Nonparametric Conditional Density Estimation

Bruce E. Hansen*
University of Wisconsin†

www.ssc.wisc.edu/~bhansen

November 2004
Preliminary and Incomplete

*Research supported by the National Science Foundation.
†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706

1 Introduction

Conditional density functions are a useful way to display uncertainty. This paper investigates nonparametric kernel methods for their estimation. The standard estimator is the ratio of the joint density estimate to the marginal density estimate. Our proposal is to instead use a two-step estimator, where the first step consists of estimation of the conditional mean, and the second step consists of estimating the conditional density of the regression error. If most of the dependence is captured by the conditional mean, the second step will require less smoothing, thereby reducing estimation variance.

Conditional density estimation was introduced by Rosenblatt (1969). A bias correction was proposed by Hyndman, Bashtannyk and Grunwald (1996). Fan, Yao and Tong (1996) proposed a direct estimator based on local polynomial estimation; see also Section 6.5 of Fan and Yao (2003). Bandwidth selection rules have been proposed by Bashtannyk and Hyndman (2001), Fan and Yim (2004), and Hall, Racine and Li (2004). The related problem of conditional distribution estimation is examined in Hall, Wolff and Yao (1999). Other papers have used conditional density estimates as an input to other problems, including Robinson (1991), Tjostheim (1994), Polonik and Yao (2000) and Hyndman and Yao (2002).

Our two-step conditional density estimator is partially motivated by the two-step conditional variance estimator of Fan and Yao (1998). They showed that two-step estimation is asymptotically efficient since the first-step conditional mean estimate does not affect the asymptotic distribution of the second-step variance estimator. We show here that this property also applies to conditional density estimation.

Our analysis is confined to the case of a real-valued conditioning variable. The generalization to the case of vector-valued conditioning variables should be straightforward, so long as the conditioning set for the conditional mean and conditional density are identical. However, if the conditional density of the regression error has a reduced conditioning set relative to the conditional mean, the analysis changes. (For example, if the conditional mean has two variables and the conditional error density only one.) In this case the second-step estimator may not be asymptotically independent of the first-step. More importantly, it appears that the two-step estimator may achieve an improved convergence rate relative to the conventional direct estimator. This analysis is more involved and remains to be completed.

Our two-step estimator could also be generalized to three steps, where an intermediate step estimates the conditional variance. We expect the qualitative analysis to be similar, and conjecture that there will be further improvements in estimation efficiency. This work remains to be completed.

Furthermore, our discussion is based on local average estimates. Alternatively, the mean, variance, or density can be estimated using local linear estimators. This should be explored, as local linear estimators have better bias properties than local averages (and thus have improved efficiency) when there is non-trivial dependence. Other than changes in the bias expressions, however, we expect that no important changes will arise in the theory. Again, this work remains to be completed.

The organization of the remainder of the paper is as follows. Section 2 introduces the framework, Section 3 the one-step estimator, Section 4 the new two-step estimator, and Section 5 compares their asymptotic biases. Section 6 discusses cross-validation for bandwidth selection. Section 7 presents simulation evidence, and Section 8 an application to U.S. GDP. Proofs are presented in the Appendix.

2 Framework

The observables $\{Y_i, X_i\}$ in $R \times R$ are strictly stationary and strong mixing. Let $f(y,x)$ and $f(y \mid x)$ denote the joint and conditional density functions, and let $f(x)$ denote the marginal density of X_i. The goal is estimation of $f(y \mid x)$.

Our estimators will be based on kernel regression. Let $K(x) : R \to R$ denote a bounded symmetric kernel function and set $\sigma_K^2 = \int_R u^2 K(u) du$ and $R(K) = \int_R K(u)^2 du$. For a bandwidth h let $K_h(u) = h^{-1} K(u/h)$. Define the derivatives

$$\begin{aligned} f^{(r)}(x) &= \frac{\partial^r}{\partial x^r} f(x) \\ f_{(s)}^{(r)}(y \mid x) &= \frac{\partial^{r+s}}{\partial y^r \partial x^s} f(y \mid x). \end{aligned}$$

3 One-Step Estimator

Let h_1 and h_2 be bandwidths. Standard kernel estimators of $f(y,x)$, $f(x)$ and $f(y \mid x)$ are

$$\begin{aligned} \tilde{f}(y,x) &= \frac{1}{n} \sum_{i=1}^{n} K_{h_2}(x - X_i) K_{h_1}(y - Y_i) \\ \tilde{f}(x) &= \frac{1}{n} \sum_{i=1}^{n} K_{h_2}(x - X_i) \end{aligned}$$

and

$$\tilde{f}(y \mid x) = \frac{\tilde{f}(y,x)}{\tilde{f}(x)} = \frac{\sum_{i=1}^{n} K_{h_2}(x - X_i) K_{h_1}(y - Y_i)}{\sum_{i=1}^{n} K_{h_2}(x - X_i)}.$$

Asymptotic approximations show that it it optimal for estimation of $f(y \mid x)$ to set $h_1 = c_1 n^{-1/6}$ and $h_2 = c_2 n^{-1/6}$ for $c_1 > 0$ and $c_2 > 0$. Under standard regularity conditions the conditional density estimator has the asymptotic distribution

$$n^{-2/6} \left(\tilde{f}(y \mid x) - f(y \mid x) \right) \to^d N\left(\theta_1, \sigma_1^2\right)$$

where

$$\theta_1 = \frac{\sigma_K^2}{2\sqrt{c_1 c_2}} \left(c_1^2 f^{(2)}(y \mid x) + c_2^2 f_{(2)}(y \mid x) + 2 c_2^2 f_{(1)}(y \mid x) f^{(1)}(x) \right)$$

and

$$\sigma_1^2 = \frac{R(K)^2 f(y \mid x)}{c_1 c_2 f(x)}.$$

Observe that the rate of convergence is $O\left(n^{-1/3}\right)$, the same as for bivariate density estimation. It slower than the $O\left(n^{-2/5}\right)$ rate obtained for univariate density estimation and bivariate regression.

4 Two-Step Estimator

Define the conditional mean

$$m(x) = E\left(Y_i \mid X_i = x\right)$$

so that

$$Y_i = m\left(X_i\right) + e_i$$

and e_i is a regression error. Letting $g\left(e \mid x\right)$ denote the conditional density of e_i given $X_i = x$, we have the equivalence

$$f\left(y \mid x\right) = g\left(y - m(x) \mid x\right).$$

From this equation we can see that an alternative method for estimation of f is through estimation of g and m.

Let b_0, b_1 and b_2 be bandwidths The Nadaraya-Watson estimator of $m(x)$ is

$$\hat{m}\left(x\right) = \frac{\sum_{i=1}^{n} K_{b_0}\left(x - X_i\right) Y_i}{\sum_{i=1}^{n} K_{b_0}\left(x - X_i\right)}$$

with residuals

$$\hat{e}_i = Y_i - \hat{m}\left(X_i\right).$$

A second-stage estimator of g is

$$\hat{g}\left(e \mid x\right) = \frac{\sum_{i=1}^{n} K_{b_2}\left(x - X_i\right) K_{b_1}\left(e - \hat{e}_i\right)}{\sum_{i=1}^{n} K_{b_2}\left(x - X_i\right)}.$$

Together we obtain the two-step estimator

$$\begin{aligned} \hat{f}\left(y \mid x\right) &= \hat{g}\left(y - \hat{m}(x) \mid x\right) \\ &= \frac{\sum_{i=1}^{n} K_{b_2}\left(x - X_i\right) K_{b_1}\left(y - \hat{m}(x) - \hat{e}_i\right)}{\sum_{i=1}^{n} K_{b_2}\left(x - X_i\right)}. \end{aligned}$$

Assume that $b_0 = a_0 n^{-1/5}$, $b_1 = a_1 n^{-1/6}$ and $b_2 = a_2 n^{-1/6}$

Theorem 1

$$n^{-2/6}\left(\hat{f}\left(y \mid x\right) - f\left(y \mid x\right)\right) \to^{d} N\left(\theta_3, \sigma_3^2\right)$$

where

$$\theta_2 = \frac{\sigma_k^2}{2\sqrt{a_1 a_2}}\left(a_1^2 g^{(2)}(e \mid x) + a_2^2 g_{(2)}(e \mid x) + 2a_2^2 g_{(1)}(e \mid x) f^{(1)}(x)\right)$$

with $e = y - m(x)$ *and*

$$\sigma_2^2 = \frac{R(K)^2 f(y \mid x)}{a_1 a_2 f(x)}.$$

This result states that the asymptotic distribution of the two-step estimator is unaffected by the first estimation step. The bandwidth b_0 does not enter the first-order approximation, and the distribution is the same as when the mean $m(x)$ and errors e_i are known without estimation. This occurs because the conditional mean estimator $\hat{m}(x)$ converges at the faster rate of $O\left(n^{-2/5}\right)$.

In the special case that $g(e \mid x) = g(e)$ does not depend on x, then it is optimal to set $b_2 = \infty$. In this case we find that the convergence rate improves to $O\left(n^{-2/5}\right)$.

$$n^{-2/5}\left(\hat{f}(y \mid x) - f(y \mid x) - \theta_2\right) \to^d N\left(\theta_2, \sigma_2^2\right)$$

5 Bias Comparison

Note that the scaled $\hat{f}$ and $\tilde{f}$ have differing biases. We can compare the latter by observing that

$$f^{(2)}(y \mid x) = g^{(2)}(e \mid x)$$

$$f_{(1)}(y \mid x) = g_{(1)}(e \mid x) - g^{(1)}(e \mid x) m^{(1)}(x)$$

$$f_{(2)}(y \mid x) = g_{(2)}(e \mid x) - g^{(1)}(e \mid x) m^{(2)}(x) + g^{(2)}(e \mid x)\left(m^{(1)}(x)\right)^2.$$

Therefore

$$\begin{aligned}\theta_1 &= \frac{\sigma_k^2 c_1^2}{2\sqrt{c_1 c_2}} g_{(2)}(e \mid x) + \frac{\sigma_k^2 c_2^2}{\sqrt{c_1 c_2}}\left(g_{(1)}(e \mid x) - g^{(1)}(e \mid x) m^{(1)}(x)\right) f^{(1)}(x) \\ &\quad + \frac{\sigma_k^2 c_2^2}{2\sqrt{c_1 c_2}}\left(g^{(2)}(e \mid x) - g^{(1)}(e \mid x) m^{(2)}(x) + g^{(2)}(e \mid x)\left(m^{(1)}(x)\right)^2\right)\end{aligned}$$

Unless $m^{(1)}(x) = 0$, θ_1 has more components than θ_2, and will typically be larger (for equal bandwidths). Thus $\hat{f}$ has lower bias than $\tilde{f}$, enabling the selection of a larger bandwidth scale a_2 for $\hat{f}$ than b_2 for $\tilde{f}$, reducing variance and mean-squared-error.

6 Bandwidth Selection

Fan and Yim (2004) and Hall, Racine and Li (2004) have proposed a cross-validation method appropriate for nonparametric conditional density estimators. In this section we describe this method and its

application to our estimators. For an estimator $\tilde{f}(y \mid x)$ of $f(y \mid x)$ define the integrated squared error

$$\begin{aligned} I &= \int\int \left(\tilde{f}(y \mid x) - f(y \mid x)\right)^2 f(x) dy dx \\ &= \int\int \tilde{f}(y \mid x)^2 f(x) dy dx - 2\int\int \tilde{f}(y \mid x) f(y \mid x) f(x) dy dx + \int\int f(y \mid x)^2 f(x) dy dx \\ &= I_1 - 2I_2 + I_3. \end{aligned}$$

Note that I_3 does not depend on the bandwidths and is thus irrelvant.

Ideally, we would like to pick the bandwidths to minimize I, but this is infeasible as the function I is unknown. Cross-validation replaces it with an estimate based on the leave-one-out principle. Let $\tilde{f}_{-i}(y \mid x)$ denote the estimator $\tilde{f}(y \mid X)$ with observation i omitted. The cross-validation estimators of I_1 and I_2 are

$$\begin{aligned} \hat{I}_1 &= \frac{1}{n}\sum_{i=1}^{n} \int \tilde{f}_{-i}(y \mid X_i)^2 dy \\ \hat{I}_2 &= \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{-i}(Y_i \mid X_i). \end{aligned}$$

We then define the cross-validation function as

$$\hat{I} = \hat{I}_1 - 2\hat{I}_2.$$

The cross-validated bandwidths are those which jointly minimize $\hat{I}$.

For the one-step estimator these equal components equal

$$\hat{I}_2 = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j\neq i} K_{h_2}(X_i - X_j) K_{h_1}(Y_i - Y_j)}{\sum_{j\neq i} K_{h_2}(X_i - X_j)}$$

and

$$\begin{aligned} \hat{I}_1 &= \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j\neq i}\sum_{k\neq i} K_{h_2}(X_i - X_j) K_{h_2}(X_i - X_k) \int K_{h_1}(y - Y_j) K_{h_1}(y - Y_k) dy}{\left(\sum_{j\neq i} K_{h_2}(X_i - X_j)\right)^2} \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j\neq i}\sum_{k\neq i} K_{h_2}(X_i - X_j) K_{h_2}(X_i - X_k) K_{\sqrt{2}h_1}(Y_k - Y_j)}{\left(\sum_{j\neq i} K_{h_2}(X_i - X_j)\right)^2}, \end{aligned}$$

the second equality when $K(u) = \phi(u)$, the Gaussian kernel.

For the two-step estimator we suggest selecting the bandwidths in two steps. First, the bandwidth b_0 may be selected by least-squares cross-validation. Second, (b_1, b_2) may be selected by using the method outlined above.

7 Simulation Evidence

The performance of the nonparametric estimators were compared in a simple stochastic setting. The data are generated by the process

$$\begin{aligned} x_i &\sim N(0,1) \\ y_i \quad | \quad x_i &\sim N\left(\beta_1 x_i, \frac{1+\beta_2 x_i^2}{1+\beta_2}\right). \end{aligned}$$

1000 samples of size $n = 100$ were generated. We vary β_1 among 0.1, 1, and 2, and β_2 among 0.1 and 1.

On each sample, the one-step estimator $\tilde{f}(y \mid x)$ and two-step estimator $\hat{f}(y \mid x)$ were calculated, using a Gaussian kernel. We measure accuracy by mean integrated squared error

$$I(\tilde{f}) = 100 \times E \int\int \left(\tilde{f}(y \mid x) - f(y \mid x)\right)^2 f(x) dy dx$$

where the integrals are approximated by a 50×50 grid on (y, x).

The estimators depend critically on the bandwidths $h = (h_1, h_2)$ and $b = (b_0, b_1, b_2)$. For our first comparison, we use the infeasible oracle bandwidth. This is the bandwidth which minimizes the finite sample MISE. This enables a comparison of the estimation methods free of dependence on bandwidth selection methods.

For the two estimators Table 1 reports the MISE and the oracle bandwidths. The results are as expected. For the case of small conditional mean effect ($\beta_1 = 0.1$), then the two estimators perform similarly in terms of MISE. However, if the conditional mean effect is non-trivial, then the two-step estimator $\hat{f}$ has much smaller MISE. The reduction in MISE is as much as 50%.

Table 1
Mean Integrated Squared Error Using Oracle Bandwidth
$n = 100$

β_1	β_2	$I(\tilde{f})$	$I(\hat{f})$	h_1	h_2	b_0	b_1	b_2
0.1	0.1	0.59	0.56	.46	1.64	1.16	.47	2.61
1.0	0.1	1.46	0.82	.57	.35	.38	.48	1.34
2.0	0.1	2.15	1.05	.64	.21	.29	.50	1.09
0.1	1.0	1.18	1.18	.44	.66	8.49	.44	.66
1.0	1.0	2.01	1.39	.49	.32	.51	.45	.58
2.0	1.0	2.88	1.57	.56	.20	.37	.46	.53

For our second comparison, we use data-dependent bandwidths. For the one-step estimator $\tilde{f}$ we use the cross-validated bandwidth. For the two-step estimator $\hat{f}$ we use sequential bandwidths. The bandwidth $\hat{b}_0$ is selected by least-squares cross-validation for the mean, and $(\hat{b}_1, \hat{b}_2)$ are selected by conditional density cross-validation using the estimated residuals.

Table 2 reports the MISE for the two estimators. It also reports the median data-dependent bandwidths. The qualitative results are similar to those for the optimal bandwidths, with the notable change that the improvement of the two-step estimator relative to the one-step estimator has been reduced. For the cases with small conditional mean effect ($\beta_1 = 0.1$) the MISE is even somewhat higher for $\hat{f}$ than for $\tilde{f}$, but in the other cases $\hat{f}$ has much lower MISE. This suggests that further investigation into bandwidth selection may yield further improvements.

Table 2
Mean Integrated Squared Error Using Data-Dependent Bandwidths
$n = 100$

β_1	β_2	$I(\tilde{f})$	$I(\hat{f})$	$\hat{h}_1$	$\hat{h}_2$	$\hat{b}_0$	$\hat{b}_1$	$\hat{b}_2$
0.1	0.1	1.07	1.26	.48	1.48	1.27	.46	4.37
1.0	0.1	1.96	1.34	.61	.39	.38	.44	4.02
2.0	0.1	2.63	1.84	.69	.23	.27	.43	2.21
0.1	1.0	1.71	1.96	.44	.77	1.27	.42	.89
1.0	1.0	2.59	2.28	.53	.38	.40	.41	.97
2.0	1.0	3.44	2.36	.60	.22	.30	.41	.94

8 Application to U.S. GDP Growth

Our first illustration is a time-series application. Let Y_t denote U.S. quarterly real GDP and let $y_t = 100(\ln(Y_t) - \ln(Y_{t-1}))$ denotes its growth rate. We are interested in estimation of the one-step ahead conditional density $f(y_t \mid y_{t-1})$. Due to strong evidence of a shift in variance in the early 1980s, we use the sample period 1983:1-2004:3 which results in a small sample.

First, for a baseline we take the linear Gaussian model, for which least-squares yield the estimate

$$\hat{f}_0(y_t \mid y_{t-1}) = \phi_{0.5}\left(y_t - .5 - .4y_{t-1}\right).$$

Second, we estimate $f(y_t \mid y_{t-1})$ using the one-step estimator with cross-validated bandwidth, and let this estimate be denoted as $\hat{f}_1(y_t \mid y_{t-1})$. The cross-validated bandwidths are $h_1 = .26$ and $h_2 = .20$.

Third, we estimate the conditional density using the two-step estimator with sequential cross-validated bandwidth, and denote this estimator as $\hat{f}_2(y_t \mid y_{t-1})$. The cross-validated bandwidths are $h_0 = .15$, $h_1 = .21$ and $h_2 = 592$. The latter value of h_2 means that cross-validation eliminates the conditional smoothing in the second step, so the estimated conditional density only depends on y_{t-1} through the estimated conditional mean. This is not surprising due to our application to a small sample. This also highlights an important distinction between the one-step and two-step estimators, as the former does not have this flexibility.

Figures 1 through 4 display the three density estimates as a function of y_t for four fixed values of y_{t-1}. In general, the three estimators differ from one another. In particular, the inefficient one-step estimator appears to be mis-centered and over-dispersed in Figure 1 ($y_{t-1} = .2$), and in all cases has a thicker right tail than the two-step estimator.

9 Application to Wage Distrbiution

Our second illustration is a cross-section application, the conditional density of log-wages given age. For individual i let Y_i denote log wages and X_i denote Age. We take our data from the 1995 Current Population Survey, March Supplement. Our sample consists of the 2128 men aged 18 to 65 who are working (not self-employed) with positive reported earnings who have a high school diploma but no college education.

Again, we estimate $f(y \mid x)$ using the one-step and two-step estimators with cross-validated bandwidths, denoted as $\hat{f}_1(y \mid x)$ and $\hat{f}_2(y \mid x)$. The one-step cross-validated bandwidths are $h_1 = 0.104$ and $h_2 = 2.401$. The two-step cross-validated bandwidths are $b_0 = 2.07$, $b_1 = 0.082$ and $b_2 = 4.87$. As in the prior application, $b_2 > h_2$, meaning that less smoothing is done in the second step than by the one-step estimator.

Figures 5 through 8 display the two density estimates as a function of y for four fixed age levels, $x = 25$, 35, 45 and 55. The two estimators differ from one another, with the two-step estimator typically more peaked.

References

[1] Bashtannyk, D.M. and Rob J. Hyndman (2001): "Bandwidth selection for kernel conditional density estimation," *Computational Statistics and Data Analysis*, 36, 279-298.

[2] Fan, Jianqing and Qiwei Yao (1998): "Efficient estimation of conditional variance functions in stochastic regression, *Biometrika*, 85, 645-660.

[3] Fan, Jianqing and Qiwei Yao (2003): *Nonlinear Time Series: Nonparametric and Parametric Methods*. New York: Springer-Verlag

[4] Fan, Jianqing, Qiwei Yao, and Howell Tong (1996): "Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems," *Biometrika*, 83, 189-206.

[5] Fan, Jianqing and Tsz Ho Yim (2004): "A cross-validation method for estimating conditional densities," *Biometrika*, forthcoming.

[6] Hall, Peter, Jeff Racine and Qi Li (2004): "Cross-validation and the estimation of conditional probability densities," working paper.

[7] Hall, Peter, R.C.L. Wolff and Qiwei Yao (1999): "Methods for estimating a conditional distribution function," *Journal of the American Statistical Association*, 94, 154-163.

[8] Hansen, Bruce E. (2004): "Uniform Convergence Rates for Kernel Estimation," working paper.

[9] Hyndman, Rob J. and Qiwei Yao (2002): "Nonparametric estimation and symmetry tests for conditional density functions," *Nonparametric Statistics*, 14, 259-278.

[10] Hyndman, Rob J., D.M. Bashtannyk and G.K. Grunwald (1996): "Estimating and visualizing conditional densities," *Journal of Computational and Graphical Statistics*, 5, 315-336.

[11] Polonik, W. and Qiwei Yao (2000): "Conditional minimum volume predictive regions for stochastic processes," *Journal of the American Statistical Association*, 95, 509-519.

[12] Robinson, Peter M. (1991): "Consistent nonparametric entropy-based testing," *Review of Economic Studies*, 58, 437-453.

[13] Rosenblatt, M. (1969): "Conditional probability density and regression estimates," in *Multivariate Analysis II*, Ed. P.R. Krishnaiah, pp. 25-31. New York: Academic Press.

[14] Tjostheim, D. (1994): "Non-linear time series: A selective review," *Scandinavian Journal of Statistics*, 21, 97-130.

10 Appendix

The proofs contained here are incomplete sketches, and omit regularity conditions.

We first state two results from Hansen (2004).

Lemma 1 *Let*

$$\hat{G}(x,z) = \frac{1}{h_1 h_2 n}\sum_{i=1}^{n}\psi\left(Y_i, X_i, Z_i\right) G_1\left(\frac{x-X_i}{h_1}\right) G_2\left(\frac{z-Z_i}{h_2}\right).$$

Under regularity conditions

$$\sup_{x\in R, z\in R}\left|\hat{G}(x,z) - E\hat{G}(x,z)\right| = O_p\left(\left(\frac{\log n}{h_1 h_2 n}\right)^{1/2}\right).$$

Let $\delta_n = (\log n)^{-1/2}$ and define the set

$$S_n = \left\{x \in R : f(x) \geq \delta_n \text{ and } \left|\frac{d^3}{dx^3} m(x)\right| \leq \delta_n\right\}.$$

Lemma 2 *Uniformly for* $x \in S_n$

$$\hat{m}(x) - m(x) = f(x)^{-1}\frac{1}{n}\sum_{i=1}^{n} K_{b_0}\left(x - X_i\right) e_i - b_0^2 \sigma_k^2 f(x)^{-1}\left(f^{(1)}(x) m^{(1)}(x) + m^{(2)}(x)\right) + O_p\left(\left(\log n\right) n^{-3/5}\right).$$

Define

$$\hat{g}^*\left(e \mid x\right) = \frac{\sum_{i=1}^{n} K_{b_2}\left(x - X_i\right) K_{b_1}\left(e - e_i\right)}{\sum_{i=1}^{n} K_{b_2}\left(x - X_i\right)}.$$

Lemma 3 *Uniformly for* $e, x \in R \times S_n$

$$\hat{g}\left(e \mid x\right) - \hat{g}^*\left(e \mid x\right) = O_p((\log n)^{1/2} n^{-2/5})$$

Proof. Observe that

$$\begin{aligned}
\hat{g}\left(e \mid x\right) - \hat{g}^*\left(e \mid x\right) &= B_n^{-1} A_n \\
A_n &= \frac{1}{n}\sum_{i=1}^{n} K_{b_2}\left(x - X_i\right)\left(K_{b_1}\left(e - \hat{e}_i\right) - K_{b_1}\left(e - e_i\right)\right) \\
B_n &= \frac{1}{n}\sum_{i=1}^{n} K_{b_2}\left(x - X_i\right).
\end{aligned}$$

Since

$$EK_{b_2}\left(x - X_i\right) = f(x) + O(b_2^2) = f(x) + O\left(n^{-1/3}\right)$$

then using Lemma 1, uniformly in $x \in R$

$$B_n = f(x) + O\left(n^{-1/3}\right) + O_p\left(\left(\frac{\log n}{b_2 n}\right)^{1/2}\right) = f(x) + O_p\left(n^{-1/3}\right)$$

and by a Taylor expansion, uniformly for $x \in S_n$

$$B_n^{-1} - f(x)^{-1} = O_p\left(\delta_n^{-2} n^{-1/3}\right).$$

Next, to decompose A_n, first observe that by a Taylor expansion

$$\begin{aligned} K_{b_1}(e - \hat{e}_i) - K_{b_1}(e - e_i) &\simeq K_{b_1}^{(1)}(e - e_i)(e_i - \hat{e}_i) \\ &= \frac{1}{b_1^2} K^{(1)}\left(\frac{e - e_i}{b_1}\right)\left(\hat{m}(X_i) - m(X_i)\right). \end{aligned}$$

Second, by Lemma 2, uniformly in i

$$\begin{aligned} \hat{m}(X_i) - m(X_i) &= f(X_i)^{-1} \frac{1}{n b_0} \sum_{j=1}^{n} K\left(\frac{X_i - X_j}{b_0}\right) e_j \\ &\quad - b_0^2 \sigma_k^2 f(X_i)^{-1}\left(f^{(1)}(X_i) m^{(1)}(X_i) + m^{(2)}(X_i)\right) \\ &\quad + O_p\left((\log n) n^{-3/5}\right). \end{aligned}$$

Together

$$\begin{aligned} A_n &\simeq \frac{1}{n}\sum_{i=1}^{n} K_{b_2}(x - X_i) \frac{1}{b_1^2} K^{(1)}\left(\frac{e - e_i}{b_1}\right) \\ &\quad \left[f(X_i)^{-1} \frac{1}{n b_0} \sum_{j=1}^{n} K\left(\frac{X_i - X_j}{b_0}\right) e_j - b_0^2 \sigma^2(K) f(X_i)^{-1} f^{(1)}(X_i) m^{(1)}(X_i) + O_p\left((\log n) n^{-3/5}\right) \right] \\ &= \frac{1}{n^2 b_0 b_1^2 b_2} \sum_{1 \le i \ne j \le n} K\left(\frac{x - X_i}{b_2}\right) K^{(1)}\left(\frac{e - e_i}{b_1}\right) K\left(\frac{X_i - X_j}{b_0}\right) f(X_i)^{-1} e_j \\ &\quad + \frac{K(0)}{n^2 b_0 b_1^2 b_2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b_2}\right) K^{(1)}\left(\frac{e - e_i}{b_1}\right) f(X_i)^{-1} e_i \\ &\quad - \frac{b_0^2 \sigma_k^2}{n b_1^2 b_2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b_2}\right) K^{(1)}\left(\frac{e - e_i}{b_1}\right) f(X_i)^{-1} f^{(1)}(X_i) m^{(1)}(X_i) \\ &\quad - \frac{b_0^2 \sigma_k^2}{n b_1^2 b_2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b_2}\right) K^{(1)}\left(\frac{e - e_i}{b_1}\right) f(X_i)^{-1} m^{(2)}(X_i) \\ &\quad + \frac{1}{n b_1^2 b_2} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b_2}\right) K^{(1)}\left(\frac{e - e_i}{b_1}\right) O_p\left((\log n) n^{-3/5}\right) \\ &= A_{1n} + A_{2n} + A_{3n} + A_{4n} + A_{5n} \end{aligned}$$

say. We now examine the four terms on the right-hand-side, in reverse

First, observe that

$$\begin{aligned} E\frac{1}{b_1^2 b_2}\left(K\left(\frac{x-X_i}{b_2}\right)K^{(1)}\left(\frac{e-e_i}{b_1}\right)\right) &= \frac{1}{b_1^2 b_2}\int\int K\left(\frac{x-u}{b_2}\right)K^{(1)}\left(\frac{e-v}{b_1}\right)g\left(v \mid u\right)f(u)dvdu \\ &= \frac{1}{b_1}\int\int K\left(u\right)K^{(1)}\left(v\right)g\left(e-b_1 v \mid x-b_2 u\right)f\left(x-b_2 u\right)dvdu \\ &= -\int K^{(1)}\left(v\right)vg^{(1)}\left(e \mid x\right)f(x)dv + O\left(b_1^2\right) + O\left(b_2^2\right) \\ &= g^{(1)}\left(e \mid x\right)f(x) + O\left(n^{-1/3}\right) \end{aligned}$$

Thus using Lemma 1

$$\begin{aligned} A_{5n} &= \left(g^{(1)}\left(e \mid x\right)f(x) + O\left(n^{-1/3}\right) + \frac{1}{b_1}O_p\left(\left(\frac{\log n}{b_1 b_2 n}\right)^{1/2}\right)\right)O_p\left(\left(\log n\right)n^{-3/5}\right) \\ &= O_p\left(\left(\log n\right)n^{-3/5}\right) \end{aligned}$$

Second,

$$\begin{aligned} & E\frac{b_0^2}{b_1^2 b_2}K\left(\frac{x-X_i}{b_2}\right)K^{(1)}\left(\frac{e-e_i}{b_1}\right)f(X_i)^{-1}f^{(1)}(X_i)m^{(1)}(X_i) \\ = & \frac{b_0^2}{b_1^2 b_2}\int\int K\left(\frac{x-u}{b_2}\right)K^{(1)}\left(\frac{e-v}{b_1}\right)f^{(1)}(u)m^{(1)}(u)g\left(v \mid u\right)dvdu \\ = & b_0^2 f^{(1)}(x)m^{(1)}(x)g^{(1)}\left(e \mid x\right) + O\left(n^{-2/3}\right) \\ = & O(n^{-2/5}) \end{aligned}$$

so by Lemma 1

$$A_{3n} = O(n^{-2/5}) + \frac{b_0^2}{b_1}O_p\left(\left(\frac{\log n}{b_1 b_2 n}\right)^{1/2}\right) = O_p(n^{-2/5})$$

Third, similarly,

$$E\frac{1}{b_0 b_1^2 b_2}K\left(\frac{x-X_i}{b_2}\right)K^{(1)}\left(\frac{e-e_i}{b_1}\right)f\left(X_i\right)^{-1}e_i = O\left(\frac{1}{b_0}\right)$$

Thus

$$EA_{2n} = O\left(\frac{1}{nb_0}\right) = O\left(n^{-5/6}\right)$$

and

$$A_{2n} = O\left(n^{-5/6}\right) + \frac{1}{nb_0 b_1}O_p\left(\left(\frac{\log n}{b_1 b_2 n}\right)^{1/2}\right) = O\left(n^{-5/6}\right).$$

Finally, we turn to A_{1n}. Note that $EA_{1n} = 0$. A tedious argument [to be completed] bounds $E(A_{1n}^2)$.

Together, we have $A_n = O_p(n^{-2/5})$ and hence

$$\begin{aligned}\hat{g}(e \mid x) - \hat{g}^*(e \mid x) &= \left(f(x)^{-1} + O_p\left(\delta_n^{-2} n^{-1/3}\right)\right) O_p(n^{-2/5}) \\ &= O_p((\log n)^{1/2} n^{-2/5})\end{aligned}$$

■

Proof of Theorem 1. By Lemma 3,

$$\begin{aligned}\hat{f}(y \mid x) &= \hat{g}(y - \hat{m}(x) \mid x) \\ &= \hat{g}^*(y - \hat{m}(x) \mid x) + O_p((\log n)^{1/2} n^{-2/5}).\end{aligned}$$

By a Taylor expansion

$$\begin{aligned}|\hat{g}^*(y - \hat{m}(x) \mid x) - \hat{g}^*(y - m(x) \mid x)| &\leq \sup_{e,x} \left|\frac{\partial}{\partial e}\hat{g}^*(e \mid x)\right| |\hat{m}(x) - m(x)| + O_p((\log n)^{1/2} n^{-2/5}) \\ &= O_p((\log n)^{1/2} n^{-2/5})\end{aligned}$$

Hence

$$\hat{f}(y \mid x) = \hat{g}^*(e \mid x) + O_p((\log n)^{1/2} n^{-2/5})$$

with $e = y - m(x)$ and therefore

$$\begin{aligned}n^{-2/6}\left(\hat{f}(y \mid x) - f(y \mid x)\right) &= n^{-2/6}\left(\hat{g}^*(e \mid x) - g(e \mid x) - \theta_2\right) + O_p((\log n)^{1/2} n^{2/6-2/5}) \\ &\to^d N\left(\theta_2, \sigma_2^2\right)\end{aligned}$$

as the asymptotic distribution of $\hat{g}^*$ is well known. ■

Figure 1
Conditional Density Estimates, y_{t-1}=0.2

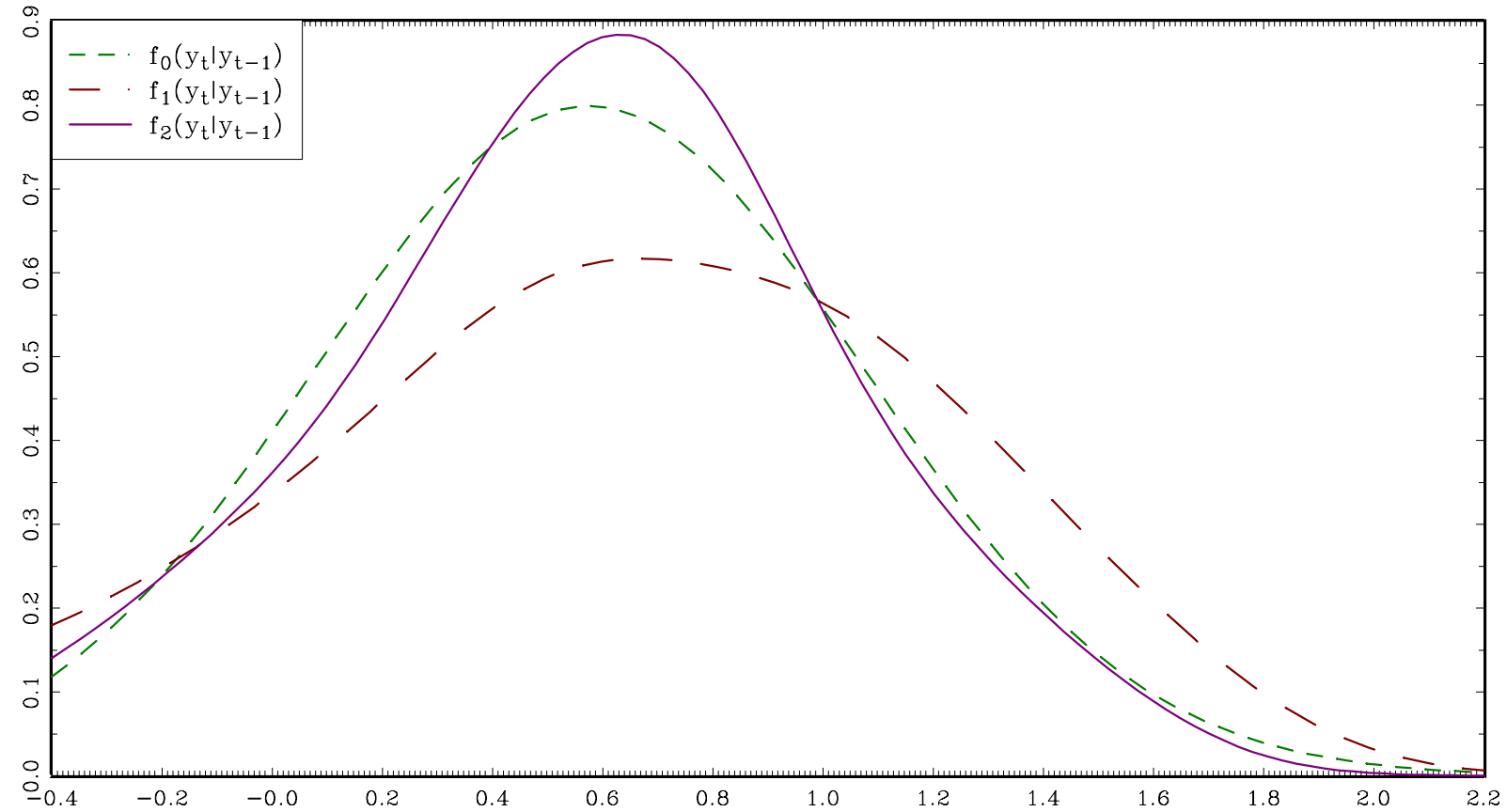

Figure 2
Conditional Density Estimates, y_{t-1}=0.6

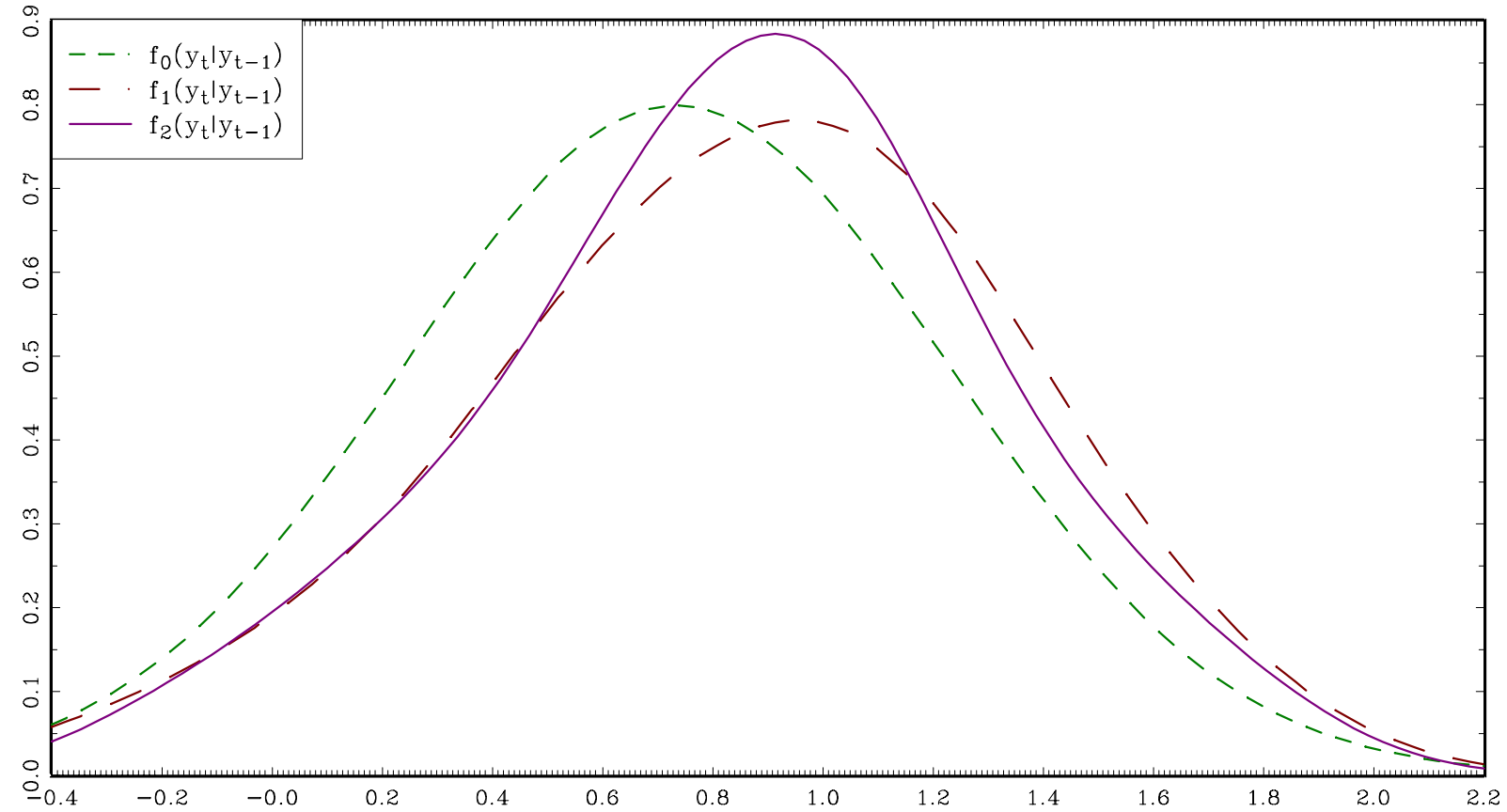

Figure 3
Conditional Density Estimates, y_{t-1}=1.0

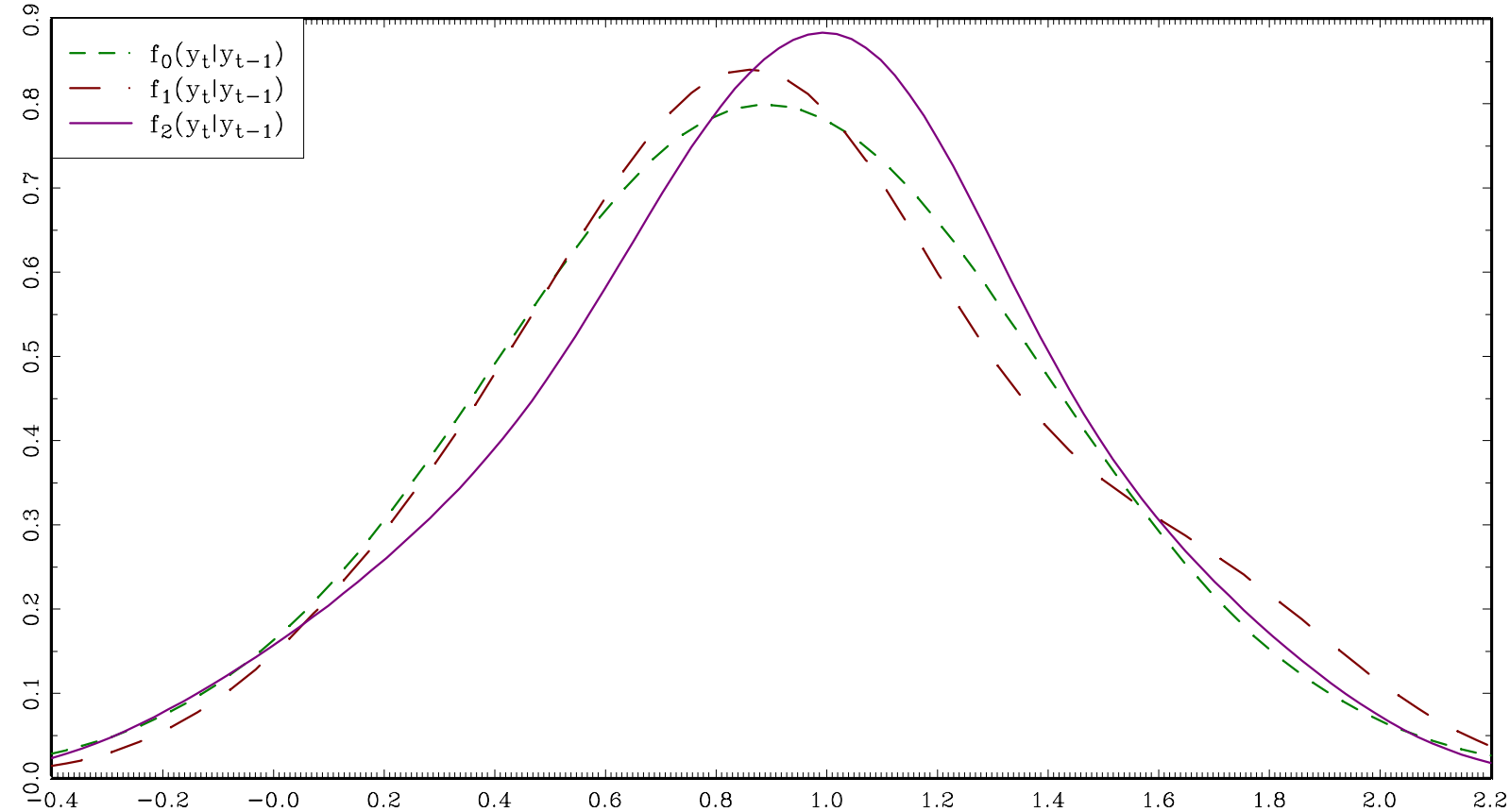

Figure 4
Conditional Density Estimates, y_{t-1}=1.4

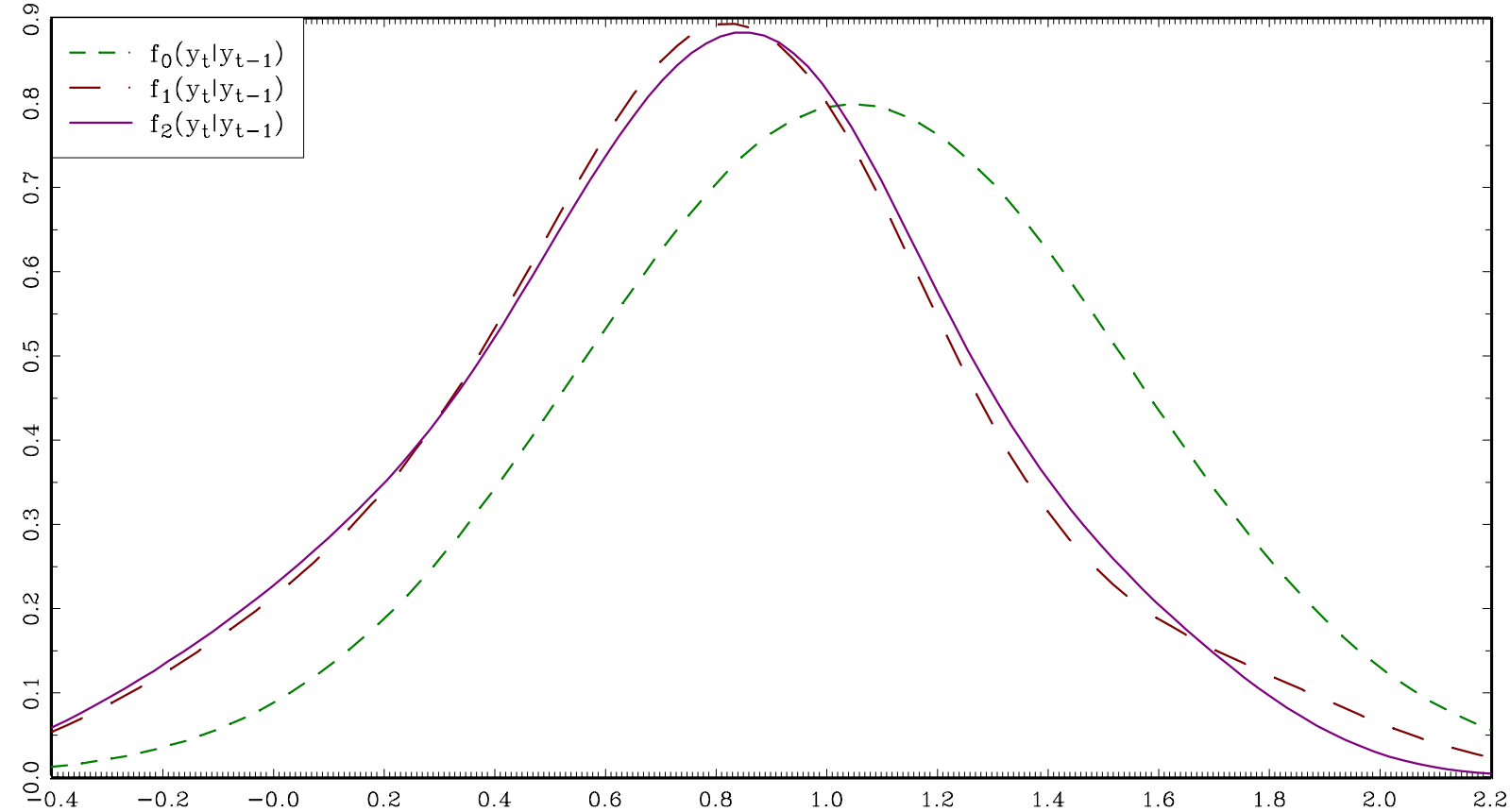

Figure 5
Log(Wage) Conditional Density, Age=25

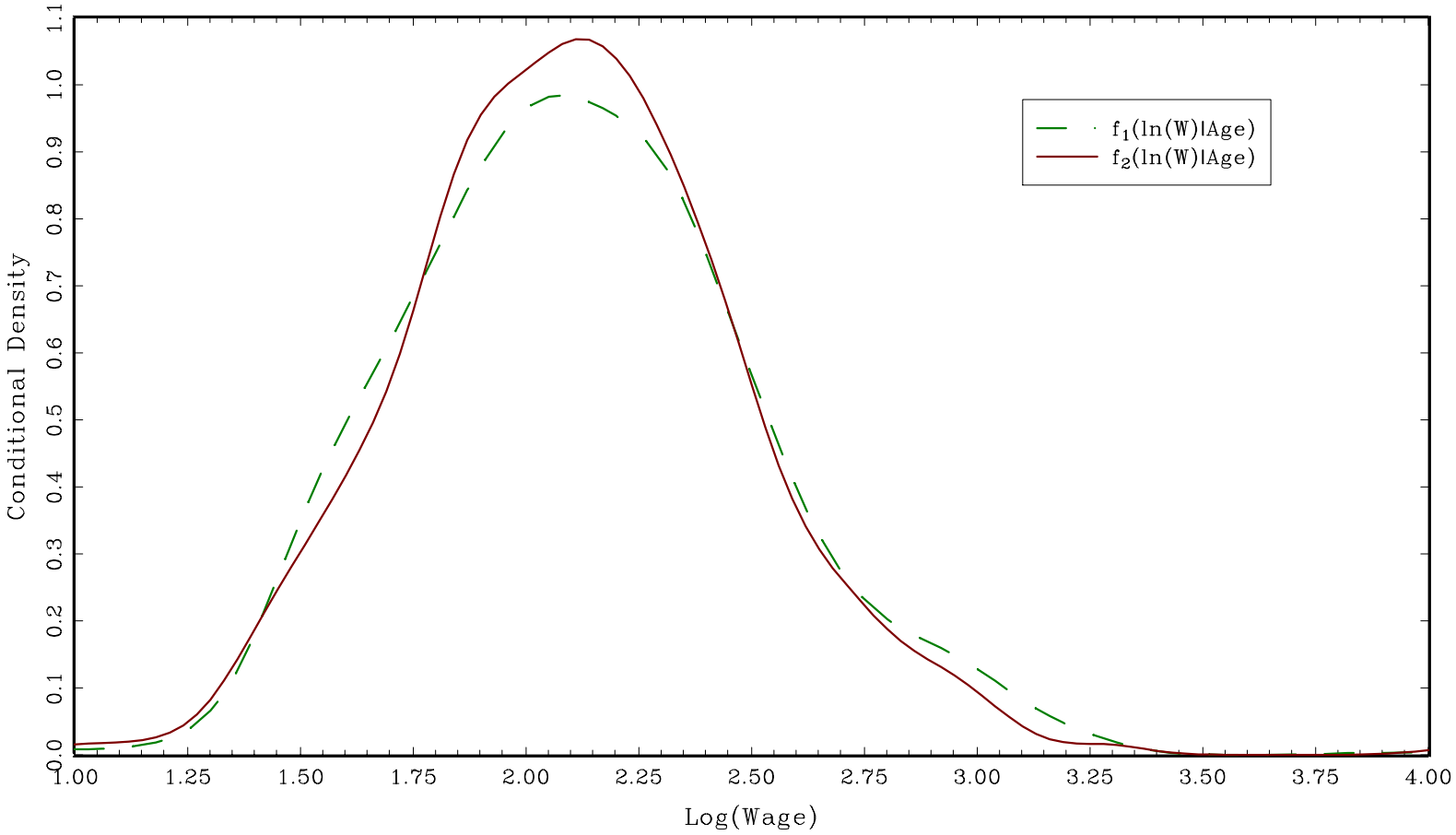

Figure 6
Log(Wage) Conditional Density, Age=35

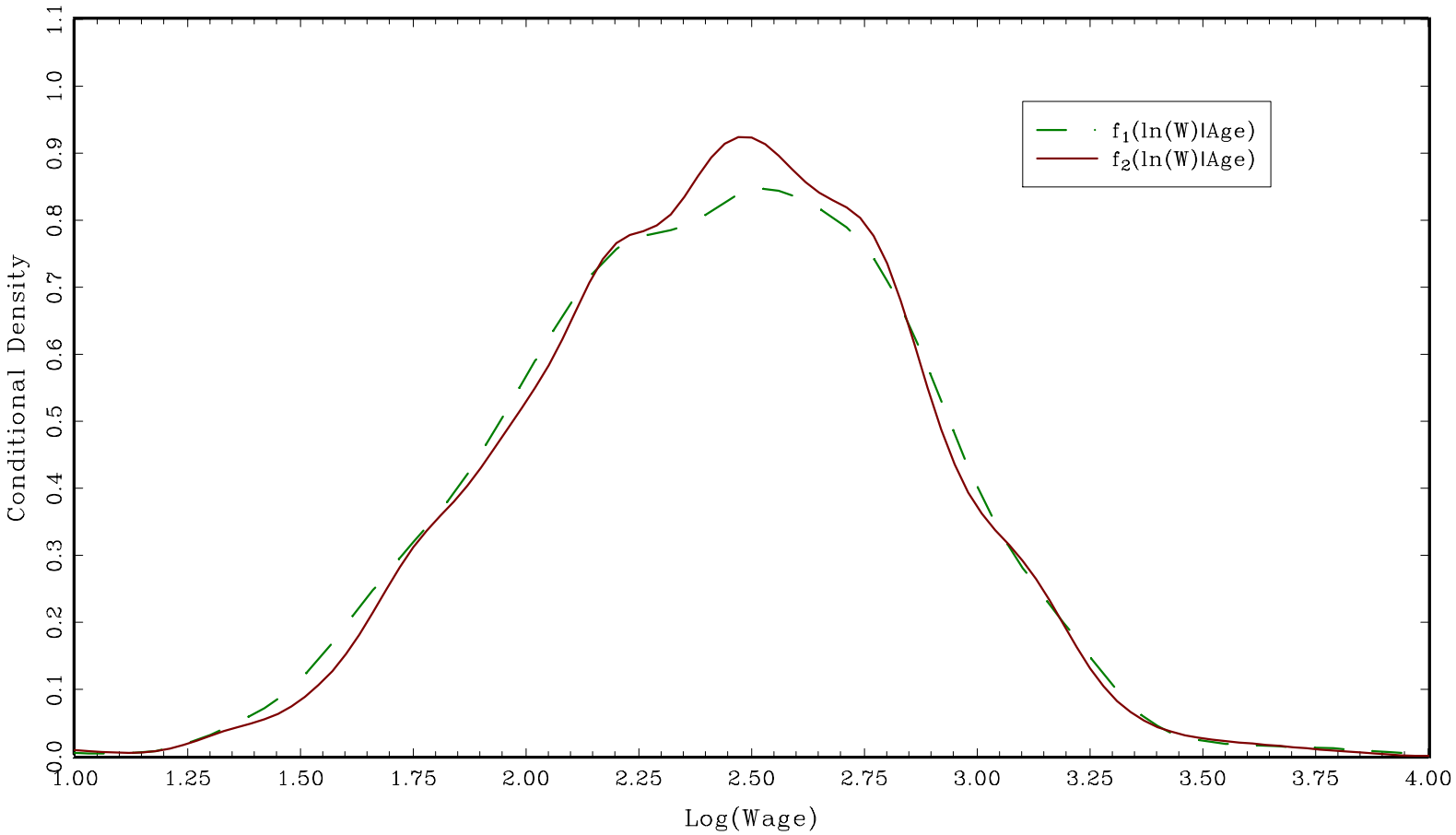

Figure 7
Log(Wage) Conditional Density, Age=45

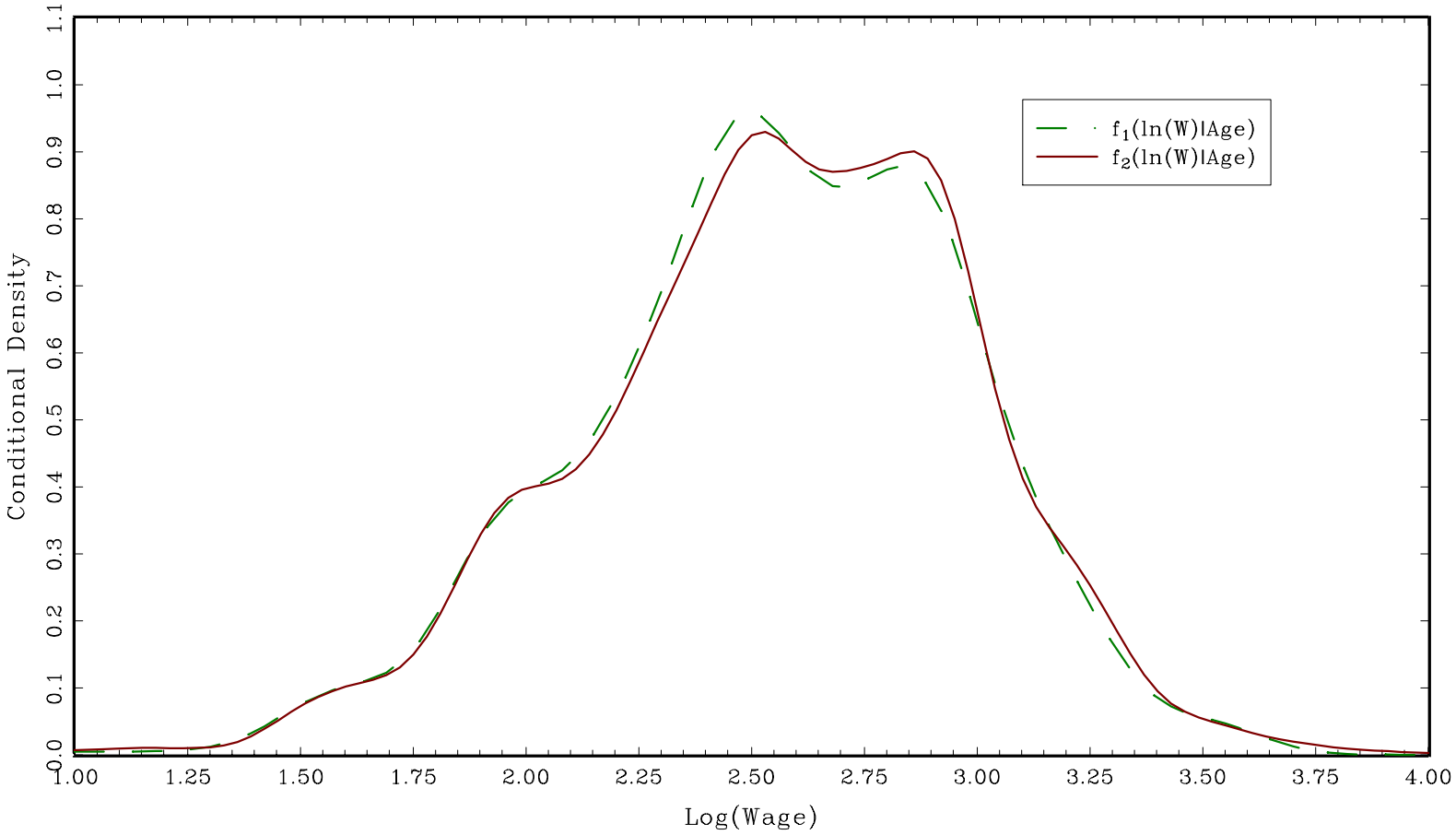

Figure 8
Log(Wage) Conditional Density, Age=55