

Predictive Density Estimation for Nonparametric Regression models

Xinyi Xu
The Ohio State University

Joint work with Feng Liang
University of Illinois at Urbana-Champaign

June 8, 2009

Outline

- Introduction
- The Problem Setup
- Asymptotic Minimax Risk
- Examples
- Concluding Remarks

Outline

Introduction

The Problem Setup

Asymptotic Minimax Risk

Examples

Concluding Remarks

Predictive Density Estimation

- Predictive analysis is one of the most fundamental problems in statistics.
- Two ways to report prediction results:
 - A point prediction accompanied with an error bound
 - A predictive density estimate that assigns probabilities to all possible outcomes of a future outcome

We will focus on the second way to provide a comprehensive description on the future uncertainty

- Some other uses of predictive density estimates:
 - Bayesian model checking and model diagnostics
 - Missing data analysis
 - Data compression and information theory

The Nonparametric Regression Model

- Suppose that the variable of interest Y is connected to a group of predictors $\mathbf{X} = (X_1, \dots, X_p)$ through a nonparametric regression model

$$Y_i = f(\mathbf{X}_i) + \sigma \varepsilon_i, \quad i = 1, \dots, n,$$

- f is an unknown function in $\mathcal{L}^2[0, 1]$, satisfying certain smoothness conditions
- ε_i 's are i.i.d. standard Gaussian random variables
- Assume the noise level σ is known, and set $\sigma = 1$ without loss of generality

- A huge body of literature has been devoted to estimating f or various functionals of f . However, much less work has been done for predicting future observations from the same process that generated $Y(t)$.
- Suppose the interest is in extracting information from past data to predict future outcomes \tilde{Y} . To obtain a comprehensive description of future uncertainty, we'd like to construct a predictive density estimator $\hat{p}(\tilde{y}|y)$
- The traditional plug-in approach, which simply substitutes f by a reasonable estimator, is unsatisfactory because it ignores the uncertainty inherent in the function estimation process

The Bayes Approach

The Bayes approach is more appealing for several reasons.

- It produces a posterior predictive density.
- It naturally takes account of the model uncertainty by combining the predictions from individual models in a weighted average, where the weights are determined by the posterior model probabilities.
- The recent developments in numerical and simulation methods such as Markov Chain Monte Carlo and the rapid growth in computer power have made the implementation of Bayesian methods feasible even in rather complicated settings.

The Bayes Approach (Cont.)

Bayesian predictive densities are desirable in estimating predictive densities for linear regression models under the Kullback-Leibler (KL) loss

- The Bayesian predictive density \hat{p}_U under the uniform prior dominates the plug-in predictive estimate $\hat{p}_{plug-in}(\tilde{y}|\hat{\beta}_y)$, where $\hat{\beta}_y$ is the MLE of β based on the observed data (Aitchison 1975)
- \hat{p}_U is best invariant and minimax with constant risk (Murray 1977, Ng 1980, Liang and Barron 2004)
- Some other Bayesian predictive densities further dominate \hat{p}_U when the number of predictors $p \geq 3$ (George and Xu 2008)

Predictive Estimation for Nonparametric Models

To explore optimal predictive estimation for nonparametric regression models, the following theoretical issues needs to be addressed:

- How accurate can we estimation the predictive density $p(\tilde{y}|f)$?
- What estimators can achieve this optimal accuracy?

This work is an attempt to answer these questions.

Outline

Introduction

The Problem Setup

Asymptotic Minimax Risk

Examples

Concluding Remarks

The Problem Setup

- Consider the canonical nonparametric regression setup

$$Y(t_i) = f(t_i) + \sigma \varepsilon_i, \quad i = 1, \dots, n,$$

- $\tilde{Y} = (\tilde{Y}(u_1), \dots, \tilde{Y}(u_m))'$: future observations arising at $\{u_j\}_{j=1}^m$
 $\hat{p}(\tilde{y}|y)$: an estimator of the predictive density $p(\tilde{y}|f(u))$
- For each value of y , we measure the closeness of $\hat{p}(\tilde{y}|y)$ to $p(\tilde{y}|f)$ by the well-known Kullback-Leibler (KL) loss

$$L(f, \hat{p}(\tilde{y}|y)) = \int p(\tilde{y}|f) \log \frac{p(\tilde{y}|f)}{\hat{p}(\tilde{y}|y)} d\tilde{y}.$$

The Problem Setup (Cont.)

- To evaluate the performance of density prediction across the whole curve, we assume that $m \geq n$ is a large number and that u_j 's are equally spaced points in $[0, 1]$.
- The overall performance of the estimator $\hat{p}(\tilde{y}|y)$ is then summarized by the averaged KL risk

$$R(f, \hat{p}) = \frac{1}{m} E_{Y, \tilde{Y}|f} \log \frac{p(\tilde{Y}|f)}{\hat{p}(\tilde{Y}|Y)}.$$

Note: Here the densities of $(\tilde{Y}_1, \dots, \tilde{Y}_m)$ are estimated simultaneously

An Alternative Setup (Individual Prediction)

- An alternative approach is to estimate the densities individually by $\{\hat{p}(\tilde{y}_j | y)\}_{j=1}^m$ with risk

$$\frac{1}{m} E_{Y, \tilde{Y} | f} \sum_{j=1}^m \log \frac{p(\tilde{Y}_j | f(u_j))}{\hat{p}(\tilde{Y}_j | Y)}.$$

- When the u_j 's are equally spaced and m goes to infinity, the risk above converges to

$$\int_0^1 E_{Y, \tilde{Y} | f} \log \frac{p(\tilde{Y} | f(u))}{\hat{p}(\tilde{Y} | Y)} du,$$

which can be interpreted as the integrated risk of prediction at a random location u in $[0, 1]$.

- This individual prediction problem can be studied in our simultaneous prediction framework with $\hat{p}(\tilde{y}|y)$ restricted to a product form, i.e., $\hat{p}(\tilde{y}|y) = \prod_{j=1}^m \hat{p}(\tilde{y}_j|y)$.
- In general, simultaneous prediction considers a broader class of \hat{p} than the one considered by individual prediction.
- This is distinct from estimating f itself under the \mathcal{L}^2 loss where, due to the additivity of the \mathcal{L}^2 loss, simultaneous estimation and individual estimation are equivalent.

The Minimax Risk for Predictive Density Estimation

- The asymptotical optimality of a nonparametric estimator is usually associated with the convergence rate of the minimax risk (Belitser and Levit 1996).
- Assuming that f belongs to a function space \mathcal{F} , such as a Sobolev space, we are interested in the minimax risk

$$R(\mathcal{F}) = \min_{\hat{p}} \max_{f \in \mathcal{F}} R(f, \hat{p}).$$

Connection to the Gaussian Sequence Model

- To study the asymptotics of the minimax risk for predictive density estimation, we first convert the nonparametric regression model to a Gaussian sequence model
- Let $\{\phi_i\}_{i=1}^{\infty}$ be the orthonormal trigonometric basis of $\mathcal{L}^2[0, 1]$. Then f can be expanded as $f = \sum_{i=1}^{\infty} \theta_i \phi_i$.
- The smoothness assumption on the functional space \mathcal{F} determines a constraint on the parameter space θ . We consider spaces with ellipsoidal constraints, i.e.,

$$\Theta(C) = \left\{ \theta : \sum_{i=1}^{\infty} a_i^2 \theta_i^2 \leq C \right\},$$

where $a_1 \leq a_2 \leq \dots$ and $a_n \rightarrow \infty$ as $n \rightarrow \infty$.

- Approximating f by a finite summation $f_n = \sum_{i=1}^n \theta_i \phi_i$ incurs a bias

$$\text{Bias}(f, f_n) = \frac{1}{m} E_{\tilde{Y}|f} \log \frac{p(\tilde{Y}|f)}{p(\tilde{Y}|f_n)} = \frac{1}{2m} \sum_{i=n+1}^{\infty} \theta_i^2,$$

which is often negligible compared to the prediction risk.

Therefore, we set $f = f_n$.

- Now $Y|\theta \sim N(\Phi_A \theta, I_n)$ and $\tilde{Y}|\theta \sim N(\Phi_B \theta, I_m)$, where Φ_A and Φ_B are determined by the trigonometric basis. So

$$Z = \frac{1}{n} \Phi_A^t Y \quad \text{and} \quad \tilde{Z} = \frac{1}{m} \Phi_B^t \tilde{Y}$$

satisfies $Z|\theta \sim N(\theta, \sigma_n I_n)$ and $\tilde{Z}|\theta \sim N(\theta, \sigma_m I_m)$, where $\sigma_m = 1/m$ and $\sigma_n = 1/n$.

- The problem of estimating the predictive density $p(\tilde{y}|f)$ based on Y is equivalent to the problem of estimating the predictive density of $\tilde{Z}|\theta \sim N_n(\theta, \sigma_m I_n)$ based on $Z|\theta \sim N_n(\theta, \sigma_n I_n)$.
- The minimax risk on an ellipsoidal space $\Theta(C)$ is

$$R(\Theta) = \inf_{\hat{P}} \sup_{\theta \in \Theta(C)} R(\theta, \hat{P}).$$

- Note: However, this equivalence does not hold for the *individual* estimation approach, because the product form of the density estimators, i.e., $\hat{p}(\tilde{y}|y) = \prod_j \hat{p}(\tilde{y}_j|y)$, is not retained under the transformation.

Outline

Introduction

The Problem Setup

Asymptotic Minimax Risk

Examples

Concluding Remarks

Evaluation of the Minimax Evaluation

The evaluation of this minimax risk is not directly tractable because of the constraint. Therefore, we propose the following bypass consisting of two steps:

- **Step 1:** Find the minimax risk $R_L(\Theta)$ over a small subclass that consists of predictive densities under Gaussian priors on the unconstrained parameter space \mathbb{R}^n .
- **Step 2:** Compare this risk $R_L(\Theta)$ with the overall minimax risk $R_N(\Theta)$ and evaluate their difference.

Linear Predictive Estimators

- In the problem of estimating the mean of a Gaussian sequence model under \mathcal{L}^2 loss, diagonal linear estimators of the form $\hat{\theta}_i = c_i x_i$ play important roles
- If such a diagonal linear estimator is a Bayes rule under some prior $\pi(\theta)$, then the prior π has to be a Gaussian prior with a diagonal covariance matrix (Diaconis and Ylvisaker 1979)
- To investigate the minimax risk of predictive density estimation, we first restrict our attention to predictive densities under Gaussian priors $N(0, S)$, where S is a diagonal matrix
- We call these predictive densities *linear* predictive densities and call the minimax risk over this class the *linear* minimax risk due to the above connection

- Under a Gaussian prior $\pi_S(\theta) = N(0, S)$ with a diagonal covariance matrix S , the linear predictive density \hat{p}_S is

$$\hat{p}_S(\tilde{x} | x) = \int_{\mathbb{R}^n} p(\tilde{x} | \theta) \pi_S(\theta | x) d\theta.$$

- The average Kullback-Leibler risk of \hat{p}_S is given by

$$R(\theta, \hat{p}_S) = \frac{n}{2m} \log \frac{v_n}{v_{n+m}} + \frac{1}{2m} \sum_{i=1}^n \left[\log \frac{v_{n+m} + s_i}{v_n + s_i} + \frac{v_{n+m} + \theta_i^2}{v_{n+m} + s_i} - \frac{v_n + \theta_i^2}{v_n + s_i} \right],$$

where $v_{n+m} = 1/(n+m)$.

- The linear minimax risk over all \hat{p}_S is

$$R_L(\Theta) = \inf_S \sup_{\theta \in \Theta(C)} R(\theta, \hat{p}_S).$$

Linear Minimax Risk

- This linear minimax risk is still not directly tractable because the inside maximization is over a constrained space $\Theta(C)$. So we first show that the order of inf and sup can be switched, i.e.,

$$R_L(\Theta) = \inf_S \sup_{\theta \in \Theta(C)} R(\theta, \hat{p}_S) = \sup_{\theta \in \Theta(C)} \inf_S R(\theta, \hat{p}_S).$$

- Then we use the Lagrangian

$$\mathcal{L} = \sum_{i=1}^n \log \frac{v_{n+m} + \theta_i^2}{v_n + \theta_i^2} - \frac{1}{\lambda} \left(\sum_{i=1}^n a_i^2 \theta_i^2 - C \right),$$

to obtain the *linear* minimax risk.

Theorem. The linear minimax risk is

$$R_L(\Theta) = \frac{n}{2m} \log \frac{v_n}{v_{n+m}} + \frac{1}{2m} \sum_{i=1}^n \log \frac{v_{n+m} + \tilde{\theta}_i^2}{v_n + \tilde{\theta}_i^2}.$$

where $\tilde{\theta}_i^2 = \frac{1}{2} \left[(v_n - v_{n+m}) \sqrt{1 + \frac{4\tilde{\lambda}/a_i^2}{v_n - v_{n+m}}} - (v_n + v_{n+m}) \right]_+$, and

$\tilde{\lambda}(C, v_n, v_{n+m})$ is a solution of the equation

$$\sum_{i=1}^n a_i^2 \left[(v_n - v_{n+m}) \sqrt{1 + \frac{4\tilde{\lambda}/a_i^2}{v_n - v_{n+m}}} - (v_n + v_{n+m}) \right]_+ = 2C.$$

The linear minimax estimator $\hat{p}_{\tilde{V}}$ is the Bayes predictive density under a Gaussian prior

$$\pi_{\tilde{V}}(\theta) = N(0, \tilde{V}), \text{ where } \tilde{V} = \text{diag}(\tilde{\theta}_1^2, \tilde{\theta}_2^2, \dots, \tilde{\theta}_n^2).$$

A Lower Bound for the Minimax Risk

- By definition, $R_L(\Theta)$ is greater than or equal to $R(\Theta)$, the minimax risk over all possible predictive densities. We extend the approach in Belitser and Levit (1995) to evaluate the difference between $R_L(\Theta)$ and $R(\Theta)$ as n goes to infinity.
- If for some $\alpha > 0$, $\sum_{i=1}^n a_i^2 s_i^2 + \left[-8\alpha \left(\sum_{i=1}^n a_i^4 s_i^4 \right) \log v_n \right]^{1/2} \leq C$.
Then
 - $\pi_S(\Theta^c) \leq v_n^{2\alpha}$, i.e., π_S has most of its mass inside Θ
 - As $n \rightarrow \infty$,

$$R(\Theta) \geq \frac{n}{2m} \log \frac{v_n}{v_{n+m}} + \frac{1}{2m} \sum_{i=1}^n \log \frac{v_{n+m} + s_i^2}{v_n + s_i^2} + O(v_n^\alpha).$$

Asymptotic Minimax Risk

Theorem.

If $m = O(n)$ and $\log(1/v_n) \sum_{i=1}^n a_i^4 \tilde{\theta}_i^4 = o(1)$ as $v_n \rightarrow 0$, then

$$\lim_{n \rightarrow \infty} \frac{R(\Theta)}{R_L(\Theta)} = 1.$$

Outline

Introduction

The Problem Setup

Asymptotic Minimax Risk

Examples

Concluding Remarks

Example 1 (\mathcal{L}^2 Balls)

- Suppose $m = n$ and θ is restricted in a \mathcal{L}^2 ball

$$\Theta(C) = \{\theta : \sum_{i=1}^n \theta_i^2 \leq C\}.$$

- The linear minimax risk $R_L(\Theta_n(C))$ is asymptotically equivalent to the overall minimax risk $R_N(\Theta_n(C))$, and

$$\liminf_{n \rightarrow \infty} R_N(\Theta_n(C))/n = \frac{1}{2} \log \frac{(m+n)C + \sigma^2}{nC + \sigma^2}.$$

- This minimax risk is strictly smaller than the minimax risk over the class of plug-in estimators, which is $C/(1+C)$ by Pinsker's theorem.

Example 2 (Sobolev ellipsoids)

- Suppose $m = n$ and θ is restricted in a Sobolev ellipsoid

$$\Theta(C, \alpha) = \left\{ \theta : \sum_{i=1}^{\infty} a_i^2 \theta_i^2 \leq C \right\},$$

where $a_{2i} = a_{2i-1} = (2i)^\alpha$ ($\alpha > 0$) for $i = 1, 2, \dots$.

- The linear minimax risk $R_L(\Theta_n(C))$ is asymptotically equivalent to the overall minimax risk $R_N(\Theta_n(C))$, which has the same convergence rate $n^{-2\alpha/(2\alpha+1)}$ as the minimax risk over the class of plug-in estimators but has a strictly smaller convergence constant.

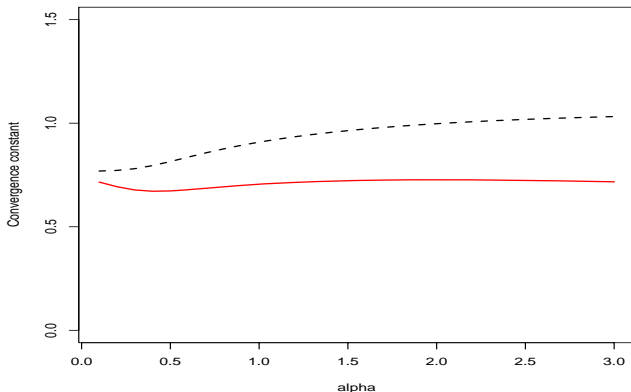


Figure: Convergence constants of the overall minimax risk (the bottom red line) and the minimax risk over plug-in estimators (the top red line). The sample size $n = 10,000,000$ and $C = 1$.

Outline

Introduction

The Problem Setup

Asymptotic Minimax Risk

Examples

Concluding Remarks

Concluding Remarks

- In this work we study the problem of estimating the predictive density of future observations from a nonparametric regression model and establish the exact asymptotics of minimax risk
- The plug-in estimators are not satisfactory because they ignore the uncertainty inherent in the function estimation process
- The Bayesian predictive densities under Gaussian priors could reach the asymptotic minimax risk
- Future work:
 - Empirical Bayes and James-Stein type shrinkage predictive densities
 - Adaptive minimaxity over ellipsoids

References

- Aitchison, J. (1975). Goodness of Prediction Fit. *Biometrika*, 62, 547-554.
- Belitser, E.N. and Levit, B.Y. (1995). On minimax filtering over ellipsoids. *Mathematical Methods of Statistics*, 3, 259-273.
- Belitser, E. and Levit, B. (1996). Asymptotically minimax nonparametric regression in L_2 . *Statistics* 28, 105-122.
- Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Annals of Statistics*, 7, 269-281.
- George, E.I. and Xu, X. (2008). Predictive Density Estimation for Multiple Regression. *Econometric Theory*, 24, 528-544.

References (Cont.)

- Liang, F. and Barron, A. (2004). Exact Minimax Strategies for Predictive Density Estimation, Data Compression and Model Selection. *IEEE Trans. Inform. Theory*, 50, 2708-2726.
- Murray, G.D. (1977). A Note on the Estimation of Probability Density Functions. *Biometrika*, 64, 150-152.
- Ng, V.M. (1980). On the Estimation of Parametric Density Functions. *Biometrika*, 67, 505-506.