plug-in density $\widehat{p}[\, r, \mathrm{T}, 0, \mathrm{U}\,]$ (in blue) (ii) unshared predictive density $\widehat{p}[\, r, \mathrm{T}, \pi[\eta, r, 2], \mathrm{U}\,]$(in violet) (iii) cluster prior based diversified density $\widehat{p}[\, r, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}\,]$ (in green). The horizontal line denotes the theoretical minimax risk. From top-left, in clockwise direction, the plots corresponds to r=1.5, 1.0, 0.5, 0.3, 0.2 and 0.1.

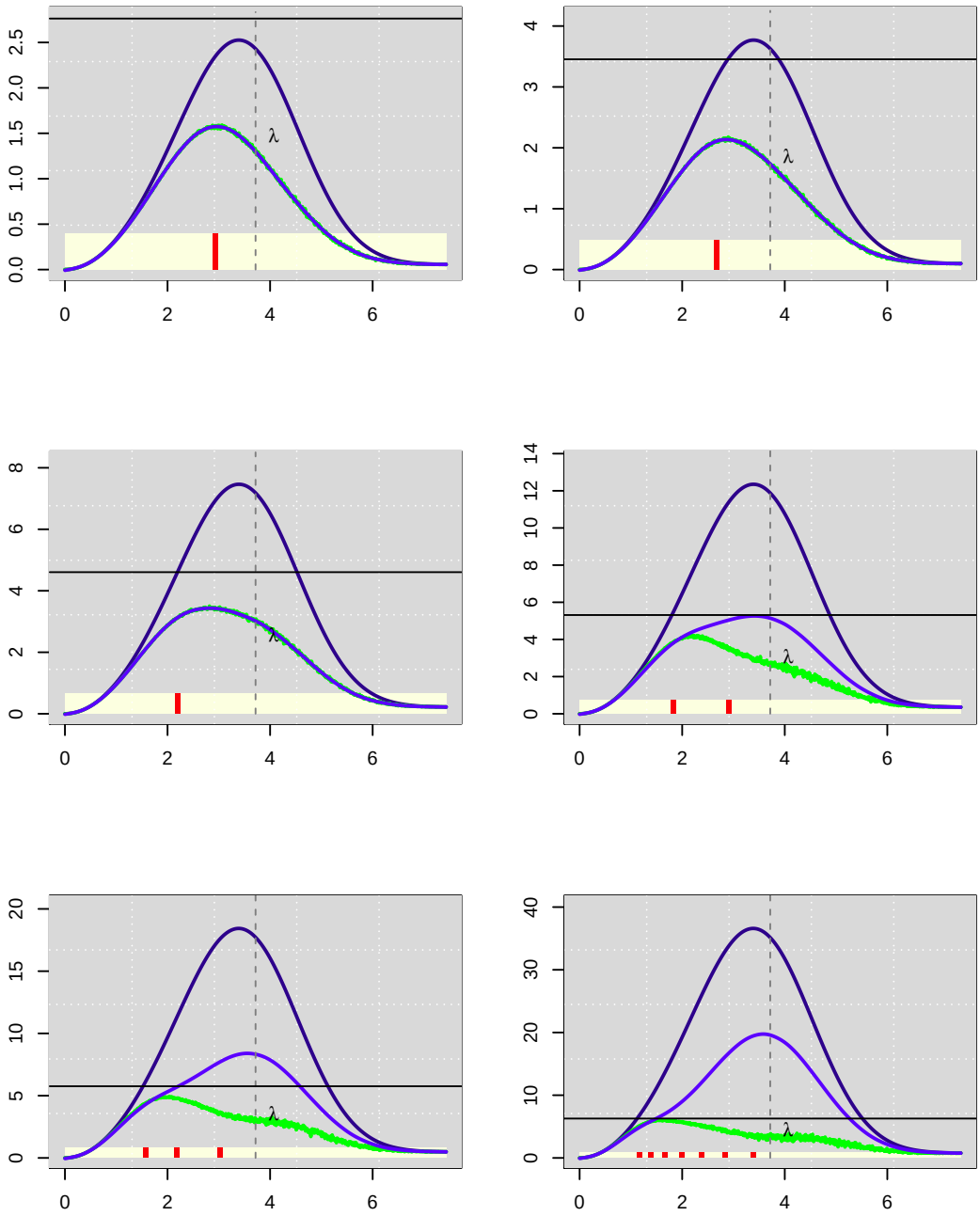

Figure 4.10: *Risk plots under high sparsity,* $\eta = 0.001$: As the parameter θ varies over $\mathbb{R}^+$ these plots show the risk $\rho(\theta, \cdot)$ for the 3 different univariate predictive densities (i) hard threshold plug-in density $\widehat{p}[\, r, \mathrm{T}, 0, \mathrm{U}\,]$ (in blue) (ii) unshared predictive density $\widehat{p}[\, r, \mathrm{T}, \pi[\eta, r, 2], \mathrm{U}\,]$(in violet) (iii) cluster prior based diversified density $\widehat{p}[\, r, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}\,]$ (in green). The horizontal line denotes the theoretical minimax risk. From top-left, in clockwise direction, the plots corresponds to r=1.5, 1.0, 0.5, 0.3, 0.2 and 0.1.

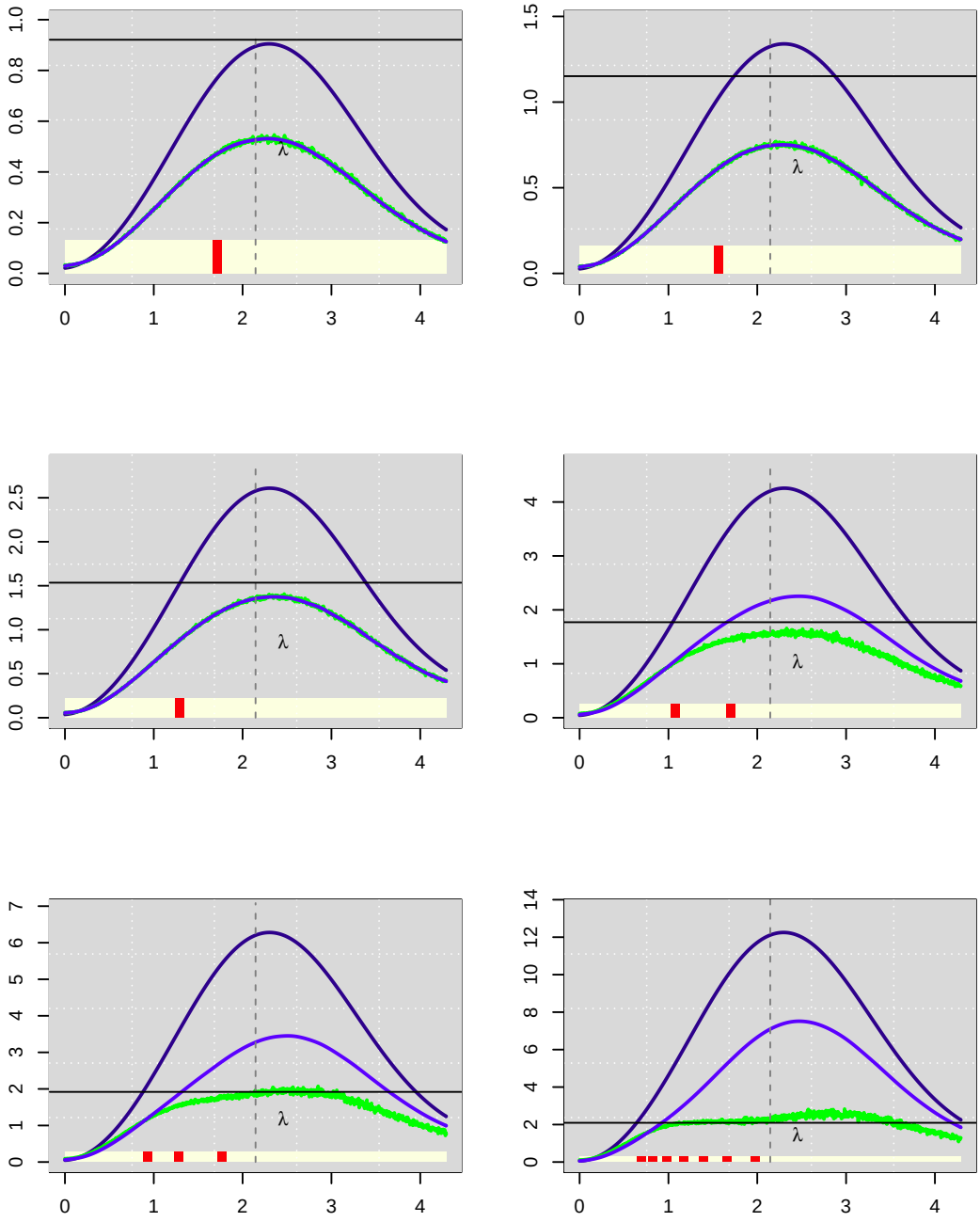

Figure 4.11: *Risk plots under moderate sparsity,* $\eta = 0.1$: As the parameter θ varies over $\mathbb{R}^+$ these plots show the risk $\rho(\theta, \cdot)$ for the 3 different univariate predictive densities (i) hard threshold plug-in density $\widehat{p}[\, r, \mathrm{T}, 0, \mathrm{U}\,]$ (in blue) (ii) unshared predictive density $\widehat{p}[\, r, \mathrm{T}, \pi[\eta, r, 2], \mathrm{U}\,]$(in violet) (iii) cluster prior based diversified density $\widehat{p}[\, r, \mathrm{T}, \mathrm{CL}^+, \mathrm{U}\,]$ (in green). The horizontal line denotes the theoretical minimax risk. From top-left, in clockwise direction, the plots corresponds to r=1.5, 1.0, 0.5, 0.3, 0.2 and 0.1.

Maximum Univariate Predictive Risk Under Sparsity							
Sparsity (η)	r	Theory	Plug-in	Unshared	Cluster	Equispaced	Bayes
10^{-10}	1.00	1.15e-09	1.43e-09	8.86e-10	8.83e-10	8.89e-10	8.80e-10
	0.50	1.54e-09	2.85e-09	1.30e-09	1.30e-09	1.30e-09	1.31e-09
	0.25	1.84e-09	5.68e-09	2.28e-09	1.69e-09	1.72e-09	1.69e-09
	0.10	2.04e-09	1.09e-08	5.26e-09	1.98e-09	2.02e-09	NaN
0.001	1.00	0.00345	0.00377	0.00220	0.00220	0.00217	0.00206
	0.50	0.00461	0.00746	0.00352	0.00351	0.00349	0.00335
	0.25	0.00553	0.01480	0.00667	0.00449	0.00465	0.00435
	0.10	0.00628	0.03660	0.02010	0.00618	0.00639	0.00610
0.1	1.00	0.115	0.134	0.0787	0.0779	0.0786	0.0603
	0.50	0.154	0.261	0.1440	0.1430	0.1430	0.1100
	0.25	0.184	0.507	0.2860	0.2160	0.1710	0.1580
	0.10	0.209	1.230	0.7700	0.2980	0.2850	0.2060

Table 4.2: Numerical evaluation of the maximum risk under ℓ_0 sparsity for the different univariate predictive densities as the degree of sparsity (η) and predictive difficulty r varies.

Maxima of the risk plots: $\max_{\theta \in \mathbb{R}^+} \rho(\theta, \cdot)$									
η	r	λ_f	λ_e	Theory	Plug-in	Unshared	Cluster	Equispaced	Bayes
10^{-10}	1.00	4.80	6.79	11.5	14.3	8.82	8.8	8.84	8.79
	0.50	3.92	6.79	15.4	28.5	13.00	13.0	13.10	13.00
	0.25	3.03	6.79	18.4	56.8	22.70	17.0	17.30	16.90
	0.10	2.30	6.79	20.4	109.0	52.60	19.7	20.20	NaN
0.001	1.00	2.63	3.72	3.45	3.77	2.19	2.18	2.18	2.07
	0.50	2.15	3.72	4.61	7.46	3.53	3.53	3.55	3.34
	0.25	1.66	3.72	5.53	14.80	6.61	4.55	4.67	4.37
	0.10	1.12	3.72	6.28	36.60	20.10	6.18	6.43	6.08
0.1	1.00	1.520	2.15	1.15	1.34	0.783	0.791	0.791	0.612
	0.50	1.240	2.15	1.54	2.61	1.430	1.440	1.440	1.110
	0.25	0.960	2.15	1.84	5.07	2.810	2.110	1.740	1.550
	0.10	0.647	2.15	2.09	12.30	7.810	3.000	2.920	2.040

Table 4.3: Numerical evaluation of the maxima of the risk plots for the different univariate predictive densities as the degree of sparsity (η) and predictive difficulty r varies.

Figures 4.11, 4.10 and 4.9 show that that the fundamental features of the risk plots are unchanged even under moderate sparsity. For all the different regimes, risk-diversified density estimates have better performances than the unshared and plugin estimates and its maximum risk is close to the evaluation of the minimax risk based on the asymptotic formula in Theorem 4.3.4.

In Table 4.3 we report the maximum value of the risk-plots of these predictive strategies as well as those of the risk-diversified Equispaced predictive density $\widehat{p}[\, r, \mathrm{T}, \pi[\eta, r, E], \mathrm{U}\,]$ (denoted by Equispaced) and that of the Bayes predictive density estimate based on the infinite support prior $\pi[\eta, r, \mathrm{INF}]$ (denoted by Bayes). In Table 4.2 we provide numerical evaluations of their respective maximum predictive risk under the ℓ_0 sparsity restriction of having at least $(1-\eta)$ probability at the origin. In the tables, by Theory we denote evaluation of the minimax risk by the expression $\eta \log \eta^{-1}/(1+r)$ which is given by Theorem 4.3.4. We see that across all regimes the risk diversified strategies perform better than the unshared and plug-in schemes. All the risk-diversified schemes have very similar maximum risks though the Bayes predictive density based on $\pi[\eta, r, \mathrm{INF}]$ has optimal performance among them. In Figure 4.12, we have plots of the risk of $(1-\eta)$-sparse 2-point priors at the non-origin support point. The maxima and maximum values based on Lemma 4.3.2 are also shown. From the figures and the tables, it is seen that the differences between our theory and the numerical evaluations increase as r increases or η decreases. And, in those cases our theory results overestimate the minimax risks. Second order optimality calculations as done in Johnstone (1994*b*) for point estimation case can be helpful in those regimes.

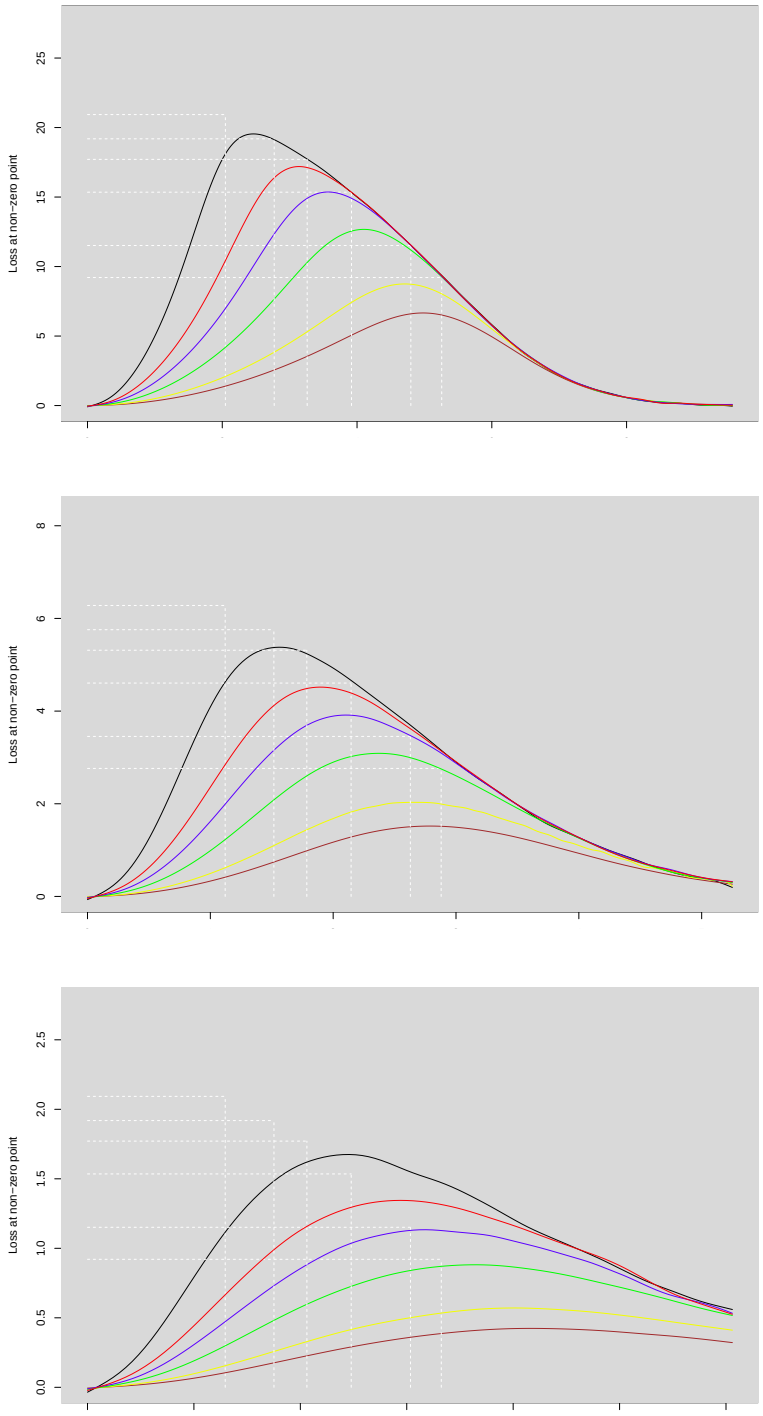

Figure 4.12: *Non-origin risk of* $(1-\eta)$*-sparse* 2*–point priors*: Each of these curves is the plot of $\rho(\nu, \widehat{p}[\pi_{2\text{pt}}(\eta, \nu)])$ as ν varies for a fixed r and η. In brown, yellow, green, blue, red and black respectively are curves corresponding to $r = 1.5, 1, 0.5, 0.3, 0.2$ and 0.1. The white lines correspond to the theoretical maxima and maximum values respectively. From top to bottom we have 3 different levels for $\eta : 10^{-10}, 0.001$ and 0.1 respectively.

DISCUSSIONS

Our results on ℓ_0 sparsity can be extended to the orthogonal sequence model M.2 with approximate sparsity restrictions on the parameter space. The constrained minimax results can be generalized to a wide range of parameter class indexed by ℓ_p balls with shape parameter p and normalized mean radius τ_n both varying in $(0, \infty)$:

$$\ell_p \text{ balls: } \Theta_{n,p}(C_n) = \left\{ \sum_{i=1}^{n} |\theta_i|^p \leq C_n^p \right\} \text{ with mean radius } \tau_n = n^{-1/p} \sigma_p^{-1} C_n.$$

Figure 1.2 shows the relation between exact and approximate or ℓ_p–sparsity. In the context of the curve prediction problem discussed in Section 1.3.3, here we assume that the future and past observation vector are sampled with noise from an unknown function at the same set of n equispaced points (i.e. $m_1 = m_2 = n$). However, the past and future noise variability can vary but are known to be σ_p^2 and σ_f^2 respectively. And so, the predictive difficulty is given by $r = \sigma_f^2/\sigma_p^2$.

Approximate Sparsity and ℓ_p constrained parameter spaces

As $n \to \infty$, we derive expressions for the first order minimax risk over ℓ_p balls for all choices of p and τ_n. Asymptotically least favorable priors are exchangeable priors and first order minimax optimal rules are co-ordinate wise rules. Table 4.4 shows the

results in the two extreme regimes

- Very High Signal to Noise Ratio (SNR): $\tau_n \to \infty$.
- Very Low Signal to Noise Ratio(SNR: $\tau_n \to 0$.

In high SNR, the class of plug-in predictive density estimates is sub-optimal but the class of linear estimates attain minimax risk. As such, the best invariant density estimate $\widehat{p}_U$ is minimax optimal in high SNR. In low SNR, as in point estimation theory, here also phrase transition in the behavior of the asymptotic minimax risk is witnessed at $p = 2$. For $p \geq 2$, the zero-density estimate is first order minimax. For $p < 2$, linear sub-optimality is ∞ but the class of plug-in estimates ($\mathcal{P}$) perform much better. However, in these regimes there is no gain in minimax risk if we consider the wider class of all Gaussian density estimates instead of $\mathcal{P}$. In the non-extreme regimes where $\tau_n \to \tau \in (0, \infty)$, informative upper and lower bounds can be derived on the minimax predictive risk. Following Johnstone (1994*a*), these results can be generalized to weak-ℓ_p sparsity.

Mean Radius	p	$\mathcal{P}$	$\mathcal{L}$	$\mathcal{G}$	ρ **minimax**
$\tau_n \to \infty$	$[0, \infty)$	$r^{-1}\{\log(1 + r^{-1})\}^{-1}$	1	1	$2^{-1}\, n\, \log(1 + r^{-1})$
$\tau_n \to 0$	$[2, \infty)$	1	1	1	$(2r)^{-1}\, n\, \tau_n^2$
	$[0, 2)$	$(1 + r^{-1})^{1-p/2}$	∞	$(1 + r^{-1})^{1-p/2}$	$\kappa_{p,r}\, n\, \tau_n^p\, \{\log \tau_n^{-p}\}^{1-p/2}$

Table 4.4: The sub-optimality coefficients of the classes of Plug-in ($\mathcal{P}$), Linear ($\mathcal{L}$) and Gaussian ($\mathcal{G}$) density estimates as the parameter spaces lie in ℓ_p balls with mean radius reflecting the two extreme signal-to-noise regimes. Here, $\kappa_{p,r} = (2r)^{-p/2}(1 + r)^{(p-2)/2}$.

Example: Simultaneous probability forecasts of Wind Speed

Next, we illustrate the approximate sparsity results through the motivating example of simultaneous estimation of predictive densities for wind-speeds over a series of time points at a particular meteorological station. It is important to predict the occurrence of extreme wind speeds with high accuracy and warnings for weather hazards

are based on the probability estimates of these outliers. Simultaneous probability forecasts of wind speeds are used in weather forecasting, aircraft and maritime operations, atmospheric dispersion modeling and growth & metabolism rate estimation of many plant species. Often, objects of predictive interest are based on the log wind profiles which describe the vertical distribution of horizontal mean wind speeds.

We consider the data set where the average wind speeds at every 4 hours are recorded over the course of 10 years starting from January 1, 2003 to December 31, 2012. The average wind speed (in miles per hours) data is recorded at AgriMet Station (BKVO) at Baker Valley (44.78°N, 117.85°W, Elevation 3483ft) in Oregon and is available from `http://www.usbr.gov/pn/agrimet/`.

Using the first 5 years of data from January 1, 2003 to December 31, 2007 we would like to make simultaneous predictive inferences on the wind speeds in the succeeding years. We assume that each year's the wind speed data $\{W[j] : 1 \leq j \leq J\}$ indexed sequentially at the set J of 4-hours intervals can be assumed to be coming from the same smooth curve f:

$$W[j] = f[j] + \sigma \cdot \epsilon[j], \quad j \in J.$$

We would like to estimate the simultaneous predictive densities of wind speeds over the set J for the year 2008. Note that in this case the predictive difficulty of the problem is low as $r = 5$. If we want to estimate the simultaneous predictive densities of wind speeds averaged over 2, 3, 4 and 5 years following 2007 then the values of r are 2.5, 1.67, 1.25 and 1 respectively. Here, we use two different choices of predictive densities:

- The best invariant linear predictive density estimate $\widehat{p}_U$.
- The hard threshold wavelet transformed plug-in density estimate $\widehat{p}[\text{wavelet}, \text{H}]$. Using the *waved* R-package of Raimondo & Stewart (2011) we fit a wavelet location estimate $\widehat{\theta}[\text{wavelet}, \text{H}]$ for each point in J. $\widehat{\theta}[\text{wavelet}, \text{H}]$ is hard thresholded in the wavelet basis based on the our theoretical threshold choice of 0.104 (we used Meyer wavelet of level 10). Figure 4.13 shows that $\widehat{\theta}[\text{wavelet}, \text{H}]$ is able to capture the the seasonal variability in the mean wind speed. Usually, the

autumn months are comparatively calmer than the summer (when the planetary boundary layer is more turbulent) and winter months (when there is snow on the ground). We consider the plugin predictive density $\widehat{p}[\text{wavelet}, \text{H}]$ around $\widehat{\theta}[\text{wavelet}, \text{H}]$.

In Figure 4.14, we produce point-wise 90% prediction interval based on the above two estimates. Based on data from 2008 to 2012, Table 4.5 shows their respective coverage as well as the percentage of time-points when the the linear prediction interval contains the plug-in prediction interval. We observe that the intervals based on both the predictive densities have similar coverage which is close to 90%. In most cases the intervals from $\widehat{p}_U$ contain those from $\widehat{p}[\text{wavelet}, \text{H}]$ and this proportion of inclusion increases with increase in predictive difficulty (r^{-1}). Our theoretical results support these observations. Though our results do not specifically address optimality (in terms of width) of prediction intervals, they provide predictive performances of the density estimates the in terms of the average likelihood ratio. Attributes such as width and coverage of the prediction intervals are functions of the predictive likelihood ratio.

Here, in the wavelet basis we have a sparse predictive regime with $\tau_n = 5 \cdot 10^{-4}$ and $n = 2^{11}$ and theoretically the sub-optimality of the entire class of linear estimates

$$\lim_{\tau_n \to 0} S_r(p, \tau_n, \mathcal{L}) \to \infty \text{ for any } p \in (0, 2)$$

where as sub-optimality of $\widehat{p}[\text{wavelet}, \text{H}]$ in this low SNR regime is less than $1 + r^{-1}$. So, for all the different choices of r that we have here, $\widehat{p}[\text{wavelet}, \text{H}]$ is at least 50% efficient with respect to the optimal minimax predictive risk whereas linear estimates are extremely inefficient and it is reflected in the greater width of its associated prediction intervals.

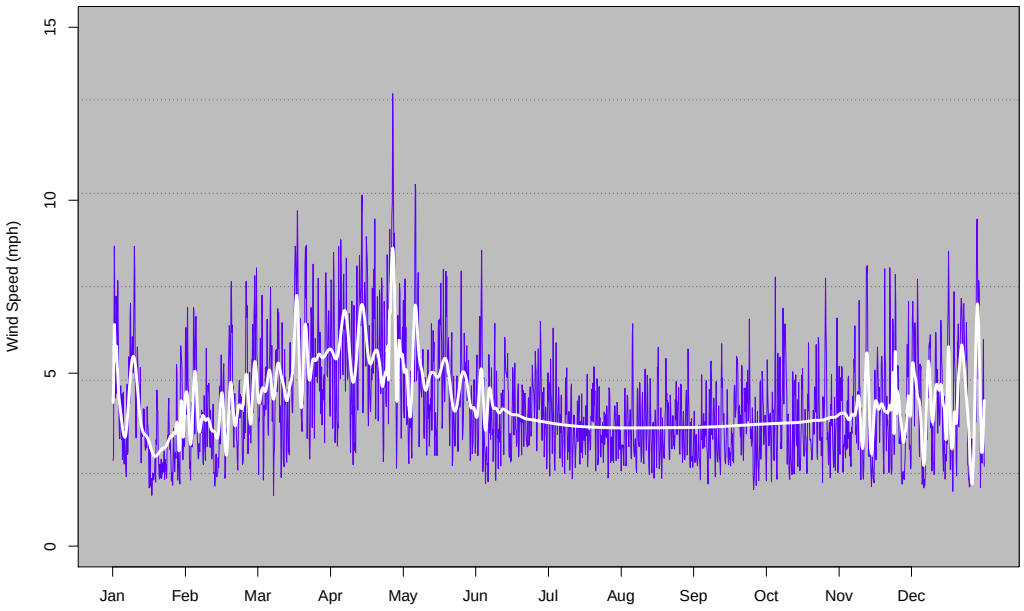

Figure 4.13: *Mean Wind Speed at* 4 *hours interval:* In blue, we have the mean wind-speed (in miles per hour) at every 4 hours interval averaged over the 5 years 2003-2008. The white line is the hard-threshold based wavelet transformed location estimate $\widehat{\theta}[\text{wavelet}, \text{H}]$.

r	Coverage of $\widehat{p}_U$	Coverage of $\widehat{p}[\text{wavelet}, \text{H}]$	Relative Width
5.00	90.92	91.55	64.94
2.50	88.77	90.58	77.83
1.67	87.06	87.45	89.65
1.25	88.92	88.62	98.19
1.00	88.43	87.74	99.90

Table 4.5: Percentage of coverage in 2008-2012 of 90% point-wise prediction intervals which are constructed based on 2003-2007 data by using predictive densities estimates $\widehat{p}_U$ and $\widehat{p}[\text{wavelet}, \text{H}]$. By relative width, we denote the proportion of time-points in which the prediction interval constructed from $\widehat{p}[\text{wavelet}, \text{H}]$ is entirely contained in those built from $\widehat{p}_U$.

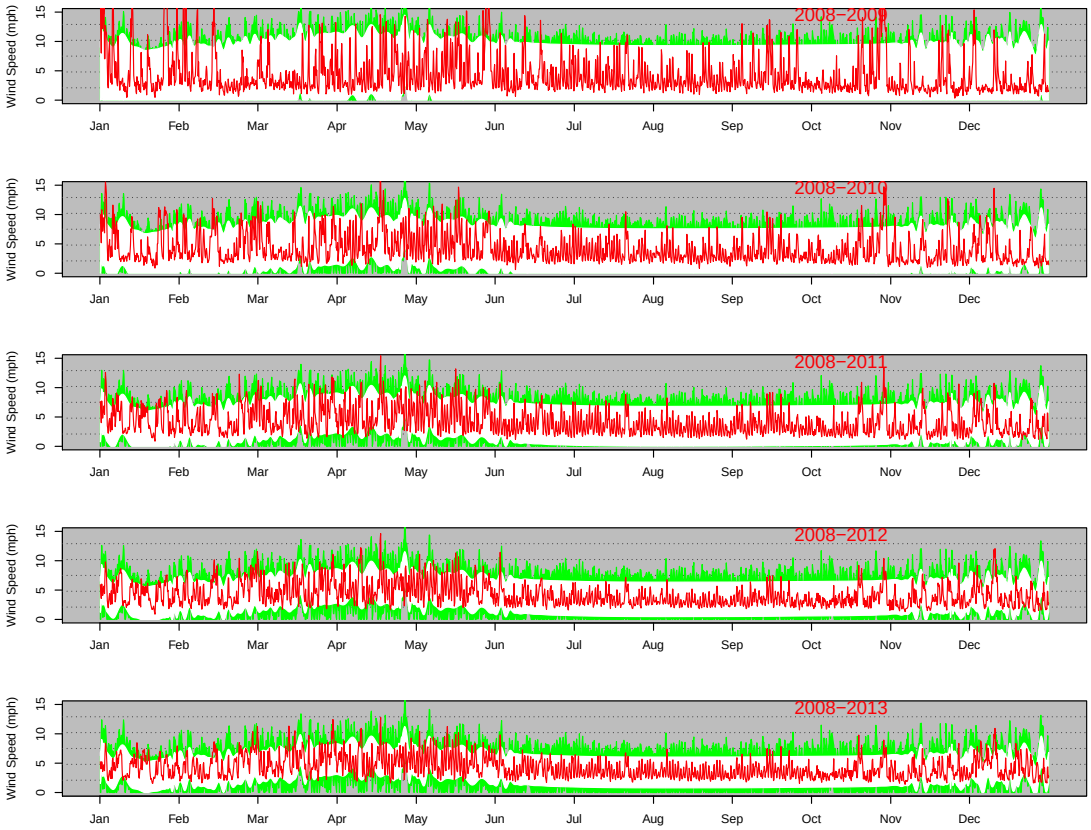

Figure 4.14: *90% point-wise prediction interval.* The white region represents the 90% prediction interval at each time point from $\widehat{p}[\mathrm{H}, \lambda]$ – the hard threshold plug-in density estimate. In green, we have the point-wise 90% prediction interval from the best invariant linear predictive density $\widehat{p}_U$. These prediction intervals are constructed based on the data from January 1, 2003 to December 31, 2007. Superimposed on them, in red we have the wind speed averaged across 5, 4, 3, 2 and 1 year (bottom to top) starting from January 1, 2008.

4.8 Appendix

Mills ratio and Gaussian tails (Johnstone 2012, Exercise 8.1)

The function $M(u) = \widetilde{\Phi}(u)/\phi(u)$ is called Mills Ratio. The following inequalities will provide an approximation to the Mills ratio which will be very helpful for our calculations with truncated Gaussian random variable. As such:

$$\text{For any } u \geq 0 \text{ we have } \frac{\phi(u)}{u}\left(1 - \frac{1}{u^2}\right) \leq \tilde{\Phi}(u) \leq \frac{\phi(u)}{u}. \tag{4.22}$$

And so for large u, which will typically be the case, the approximation $\tilde{\Phi}(u) \sim u^{-1}\phi(u)$ is quite sharp.

Multivariate Minimax Risk

Lemma 4.8.1.

For any fixed n, s, r we have $B(r, \mathcal{M}(n, s_n)) = n\beta(s/n, r)$.

Proof. For any prior π on $\mathbb{R}^n$ and its marginals $\{\pi_i : i = 1, \ldots, n\}$ we have $\boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\pi_1 \times \pi_2 \times \ldots \pi_n}) = \sum_{i=1}^n \rho(\theta_i, \widehat{p}_{\pi_i})$. So,

$$B(\pi) = \int \pi(\theta)\boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_\pi)\, d\boldsymbol{\theta} \leq \int \pi(\boldsymbol{\theta})\boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\pi_1 \times \pi_2 \times \ldots \pi_n}) = \sum_{i=1}^n \int \pi(\theta_i)\rho(\theta_i, \widehat{p}_{\pi_i})\, d\theta_i$$

Again, if $\pi \in \mathcal{M}(n, s_n)$ then $\bar{\pi} = \pi_1 \times \pi_2 \times \ldots \pi_n \in \mathcal{M}_p(n, s_n)$ and $\mathcal{M}_p(n, s_n) \subset \mathcal{M}(n, s_n)$. So, $B(r, \mathcal{M}(n, s_n)) = B(r, \mathcal{M}_p(n, s_n))$ and due to decomposability of the Bayes risk for product priors we have

$$B(r, \mathcal{M}_p(n, s_n)) = \sup\left\{\sum_{i=1}^n \beta(\tau_i, r) : \sum_{i=1}^n \tau_i \leq s_n\right\}.$$

Now as $\beta(\tau, r)$ is concave function of τ the supremum in the above expression occurs when $\tau_i = s/n \,\forall\, i$. This completes the proof. □

Note that for any $\epsilon \in (0,1)$ the parametric space $\Theta(n, \epsilon\, s)$ as well as the prior space $\mathcal{M}(n, \epsilon\, s_n)$ is equivariant in the sense $\Theta(n, \epsilon s) = \epsilon \cdot \Theta(n, s)$ and $\mathcal{M}(n, \epsilon s_n) = \epsilon \cdot \mathcal{M}(n, s_n)$.

Proof of Lemma 4.6.2. From definition of Bayes risk it follows

$$\begin{aligned} B(\pi_{n,\epsilon}) &= \int \pi_{n,\epsilon}(\boldsymbol{\theta})\, \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\pi_{n,\epsilon}})\, d\theta \leq \int \pi_{n,\epsilon}(\boldsymbol{\theta})\, \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\nu_{n,\epsilon}})\, d\boldsymbol{\theta} \\ &= \left\{ \int_{\Theta(n,s)} \nu_{n,\epsilon}(\boldsymbol{\theta}) \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\nu_{n,\epsilon}})\, d\boldsymbol{\theta} \right\} \cdot \pi_{n,\epsilon}(\Theta_n) + \int_{\Theta^c(n,s)} \pi_{n,\epsilon}(\boldsymbol{\theta}) \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\nu_{n,\epsilon}})\, d\boldsymbol{\theta} \\ &= \pi_{n,\epsilon}\big(\Theta(n,s)\big) B(\nu_{n,\epsilon}) + \int_{\theta \in \Theta^c(n,s)} \pi_{n,\epsilon}(\boldsymbol{\theta})\, \boldsymbol{\rho}(\boldsymbol{\theta}, \widehat{p}_{\nu_{n,\epsilon_n}})\, d\boldsymbol{\theta} \\ &\leq \pi_{n,\epsilon}(\Theta(n,s)) R(n,s,r) + \mathrm{o}\left(B(r, \mathcal{M}(n, s_n))\right) \end{aligned}$$

as support of ν_n is contained in $\Theta(n,s)$, so we have $B(\nu_{n,\epsilon}) \leq R(n,s,r)$ and we use property (c) on the second sum.

Now, using Condition (b) of the lemma, we have $R(n,s,r) \geq \epsilon\, B(r, \mathcal{M}(n, \epsilon s_n)) - \mathrm{o}\left(B(r, \mathcal{M}(n, s_n))\right)$ and the result follows by using the following Lemma 4.8.2. □

Lemma 4.8.2.

For any fixed $r \in (0, \infty)$ we have

$$\lim_{\epsilon \uparrow 1} \liminf_{n \to \infty} \frac{B(r, \mathcal{M}(n, \epsilon s_n))}{B(r, \mathcal{M}(n, s_n))} = 1.$$

Proof. The proof is similar to Exercise 4.7 in Johnstone (2012). □

Minimax Theorem

We consider the Gaussian predictive sequence model

$$x_i = \theta_i + \sigma_p\, \epsilon_{1,i} \text{ and } y_i = \theta_i + \sigma_p\, \epsilon_{2,i} \tag{4.23}$$

for $i \in I \subset \mathbb{N}$, with $\epsilon_{1,i}$ and $\epsilon_{2,i}$ are i.i.d. $N(0,1)$ random variables. The parameter space is a collection of $\boldsymbol{\theta}$ for which $\sum_i \theta_i^2 < \infty$ and is denoted by $\ell_2(\mathbb{N})$. The action set is given by

$$\mathcal{A}_\infty = \left\{ p : \mathbb{R}^\infty \to \mathbb{R} \text{ such that } \int_{\mathbb{R}^\infty} p(\mathbf{y})\, d\mathbf{y} = 1 \text{ and } p(\mathbf{y}) \geq 0 \text{ for all } \mathbf{y} \right\}.$$

For each $n \in \mathbb{N}$ consider all sub-probabilities in $\mathbb{R}^n$ bounded by c_n by extending $\mathcal{A}(n, c_n)$ to its closure

$$\bar{\mathcal{A}}(n, c_n) = \left\{ p : \mathbb{R}^n \to \mathbb{R} \text{ such that } \int_{\mathbb{R}^n} p\,(\mathbf{y})\, d\mathbf{y} \leq 1 \text{ and } p \in [0, c_n] \right\}.$$

$\bar{\mathcal{A}}(n, c_n)$ is a sub-set of the Banach space $\mathcal{L}_\infty(\mathbb{R}^n, \mathbb{R})$ – all bounded functionals in $\mathbb{R}^n$. Now, expanding the discussions in Section 2.4, we consider the topology on $\bar{\mathcal{A}}(n, c_n)$ induced by the weak* topology on $\mathcal{L}_\infty(\mathbb{R}^n, \mathbb{R})$.

In the predictive sequence model consider the experiment $(\Omega, \mathcal{B}, \{\mathcal{P}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \ell_2(\mathbb{N})\})$ where the sample space $\Omega = \{\otimes_{i \in \mathbb{N}}(x_i, y_i) : x_i, y_i \in \mathbb{R}\}$ and $\mathcal{B}$ is the associated Borel sigma field. As the parameter space is ℓ_2, we have a dominated experiment here with

$$\frac{d\mathcal{P}_{\boldsymbol{\theta}}}{d\mathcal{P}_0} = \exp\left[\sigma_p^{-2} \left\{ \left\langle \boldsymbol{\theta}, \mathbf{x} + r^{-1}\mathbf{y} \right\rangle - \frac{1}{2}\left(1 + r^{-1}\right) ||\boldsymbol{\theta}||^2 \right\} \right].$$

Also, for each $n \in \mathbb{N}$, the restricted, closed action set $\bar{\mathcal{A}}(n, c_n)$ is weak* compact and the loss function $\mathrm{L}(\boldsymbol{\theta}, p)$ is

- strictly convex in $p \in \bar{\mathcal{A}}(n, c_n, +)$ for any $\boldsymbol{\theta} \in \mathbb{R}^n$ where $\bar{\mathcal{A}}(n, c_n, +) = \{p \in \bar{\mathcal{A}}(n, c_n)$ such that $\mathrm{L}(\boldsymbol{\theta}, p) < \infty$ for all $\boldsymbol{\theta} \in \mathbb{R}^n\}$ and
- lower semi-continuous in p on $\bar{\mathcal{A}}(n, c_n)$ for any fixed $\boldsymbol{\theta} \in \mathbb{R}^n$.

And so, following the lines of Johnstone (2012, Appendix 1) and Brown (1974) we can arrive at a version of the minimax theorem for the predictive setting. It is stated below.

Theorem 4.8.3.

Consider the predictive density estimation problem in the Gaussian predictive sequence model (4.23) with the parameter space $\ell_2(\mathbb{N})$. For any convex set $\mathcal{M}$ of probability measures on $\ell_2(\mathbb{N})$ we have

$$\inf_{\widehat{p}\in\mathcal{A}_\infty}\sup_{\pi\in\mathcal{M}} B\left(\pi,\widehat{p}\right) \;=\; \sup_{\pi\in\mathcal{M}}\inf_{\widehat{p}\in\mathcal{A}_\infty} B\left(\pi,\widehat{p}\right)$$

BIBLIOGRAPHY

Aitchison, J. (1975), 'Goodness of prediction fit', *Biometrika* **62**(3), 547–554.

Aslan, M. (2006), 'Asymptotically minimax Bayes predictive densities', *Ann. Statist.* **34**(6), 2921–2938.

Barndorff-Nielsen, O. E. & Cox, D. R. (1996), 'Prediction and asymptotics', *Bernoulli* **2**(4), 319–340.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D. & Loris, I. (2009), 'Sparse and stable markowitz portfolios', *Proceedings of the National Academy of Sciences* **106**(30), 12267–12272.

Brown, L. (1974), 'Lecture notes on statistical decision theory'. Available at `"http://www-stat.wharton.upenn.edu/~lbrown"`.

Brown, L. D. (1971), 'Admissible estimators, recurrent diffusions, and insoluble boundary value problems', *Ann. Math. Statist.* **42**, 855–903.

Brown, L. D., George, E. I. & Xu, X. (2008), 'Admissible predictive density estimation', *Ann. Statist.* **36**(3), 1156–1170.

Brown, L. D. & Hwang, J. T. (1982), A unified admissibility proof, *in* 'Statistical decision theory and related topics, III, Vol. 1 (West Lafayette, Ind., 1981)', Academic Press, New York, pp. 205–230.

Candès, E. J. (2006), Compressive sampling, *in* 'International Congress of Mathematicians. Vol. III', Eur. Math. Soc., Zürich, pp. 1433–1452.

Candès, E. J., Romberg, J. & Tao, T. (2006), 'Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information', *IEEE Trans. Inform. Theory* **52**(2), 489–509.

Candes, E. & Tao, T. (2007), 'The Dantzig selector: statistical estimation when p is much larger than n', *Ann. Statist.* **35**(6), 2313–2351.

Dicker, L. H. (2012), 'Optimal estimation and prediction for dense signals in high-dimensional linear models'. arXiv:1203.4572.

Donoho, D., Johnstone, I. & Montanari, A. (2011), 'Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising'. arXiv:1111.1041.

Donoho, D. L. (2006), 'Compressed sensing', *IEEE Trans. Inform. Theory* **52**(4), 1289–1306.

Donoho, D. L. & Johnstone, I. M. (1994*a*), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**(3), 425–455.

Donoho, D. L. & Johnstone, I. M. (1994*b*), 'Minimax risk over l_p-balls for l_q-error', *Probab. Theory Related Fields* **99**(2), 277–303.

Donoho, D. L., Johnstone, I. M., Hoch, J. C. & Stern, A. S. (1992), 'Maximum entropy and the nearly black object', *J. Roy. Statist. Soc. Ser. B* **54**(1), 41–81. With discussion and a reply by the authors.

Donoho, D. L., Maleki, A. & Montanari, A. (2011), 'The noise-sensitivity phase transition in compressed sensing', *IEEE Trans. Inform. Theory* **57**(10), 6920–6941.

Fan, J., Lv, J. & Qi, L. (2011), 'Sparse high dimensional models in economics', *Annual review of economics* **3**, 291.

Foster, D. P. & George, E. I. (1994), 'The risk inflation criterion for multiple regression', *Ann. Statist.* **22**(4), 1947–1975.

Fourdrinier, D., Marchand, É., Righi, A. & Strawderman, W. E. (2011), 'On improved predictive density estimation with parametric constraints', *Electron. J. Stat.* **5**, 172–191.

George, E. I., Liang, F. & Xu, X. (2006), 'Improved minimax predictive densities under Kullback-Leibler loss', *Ann. Statist.* **34**(1), 78–91.

George, E. I., Liang, F. & Xu, X. (2012), 'From minimax shrinkage estimation to minimax shrinkage prediction', *Statist. Sci.* **27**(1), 82–94.

Ghosh, M., Mergel, V. & Datta, G. S. (2008), 'Estimation, prediction and the Stein phenomenon under divergence loss', *J. Multivariate Anal.* **99**(9), 1941–1961.

Hartigan, J. A. (1998), 'The maximum likelihood prior', *Ann. Statist.* **26**(6), 2083–2103.

Huber, N. & Leeb, H. (2012), 'Shrinkage estimators for prediction out-of-sample: Conditional performance'. arXiv:1209.0899.

Johnstone, I. M. (1994*a*), Minimax Bayes, asymptotic minimax and sparse wavelet priors, *in* 'Statistical decision theory and related topics, V (West Lafayette, IN, 1992)', Springer, New York, pp. 303–326.

Johnstone, I. M. (1994*b*), 'On minimax estimation of a sparse normal mean vector', *Ann. Statist.* **22**(1), 271–289.

Johnstone, I. M. (2012), Gaussian estimation: Sequence and wavelet models. Available at: `"http://www-stat.stanford.edu/~imj"`.

Johnstone, I. M. & Silverman, B. W. (2004), 'Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences', *Ann. Statist.* **32**(4), 1594–1649.

Komaki, F. (1996), 'On asymptotic properties of predictive distributions', *Biometrika* **83**(2), 299–313.

Komaki, F. (2001), 'A shrinkage predictive distribution for multivariate normal observables', *Biometrika* **88**(3), 859–864.

Komaki, F. (2004), 'Simultaneous prediction of independent Poisson observables', *Ann. Statist.* **32**(4), 1744–1769.

Leeb, H. (2009), 'Conditional predictive inference post model selection', *Ann. Statist.* **37**(5B), 2838–2876.

Lustig, M., Donoho, D. & Pauly, J. (2007), 'Sparse mri: The application of compressed sensing for rapid mr imaging', *Magnetic Resonance in Medicine* pp. 1182–1195.

Murray, G. D. (1977), 'A note on the estimation of probability density functions', *Biometrika* **64**(1), 150–152.

Ng, V. M. (1980), 'On the estimation of parametric density functions', *Biometrika* **67**(2), 505–506.

Nussbaum, M. (1996), 'Asymptotic equivalence of density estimation and Gaussian white noise', *Ann. Statist.* **24**(6), 2399–2430.

Raimondo, M. & Stewart, M. (2011), *waved: Wavelet Deconvolution*. R package version 1.1-1.

Raskutti, G., Wainwright, M. & Yu, B. (2011), 'Minimax rates of estimation for high-dimensional linear regression over ℓ_q balls', *IEEE Transactions of Information Theory* **57**(10), 6976–6994.

Stein, C. (1974), Estimation of the mean of a multivariate normal distribution, *in* 'Proceedings of the Prague Symposium on Asymptotic Statistics (Charles Univ., Prague, 1973), Vol. II', Charles Univ., Prague, pp. 345–381.

Strawderman, W. E. (1971), 'Proper Bayes minimax estimators of the multivariate normal mean', *Ann. Math. Statist.* **42**(1), 385–388.

Tibshirani, R. (2011), 'Regression shrinkage and selection via the lasso: a retrospective', *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**(3), 273–282.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2002), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proceedings of the National Academy of Sciences* **99**(10), 6567–6572.

Xu, X. & Liang, F. (2010), 'Asymptotic minimax risk of predictive density estimation for non-parametric regression', *Bernoulli* **16**(2), 543–560.

Xu, X. & Zhou, D. (2011), 'Empirical bayes predictive densities for high-dimensional normal models', *J. Multivariate Analysis* **102**(10), 1417–1428.

Zhang, C.-H. (2005), 'General empirical Bayes wavelet methods and exactly adaptive minimax estimation', *Ann. Statist.* **33**(1), 54–100.

Zhang, C.-H. (2010), 'Nearly unbiased variable selection under minimax concave penalty', *Ann. Statist.* **38**(2), 894–942.