

Bayesian estimation of bandwidths in a nonparametric regression model with an unknown error density

Xibin (Bill) Zhang, Maxwell L. King, Han Lin Shang

Department of Econometrics and Business Statistics



MONASH University

xibin.zhang@monash.edu

Outline

- 1 Motivation
- 2 Introduction
- 3 Bayesian estimation
- 4 A Simulation Study
- 5 Applications
- 6 Conclusion

Motivation

- 1 Nonparametric regression is an important tool for exploring unknown relationships between variables.
- 2 It is of great importance to investigate the distribution of the response, which is implied by the error density.
- 3 Any assumption on the analytical form of the error density is only an approximation to the unknown true error density.
- 4 We propose to approximate the true error density by a mixture of n normal densities, which have a common variance and individual means at the errors.
- 5 We investigate the likelihood and posterior under this mixture error density.

Nonparametric regression model

- Let \mathbf{y} denote the response and $\mathbf{x} = (x_1, x_2, \dots, x_d)'$ a set of regressors. Given observations (y_i, \mathbf{x}_i) , for $i = 1, 2, \dots, n$, the nonparametric regression model is

$$y_i = m(\mathbf{x}_i) + \varepsilon_i, \quad (1)$$

where ε_i , for $i = 1, 2, \dots, n$, are assumed to be iid with an unknown density denoted as $f(\varepsilon)$.

- We assume that $f(\varepsilon)$ is approximated by a mixture density:

$$f_b(\varepsilon) = \frac{1}{n} \sum_{i=1}^n \frac{1}{b} \phi\left(\frac{\varepsilon - \varepsilon_i}{b}\right), \quad (2)$$


where $\phi(\cdot)$ is the Gaussian PDF, and the component Gaussian densities have a common variance b^2 and different means at ε_i , for $i = 1, 2, \dots, n$.

Mixture of normal densities

- From the view of kernel smoothing, this mixture error density is a kernel density estimator of the errors (rather than residuals), and b is the bandwidth. We call $f_b(\varepsilon)$ either the mixture normal or the kernel-form error density, where b is referred to as either the standard deviation or bandwidth.
- If $m(\mathbf{x})$ were known, the density of y_i would be

$$y_i \sim f_b(y_i - m(\mathbf{x}_i)) = \frac{1}{n} \sum_{j=1}^n \frac{1}{b} \phi \left(\frac{\{y_i - m(\mathbf{x}_i)\} - \{y_j - m(\mathbf{x}_j)\}}{b} \right),$$

for $i = 1, 2, \dots, n$.

- The validity of this mixture density as a density of the regression errors was investigated by [Yuan and de Gooijer \(2007\)](#) in a class of nonlinear regression models and [Zhang and King \(2010\)](#) in univariate GARCH models. 

Nadaraya-Watson (NW) kernel estimator

- In nonparametric regression, $m(\mathbf{x})$ is unknown and often estimated by the Nadaraya-Watson (NW) kernel estimator. To derive the likelihood, we use the leave-one-out NW estimator.

$$\hat{m}_i(\mathbf{x}_i; \mathbf{h}) = \frac{(n-1)^{-1} \sum_{j=1; j \neq i}^n K((\mathbf{x}_i - \mathbf{x}_j) ./ \mathbf{h}) y_j}{(n-1)^{-1} \sum_{j=1; j \neq i}^n K((\mathbf{x}_i - \mathbf{x}_j) ./ \mathbf{h})},$$

where “./” is division by elements, and $\mathbf{h} = (h_1, h_2, \dots, h_d)'$.

- The error density is then approximated by $f_b(\varepsilon)$ with plugged-in NW regression estimator

$$\varepsilon_i \sim \hat{f}_b(\hat{\varepsilon}_i) = \frac{1}{n} \sum_{j=1}^n \frac{1}{b} \phi \left(\frac{\{y_i - \hat{m}_i(\mathbf{x}_i)\} - \{y_j - \hat{m}_j(\mathbf{x}_j)\}}{b} \right),$$

where $\hat{\varepsilon}_i = y_i - \hat{m}_i(\mathbf{x}_i)$, for $i = 1, 2, \dots, n$.

Performance of the two kernel estimators

- 1 The performance of the NW estimator is mainly determined by the bandwidth vector \mathbf{h} .
- 2 The accuracy of the kernel-form error density is completely determined by the bandwidth, or the standard deviation shared by the component normal densities.
- 3 In our investigation, we treat both types of bandwidths as parameters. One typical example of such a treatment is the likelihood cross-validation for kernel density estimation of directly observed data, where bandwidths are regarded as parameters. There are many other examples as well.
- 4 We aim to derive an approximate likelihood of \mathbf{y} for given (\mathbf{h}, b) , and then the posterior of (\mathbf{h}, b) .
- 5 The benefit is we can estimate \mathbf{h} and b simultaneously.

Related investigations in nonparametric regression

- [Efromovich \(2005\)](#) emphasized the importance for being able to estimate error density. He proposed an error density estimator, which was shown to be asymptotically as accurate as an oracle that knows the true errors.
- [Linton and Xiao \(2007\)](#) proposed a kernel density estimator based on the local polynomial fitting under an unknown error density. They showed that their estimator is adaptive.
- [Zhang, Brooks and King \(2009\)](#) derived the posterior of bandwidths under Gaussian errors. They used the MCMC simulation technique to estimate bandwidths and error variance.
- We continue the work of [Zhang, Brooks and King \(2009\)](#) by assuming a kernel-form, or a mixture normal density of errors. Our purpose to is to derive the posterior of bandwidths.

Gaussian error distribution

- 1 Zhang, Brooks and King (2009) considered the same nonparametric regression model, where errors are iid $N(0, \sigma^2)$ with σ^2 an unknown parameter.

- 2 The error assumption implies that

$$\frac{y_i - m(\mathbf{x}_i)}{\sigma} \sim N(0, 1),$$

- 3 As $m(\mathbf{x}_i)$ is unknown, they suggested plugging-in the leave-one-out NW kernel estimator of the regression function, and consequently,

$$y_i \sim N(\widehat{m}_i(\mathbf{x}_i; \mathbf{h}), \sigma^2), \quad \text{approximately,}$$

for $i = 1, 2, \dots, n$.

- 4 Therefore, they obtained the likelihood of \mathbf{y} for given \mathbf{h} and σ^2 .

Posterior under Gaussian errors

- Treating (\mathbf{h}, σ^2) as parameters, they derived the likelihood of $\mathbf{y} = (y_1, y_2, \dots, y_n)'$:

$$\begin{aligned} \ell_G(y_1, y_2, \dots, y_n | \mathbf{h}, \sigma^2) \\ = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n \{y_i - \hat{m}_i(\mathbf{x}_i, \mathbf{h})\}^2\right) \end{aligned}$$

- The prior of each squared bandwidth, as well as the prior of σ^2 is assumed to be an inverse Gamma density:

$$\begin{aligned} \pi(h_k^2) &= \frac{(\beta_h)^{\alpha_h}}{\Gamma(\alpha_h)} \left(\frac{1}{h_k^2}\right)^{\alpha_h+1} \exp\left\{-\frac{\beta_h}{h_k^2}\right\}, \quad k = 1, 2, \dots, d, \\ \pi(\sigma^2) &= \frac{(\beta_\sigma)^{\alpha_\sigma}}{\Gamma(\alpha_\sigma)} \left(\frac{1}{\sigma^2}\right)^{\alpha_\sigma+1} \exp\left\{-\frac{\beta_\sigma}{\sigma^2}\right\}, \end{aligned}$$

where α_h , β_h , α_σ and β_σ are hyperparameters.

Kernel density estimator as a mixture density

- Let $\{z_1, z_2, \dots, z_n\}$ denote a sample of independent observations drawn from an unknown density $g_0(z; \kappa)$ with an unbounded support, where κ is the parameter vector. To make inference based on the sample, one has to make assumptions about the analytical form of $g_0(z; \kappa)$.
- Any specification of the true density is only an approximation to $g(x; \kappa)$. One such approximation is given by

$$\tilde{g}(z; b) = \frac{1}{n} \sum_{i=1}^n \frac{1}{b} \phi((z - z_i)/b),$$

which is a mixture density of n Gaussian components with the same variance b^2 and different means at individual observations. This mixture density is also known as the kernel estimator of $g_0(z; \kappa)$.

Mixture normal density of the errors

- This mixture density is a well-defined density function. If $\mathbf{z} \sim \tilde{g}(\mathbf{z}; b)$, we have $E(\mathbf{z}) = \bar{z}$ and $Var(\mathbf{z}) = b^2 + s_z^2$, where $\bar{z} = 1/n \sum_{i=1}^n z_i$ and $s_z^2 = 1/n \sum_{i=1}^n (z_i - \bar{z})^2$.
- Our investigate is focused on how this mixture can be used as an approximation to the unknown error density.
- When the error density of a nonparametric regression model is assumed to the mixture of n normal densities:

$$f_b(\varepsilon) = \frac{1}{n} \sum_{i=1}^n \frac{1}{b} \phi\left(\frac{\varepsilon - \varepsilon_i}{b}\right),$$

we have

$$E(\varepsilon) = \bar{\varepsilon}, \quad Var(\varepsilon) = h^2 + s_\varepsilon^2,$$

where $\bar{\varepsilon} = 1/n \sum_{i=1}^n \varepsilon_i$, and $s_\varepsilon^2 = 1/n \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2$

Posterior under kernel-form error density

- We treat $\mathbf{h}^2 = (h_1^2, h_2^2, \dots, h_d^2)'$ and b^2 as parameters.
- The density of y_i is approximately

$$\hat{f}_y(y_i; \mathbf{h}^2, b) = \frac{1}{n-1} \sum_{j=1; j \neq i}^n \frac{1}{b} \phi\left(\frac{\{y_i - \hat{m}_i(\mathbf{x}_i)\} - \{y_j - \hat{m}_j(\mathbf{x}_j)\}}{b}\right),$$

where the purpose of leave-one-out is to exclude $\phi(0)/b$.

- The likelihood of \mathbf{y} for given \mathbf{h}^2 and b^2 is

$$\ell(\mathbf{y}|\mathbf{h}^2, b) = \prod_{i=1}^n \hat{f}_y(y_i; \mathbf{h}^2, b).$$

- The prior of each element of \mathbf{h}^2 is an inverse Gamma, and the prior of b^2 is

$$\pi(b^2) = \frac{(\beta_b)^{\alpha_b}}{\Gamma(\alpha_b)} \left(\frac{1}{b^2}\right)^{\alpha_b+1} \exp\left\{-\frac{\beta_b}{b^2}\right\}$$

Posterior and MCMC simulation

- The posterior of $(\mathbf{h}^2, b^2)'$ is approximately

$$\begin{aligned}\pi(\mathbf{h}^2, b^2 | \mathbf{y}) &\propto \pi(\mathbf{h}^2) \times \pi(b^2) \times \ell(\mathbf{y} | \mathbf{h}^2, b^2), \\ &= \left\{ \prod_{k=1}^d \pi(h_k^2) \right\} \times \pi(b^2) \times \prod_{i=1}^n \left\{ \frac{1}{n-1} \sum_{j=1; j \neq i}^n \frac{1}{b} K \left(\frac{\hat{\varepsilon}_i - \hat{\varepsilon}_j}{b} \right) \right\}\end{aligned}$$

- We used the random-walk Metropolis algorithm to sample b^2 and the elements of \mathbf{h}^2 .

Simulation of one data set

- 1 Consider the relationship between \mathbf{y} and $\mathbf{x} = (x_1, x_2, x_3)'$ given by

$$m(\mathbf{x}) = \sin(2\pi x_1) + 4(1 - x_2)(1 + x_2) + \frac{2x_3}{1 + 0.8x_3^2}.$$

- 2 A sample was generated by drawing x_1, x_2, x_3 independently from $U(0, 1)$, and ε_i from a mixture of two Gaussian densities defined by $0.7N(0, 0.7^2) + 0.3N(0, 1.5^2)$. The sample size is $n = 1000$.
- 3 We consider three assumptions of the error density: Our proposed mixture of n normal densities; the normal density with mean zero and variance σ^2 ; and the Student t density with ν degrees of freedom.

Table 1: Parameters estimated through Bayesian sampling under assumptions of kernel-form, Student t and Gaussian error densities.

Error density	Parameter	Estimate	95% Bayesian credible intervals	SIF
Mixture	b	0.1886	(0.1236, 0.2736)	4.88
	h_1	0.0854	(0.0706, 0.1015)	13.33
	h_2	0.0925	(0.0783, 0.1095)	17.61
	h_3	0.2184	(0.1652, 0.2681)	18.51
	log marginal likelihood	-1381.21		
Student t	ν	13.4875	(9.1561, 19.2255)	4.37
	h_1	0.0875	(0.0691, 0.1078)	9.69
	h_2	0.0962	(0.0782, 0.1152)	12.48
	h_3	0.2219	(0.1638, 0.2901)	12.65
	log marginal likelihood	-1409.95		
Gaussian	σ	0.9912	(0.9478, 1.0353)	1.00
	h_1	0.0866	(0.0688, 0.1097)	11.94
	h_2	0.0957	(0.0790, 0.1135)	12.75
	h_3	0.2236	(0.1691, 0.2805)	18.84
	log marginal likelihood	-1423.12		

Mixing performance of the sampler

- We computed the batch-mean standard deviation and simulation inefficiency factor (SIF) to monitor the mixing performance.
- All simulated chains obtained under each assumption of the error density have achieved reasonable mixing performance.
- We computed the marginal likelihood under each assumption of the error density. We found that the log marginal likelihood derived under the kernel-form or the Student t densities of the errors are obviously larger than that derived under the normal errors.
- This is not surprising, because the true error density is a mixture of two normals. The Student t and kernel-form error densities capture the distributional properties of errors well.

Bayes factor for model comparison

- Bayes factor is a ratio of the marginal likelihoods derived under a model of interest and its competing model.
- Let θ denote the parameter vector under model A . The marginal likelihood under model A is (Chib, 1995)

$$p_A(\mathbf{y}) = \frac{\ell_A(\mathbf{y}|\theta)\pi_A(\theta)}{\pi_A(\theta|\mathbf{y})},$$

which is computed at the posterior estimate of θ . $\ell_A(\mathbf{y}|\theta)$ and $\pi_A(\theta)$ are the likelihood and prior under model \mathcal{A} .

- The Bayes factor of model \mathcal{A} against model \mathcal{B} is

$$\text{BF} = \frac{p_A(\mathbf{y})}{p_B(\mathbf{y})}.$$

- We used the methods proposed by Chib (1995) and Geweke (1999) to compute marginal likelihood.

Monte Carlo simulation with 1000 generated samples

- We simulate 1000 independent samples under each of the following three assumptions of the errors.
 - ① Gaussian density with mean zero and variance $\sigma^2 = 0.9^2$;
 - ② Student t with $\nu = 4$ degrees of freedom; and
 - ③ A mixture of two normal densities:
 $0.7N(0, 0.7^2) + 0.3N(0, 1.5^2)$.

Simulation results with errors simulated from $N(0, 0.9^2)$

Table 2: Relative frequencies for the normal assumption (\mathcal{A}) to be supported against the other two assumptions.

Assumption of error density	Gaussian \mathcal{A}	Student t \mathcal{B}	Mixture \mathcal{B}
Very strong evidence in favor of \mathcal{A}		92.2%	31.1%
Strong evidence in favor of \mathcal{A}		5.6%	31.9%
Positive evidence in favor of \mathcal{A}		1.7%	22.6%
Not worth more than a bare mention		0.3%	7.1%
Not worth more than a bare mention		0.0%	4.3%
Positive evidence in favor of \mathcal{B}		0.1%	2.1%
Strong evidence in favor of \mathcal{B}		0.1%	0.6%
Very strong evidence in favor of \mathcal{B}		0.0%	0.3%

- The normal assumption is supported against the Student t in 99.5% of the simulated samples, and against the mixture normal assumption in 85.6% of the samples.
- The mixture normal assumption is supported against the normal in 3% of the simulated samples.

Simulation results with errors simulated from t_4

- Table 3: Relative frequencies for the Student t assumption (\mathcal{A}) to be supported against the other two assumption.

Assumption of error density	Gaussian \mathcal{B}	Student t \mathcal{A}	Mixture \mathcal{B}
Very strong evidence in favor of \mathcal{A}	100.0%		84.4%
Strong evidence in favor of \mathcal{A}	0.0%		5.9%
Positive evidence in favor of \mathcal{A}	0.0%		3.5%
Not worth more than a bare mention	0.0%		1.9%
Not worth more than a bare mention	0.0%		1.8%
Positive in favor of model \mathcal{B}	0.0%		0.9%
Strong evidence in favor of \mathcal{B}	0.0%		0.9%
Very strong evidence in favor of \mathcal{B}	0.0%		0.7%

- In 93.8% of the simulated samples, the Student t assumption is supported against the mixture normal. In 2.6% of the simulated samples, the mixture normal assumption is supported against the Student t .

Error simulated from the mixture of Gaussian densities

- Table 4: Relative frequencies for the mixture normal assumption (\mathcal{A}) to be supported against the other two assumptions.

Assumption of error density	Gaussian \mathcal{B}	Student t \mathcal{B}	Mixture \mathcal{A}
Very strong evidence in favor of \mathcal{A}	97.1%	37.7%	
Strong evidence in favor of \mathcal{A}	1.6%	11.4%	
Positive evidence in favor of \mathcal{A}	0.5%	11.5%	
Not worth more than a bare mention	0.4%	8.0%	
Not worth more than a bare mention	0.1%	5.7%	
Positive evidence in favor of \mathcal{B}	0.2%	9.7%	
Strong evidence in favor of \mathcal{B}	0.1%	5.1%	
Very strong evidence in favor of \mathcal{B}	0.0%	10.9%	

- The mixture normal assumption is supported against the normal in 99.2% of the simulated samples, and against the Student t in 60.6% of the samples.
- The Student t is supported against the mixture in 25.7% samples.

Regression of AORD returns on FTSE and S&P 500 returns

- 1 Data consist of All Ordinaries (AORD), FTSE and S&P 500 closing indices from 03 Jan 2007 to 30 Dec 2011, excluding non-trading days.
- 2 Daily AORD was matched with the overnight FTSE and S&P 500 returns.
- 3 Multivariate kernel regression model is given by

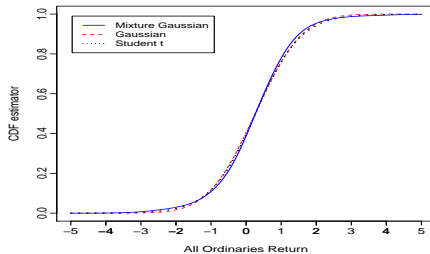
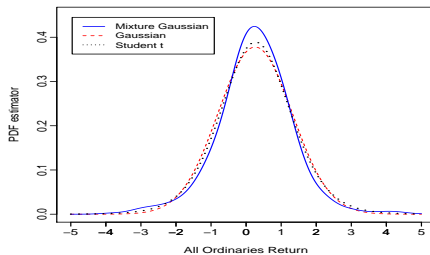
$$y_i = m(x_{1,i}, x_{2,i}) + \varepsilon_i,$$

where y_i is the return of AORD, and $\varepsilon_i, i = 1, 2, \dots, n$ are assumed to be iid and follow either Gaussian, Student t or mixture Gaussian.

Bayesian bandwidth estimation under 3 assumptions of errors

Error density	Parameter	Estimate	95% Bayesian credible interval	SIF
Mixture	b	0.3108	(0.2194, 0.4077)	15.53
	h_1	0.4902	(0.3516, 0.5977)	20.44
	h_2	0.7462	(0.6291, 0.8796)	11.76
	LML	-1484.44		
Gaussian	σ	1.0557	(1.0115, 1.1069)	0.78
	h_1	0.6244	(0.5414, 0.7125)	19.43
	h_2	0.7951	(0.6941, 0.9160)	19.03
	LML	-1514.79		
Student t	ν	9.9097	(7.0205, 13.8122)	6.19
	h_1	0.5960	(0.4942, 0.7050)	27.91
	h_2	0.7201	(0.5995, 0.8404)	24.49
	LML	-1485.24		

PDF and CDF of the AORD returns

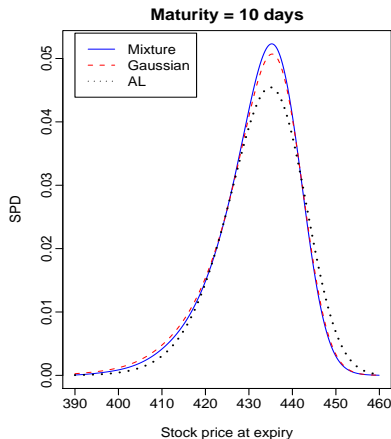
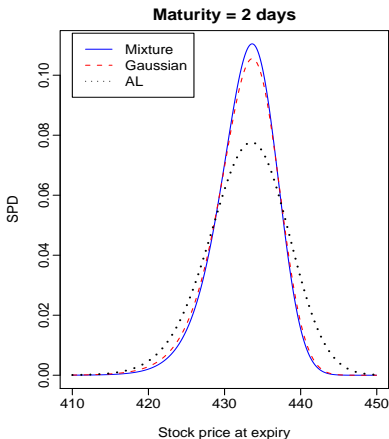


Nonparametric regression involved in state-price density estimation

- ➊ [Aït-Sahalia and Lo \(1998\)](#) suggested estimating the state-price density of a security through the BS formula with plugged-in estimated volatility, which is estimated by a nonparametric regression of implied volatility on strike price, futures price and maturity.
- ➋ They used the rule-of-thumb to choose bandwidths.
- ➌ [Zhang, Brooks and King \(2009\)](#) used Bayesian sampling techniques to estimate bandwidths under Gaussian errors.
- ➍ We assume the error density is a mixture of n normal densities.
- ➎ Sample: S&P 500 options data from the 3rd Jan to 31st Dec 1993; and the sample size $n=14431$.

Graphs of estimated state-price density

- The Bayes factor of the kernel-form error density against the normal error density is $\exp(4210.3)$.



Conclusion

- We propose using a mixture of n normal densities as the error density for the nonparametric regression model for the purpose of estimating bandwidths. This error density has the form of a kernel density estimator of the errors.
- We derived an approximate likelihood and posterior, where the bandwidths in the NW estimator and the kernel-form error density are treated as parameters.
- According to Bayes factors, the kernel-form error density is not supported against the true error density, but is supported against a wrong specification of the error density.

- A benefit of the kernel-form error density is to forecast the density of the response, such as asset return density.
- This error density is supported against Gaussian in the nonparametric regression of realized volatility on strike price, futures prices and maturity in the sample of S&P 500 options data. The resulting SPD estimator is different from that under Gaussian errors for short maturities.