Bayesian Semiparametric Density Estimation and Model Verification Using a Logistic–Gaussian Process

Peter J. LENK

This article proposes a semiparametric model, which consists of parametric and nonparametric components, for density estimation. The parametric component represents the researcher's a priori beliefs about a likely family of density functions. The nonparametric component, which is modeled by a logistic–Gaussian process, allows the predictive distribution to deviate from the parametric family if it is inadequate. Bayesian hypothesis testing is used to examine the adequacy of the parametric model relative to the flexible alternative provided by the semiparametric model. The article presents a Markov chain Monte Carlo algorithm that efficiently handles the large number of parameters.

Key Words: Bayesian inference; Density estimation; Model selection; Smoothing,

1. INTRODUCTION

Suppose that a researcher observes a random sample from an absolutely continuous density f with respect to a known, dominating measure G on the support $\mathcal{Y}$. He or she tentatively specifies an exponential family with density

$$f_0(y|\beta) = \frac{\exp\left[\vec{h}(y)'\beta\right]}{\int_{\mathcal{Y}} \exp\left[\vec{h}(x)'\beta\right] dG(x)} \quad \text{for} \quad y \in \mathcal{Y} \tag{1.1}$$

where $\vec{h}(y) = [h_1(y), \ldots, h_m(y)]'$ is a vector of m, nonconstant functions on $\mathcal{Y}$; and β is a m vector of unknown coefficients. This article assumes that the choice of parametric family, $\vec{h}$ and G, is based on theoretical or scientific considerations. The researcher has a prior distribution p_0 for β. After observing data, the predictive density, $E[f_0(y|\beta)|\text{Data}]$, is constrained by the choice of the parametric family and prior. Two questions of considerable

---

Peter J. Lenk is Associate Professor, The University of Michigan Business School, Ann Arbor, MI 48109–1234 (E-mail: plenk@umich.edu).

Journal of Computational and Graphical Statistics, Volume 12, Number 3, Pages 548–565
DOI: 10.1198/1061860032021

statistical interest arise. Is the parametric family adequate? If not, how should one coherently update the predictive density in light of this information?

This article proposes a flexible alternative model for Bayesian testing of the parametric model by embedding the exponential family in a semiparametric model (Leonard 1978; Lenk 1988):

$$f_\infty(y|\beta, Z) = \frac{\exp\left[\vec{h}(y)'\beta + Z(y)\right]}{\int_{\mathcal{Y}} \exp\left[\vec{h}(x)'\beta + Z(x)\right] dG(x)} \quad \text{for} \quad y \in \mathcal{Y}, \tag{1.2}$$

where Z is a zero mean, second-order Gaussian process with bounded, continuous covariance function: $E[Z(x), Z(y)] = \sigma(x, y)$, and $\int_{\mathcal{Y}} Z\, dG = 0$ almost surely to identify the model. The semiparametric model (1.2) is proportional to the product $f_0(y|\beta) f_\infty(y|0, Z)$ where the second factor is a logistic transform of Z or the "logistic–Gaussian" process. Allenby and Lenk (1994) used a logistic–normal model for polychotomous regression.

The variance of Z expresses the uncertainty about the parametric model. If the variance is close to zero, then the sample paths of $f_\infty(\bullet|0, Z)$ tend to be nearly constant, and the parametric density f_0 will dominate. If the variance is large, $f_\infty(\bullet|\beta, Z)$ tends to deviate substantially from f_0. The covariance of Z determines the smoothness of the density estimator (Whittle 1958).

Using the Karhunen–Loève Expansion (Grenander 1981), Z can be expressed as an infinite series with random coefficients

$$Z(y) = \sum_{k=1}^{\infty} \theta_k \phi_k(y) \quad \text{almost surely,}$$

where $\{\phi_k\}$ is an orthogonal basis of $\mathcal{L}_2(G, \mathcal{Y})$, the square integrable functions on $\mathcal{Y}$ with respect to G. This article will consider only real-valued basis functions. The random Fourier coefficients of the expansion are:

$$\theta_k = \int_{\mathcal{Y}} Z(y)\phi_k(y)\, dG(y) \quad \text{for} \quad k = 1, 2, \ldots.$$

The basis functions must satisfy $\int_{\mathcal{Y}} \phi_k\, dG = 0$ for the sample paths of Z to integrate to zero. That is, the basis elements are orthogonal to the constant function, which is excluded from the expansion. Because Z is a zero-mean, Gaussian process, $\{\theta_k\}$ are mutually independent, normally distributed with mean 0. Their standard deviations $\{\upsilon_k\}$ are determined by the expansion of the covariance function:

$$\sigma(x, y) = \sum_{k=1}^{\infty} \upsilon_k^2 \phi_k(x)\phi_k(y),$$

assuming that $\sum \upsilon_k^2 < \infty$.

This article uses the cosine basis (Kreider, Kuller, Ostberg, and Perkins 1966)

$$\phi_k(y) = \sqrt{2}\cos[k\pi G(y)] \quad \text{for} \quad y \in \mathcal{Y} \quad \text{for} \quad k = 1, 2, \ldots,$$

though other basis functions could be used. Piecewise continuous, nonperiodic functions have expansions with respect to this basis, while both sine and cosine functions would force the density to be periodic.

Cosine terms have a natural ordering that relates to the smoothness of the semiparametric density: smooth f_∞ will not have high frequency components. The rate of decay of the variances $\{v_k^2\}$ controls the smoothness of the sample paths of Z. If the Fourier coefficients for a function decays at rate $o(k^{-J})$ for some $J > 1$, then the function is J times differentiable almost everywhere (Katznelson 1976). Two, possible parameterizations of the variances as functions of k are

$$\text{var}(\theta_k|\tau,\xi) = v_k^2 = \tau^2 \exp(-d_k\xi) \quad \text{for} \quad \tau > 0 \quad \text{and} \quad \xi > 0, \tag{1.3}$$

where $d_k = k$ for the geometric smoother, and $d_k = \ln(k+1)$ for the algebraic smoother. The support for the geometric smoother is piecewise, analytic functions, while the algebraic smoother puts prior mass on piecewise-differentiable functions of order less than the integer part $\xi/2$. Abramovich, Sapatinas, and Silverman (1998) related the choice of the variance parameters for wavelet coefficients to the support of the model in Besov space, and Wahba (1978, 1983) used Gaussian processes on Sobolev spaces. Lenk (1988) provided moment conditions on Z and the smoothness of its sample paths.

In practice, the infinite series is truncated to a finite κ. For a given dataset, Fourier coefficients of frequencies above an upper bound are not identified and are aliased with lower frequency coefficients. Truncating the infinite series implies that the sample paths of the truncated process $Z_\kappa(y) = \sum_{k=1}^{\kappa} \theta_k \phi_k(y)$ are analytic. The article considers two models for the model order: κ fixed to a large number K, and κ random from 0 to K.

Substituting the finite series expansion into Equation (1.2) results in a model similar to Efron and Tibshirani (1996) and Stone, Hansen, Kooperbeg, and Truong (1997):

$$f_\kappa(y|\beta,\Theta_\kappa) \propto \exp\left[\vec{h}(y)'\beta + \vec{\phi}_\kappa(y)'\Theta_\kappa\right] \quad \text{for} \quad y \in \mathcal{Y}, \tag{1.4}$$

where $\vec{\phi}_\kappa(y) = [\phi_1(y),\ldots,\phi_\kappa(y)]'$, and $\Theta_\kappa = (\theta_1,\ldots,\theta_\kappa)'$. The article assumes that $\vec{\phi}_\kappa$ does not contain any of the components of $\vec{h}$.

The reduced, exponential model in Equation (1.1) corresponds to the researcher's "best guess" of the density before observing the data. The full, semiparametric model in Equation (1.4) allows the predictive density to deviate from the reduced model. If the exponential family is true, the semiparametric predictive density is similar to that of the parametric model, although it is less efficient. If the exponential family is inadequate, the semiparametric predictive density coherently adapts to the data. Bayesian hypothesis testing can be used to decide between the two models (Carlin and Chib 1995; DiCiccio, Kass, Raftery, and Wasserman 1997; Kass and Raftery 1995).

Stone (1990) detailed large-sample properties of the maximum likelihood estimator for log-spline models as the sample size and κ go to infinity. The main idea is that the sample size and κ must grow at the appropriate rates to obtain consistent estimates. If one includes too many basis functions for a given sample size, then one will obtain spurious estimates of higher order terms, and the resulting density estimator will follow the data too

closely or "under-smooths." On the other hand, if too few terms are included, the estimator over-smooths and will miss important features, for the support for the model is not large enough. Ghosal, Ghosh, and Van der Vaart (2000) gave rates of convergence of the posterior distributions for the log-spline model.

This article differs from maximum likelihood in that it uses a smoothing prior for the Fourier coefficients. Instead of selecting κ based on large sample behavior or other considerations, it retains "too many" coefficients but shrinks them toward zero. This approach can be viewed as a Bayesian version of penalized maximum likelihood (Good and Gaskins 1971, 1980; Silverman 1982) where the prior acts as the penalty term. In a sense, maximum likelihood procedures use an extreme form of shrinkage: no shrinkage up to κ, and shrinkage to zero after κ. The smoothing prior attenuates the effect of sampling variation of high-frequency terms. The predictive density, conditional on the smoothing parameters τ and ξ, is sensitive to their choice. They will be treated as unknown parameters, and the predictive density will be integrated over their posterior distributions.

The posterior analysis is accomplished by Markov chain Monte Carlo (MCMC). When κ is large, a random-walk Metropolis algorithm (see Hastings 1970; Chib and Greenberg 1995) tends to transverse the parameter space very slowly. Using a Laplace approximation of the posterior distribution that matches the covariance matrix of the proposal distribution to the Fisher's information speeds the algorithm and works well for moderate values of κ. However, computing and inverting the Hessian becomes numerically intensive and possibly unstable for large values of κ. Although the basis functions are orthogonal, the off-diagonal elements of the Fisher's information is nonzero in the density estimation problem.

A MCMC method that rapidly transverses the parameters spaces and does not require extensive computations at each iteration is described. First, the data are binned over a fine mesh, and logits for the binned data are generated by "slice sampling" (Damien, Wakefield, and Walker 1999). This procedure requires generating a univariate, truncated normal random deviation for each logit. Second, given the logits, the semiparametric regression method in Lenk (1999) is employed to estimate β, the Fourier coefficients, and their smoothing parameters. By transforming the problem from density estimation to linear regression, the algorithm exploits the orthogonal basis to generate candidate values for the Fourier coefficients from independent, univariate distributions in one block. Third, a Metropolis step is used to adjust for the discrete data. Lenk (1993) implemented a MCMC algorithm using rejection sampling for binned data and did not adjust for the discrete approximation. Koo and Kooperberg (2000) investigated a log-spline density estimator for binned data using maximum incomplete likelihood estimation. Finally, if the model order is random, it is generated with a reversible jump step (Green 1995).

The next section presents the prior distributions for the full and reduced models. Section 3 details the MCMC procedure. A rough outline of the steps are given in the Appendix. Section 4 reports a simulation experiment and two applications: one where the reduced model is adequate, and the other where the full model is needed. Section 5 concludes the article.

2. PRIOR DISTRIBUTIONS

Bracket notation (Gelfand and Smith 1990) will be used to indicate the density of random variables where the arguments define the random variable: $[X, Y]$ is the joint density of X and Y, and $[X|Y]$ is the conditional density of X given Y. By construction of the semiparametric model (1.2), the distribution of the kth Fourier coefficient is normal: $[\theta_k|\tau, \xi] = N[\theta_k|0, \tau^2 \exp(-d_k\xi)]$ for $d_k = k$ or $\ln(k+1)$. Define Ψ_κ to be a $\kappa \times \kappa$ diagonal matrix with (k, k) element equal to $\exp(-d_k\xi)$. Then $[\Theta_\kappa|\tau^2, \Psi_\kappa] = N_\kappa(\Theta_\kappa|0, \tau^2\Psi_\kappa)$, the κ-dimensional, normal distribution.

The prior distribution for $[\tau^2]$ is the inverse gamma density $\mathrm{IG}(\tau^2|r_0/2, s_0/2)$ with mean $s_0/(r_0-2)$ if $r_0 > 2$. The prior distribution for ξ is the gamma or exponential density $\mathrm{G}(\xi|1, q_0)$ with mean q_0^{-1}. The coefficients β have the same prior distributions under both models (1.1) and (1.4). In the empirical section, $[\beta] = N_m(\beta|\nu_0, \Upsilon_0)$, where the mean ν_0 is a vector of zeros, and the covariance Υ_0 is an identity matrix times a large constant. The normal priors are not essential to the analysis, and the MCMC algorithm accommodates nonnormal priors for β.

The article considers two models for κ the number of terms in the Fourier series representation: κ fixed to a large integer K, and κ random on the integers 0 to K. The prior distribution for random κ is uniform on the integers 0 to K. The difference between the two models can be seen from the unconditional prior variances of the Fourier coefficients. Provided that $r_0 > 2$, integrating Equation (1.3) over the smoothing priors gives:

$$\mathrm{var}(\theta_k) = \begin{cases} \left(\dfrac{s_0}{r_0-2}\right)\left(\dfrac{q_0}{d_k+q_0}\right) & \kappa = K \\ \left(\dfrac{s_0}{r_0-2}\right)\left(\dfrac{q_0}{d_k+q_0}\right)\left(\dfrac{K-k+1}{K+1}\right) & \kappa \text{ is random,} \end{cases} \tag{2.1}$$

where $d_k = k$ or $\ln(k+1)$. The variance under the random κ model is less than that for fixed κ and results in a more informative model. For large K, the variances of low-frequency Fourier coefficients are nearly identical.

The support of the model using the geometric smoother, $d_k = k$, is actually larger than it first appears. Conditional on the smoothing parameters τ and ξ, the Fourier coefficients decay exponentially in probability. Unconditionally, their variance decay at the rate k^{-1} in Equation (2.1). The unconditional model is substantially more flexible than one using fixed smoothing parameters, as demonstrated in Figure 1. Fifty and 5,000 observations were generated from a truncated normal distribution. The Fourier coefficients of the log-density decay at rate k^{-2}, which is not in the support of the model using geometric smoothing with fixed τ and ξ. Fifty basis functions were used, and the model did not have a parametric component. The plots of the coefficients in Figures 1(b) and 1(d) indicate the bias in the posterior means of the Fourier coefficients due to the smoothing priors. The amount of shrinkage decreases with increasing sample size. Figures 1(a) and 1(c) plot the posterior mean of f and the optimal kernel estimator, which is given in Equation (4.1). The unconditional model that mixes over smoothing parameters seems to be able to recover the true density, even though the true density does not belong to the support of the model for fixed τ

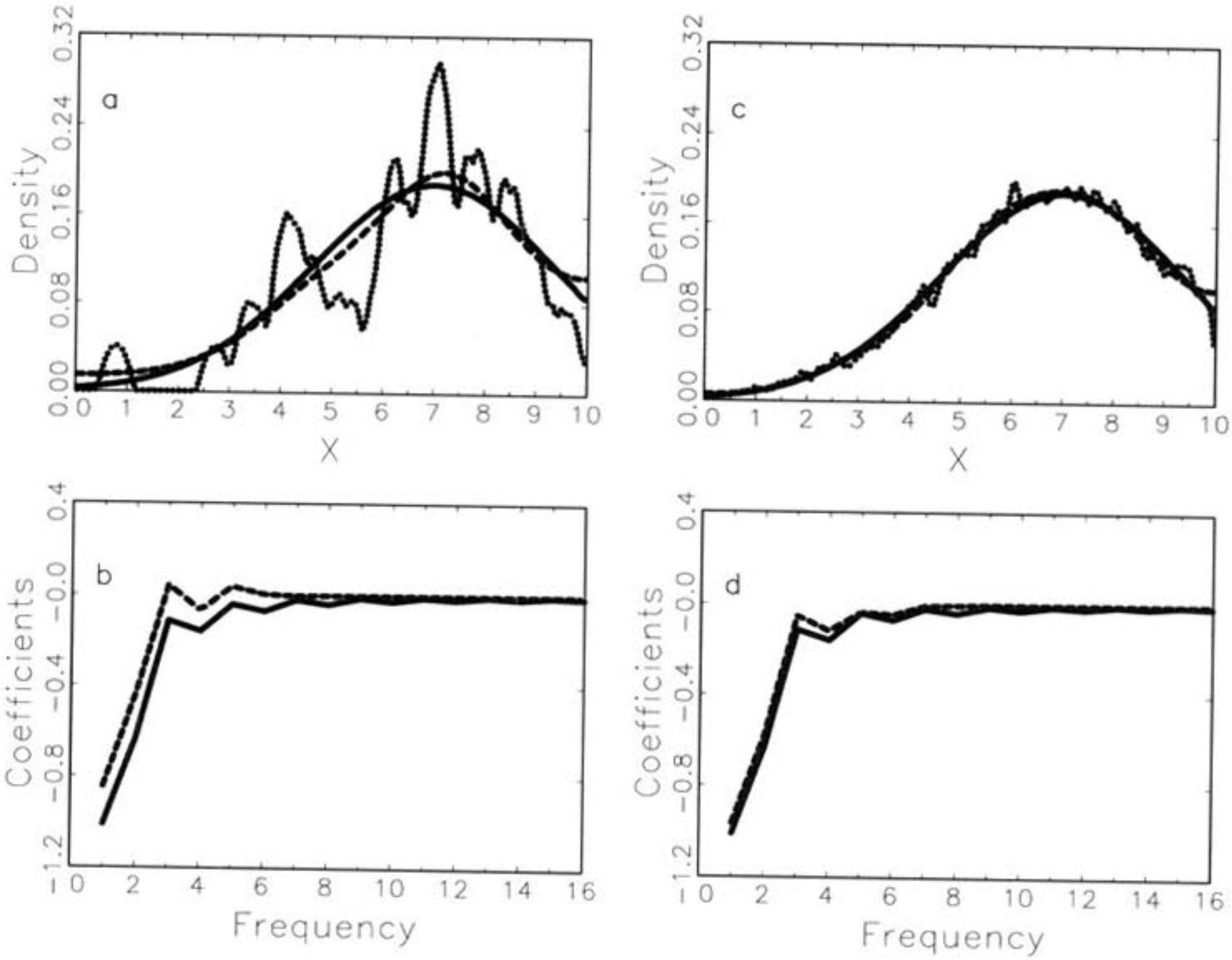

Figure 1. Truncated normal distribution with 50 observations in 1(a) and 1(b) and 5,000 observations in 1(c) and 1(d). Figures 1(a) and 1(c): solid line = true density; dashed line = posterior mean; dotted line = optimal kernel. Figures 1(b) and 1(d): solid line = true coefficients, and dashed line = posterior mean.

and ξ. In contrast, the conditional posterior analysis is very sensitive to the choice of τ and ξ if these are treated as known parameters. The posterior mean of f will not be consistent if κ is fixed at 50. However, not much is lost by truncating the Fourier series at 50 because the true density has low power at higher frequencies. Lenk (1991) provided error bounds for truncation.

3. BAYESIAN INFERENCE

If $y_1, \ldots, y_n$ are a random sample, then the log-likelihood for the semiparametric or full model (1.4) is:

$$L(\beta, \Theta_\kappa) = \sum_{i=1}^{n} \vec{h}(y_i)'\beta + \vec{\phi}_\kappa(y_i)'\Theta_\kappa - n \ln \left\{ \int_{\mathcal{Y}} \exp\left[\vec{h}(x)'\beta + \vec{\phi}_\kappa(x)'\Theta_\kappa\right] \, dG(x) \right\}.$$

The model with $\kappa = K$ will be described before considering the random κ case. The candidate distribution for (β, Θ_κ) is defined in four steps.

1. Compute a discrete version of the likelihood over a fine mesh on $\mathcal{Y}$.

2. Use "slice sampling" (Damien, Wakefield, and Walker 1999) to generate logits for the discrete problem.
3. Generate candidate values from the posterior distribution from a semiparametric regression model (Lenk 1999).
4. Use Metropolis to compensate for the discrete approximation. For random κ the algorithm adds a reversible jump step (Green 1995).

Appendix B provides a rough outline of the steps.

Without loss of generality, suppose that $\mathcal{Y} = (a, b)$ with $a < b$, and G is Lebesgue measure. If not, consider the transformed problem $G(Y)$. The discrete problem is defined as follows. Consider a grid of equally spaced points on (a, b):

$$x_j \quad = \quad a + \left(\frac{2j-1}{2}\right)\left(\frac{b-a}{J}\right) \quad \text{for} \quad j = 1, \ldots, J. \tag{3.1}$$

This grid was selected because the basis functions are orthogonal on this grid, which greatly simplifies the algorithm. The grid has a dual role: it is used in the MCMC, and it is used for plotting the density. The probability of the interval, centered at the grid point x_j, is

$$P\left(x_j - \frac{b-a}{J} < Y \leq x_j + \frac{b-a}{J}\right) = \frac{\exp(V_j)}{\sum_{k=1}^{J} \exp(V_k)},$$

where the logits $V = (V_1, \ldots, V_J)'$ are mutually independent and normally distributed with means $E(V_j) = \mu_j = \vec{h}(x_j)'\beta + \vec{\phi}_\kappa(x_j)'\Theta_\kappa$ and standard deviation σ. The design matrices for β and Θ_κ for the discrete problem are

$$H = \begin{bmatrix} \vec{h}(x_1)' \\ \vdots \\ \vec{h}(x_J)' \end{bmatrix} \quad \text{and} \quad \Phi_\kappa = \begin{bmatrix} \vec{\phi}_\kappa(x_1)' \\ \vdots \\ \vec{\phi}_\kappa(x_J)' \end{bmatrix}.$$

Then $[V|\beta, \Theta_\kappa, \sigma] = N_J(V|H\beta + \Phi_\kappa\Theta_\kappa, \sigma^2 I_J)$, where I_J is the $J \times J$ identity matrix. By construction of the grid, Φ_κ is orthogonal: $\Phi_\kappa'\Phi_\kappa = JI_\kappa$.

The generating distributions for β and Θ_κ are developed from the binned data as though they have a Bayesian model. In the discrete problem, the "prior" distribution for Θ_κ is the same as the original problem: $N_\kappa(\Theta_\kappa|0, \tau^2\Psi_\kappa)$; the "prior" distribution of β is $N_m(\beta|\nu_0, \Upsilon_0)$; and the "prior" distribution for σ^2 is $\text{IG}(\sigma^2|a0/2, b0/2)$. The log-likelihood for the logits is:

$$L_J(V) = \sum_{j=1}^{J} n_j V_j - n \ln\left[\sum_{k=1}^{J} \exp(V_k)\right],$$

where n_j is the number of observations in the interval $(x_j - \frac{b-a}{J}, x_j + \frac{b-a}{J}]$.

The full conditional distribution of V_j is

$$\begin{aligned} [V_j|\text{Rest}] \quad &\propto \quad \left[\exp(V_j) + S_j\right]^{-n} N(V_j|\mu_j + n_j\sigma^2, \sigma^2) \\ S_j \quad &= \quad \sum_{k \neq j} \exp(V_k), \end{aligned}$$

V_j is generated by "slice sampling," which introduces an auxiliary, uniform random deviate U_j such that the joint distribution on U_j and V_j is

$$[U_j, V_j] = \chi\left\{U_j < \left[\exp(V_j) + S_j\right]^{-n}\right\} N(V_j|\mu_j + n_j\sigma^2, \sigma^2),$$

where $\chi(\bullet)$ is the indicator function. Given the current value $V_j^{(i-1)}$ on iteration $i-1$ of the MCMC, $U_j^{(i)}$ is generated from a uniform distribution for 0 to $\left[\exp\left(V_j^{(i-1)}\right) + S_j\right]^{-n}$. Given $U_j^{(i)}$, the new value of V_j is generated from a univariate, truncated normal distribution:

$$\left[V_j|U_j^{(i)}, \text{Rest}\right] \propto N(V_j|\mu_j + n_j\sigma^2, \sigma^2)\chi\left\{V_j < \ln\left[\left(U_j^{(i)}\right)^{1/n} - S_j\right]\right\}. \tag{3.2}$$

The full conditional distribution of σ is

$$\begin{aligned}[\sigma^2|\text{ Rest }] &= \text{IG}\left(\sigma^2|\frac{a_J}{2}, \frac{b_J}{2}\right) \qquad (3.3)\\ a_J &= a_0 + J \quad \text{and} \quad b_J = b_0 + \sum_{j=1}^{J}(V_j - \mu_j)^2.\end{aligned}$$

The design matrices, H and Φ_κ may not be orthogonal, so candidate values of β and Θ_κ are generated in separate blocks. The candidate β^c is generated from a normal distribution:

$$\begin{aligned}[\beta^c|\text{ Rest }] &= N_m\left(\beta^c|\nu_J, \Upsilon_J\right) \qquad (3.4)\\ \Upsilon_J &= \left(H'H/\sigma^2 + \Upsilon_0^{-1}\right)^{-1}\\ \nu_J &= \Upsilon_J\left[H'(V - \Phi_\kappa\Theta_\kappa)/\sigma^2 + \Upsilon_0^{-1}\nu_0\right].\end{aligned}$$

Given β^c, the candidate Θ_κ^c is generated from a normal distribution:

$$\begin{aligned}[\Theta_\kappa^c|\text{ Rest }] &= N_\kappa\left(\Theta_\kappa^2|\zeta_J^c, \Omega_J\right) \qquad (3.5)\\ \Omega_J &= \left(JI_\kappa/\sigma^2 + \Psi_\kappa^{-1}/\tau^2\right)^{-1}\\ \zeta_J^c &= \Omega_J\Phi_\kappa'(V - H\beta^c)'/\sigma^2.\end{aligned}$$

Ω_J is a diagonal matrix, so Θ_κ^c consists of κ, univariate normal deviates, which greatly speeds the algorithm for large κ.

The candidates β^c and Θ_κ^c are accepted with log-probability

$$\begin{aligned}&\min\left[0, L(\beta^c, \Theta_\kappa^2) - L(\beta, \Theta_\kappa) + \ln\left([\beta^c]/[\beta]\right)\right. \qquad (3.6)\\ &\quad + \frac{1}{2}\left(\beta^c - \nu_J\right)'\Upsilon_J^{-1}\left(\beta^c - \nu_J\right) + \frac{1}{2}\left(\Theta_\kappa^c - \zeta_J^c\right)'\Omega_J^{-1}\left(\Theta_\kappa^c - \zeta_J^c\right)\\ &\quad \left. - \frac{1}{2}\left(\beta - \nu_J^c\right)'\Upsilon_J^{-1}\left(\beta - \nu_J^c\right) - \frac{1}{c}\left(\Theta_\kappa - \zeta_J\right)'\Omega_J^{-1}\left(\Theta_\kappa - \zeta_J\right)\right],\end{aligned}$$

where $[\beta]$ is the prior density for β; ν_J^c is ν_J with the current Θ_κ replaced by the candidate Θ_κ^c, and ζ_J is ζ_J^c with candidate β^c replaced by the current β. The log-likelihood for the original model does not depend on the logits $\{V_j\}$ or σ, which are accepted on each iteration of the MCMC.

Finally, the smoothing parameters τ and ξ are generated for the semiparametric model. The full conditional distribution of τ^2 is

$$[\tau^2|\text{Rest}] = \text{IG}(\tau^2|r_\kappa/2, s_\kappa/2)$$

where

$$r_\kappa = r_0 + \kappa$$

and

$$s_\kappa = s_0 + \sum_{k=1}^{\kappa} \theta_k^2 \exp(d_k \xi) \tag{3.7}$$

where $d_k = k$ or $\ln(k+1)$.

The full conditional for ξ is

$$[\xi|\text{Rest}] \propto \exp(q_\kappa \xi) \prod_{k=1}^{\kappa} \exp\left[-\frac{1}{2\tau^2}\theta_k^2 \exp(d_k\xi)\right], \tag{3.8}$$

where $q_\kappa = \kappa(\kappa+1)/4 - q_0$. ξ is generated with slice sampling, which introduces κ uniform random variables $\{U_k\}$ such that their joint distribution with ξ is

$$\begin{aligned}[U,\xi] &\propto \prod_{k=1}^{\kappa} \chi[0 \le U_k \le a_k] \exp(q_\kappa \xi) \\ a_k &= \exp\left[-\frac{\theta_k^2}{2\tau^2}\exp(d_k\xi)\right].\end{aligned}$$

The marginal density of ξ is proportional to (3.8). The condition distribution of U_k given ξ is uniform on $[0, a_k]$. The conditional distribution of ξ is:

$$\begin{aligned}[\xi|U] &\propto \prod_{k=1}^{K} \chi[\xi \le b_k] \exp(q_\kappa \xi) \\ &\propto \exp(q_\kappa \xi) \text{ for } 0 \le \xi \le b_{\min} = \min_{k=1\ldots\kappa}(b_k) \\ b_k &= \frac{1}{d_k} \ln\left[-\frac{2\tau^2}{\theta_k^2}\ln(U_k)\right],\end{aligned}$$

which has cumulative distribution function

$$F(\xi) = \frac{\exp(q_\kappa \xi) - 1}{\exp(q_\kappa b_{\min}) - 1} \text{ for } 0 < \xi < b_{\min}.$$

Inverting this cdf, one generates

$$\xi = q_\kappa^{-1} \ln\left\{1 + u[\exp(q_\kappa b_{\min}) - 1]\right\}, \tag{3.9}$$

where u is a standard uniform random deviate.

The above analysis assumed that κ was fixed to K. Two modifications are needed for the random κ case. For $\kappa < k \leq K$, generate θ_k from its prior: $N[\theta_k|0, \tau^2 \exp(-d_k\xi)]$. Second, random-walk Metropolis is used to generate κ. Given the current value of κ, a candidate κ^c is generated from:

$$P(\kappa^c|\kappa) \propto (1 + |\kappa^c - \kappa|)^{-1} \quad \text{for} \quad \max(0, \kappa - C) \leq \kappa^c \leq \min(K, \kappa + C) \quad (3.10)$$

for a fixed $C > 0$. A Metropolis step is used to accept κ^c and Θ_{κ^c}.

A modification of the model is to make $\vec{\phi}_\kappa$ orthogonal to $\vec{h}$ and the constant function by using Gramm–Schmidt orthogonalization. Unlike linear models, constructing $\vec{h}$ and $\vec{\phi}$ to be orthogonal does not imply that the Fisher's information is block-diagonal nor that the posterior distributions of β and Θ_κ are independent. The predictive densities with and without the orthogonalization were essentially the same, and the article will not report results for this modification.

A similar MCMC procedure is used to generate β from the reduced model (1.1). Usually, β is a low-dimensional parameter, and many Metropolis algorithms work very well. The full and reduced models are generated with independent chains. The marginal distributions for the two models are approximated using the procedure of Gelfand and Dey (1994). These marginal distributions can be used either to compute Bayes factors or the posterior probabilities of the full and reduced models. If one was using a 0/1 loss function for model choice, the posterior probabilities, along with relative costs, would determine the correct model. If one was using squared error loss for prediction, the posterior probabilities could be used to average the predictive densities from both models. Because the full model contains the reduced model, there is relatively little benefit in averaging their predictive densities. Model choice is more meaningful, especially when the reduced model was selected on theoretical or scientific grounds.

4. EXAMPLES

4.1 SIMULATION STUDY

A simulation study was performed to test the algorithm and the models' performance. The support is (a, b) with $a = 0$ and $b = 10$, and the dominating measure is Lebesgue. The parametric model has a truncated Gamma distribution on (a, b) with $h_1(y) = \ln(y - a)$ and $h_2(y) = y$. The true values for β were $(3, -1)'$.

One hundred and one grid points were used for computing the densities, and the total number of basis functions was $K = 98$. For each dataset, the MCMC had 24,000 iterations. The first 18,000 iterations formed the transition period. Afterwards, every sixth iteration was saved for the estimation for a total of 1,000 iterates in the posterior analysis. For the semiparametric model of random order, κ was generated every sixth iteration. The usual diagnostics indicated that the MCMC had stabilized well before 18,000 iterations.

A 2×2 factorial design was used for the simulation. The first condition was the number of observations: 50 and 500. The second condition was the generating distribution of the data: truncated Gamma distribution ($\kappa = 0$) or the semiparametric model in Equation (1.4) with κ random from 5 to 98; $\tau = 4$, and $\xi = 0.5$. Fifty datasets were generated for each treatment.

The simulation compares four density estimators: the predictive densities for the reduced model (1.1) and the full model (1.4) with fixed and random orders, along with the optimal kernel estimator (Silverman 1986, pp. 40–43). The kernel estimator is a leading alternative and is included to give an indication of the performance for the Bayesian models. The kernel estimator

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - y_i}{h}\right) \tag{4.1}$$

uses the Epanechnikov kernel

$$K(z) = \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right) \quad \text{for} \quad |z| < \sqrt{5},$$

which has maximal efficiency. The optimal bandwidth h is used:

$$h_{\text{opt}} = \left[\int_{-\sqrt{5}}^{\sqrt{5}} z^2 K(z)dz\right]^{-\frac{2}{5}} \left[\int_{-\sqrt{5}}^{\sqrt{5}} K^2(z)dz\right]^{\frac{1}{5}} \left[\int_a^b f''(z)^2 dz\right]^{-\frac{1}{5}} n^{-\frac{1}{5}},$$

which requires knowledge about the mean curvature of the true density. The optimal bandwidth minimizes the asymptotic root mean integrated squared error between the kernel estimator and the true density.

Figure 2 graphs four of the 50 datasets when the true density is generated from the semiparametric model of random order, and the sample size is 500. The plots indicate that the estimators and true densities are close to each other. Table 1 reports the RISE for the estimators in Figure 2. The true density in Figure 2(a) is fairly typical: multiple modes that are asymmetric. Figures 2(b) to 2(d) are extreme points from the simulation. In Figure 2(b) the RISE for the semiparametric model was minimum among the 50 datasets. In Figure 2(c), the kernel estimator performed the worse in the 50 datasets relative to the semiparametric estimators: it has trouble recovering the mode at the endpoint, which is a well-known problem with kernel estimators. In Figure 2(d), the kernel estimator performed the best in the 50 datasets relative to the semiparametric estimators: it was slightly closer to the large mode. All three estimators missed the small bump at 1.25.

Table 2 records the means and standard deviations across the 50 datasets of the posterior mean and posterior standard deviations of the logit error standard deviation σ from the binned approximation and β. Only results for the semiparametric model of random order using the geometric smoother are tabulated; the results for the semiparametric model of fixed order or with the algebraic smoother are very similar. The logit error standard deviations σ contain information about how well the model fits the logits in the discrete problem. When the data are from the truncated Gamma distribution, the estimates of σ for the reduced

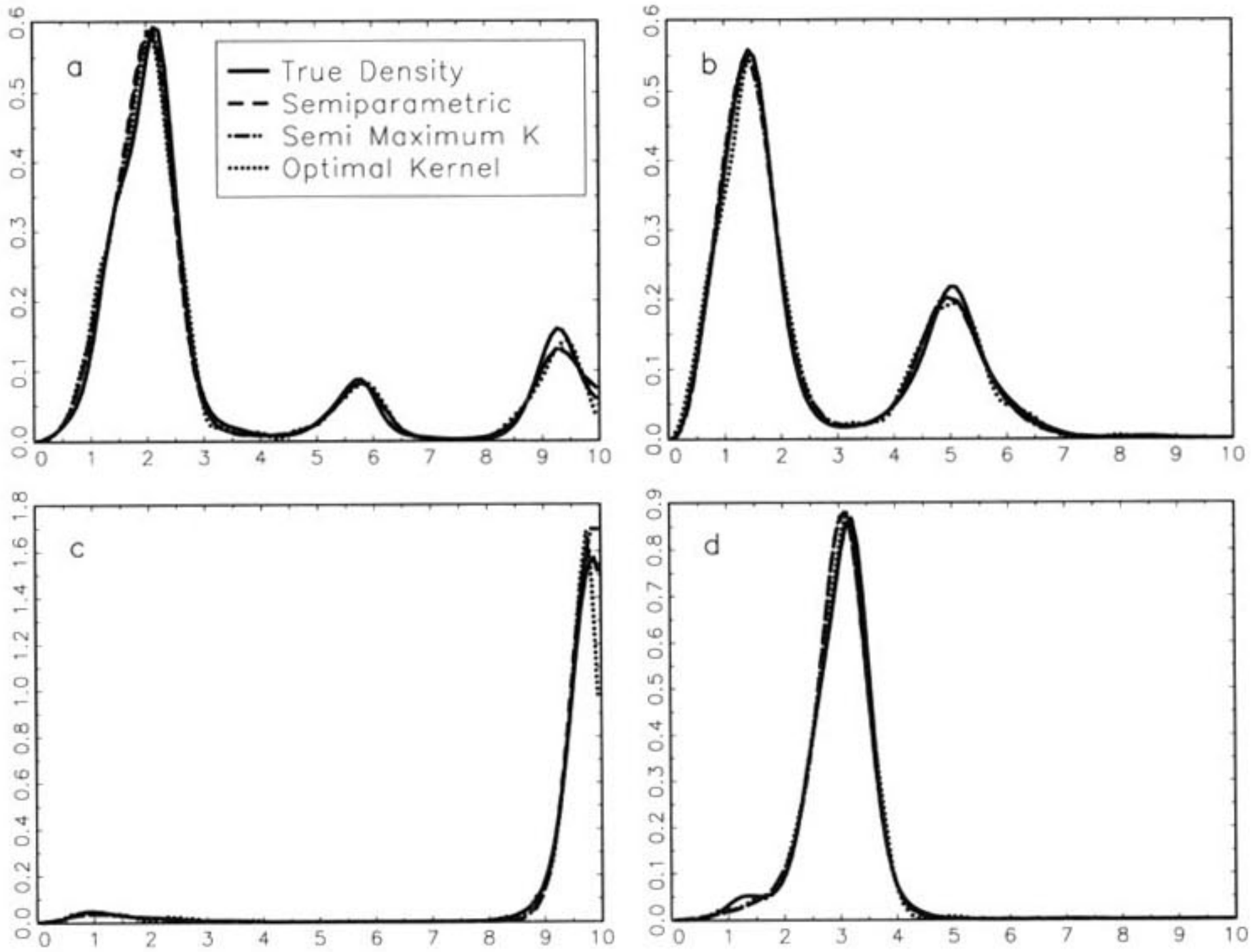

Figure 2. Selected simulations from semiparametric model with 500 observations.

and full models are nearly the same. When the data are from the semiparametric model, the posterior means of σ from the full model are smaller than those from the reduced model. When the true density is the truncated Gamma distribution, both the reduced and full models recover the true β, though the reduced model is more efficient with smaller posterior standard deviations. When the true model is generated from the semiparametric model, the estimates of β are inaccurate. Making the basis functions orthogonal to $\vec{h}$ did not improve their accuracy.

Table 3 reports the natural logarithm of the marginal distribution of the data. When the data are generated from the truncated Gamma distribution, the marginal distributions of the reduced model are larger than that of the full models in 92% to 98% of the simulations. The situation is reversed when the data are generated from the semiparametric model with the full

Table 1. RISE for the Estimators in Figure 1

		Semiparametric	
Figure	*Optimal kernel*	*Random order*	*Fixed order*
1.a	0.0593	0.0563	0.0515
1.b	0.0436	0.0209	0.0267
1.c	0.1361	0.0583	0.0686
1.d	0.0462	0.0742	0.0766

Table 2. Estimates of Parameters for Parametric Component in Simulation Study

		Parametric model				*Semiparametric model*			
		Posterior mean		*Posterior STD*		*Posterior mean*		*Posterior STD*	
	True	*Mean*†	*STD*‡	*Mean*†	*STD*‡	*Mean*†	*STD*‡	*Mean*†	*STD*‡
50 observations from truncated gamma distribution									
Logit Error STD		0.733	0.052	0.143	0.016	0.738	0.052	0.146	0.016
β_1	3	2.730	0.515	0.549	0.200	2.442	0.389	0.731	0.184
β_2	−1	−0.949	0.200	0.169	0.013	−0.873	0.184	0.261	0.047
500 observations from truncated gamma distribution									
Logit Error STD		0.420	0.016	0.047	0.002	0.425	0.016	0.048	0.002
β_1	3	2.999	0.242	0.237	0.071	2.828	0.312	0.422	0.102
β_2	−1	−1.005	0.071	0.067	0.003	−0.965	0.102	0.167	0.031
50 observations from full model									
Logit Error STD		1.265	0.735	0.268	0.125	0.732	0.054	0.151	0.018
β_1	3	2.899	2.085	0.551	0.925	1.334	0.705	1.079	0.710
β_2	−1	−0.946	0.925	0.173	0.055	−0.619	0.710	0.470	0.159
500 observations from full model									
Logit Error STD		5.170	2.292	0.404	0.147	0.564	0.432	0.105	0.225
β_1	3	4.922	3.943	0.202	1.365	1.778	1.009	0.831	0.775
β_2	−1	−1.499	1.365	0.059	0.058	−0.724	0.775	0.317	0.130

† Mean over 50 simulations
‡ Standard Deviation over 50 simulations

model having larger marginal distribution in 82% to 100% of the simulations. The relatively large standard deviations across simulations reflects the wide variety of distributions in the support of the semiparametric model and not the accuracy in estimating the log marginal distribution for a fixed density. These marginal distributions can be used to compute Bayes factors and posterior probabilities of the models in order to perform formal, Bayesian hypothesis testing or mixing the predictive densities.

Table 4 reports the root integrated squared errors (RISE) between the true density and the estimators. The RISE is smaller for the reduced model when the data are generated from the truncated Gamma distribution, and it is smaller for the kernel estimators and full models when the data are generated from the full model. The RISE of the semiparametric models was smaller than that of the optimal kernel estimator in 58% to 92% of the simulations. The RISE for the semiparametric models of random and fixed orders are not significantly different.

In the simulations, the semiparametric model of random order did not perform uniformly better than the model with fixed order, and the procedure tended to overestimate the true model order κ_0. It appears that the likelihood function is relatively noninformative about κ_0. The smoothing prior shrinks θ_k to zero for $k > \kappa_0$. Consequently, the likelihood and predictive density when $\kappa > \kappa_0$ are not much different than those when $\kappa = \kappa_0$. Also, the random κ model puts positive prior probability on all K coefficients. This points to the advantage of running parallel chains for the reduced ($\kappa = 0$) and full ($\kappa \geq 0$) models. If

Table 3. Log Marginal Distributions for Simulation Study

	Truncated gamma simulation			*Semiparametric simulation*		
	Mean	*STD*	*# simulations larger than parametric*	*Mean*	*STD*	*# simulations larger than parametric*
			50 observations			
Parametric	−99.29	5.48		−86.08	20.83	
SemiPar1†	−101.95	5.83	1	−66.37	24.75	41
SemiPar2‡	−100.97	5.43	4	−63.47	25.00	44
			500 observations			
Parametric	−986.26	14.35		−826.34	228.45	
SemiPar1†	−992.17	14.48	1	−598.93	215.47	50
SemiPar2‡	−991.46	14.77	1	−599.32	216.30	50

† Semiparametric model with random order κ
‡ Semiparametric model with fixed order κ

the reduced model is true, Bayesian hypothesis testing penalizes the full model for having a larger parameter space and tends to select the reduced model.

4.2 APPLICATIONS

The first application uses 86 times in days of treatment spells of control patients in a suicide study (Silverman 1986, p. 8, Table 2.1). The reduced model is truncated Gamma distribution on $(0, 500]$. Figure 3(a) graphs the predictive densities of the reduced model and full model of random order, along with the posterior standard deviation from the full model. The predictive densities for the two models are nearly identical. The Bayes factor in favor of the reduced model is 14,765, which clearly indicates that the reduced model is preferred. The Gamma distribution seems to describe adequately treatment spells.

The second application uses 107 eruption lengths in minutes of Old Faithful geyser (Silverman 1986, p. 8, Table 2.2). The reduced model is the truncated Gamma distribution on $(0, 8]$. The predictive density of the full model in Figure 3(b) has three modes and clearly indicates that the Gamma density is inadequate. The Bayes factor in favor of the full model is 5.09E24, which overwhelmingly favors the full model.

5. CONCLUSION

This article presents Bayesian inference of a semiparametric model that is proportional to the product of parametric and nonparametric components. The parametric component corresponds to the researcher's a priori beliefs about a likely family of densities before observing data. This article assumes that the researcher is interested only in one parametric family, though the model could be easily extended to the situation where the researcher is tentatively entertaining more than one parametric model. The nonparametric component,

Table 4. Root Mean Squared Errors between True Density and Estimates in Simulation Study

	Truncated gamma simulation			*Semiparametric simulation*		
	Mean	*STD*	*# simulations smaller than optimal kernel*	*Mean*	*STD*	*# simulations smaller than optimal kernel*
			50 observations			
Optimal Kernel	0.076	0.027		0.163	0.059	
Parametric	0.051	0.028	40	0.440	0.165	1
SemiPar1†	0.064	0.027	37	0.163	0.061	30
SemiPar2‡	0.073	0.033	33	0.162	0.062	29
			500 observations			
Optimal Kernel	0.033	0.010		0.068	0.020	
Parametric	0.015	0.008	50	0.404	0.144	0
SemiPar1†	0.023	0.010	46	0.057	0.017	41
SemiPar2‡	0.023	0.012	45	0.060	0.018	39

† Semiparametric model with random order κ

‡ Semiparametric model with fixed order κ

which is modeled by a logistic–Gaussian process, allows the predictive density to deviate from parametric family if the parametric family is inadequate. A hierarchical Bayes, smoothing prior provides a data-driven method for determining the appropriate degree of smoothing. Bayesian hypotheses testing verifies the adequacy of the parametric family relative to the flexible form of the semiparametric model.

The model uses a random Fourier series representation of the Gaussian process for the nonparametric component. Other basis, such a orthogonal polynomials, splines or wavelets, could be used. An important feature of the cosine basis is that it has a natural ordering with respect to the smoothness of the density: smooth functions have low power at high frequencies. This fact is exploited in constructing the smoothing prior for the Fourier coefficients.

The article considers both models with fixed and random number of basis functions. The predictive densities for the two cases are very similar. The smoothing prior seems to mitigate the importance of correctly identifying the correct model order. Also, the cosine basis is a global basis. Selecting the correct set of bases functions may be more important with a local basis that is trying to detect special features, such as discontinuous derivatives, in the data. With local basis, not only the degree of variation but also the locations of the basis functions matters (Abramovich, Sapatinas, and Silverman 1998; Denison, Mallick, and Smith 1998).

The article assumes that the choice of parametric family is driven by the researcher's substantive knowledge of the phenomenon. An open question is the sensitivity of the predictive distribution to this choice when the researcher does not have substantive knowledge to guide its selection.

The article focuses exclusively on the univariate case. Conceptually, the multivariate case is a straightforward generalization by using multivariate Fourier series expansions. However, the number of Fourier coefficients increases exponentially in the number of dimen-

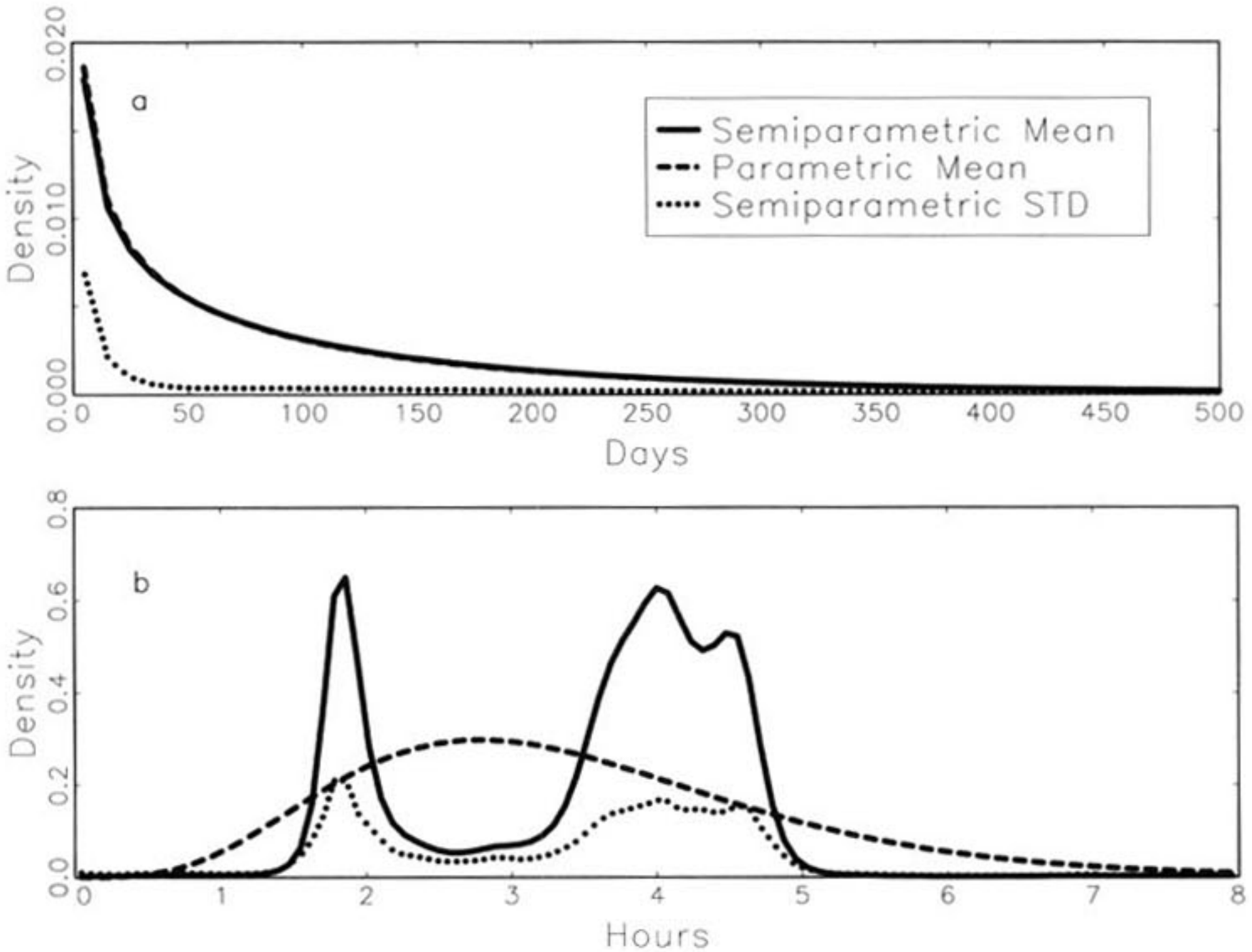

Figure 3. Examples: (a) Suicide data and (b) Old Faithful eruptions. Solid line = semiparametric model; dashed line = parametric model; dotted line = posterior standard deviation from the semiparametric model.

sions, and the computational effort renders this approach unwieldy for higher dimensional problems. An outstanding research question is to adapt the model for high-dimensional problems in such a way that it avoids the computational load of the general, multivariate Fourier series yet has sufficiently rich structure to describe nontrivial multivariate behavior. Huang, Kooperberg, Stone, and Truong (2000) and Wahba et al. (1995) proposed ANOVA decompositions of the multidimensional functions, and Kooperberg and Stone (1999) proposed a neural network model for adaptively selected linear splines of Kooperberg, Bose, and Stone (1997). Adaptations of these methods may prove useful.

APPENDIX: OUTLINE OF MCMC ALGORITHM

1. Set the maximum number of basis elements K and the mesh size J to large numbers.
2. Establish grid on $\{x_j\}$ (Equation (3.1)) on the support $\mathcal{Y} = (a, b)$. Bin the observations on grid. Evaluate $\vec{h}$ and $\vec{\phi}_K$ at the data $\{y_i\}$ and the grid $\{x_j\}$.
3. Initialize prior parameters and parameters for MCMC.
4. Loop over MCMC iterations.

(a) Generate the logits $\{V_j\}$ from truncated, univariate normal distributions (Equation (3.2)).
(b) Generate the error variance σ^2 of the logits from an inverted Gamma distribution (Equation (3.3)).
(c) Generate candidates for β (Equation (3.4)) and θ_κ (Equation (3.5)) from normal distributions. Accept these candidates with a Metropolis step with probability in Equation (3.6).
(d) Generate τ^2 from an inverted Gamma distribution (Equation (3.7)).
(e) Generate ξ using slice sampling (Equation (3.9)).
(f) If κ is random, generate a candidate value from a random walk on the integers 0 to K (Equation (3.10)). Accept this candidate with a Metropolis step.

5. Use the last M iterations to approximate posterior quantities.

[Received March 2000. Revised May 2002.]

REFERENCES

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998), "Wavelet Thresholding via a Bayesian Approach," *Journal of the Royal Statistical Society*, Series B, 60, 725–749.

Allenby, G., and Lenk, P. (1994), "Modeling Household Purchase Behavior with Logistic Normal Regression," *Journal of the American Statistical Association*, 83, 1218–1231.

Carlin, B. P., and Chib, S. (1995), "Bayesian Model Choice via Markov-chain Monte-Carlo Methods," *Journal of the Royal Statistical Society*, Series B, 57, 473–484.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis–Hastings Algorithm," *The American Statistician*, 49, 327–335.

Damien, P., Wakefield, J. C., and Walker, S. (1999), "Gibbs Sampling for Bayesian Nonconjugate and Hierarchical Models Using Auxiliary Variables," *Journal of the Royal Statistical Society*, Series B, 61, 331–344.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian Curve Fitting," *Journal of the Royal Statistical Society*, Series B, 60, 333–350.

DiCiccio, T. J., Kass, R. E., Raftery, A., and Wasserman, L. (1997), "Computing Bayes Factors by Combining Simulation and Asymptotic Approximations," *Journal of the American Statistical Association*, 92, 903–915.

Efron, B., and Tibshirani, R. (1996), "Using Specially Designed Exponential Families for Density Estimation," *The Annals of Statistics*, 24, 2431–2461.

Gelfand, A. E., and Dey, D. K. (1994), "Bayesian Model Choice: Asymptotics and Exact Calculations," *Journal of the Royal Statistical Society*, Series B, 56, 501–514.

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Ghosal, S., Ghosh, J. K., and Van der Vaart, A. W. (2000), "Convergence Rate of Posterior Distributions," *The Annals of Statistics*, 28, 500–531.

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732.

Grenander, U. (1981), *Abstract Inference*, New York: Wiley.

Good, I. J., and Gaskin, R. A. (1971), "Nonparametric Roughness Penalties for Probability Densities," *Biometrika*, 58, 255–277.

——— (1980), "Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data," *Journal of the American Statistical Association*, 75, 42–73.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Huang, J. Z., Kooperberg, C., Stone, C. J., and Truong, Y. K. (2000), "Functional ANOVA Modeling for Proportional Hazards Regression," *The Annals of Statistics*, 28, 961–999.

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Katznelson, Y. (1976), *An Introduction to Harmonic Analysis*, New York: Dover.

Kreider, D. L., Kuller, R. G., Ostberg, D. R., and Perkins, F. W. (1966), *An Introduction to Linear Analysis*, Reading, MA: Addison-Wesley.

Koo, J. Y., and Kooperberg, C. (2000), "Logspline Density Estimation for Binned Data," *Statistics and Probability Letter*, 46, 133–147.

Kooperberg, C., Bose, S., and Stone, C. J. (1997), "Polychotomous Regression," *Journal of the American Statistical Association*, 92, 117–127.

Kooperberg, C., and Stone, C. J. (1999), "Stochastic Optimization Methods for Fitting PLOYCLASS and Feed-Forward Neural Network Models," *Journal of Computational and Graphical Statistics*, 8, 169–189

Lenk, P. (1988), "The Logistic Normal Distribution for Bayesian, Nonparametric, Predictive Densities," *Journal of the American Statistical Association*, 83, 509–516.

——— (1991), "Towards a Practicable Bayesian Nonparametric Density Estimator," *Biometrika*, 78, 531–543.

——— (1993), "A Bayesian Nonparametric Density Estimator," *Journal of Nonparametric Statistics*, 3, 53–69.

——— (1999), "Bayesian Inference for Semiparametric Regression using a Fourier Representation," *Journal of the Royal Statistical Society*, Series B, 61, 863–879.

Leonard, T. (1978), "Density Estimation, Stochastic Processes, and Prior Information," *Journal of the Royal Statistical Society*, Series B, 40, 113–146.

Silverman, B. W. (1982), "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method," *The Annals of Statistics*, 10, 795–810.

——— (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Stone, C. J. (1990), "Large-Sample Inference for Log-Spline Models," *The Annals of Statistics*, 18, 717–741.

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y., K. (1997), "Polynomial Splines and their Tensor Products in Extended Linear Models," *The Annals of Statistics*, 25, 1371–1470.

Wahba, G. (1978), "Improper Priors, Spline Smoothing, and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society*, Series B, 364–372.

——— (1983), "Bayesian 'Confidence Intervals' for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society*, Series B, 45, 133–150.

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995), "Smoothing Spline ANOVA for Exponential Families, with Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy," *The Annals of Statistics*, 23, 1865–1895.

Whittle, P. (1958), "On the Smoothing of Probability Density Functions," *Journal of the Royal Statistical Society*, Series B, 20, 334–343.