Bayesian approaches to non- and semiparametric density estimation*

Nils Lid Hjort, University of Oslo

ABSTRACT. This paper proposes and discusses several Bayesian attempts at nonparametric and semiparametric density estimation. The main categories of these ideas are as follows: (1) Build a nonparametric prior around a given parametric model. We look at cases where the nonparametric part of the construction is a Dirichlet process or relatives thereof. (2) Express the density as an additive expansion of orthogonal basis functions, and place priors on the coefficients. Here attention is given to a certain robust Hermite expansion around the normal distribution. Multiplicative expansions are also considered. (3) Express the unknown density as locally being of a certain parametric form, then construct suitable local likelihood functions to express information content, and place local priors on the local parameters.

KEY WORDS: *Bayesian density estimation; Dirichlet process; Hermite expansions; local likelihood; log-linear expansions; semiparametric estimation; smoothed Dirichlet priors*

1. Introduction and summary. Lindley (1972) noted in his review of general Bayesian methodology that Bayesians up to then had been 'embarrassingly silent' in the area of nonparametric statistics. Bayesian nonparametrics has enjoyed healthy progress since then, but the sub-field of curve and surface estimation, nonparametric regression, semiparametric estimation problems, density and hazard rate estimation, and statistical pattern recognition, seems as yet to be not fully developed. This contrasts with the rapid growth and widespread routine use that can be witnessed in the frequentist corner of this area.

It is difficult to be a purist Bayesian in problems with many parameters, since setting the simultaneous prior is hard and parameter interactions can have unforeseen consequences. Such difficulties are even more prominent in the nonparametric case, where the parameter space is infinite-dimensional, and the possibilities for construction of prior distributions are so unlimited. One must therefore expect a broader range of possible solutions, as opposed to the relatively clear-cut strategies for the parametric cases. One must also expect difficulties on the technical level, in that posterior calculations quickly become complicated. Furthermore, Diaconis and Freedman (1986a, 1986b) and others have given serious warnings about lurking dangers for nonparametric Bayesian constructions in the form of large-sample inconsistency, so performance properties of the resulting Bayes estimators, once derived, should also be investigated.

The present article is about non- and semiparametric density estimation. For recent accounts of many standard methods, see Scott (1992) and Wand and Jones (1994). Again, the vast majority of these are non-Bayesian, in the sense that they do not (explicitly) utilise any prior information about what general shapes or what degree of smoothness are more likely than others. We intend to propose and discuss several Bayesian approaches to the problem.

Invited paper, to be presented at the Fifth Valencia International Meeting on Bayesian Statistics, Alicante, June 1994

One can perhaps argue that Bayesian methods are never quite as fullbloodedly nonparametric as some of the frequentist ones, in that they after all require *some* prior knowledge and prior distributions as input. Some of the estimators we discuss are indeed semiparametric in nature, in that they in various ways build on 'nonparametric uncertainty' around given parametric models. These estimators should have better performance properties than traditional nonparametric ones in a broad nonparametric neighbourhood around the parametric base model. Other methods are not geared towards any such parametric home grounds and are therefore more naturally thought of as simply nonparametric. The methods to be discussed fall into three categories. Sections 2 and 3 treat estimators that build on Dirichlet processes in various forms, Sections 4 and 5 consider placing priors on coefficients in expansions, while Sections 6 and 7 discuss non- and semiparametric methods that use locally parametric approximations. Some final remarks are offered in Section 8.

1.1. THREE GROUPS OF IDEAS. The first such group of ideas has the Dirichlet process as basic building tool. The unknown distribution can be modelled as coming from a straight Dirichlet process or from one of various related forms. The variants given attention to in Section 2 and 3 are smoothed Dirichlets, mixtures of Dirichlets, and pinned-down Dirichlets. Parts of the material of Section 2 are presumably known to workers in the field, but has been included since ready references do not seem to be available, since some of the later material in the paper builds on observations made here, and since it is of interest to see that some of the Bayes solutions also pop up in the quite different framework of Section 6. There is new material in Section 2.4 and Section 3, on attempts at smoothing and pinning down the Dirichlet.

The second general approach is to place priors on the coefficients of series expansions. In Section 4 we focus on additive orthogonal expansions for the densities themselves. These could for example be in terms of cosines or Legendre polynomials in situations where the density is supported on a finite interval. Particularly attractive from a semiparametric point of view are models of the form a normal density times an expansion in Hermite polynomials, since this allows modelling of uncertainty around the normal. In particular we discuss a special model of this sort which is more robust than the more immediate Hermite expansion. There are certain computational problems with additive expansions of densities since the likelihood function quickly becomes a very large sum of products, leading us to outline a simplifying recursive computational scheme. In Section 5 we also consider additive expansions of the log-densities, that is, multiplicative expansions of the densities. This avoids some of the obstacles that face the otherwise attractive additive expansions of Section 4, and should also be easier regarding computations.

The third general class of methods we discuss, in Sections 6 and 7, is based on using locally parametric approximations to the true density, and then placing priors on these local parameters. For a fixed x we might for example view $f(t) = a\exp\{b(t-x)\}$ as a convenient approximation to the density for t in the vicinity of x, and one can place prior distributions on local level a and local slope b, and perhaps even on the width of the local window inside which the approximation is expected to be sufficiently adequate. The problem is to establish an adequate local likelihood function that makes it possible to compute the posterior distribution for the local parameters given the local data. For this we partly rely on methods recently

developed in Hjort and Jones (1994). It is seen in Section 7 that this general locally parametric approach leads to a long list of appealing special cases.

1.2. OTHER WORK. Bayesian density estimation means placing priors on large sets of distributions, and it is only to be expected that this can be fruitfully done in many more ways than developed or mentioned in the present article. Here are some quick glances at other categories of such constructions.

Building such priors via general Pólya urn schemes was first treated systematically in Ferguson (1974). The Dirichlet again occupies a special place. Some specialisations lead to continuous densities with probability 1, but often tiny details of the construction have too much influence on the posterior distribution. There has been recent renewed interest in some of the branches of Pólya trees, see Mauldin, Sudderth and Williams (1992) and Lavine (1992). In particular Lavine shows how to construct a Pólya tree with a given predictive density, and how mixtures of them can model uncertainty around a parametric model.

Mixtures of Dirichlet processes were first studied by Antoniak (1974). Using such models along with hierarchical and otherwise generalised versions for density estimation is a current growth area; see references noted in Section 2.3. Modelling the logarithm of the density as a stochastic process is done in Section 5; see references there, and further references in Lenk (1993).

Maximum penalised likelihood and several similar methods, such as splines smoothing, can be viewed as Bayesian. See Good and Gaskins (1971, 1980) and the discussion in Silverman (1986, Section 5.4), for example, in addition to remarks given in Section 2.4 below. The estimators discussed in Rissanen, Speed and Yu (1992) based on stochastic complexity also have Bayesian overtones.

2. Dirichlet process prior with smoothing. This section discusses various approaches based on the Dirichlet process or some of its smoothed and mixed relatives.

2.1. BINNED DATA AND THE DIRICHLET SMOOTHED HISTOGRAM. Divide the interval where data fall into k cells $C_1, \ldots, C_k$, and let N_j be the number of data points falling in C_j. These form a multinomially distributed vector with parameters $(p_1, \ldots, p_k)$, where $p_j = F(C_j)$. Suppose $(p_1, \ldots, p_k)$ is given a Dirichlet prior distribution with parameters $(ap_{0,1}, \ldots, ap_{0,k})$, where $p_{0,1}, \ldots, p_{0,k}$ are 'prior guesses' for the k probabilities and a is the 'strength of belief' parameter: p_j has mean $p_{0,j}$ and variance $p_{0,j}(1-p_{0,j})/(a+1)$. The posterior is the easily updated Dirichlet $(ap_{0,1}+N_1, \ldots, ap_{0,k}+N_k)$. If the underlying probability density is viewed as approximately constant over the C_j interval, with length say h_j, then the Bayes estimate is

$$\widehat{f}(x) = \mathrm{E}\Big\{\frac{p_j}{h_j} \mid \text{data}\Big\} = \frac{1}{h_j}\frac{ap_{0,j}+N_j}{a+n} = w_n\frac{p_{0,j}}{h_j} + (1-w_n)\frac{N_j}{nh_j}, \quad x \in C_j, \quad (2.1)$$

where $w_n = a/(a+n)$.

Equation (2.1) is our first and simplest Bayesian density estimate, and its structure is typical also for more advanced methods to come. It is a convex combination of the prior guess of the density and the histogram estimate $N_j/(nh_j)$, with weights respectively $w_n = a/(a+n)$ and $1-w_n = n/(a+n)$. The estimate can perhaps

be considered parametric or semiparametric or nonparametric, depending on the fine-ness of the binning, that is, the number of cells compared to the number of data points. A smoother estimate than the simple histogram-type version above emerges when a cell is placed symmetrically around the temporarily fixed x. With such a moving cell $C(x) = [x - \frac{1}{2}h, x + \frac{1}{2}h]$ the result is

$$\widehat{f}(x) = w_n h^{-1} \int_{C(x)} f_0(t)\,\mathrm{d}t + (1 - w_n) f_n(x), \tag{2.2}$$

say, in terms of a prior guess density f_0 and where $f_n(x) = n^{-1} \sum_{i=1}^n h^{-1} I\{|x_i - x| \leq \frac{1}{2}h\}$ is the kernel estimator with a uniform kernel. And if the cell around x is determined dynamically by the requirement that it should contain at least r data points, then f_n is a r-nearest-neighbour estimate.

2.2. THE DIRICHLET PROCESS PRIOR. Ferguson (1973, 1974) introduced the Dirichlet process, which in the present context allows one to carry out analysis like in the previous subsection more easily and more generally, without having to discretise the sample space into cells. Let the distribution F which governs the data points have such a Dirichlet process prior with parameter aF_0, where F_0 is a fixed distribution and a is positive. The definition is that for each partition $B_1 \cup \cdots \cup B_k$ of the sample space, $(F(B_1), \ldots, F(B_k))$ is Dirichlet with parameters $(aF_0(B_1), \ldots, aF_0(B_k))$. A basic result is that F given the data is still a Dirichlet with updated parameter $aF_0 + \sum_{i=1}^n \delta(x_i) = aF_0 + nF_n$, where $\delta(x_i)$ is unit point mass at x_i, and F_n is the empirical distribution function.

This can be used to find a natural Bayesian density estimator. Consider $\bar{f}(x) = h^{-1} F[x - \frac{1}{2}h, x + \frac{1}{2}h]$, to be thought of, for small h, as an approximation to the density at x. Its posterior distribution is given by the result quoted, and the Bayes estimate is found to be exactly as in (2.2), with f_0 being the density of F_0. A smoother version of this argument is to use $\bar{f}(x) = \int K_h(t - x)\,\mathrm{d}F(t)$ instead, where $K_h(z) = h^{-1} K(h^{-1} z)$ and K is a given probability density symmetric around zero, referred to as a kernel function. Its posterior mean is

$$\begin{aligned} \widehat{f}(x) &= \int K_h(t - x) \frac{a\,\mathrm{d}F_0(t) + n\,\mathrm{d}F_n(t)}{a + n} \\ &= w_n \int K_h(t - x) f_0(t)\,\mathrm{d}t + (1 - w_n) f_n(x), \end{aligned} \tag{2.3}$$

where $f_n(x)$ now signifies the more general $n^{-1} \sum_{i=1}^n K_h(x_i - x)$. This is the classical nonparametric density estimator. The simpler version (2.2) corresponds to a uniform kernel on $[-\frac{1}{2}, \frac{1}{2}]$.

The parameter a of the prior is ideally set by the practising statistician, in collaboration with the experts of the relevant field of application. The form of the posterior, and of the Bayes estimates derived above, suggest that a has interpretation as 'prior sample size' or strength of belief in the prior. The likelihood function for a, based on observed data, can be derived, but it leads to a quite artificial estimator due to special features of the unconditional distribution of data sampled from a Dirichlet process. The exact number of distinct data points, D_n, is a sufficient statistic for a, and one can prove that the maximum likelihood estimator is asymptotically equivalent to $D_n / \log n$. Results like this are more helpful in certain hierarchical

constructions. Some data-based empirical Bayesian methods for setting a value of a are briefly discussed in Hjort (1991b).

How should the smoothing parameter h be chosen? This is the topic of hundreds of non-Bayesian papers in the literature. It is not obvious how one should set up a Bayesian criterion for the selection of this parameter, which in the present context at least is an algorithmic parameter of an estimation method rather than a statistical parameter of a model. Note that h plays a role both for both terms in (2.3). It should not be too large since $\bar{f}(x)$ otherwise is too far away from the real parameter of interest; in general, if F has a smooth density f, then $\bar{f}(x)$ above is equal to $\int K_h(t-x)f(t)\,\mathrm{d}t \simeq f(x) + \frac{1}{2}\sigma_K^2 h^2 f''(x)$, where σ_K^2 is the variance of the kernel. In particular the prior guess used is about equal to $f_0 + \frac{1}{2}\sigma_K^2 h^2 f_0''$ rather than the preset f_0 itself. Neither should h be too small. An explicit formula for the conditional variance of $\bar{f}(x)$ can be worked out, and is of the form

$$\mathrm{Var}\{\bar{f}(x)\,|\,\mathrm{data}\} = \frac{1}{nh}\frac{n^2}{(n+a)(n+a+1)}\frac{1}{n}\sum_{i=1}^{n} h^{-1}K(h^{-1}(x_i-x))^2 \\ + \text{smaller order terms},$$

and the average of $h^{-1}K(h^{-1}(x_i-x))^2$ is of stable size as $h \to 0$, namely about $R(K)\bar{f}(x)$, where $R(K) = \int K^2\,\mathrm{d}z$. Thus the posterior variance is essentially of order $(nh)^{-1}$ and the squared bias involved is of order h^4. Based on these facts various Bayesian criteria can be put up, leading to preferred size of order $n^{-1/5}$ for h. This agrees with standard results from the frequentist perspective.

The choice of the kernel K is generally less crucial than that of h. Minimising the approximative posterior variance plus the squared bias, hinted at above, or for that matter the approximate risk function for the estimator, gives a result proportional to $\{\sigma_K R(K)\}^{4/5}$, which is minimal for the Yepanechnikov kernel $K(z) = \frac{3}{2}(1-4z^2)_+$ (scaled here to have support $[-\frac{1}{2},\frac{1}{2}]$). West (1991) starts out from a certain marginalisation consistency criterion and shows that strict adherence to this implies that K is of double exponential form.

2.3. SMOOTHED AND MIXED DIRICHLET PRIORS AS PRIOR. Above we gave F a Dirichlet process prior and then smoothed the posterior F around a given x to produce the Bayesian density estimate (2.3). This approach makes perfect sense, as does the answer. Nevertheless it is perhaps disturbing that the density f itself has not been directly modelled, and indeed under a Dirichlet prior it does not properly exist; the random F is with probability 1 a discrete distribution (with infinitely many random jumps at an infinite collection of random locations).

This motivates another approach, which is to 'smooth the prior first', modelling f as $f(x) = \int K_h(t-x)\,\mathrm{d}G(t)$ for a Dirichlet process G, say with parameter aG_0. This assures a well-defined random and continuous density, if only K is continuous. Such a density can also be represented as a countably infinite mixture, as per the remark above; see Ferguson (1983). The posterior distribution for f, given a set of observations coming from this f, has been worked out and characterised via a mixture of Dirichlet processes by Lo (1984) and by Ferguson (1983). The exact posterior mean is however an enormous sum over all possible partitions of the data set, and its computation accordingly quite difficult for all but very small sample sizes. This problem can be dealt with, for example via a simulation-based method due to

Kuo, see Ferguson (1983) and Kuo (1986), or via an iterative resampling scheme developed by Escobar, see Escobar and West (1994). There are still difficulties with the approach. The choices of G_0, a and h are problematic, and the performance properties of the resulting estimators are less understood than those of (2.3). Further progress and more general hierarchical versions of these schemes are discussed in Florens, Mochart and Rolin (1992), West (1992), and Escobar and West (1994).

2.4. GENERALISED DIRICHLET PRIORS. In many situations there is some knowledge of the smoothness of the underlying distribution for data. This points to an inadequacy of the Dirichlet prior; it is almost 'too nonparametric' in its lack of contextual smoothness. Considering the discrete framework of Section 2.1 again, for example, it is clear that the ordering of the p_js is immaterial under a Dirichlet prior, whereas one's prior knowledge often would suggest neighbouring p_js to be close with high probability (assuming the cell widths in that setting to be the same). Again this motivates trying to construct smoothed Dirichlet distributions, to be used as more adequate priors in smoothing problems. The following is another route towards achieving this, complementing the mixtures framework indicated above.

Consider the following way of building a prior for a probability vector $\mathbf{p} = (p_1, \ldots, p_k)$: Let $(Y_1, \ldots, Y_k)$ be positive random variables, let $p_j = \exp(-Y_j)$, and condition on their sum being equal to 1. A calculation shows that the density of $(p_1, \ldots, p_{k-1})$ becomes

$$\pi(p_1, \ldots, p_{k-1}) = \text{const.}\, h(-\log p_1, \ldots, -\log p_k)\, p_1^{-1} \cdots p_k^{-1},$$

where $p_k = 1 - \sum_{j=1}^{k-1} p_j$, in terms of the density $h(y_1, \ldots, y_k)$ of the Y_is. As a special case of this construction, consider

$$Y = (Y_1, \ldots, Y_k) \sim \text{const.} \Big\{ \prod_{i=1}^{k} \alpha_i \exp(-\alpha_i y_i) \Big\}\, g_0(y_1, \ldots, y_k), \tag{2.4}$$

for a suitable positive function g_0. In this case the distribution for the p_i's becomes proportional to $p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1} g(p_1, \ldots, p_k)$, where $g(p_1, \ldots, p_k) = g_0(-\log p_1, \ldots, -\log p_k)$. In particular independent exponentials for the Y_is, corresponding to $g_0 = g = 1$, give the familiar Dirichlet distribution. Agree therefore to call this the *generalised Dirichlet distribution*, with parameters $\alpha_1, \ldots, \alpha_k$ and g, and write $\mathbf{p} = (p_1, \ldots, p_k) \sim \mathcal{GD}(\alpha_1, \ldots, \alpha_k; g)$ to indicate this.

The idea is to use particular g_0 or g functions to push the Dirichlet in certain directions, so to speak. A generally useful form is $g(\mathbf{p}) = \exp\{-\lambda \Delta(\mathbf{p})\}$, that is,

$$\pi(p_1, \ldots, p_{k-1}) = \text{const.}\, p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1} \exp\{-\lambda \Delta(p_1, \ldots, p_k)\}, \tag{2.5}$$

where small values of $\Delta(p_1, \ldots, p_k)$ means that a certain characteristic of interest is present, and where the penalty parameter λ dictates the extent to which the characteristic is manifest; a large value of λ forces realisations of $\mathbf{p}$ to have small values of $\Delta(\mathbf{p})$. Some examples of such $\Delta(\mathbf{p})$ functions are

$$\sum_{j=1}^{k-1} (p_{j+1} - p_j)^2, \quad \sum_{j=2}^{k-1} (p_{j+1} - 2p_j + p_{j-1})^2, \quad \sum_{j=1}^{k-1} (\log p_{j+1} - \log p_j)^2.$$

Using g functions with these Δ functions leads to 'smoothed Dirichlet priors', forcing successive p_js to be closer to each other with higher probability than under standard Dirichlet conditions. Even a wish for unimodality can be built into a suitable $\Delta(\cdot)$ function.

It turns out that these generalised Dirichlet distributions are still conjugate priors for multinomial models. In fact, in the setting of Section 2.1, if $(p_1, \ldots, p_k)$ is $\mathcal{GD}(ap_{0,1}, \ldots, ap_{0,k}; g)$, then $\mathbf{p}$ given data is $\mathcal{GD}(ap_{0,1} + N_1, \ldots, ap_{0,k} + N_k; g)$. For most choices of g it is not possible to find explicit expressions for expected values, and if required these would have to be found by simulation or numerical integration. The mode is however reasonably easy to compute. The posterior mode $(p_1^*, \ldots, p_k^*)$ here, which is the Bayes solution under a sharp 0–1 loss function, is the maximiser of

$$\sum_{j=1}^{k} (ap_{0,j} + N_j - 1) \log p_j - \lambda \Delta(p_1, \ldots, p_k), \tag{2.6}$$

under the sum to 1 constraint. This amounts to a further smoothing of the original Dirichlet-smoothed histogram $(ap_{0,j} + N_j - 1)/(a + n - k)$, making sure that $\Delta(p_1^*, \ldots, p_k^*)$ is not large. With the second Δ-function mentioned above this would be quite similar to splines smoothing, for example.

We have phrased the estimation problem in Bayesian terms, starting with the (2.5) prior. If the $p_{0,j}$s are equal and $a = k$, then the criterion to maximise is $\sum_{j=1}^{k} N_j \log p_j - \lambda \Delta(\mathbf{p})$, which is the penalised log-likelihood. Other authors have used this as the starting point, wishing to smooth the simple maximum likelihood estimates across cells, perhaps without being particularly Bayesian about it. The maximum penalised likelihood method can always be rephrased in Bayesian terms, as explained here. Relevant references in the present context of histogram smoothing and density estimation include Good and Gaskins (1971, 1980), Simonoff (1983) and Silverman (1986, Section 5.4).

Notice that when n is large compared to a and λ, then the Δ-term does not matter much, and the estimators become asymptotically equivalent to the ordinary histogram counts N_j/n: the scaled differences $\sqrt{n}(p_j^* - N_j/n)$ tend to zero in probability, regardless of a and the $\lambda\Delta(\mathbf{p})$ function (unless a or λ is allowed to increase with n at a $\sqrt{n}$ rate or faster). The situation is more delicate when the number of cells increases with n, as should typically happen in the density estimation context. Simonoff (1983), who essentially worked with the third $\Delta(\mathbf{p})$ function mentioned above, studied this sparse cells framework. Results will not be given here, but we mention that methods developed in Hjort and Pollard (1994) are quite well suited to study properties of estimators defined by maximisation of (2.6), also when the number of cells increases with n.

The remarks and the construction given here relate to the discrete framework with binned data. The point raised about lack of ordering information and lack of smoothness is also pertinent for the Dirichlet process case; witness its representation as the normed version of a Gamma process with independent increments. It is therefore of interest to study continuous analogues of the distributions above, to be used, if possible, in the continuous data situations of Sections 2.2, thereby by-passing the details of binning and so on. One possibility is to study limits in distributions of the discrete processes, say as the maximal binwidth tends to zero. This is not an easy

problem, and some requirements must be placed on the Δ function in order to assure a well-defined limit process. The existence problem is made easier by the fact that the generalised Dirichlet distribution, as defined above, is closed under combination of cells, in the following sense: Suppose $(p_1,\ldots,p_k)$ is $\mathcal{GD}(\alpha_1,\ldots,\alpha_k;g)$, and let q_1 be the sum of the first j_1 of the p_js, q_2 the sum of the next j_2 of the p_js, and so on, up to q_m, the sum of the last j_m of the p_js. Then, working with (2.4) or the (2.5) version, $(q_1,\ldots,q_m)$ can be shown to be $\mathcal{GD}(\sum_{r=1}^{j_1}\alpha_r,\sum_{r=j_1+1}^{j_1+j_2}\alpha_r,\ldots;g^*)$, for a certain $g^*(q_1,\ldots,q_m)$ function. This generalised Dirichlet processes topic is not pursued here, however.

2.5. FIRST SEMIPARAMETRIC FRAMEWORK. Let there be a 'background parameter' θ with some prior $\pi(\theta)$, and assume that F for given θ is a Dirichlet $aF_0(\cdot,\theta)$. The F_0 could be a normal, for example, and then this models nonparametric uncertainty around the normal model.

For given θ it follows from previous comments that F given data is a Dirichlet $aF_0(\cdot,\theta)+nF_n$, and the arguments of Section 2.2 can be repeated to give a density estimator of type

$$\widehat{f}(x,\theta)=\mathrm{E}\{\bar{f}(x)\,|\,\mathrm{data},\theta\}=w_n\int K_h(t-x)f_0(t,\theta)\,\mathrm{d}t+(1-w_n)f_n(x),$$

where $f_0(\cdot,\theta)$ is the density of $F_0(\cdot,\theta)$. The final estimate is therefore of the form

$$\widehat{f}(x)=\mathrm{E}\{\widehat{f}(x,\theta)\,|\,\mathrm{data}\}=w_n\int K_h(t-x)\widehat{f}_0(t)\,\mathrm{d}t+(1-w_n)f_n(x),\qquad(2.7)$$

where $\widehat{f}_0(t)=\mathrm{E}\{\widehat{f}_0(t,\theta)\,|\,\mathrm{data}\}$ is the predictive parametric density. What needs to be found is the distribution of θ given data.

To this end, for simplicity order the distinct data points as $x_1<\cdots<x_k$, and suppose the multiplicities are $j_1,\ldots,j_k$. Let A be the event that $X_i\in[x_1-\frac12\varepsilon,x_1+\frac12\varepsilon]$ for the first j_1 observations, that $X_i\in[x_2-\frac12\varepsilon,x_2+\frac12\varepsilon]$ for the next j_2 observations, and so on, where ε is small. Then

$$\begin{aligned}\Pr\{\theta\in[\theta_0-\tfrac12\mathrm{d}\theta,\theta_0+\tfrac12\mathrm{d}\theta],A\}&\simeq\pi(\theta_0)\,\mathrm{d}\theta\int\Pr\{A\,|\,F\}\,\mathrm{Dir}(aF_0(\cdot,\theta_0),\mathrm{d}F)\\&=\pi(\theta_0)\,\mathrm{d}\theta\,\mathrm{E}\{F[x_1-\tfrac12\varepsilon,x_1+\tfrac12\varepsilon]^{j_1}\cdots F[x_k-\tfrac12\varepsilon,x_k+\tfrac12\varepsilon]^{j_k}\,|\,\theta_0\}\\&\simeq\pi(\theta_0)\,\mathrm{d}\theta\,\frac{\Gamma(a)}{\Gamma(a+n)}\frac{\Gamma(af_0(x_1,\theta_0)\varepsilon+j_1)}{\Gamma(af_0(x_1,\theta))}\cdots\frac{\Gamma(af_0(x_k,\theta_0)\varepsilon+j_k)}{\Gamma(af_0(x_k,\theta))},\end{aligned}$$

by a formula for product moments in a Dirichlet distribution. Since $\Gamma(b+j)/\Gamma(b)=b(b+1)\cdots(b+j-1)$ this shows that the posterior for θ is

$$\pi(\theta\,|\,\mathrm{data})=\mathrm{const.}\,\pi(\theta)\prod_{\text{distinct}}f_0(x_i,\theta),\qquad(2.8)$$

the product being over the distinct data values only. If in particular the data points are distinct, as in all proper continuous cases, then the sophisticated extra nonparametric randomness does not enter the result; the posterior is then exactly the same as the traditional one under the parametric model (which also is the one corresponding to the a parameter in the Dirichlet prior being equal to infinity).

This fills in the missing ingredient of the (2.7) estimator. Explicit formulae can be worked out for the case of a normal kernel, a normal start family for F_0, and for traditionally used conjugate priors for (μ, σ).

2.6. SECOND SEMIPARAMETRIC FRAMEWORK: ESTIMATING THE RESIDUAL DENSITY. It is often useful to think of data as location plus noise, and then modelling these terms separately. This is done in regression contexts, of course, but can also be done for a homogeneous sample. Let in general $X_i = T_\theta(\varepsilon_i)$, where the ε_is are a sample from a common distribution G, say, and θ is an unknown p-dimensional parameter in the transformation T. This is taken to be a continuous increasing transformation for each given θ with inverse $\varepsilon_i = T_\theta^{-1}(X_i)$. We think of the ε_is as residuals or normalised residuals. A simple example is $X_i = \mu + \sigma\varepsilon_i$ and $\varepsilon_i = (X_i - \mu)/\sigma$. The framework now will be to have some prior density for θ and in addition letting the nonparametric G have a distribution in the space of all distributions, centred at a suitable G_0, for example the standard normal. Note that $F(x) = G(T_\theta^{-1}(x))$, and the present Bayesian semiparametric setup has priors on both the G part and the $T_\theta^{-1}(x)$ part.

Assume that G is a Dirichlet with parameter aG_0. We are interested in $\bar f(x) = \int K_h(t-x)\,\mathrm{d}F(t)$, now with $\mathrm{d}F(t) = \mathrm{d}G(T_\theta^{-1}(t))$, and can at least proceed as earlier for each given θ, since then G given data is a Dirichlet with parameter $aG_0 + \sum_{i=1}^n \delta(T_\theta^{-1}(x_i))$. One finds

$$\begin{aligned}\mathrm{E}\{\bar f(x) \mid \text{data}, \theta\} &= \int K_h(t-x)\frac{a\,\mathrm{d}G_0(T_\theta^{-1}(t)) + n\,\mathrm{d}F_n(t)}{a+n} \\ &= w_n \int K_h(t-x) f_0(t,\theta)\,\mathrm{d}t + (1-w_n) f_n(x),\end{aligned}$$

where $f_0(t,\theta) = g_0(T_\theta^{-1}(t))|\partial T_\theta^{-1}(t)/\partial t|$ is the parametric density of X_i under the idealised $G = G_0$ conditions. This gives exactly the same density estimator as in (2.7), involving the predictive density $\mathrm{E}\{f_0(t,\theta) \mid \text{data}\}$. It further turns out that the posterior distribution of θ is exactly as in (2.8). This is really because the present model, which is defined in a somewhat roundabout manner as far as the X_is are concerned, *is* the same as in Section 2.5, with $F_0(t,\theta) = G_0(T_\theta^{-1}(t))$.

One does, however, get interesting results of a different nature for the estimation of the residual density. The mean of $G(y)$ given both data and θ is a convex combination of $G_0(y)$ and $n^{-1}\sum_{i=1}^n I\{T_\theta^{-1}(x_i) \le y\}$, and the Bayes estimate is

$$\widehat G(y) = w_n G_0(y) + (1-w_n)\,n^{-1}\sum_{i=1}^n \Pr\{T_\theta^{-1}(x_i) \le y \mid \text{data}\}.$$

The point is that this is a smooth estimate with a density, in spite of the fact that G under the stated prior model does not have a continuous distribution. Typically, θ given data is approximately a normal centred at the Bayes estimator (or for that matter the maximum likelihood estimator), say of the form $\mathcal{N}_p\{\widehat\theta, \widehat V/n\}$. Thus $T_\theta^{-1}(x_i)$ given data is approximately a normal with mean equal to the estimated residual, which we for typographical reasons write as $\widehat\varepsilon_i = T^{-1}(x_i, \widehat\theta)$, and variance say $\widehat v_i^2/n$. This leads to approximating $\Pr\{T_\theta^{-1}(x_i) \le y \mid \text{data}\}$ above with $\Phi(\sqrt{n}(y - \widehat\varepsilon_i)/\widehat v_i)$, and the density of $\widehat G$ becomes

$$\widehat g(y) \simeq w_n g_0(y) + (1-w_n)\,n^{-1}\sum_{i=1}^n \frac{1}{h_i}\phi\Big(\frac{y-\widehat\varepsilon_i}{h_i}\Big) \quad \text{where } h_i = \widehat v_i/\sqrt{n}. \tag{2.9}$$

The second term uses a variable kernel density estimate for the estimated residuals with a normal kernel and $h_i = \widehat{v}_i/\sqrt{n}$ for the bandwidths. A result of similar nature is in Bunke (1987).

It is remarkable that the present natural semiparametric Bayesian framework leads to such explicit advice for both the kernel form and in particular the bandwidths. In the case of $X_i = \mu + \sigma\varepsilon_i$ with fixed σ the $\widehat{v}_i$s are equal and the estimator is a kernel estimate with smoothing parameter $\widehat{v}/\sqrt{n}$, and this also happens to be the only allowed size in a treatment of West (1991), using a certain marginalisation coherence criterion. This amounts to smoothing somewhat less than the standard recommendation $O(n^{-1/5})$ that falls out from traditional mean squared error theory. For the case of $X_i = \mu + \sigma\varepsilon_i$ with both parameters unknown, and a normal model as null model, the h_i is approximately equal to $\{1 + \frac{1}{2}(x_i - \widehat{\mu})/\widehat{\sigma}^2\}^{1/2}/\sqrt{n}$, advicing more smoothing outside the central area than in the middle.

While (2.9) used a large-sample approximation for the posterior of θ, exact calculations, for the derivative of the exact $\widehat{G}(y)$, are also available for many cases of interest, for example when $X_i = \mu + \sigma\varepsilon_i$, G_0 is the standard normal, and (μ, σ) has a conjugate prior. We also point out that the apparatus here is easily extended to regression contexts with covariate information, say with $X_i = T_\theta(\mathbf{z}_i, \varepsilon_i)$. The obvious special case is $X_i = \mathbf{b}'\mathbf{z}_i + \sigma\varepsilon_i$, with normalised residuals coming from a Dirichlet process G centred at the standard normal. There is a mildly unpleasant surprise for the situation with higher-dimensional data; see Section 8.7.

3. Pinned-down Dirichlet processes. There are other generalisations of the Dirichlet process worth studying in connection with density estimation. Such studies are potentially of interest since the final estimators are often useful, of course, but also since models using these kind of infinite-dimensional priors serve as test-beds for general Bayesian methodology. Doss (1985a, 1985b), Diaconis and Freedman (1986a, 1986b), Hjort (1986, 1987, 1991a) and others have shown that some of these constructions lead to estimators that are inconsistent, to mention one aspect of importance; see also Section 8.7 below. In the present section we look at density estimation with certain pinned-down Dirichlet priors, for the straight distribution of data themselves or for the residuals. Still other situations worth exploring, but not pursued here, include the invariance-constrained Dirichlet priors of Dalal (1979).

3.1. DENSITY ESTIMATE WITH A PINNED-DOWN DIRICHLET. Let F be a Dirichlet process aF_0, conditioned on having fixed values $F(B_j) = z_j$ on certain *control sets* $B_1, \ldots, B_k$, where these form a partition of the sample space. This typically amounts to having preset values for certain quantiles. One can prove that this pinning down of F amounts to splitting the Dirichlet process into k different independent Dirichlet processes, $F = z_j F_j$ on the set B_j, where F_j is Dirichlet with parameter $(az_j)(F_0/z_j)$; see Hjort (1986). Furthermore, F_j given the full data set has the same distribution as F_j given only the data values that fall in B_j, that is, a Dirichlet with parameter $aF_0 + \sum_{i=1}^n \delta(x_i) I\{x_i \in B_j\}$.

This makes it easy to compute the mean of F given data. If A lies within B_j, then

$$\mathrm{E}\{F(A) \mid \text{data}\} = z_j \frac{aF_0(A) + \sum_{i=1}^n I\{x_i \in A\}}{az_j + \sum_{i=1}^n I\{x_i \in B_j\}} = z_j \frac{aF_0(A) + nF_n(A)}{az_j + nF_n(B_j)}.$$

If x is an inner point of B_j, therefore, the Bayes estimate of the $\bar{f}(x) = \int K_h(t-x)\,\mathrm{d}F(t)$ parameter, which is close to the density, is

$$\widehat{f}(x) = z_j \frac{a\int K(z) f_0(x+hz)\,\mathrm{d}z + nf_n(x)}{az_j + nF_n(B_j)}. \tag{3.1}$$

If a is small compared to n this is close to $\{z_j/F_n(B_j)\}f_n(x)$, which is the traditional density estimate times a correction to account for the known location of a set of quantiles.

3.2. SEMIPARAMETRIC MODEL WITH PINNED-DOWN DIRICHLET. In the framework of Section 2.6, let there be a prior $\pi(\theta)$ for the parameter and suppose the residual distribution G for ε_i is given a Dirichlet process prior aG_0, but pinned down to have $G(B_j) = z_j$ on certain control sets $B_1, \ldots, B_k$, as in Section 3.1. Again, $f_0(x,\theta)$ is the parametric density for Xs under the ideal $G = G_0$ condition. As a simple example of this setup, envisage X_i as $\mu + \sigma\varepsilon_i$ where the normalised residuals ε_is come from a distribution G with probability mass 0.90 on $[-1.645, 1.645]$ and 0.05 on each of $(-\infty, -1.645)$ and $(1.645, \infty)$ (1.645 being the familiar upper 5% point of the standard normal). Estimating parameters in this model, rather than in the unconstrained model that makes no such restriction on G, aims at having approximately 90% of future data points in $[\widehat{\mu} - 1.645\,\widehat{\sigma}, \widehat{\mu} + 1.645\,\widehat{\sigma}]$. Thus the general control sets apparatus is useful in connection with predictive analysis.

Finding the posterior density for the parameters involves calculations that become more complicated than those of Sections 2.5–2.6. The result was worked out in Hjort (1986), and is of the form

$$\pi(\theta \mid \text{data}) = \text{const.}\,\pi(\theta) L_n(\theta) M_n(\theta), \tag{3.2}$$

where

$$L_n(\theta) = \prod_{\text{distinct}} f_0(x_i,\theta) \quad \text{and} \quad M_n(\theta) = \prod_{j=1}^{m} \frac{z_j^{C_j(\theta)}}{\Gamma(az_j + C_j(\theta))}, \tag{3.3}$$

writing $C_j(\theta) = nF_n(T_\theta(B_j))$ for the number of $\varepsilon_i = T_\theta^{-1}(x_i)$ that fall in B_j. In the example above, taking G_0 to be the normal, the posterior becomes

$$\pi(\mu,\sigma \mid \text{data}) = \text{const.}\,\pi(\mu,\sigma)\frac{1}{\sigma^n}\exp\Big\{-\tfrac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\sigma^2}\Big\} \\ \frac{0.05^{C_1(\mu,\sigma)}}{\Gamma(0.05\,a + C_1(\mu,\sigma))}\,\frac{0.90^{C_2(\mu,\sigma)}}{\Gamma(0.90\,a + C_2(\mu,\sigma))}\,\frac{0.05^{C_3(\mu,\sigma)}}{\Gamma(0.05\,a + C_3(\mu,\sigma))}$$

(assuming the data points to be distinct), where $C_j(\mu,\sigma)$ is the number of $(x_i-\mu)/\sigma$ falling in respectively $(-\infty,-1.645)$, $[-1.645, 1.645]$, $(1.645,\infty)$. A simpler special case is the model where G is taken to have median zero, with control sets $(-\infty, 0]$ and $(0,\infty)$ and $z_1 = z_2 = \frac{1}{2}$. This, with known σ, is the situation discussed in Doss (1985a, 1985b). Thus (3.2)–(3.3) is a broad generalisation of the posterior distribution found in Doss (1985a).

The posterior density (3.2) is unusual and interesting on several accounts. The $L_n(\theta)$ term is largest around the maximum likelihood estimators, while $M_n(\theta)$ is

largest in areas where $C_j(\theta)$ is close to z_j for each j. These are sometimes conflicting aims, and the Bayes estimators in effect try to push the maximum likelihood estimates so as to better achieve the $z_1, \ldots, z_m$ balance. The unusual feature is that the data do not wash out the prior when n grows; the $M_n(\theta)$ term is about equal in strength to $L_n(\theta)$; see Hjort (1986, 1987) for further results and discussion.

Let us first concentrate on estimating the density of the Xs. Let $A = [x - \frac{1}{2}\eta, x + \frac{1}{2}\eta]$ be a short interval containing a given x. If θ is such that $T_\theta^{-1}(x) \in B_j$, then $F(A) = G(T_\theta^{-1}(A)) = z_j G_j(T_\theta^{-1}(A))$, and

$$\mathrm{E}\{F(A) \,|\, \mathrm{data}, \theta\} = z_j \frac{aG_0(T_\theta^{-1}(A)) + \sum_{i=1}^n I\{T_\theta^{-1}(x_i) \in T_\theta^{-1}(A)\}}{az_j + \sum_{i=1}^n I\{T_\theta^{-1}(x_i) \in B_j\}}.$$

Hence

$$\mathrm{E}\{F(A) \,|\, \mathrm{data}\} \simeq \sum_{j=1}^m z_j \int_{D_j(x)} \frac{af_0(x,\theta)\eta + nF_n(A)}{az_j + nF_n(T_\theta(B_j))} \pi(\theta \,|\, \mathrm{data})\, \mathrm{d}\theta,$$

where $D_j(x)$ is the set of θ for which $T_\theta^{-1}(x) \in B_j$. The Bayes estimate of $\bar{f}(x) = \int K_h(t-x)\, \mathrm{d}F(t)$ is therefore at least close to

$$\widehat{f}(x) = \sum_{j=1}^m z_j \int_{D_j(x)} \frac{a\bar{f}_0(x,\theta) + nf_n(x)}{az_j + nF_n(T_\theta(B_j))} \pi(\theta \,|\, \mathrm{data})\, \mathrm{d}\theta,$$

where $\bar{f}_0(x,\theta) = \int K_h(t-x) f_0(t,\theta)\, \mathrm{d}t$. In cases where $T^{-1}(x,\widehat{\theta})$ lies safely inside a B_j set, and the posterior distribution is not too spread out, then $\pi(\theta \,|\, \mathrm{data})$ gives most of its probability mass to a single $D_j(x)$, and a further approximation is

$$\widehat{f}(x) \simeq z_j \frac{a\mathrm{E}\{\bar{f}_0(x,\theta) \,|\, \mathrm{data}\} + nf_n(x)}{az_j + nG_n(B_j)}, \quad x \in T(B_j, \widehat{\theta}),$$

where $nG_n(B_j)$ is the number of estimated residuals in B_j.

Next we attend to the problem of estimating the residual density, as in Section 2.6. For subsets A of B_j,

$$\mathrm{E}\{G(A) \,|\, \mathrm{data}, \theta\} = z_j \frac{aG_0(A) + \sum_{i=1}^n I\{T_\theta^{-1}(x_i) \in A\}}{az_j + \sum_{i=1}^n I\{T_\theta^{-1}(x_i) \in B_j\}} = z_j \frac{aG_0(A) + nF_n(T_\theta(A))}{az_j + nF_n(T_\theta(B_j))}.$$

This must then be averaged with respect to the (3.2) distribution. By results of Hjort (1987), θ is approximately a normal centred at the Bayes estimate $\widehat{\theta}$ and with a certain variance matrix of form $\widehat{V}/n$. Hence $T_\theta^{-1}(x_i)$, given data, can be represented as being approximately equal to $\widehat{\varepsilon}_i + \widehat{v}_i N/\sqrt{n}$, where N is a standard normal. This is similar in structure to what we saw in Section 2.6, but the present posterior distribution is much more complicated than there, as is the calculation of the Bayes estimate and the estimated residuals. Nevertheless, for y an inner point of B_j,

$$\begin{aligned}
\mathrm{E}\{G[y - \tfrac{1}{2}\eta, y + \tfrac{1}{2}\eta] \,|\, \mathrm{data}\} &\simeq z_j \,\mathrm{E}\Big\{ \frac{ag_0(y)\eta + \sum_{i=1}^n I\{T_\theta^{-1}(x_i) \in y \pm \frac{1}{2}\eta\}}{az_j + \sum_{i=1}^n I\{T_\theta^{-1}(x_i) \in B_j\}} \,\Big|\, \mathrm{data} \Big\} \\
&\simeq z_j \,\mathrm{E}\Big\{ \frac{ag_0(y)\eta + \sum_{i=1}^n I\{\widehat{\varepsilon}_i + \widehat{v}_i N/\sqrt{n} \in y \pm \frac{1}{2}\eta\}}{az_j + \sum_{i=1}^n I\{\widehat{\varepsilon}_i + \widehat{v}_i N/\sqrt{n} \in B_j\}} \Big\}
\end{aligned}$$

for small values of η. The random summands appearing in the denominator are mostly equal to 1 with high probability, if $\widehat{\varepsilon}_i$ is in B_j, or equal to 0 with high probability, if $\widehat{\varepsilon}_i$ is outside B_j. An approximation to the Bayes estimate of the density for the residuals is accordingly

$$\widehat{g}(y) = z_j \frac{a g_0(y) + n g_n(y)}{a z_j + n G_n(B_j)}, \tag{3.4}$$

where G_n is the empirical distribution for the estimated residuals, so that $nG_n(B_j)$ is the number of $\widehat{\varepsilon}_i \in B_j$, and where $g_n(y) = n^{-1} \sum_{i=1}^n h_i^{-1} \phi(h_i^{-1}(y - \widehat{\varepsilon}_i))$ is a variable bandwidth kernel estimate with explicitly given bandwidths, $h_i = \widehat{v}_i/\sqrt{n}$. Again, this is quite similar to (2.9), but with differently defined estimated residuals and bandwidths, as noted above. Also note that the estimator is intent on having probability mass close to z_j on B_j, as the prior requests.

3.3. GENERALISATIONS. The framework above can be generalised to regression contexts and to multidimensional data, essentially since the (3.2)–(3.3) result holds in such models. It is up to the statistician to choose control sets, for example for prediction purposes. We also point out that the general treatment leads to quantile estimates of interest, one possibility being as follows. Suppose the pth quantile $F^{-1}(p)$ is to be estimated for a distribution which is thought to be not very far from the normal, for which the exact result is $\mu + \sigma c_p$, say, where $\Phi(c_p) = p$. Use control sets $(-\infty, c_p]$ and (c_p, ∞) with $z_1 = p$ and $z_2 = 1 - p$, compute the posterior density using (3.2), and in the end use the Bayes estimate $\widehat{\mu} + \widehat{\sigma} c_p$. This will typically be closer to the correct $F^{-1}(p)$ than say the maximum likelihood solution.

4. Additive Hermite expansions. This section discusses Bayesian density estimators based on certain additive expansions. These expansions are valid for broad classes of densities, and can as such be viewed as nonparametric or semiparametric, depending on whether the number of terms used is allowed to be infinite (or very large) or moderate. The methods we give are valid for all the familiar expansions in terms of orthogonal basis functions, for example cosine expansions and Legendre polynomial expansions for densities with support on a finite interval. We focus here on expansions that use Hermite polynomials to 'correct on the normal'. These also lead to frequentist density estimates of interest in their own right; see Fenstad and Hjort (1994). The present Bayesian programme is to place priors on the coefficients in these expansions and work out posterior moments.

4.1. THE STRAIGHT HERMITE EXPANSION. The Hermite polynomials are defined via the derivatives of the standard normal density, $\phi^{(j)}(x) = (-1)^j H_j(x)\phi(x)$. Thus $H_0(x) = 1$, $H_1(x) = x$, $H_2(x) = x^2 - 1$, $H_3(x) = x^3 - 3x$, $H_4(x) = x^4 - 6x^2 + 3$, and so on. They are orthogonal with respect to the normal density, $\int H_j H_k \phi \, dx = j! \, I\{j = k\}$. For an arbitrary density f, consider approximations of the form $f_m = \phi \sum_{j=0}^m (\gamma_j/j!) H_j$. The best approximation, in the sense of minimising $\int \{f/\phi - \sum_{j=0}^m (\gamma_j/j!) H_j\}^2 \, dx$, emerges when $\gamma_j = \int H_j f \, dx$. Thus $\gamma_0 = 1$, $\gamma_1 = \mathrm{E}_f X$, $\gamma_2 = \mathrm{E}_f(X^2 - 1)$, $\gamma_3 = \mathrm{E}_f(X^3 - 3X)$, and so on.

This approximation can be expected to be most effective if f at the outset is not too far from the starting point ϕ. It therefore helps to pre-transform to

$Y = (X - \mu)/\sigma$, writing μ and σ for mean and standard deviation, develop the approximation on that scale, and back-transform. The result is

$$\begin{aligned} f_m(x) &= \phi\Big(\frac{x-\mu}{\sigma}\Big)\frac{1}{\sigma}\sum_{j=0}^{m}\frac{\gamma_j}{j!}H_j\Big(\frac{x-\mu}{\sigma}\Big) \\ &= \phi\Big(\frac{x-\mu}{\sigma}\Big)\frac{1}{\sigma}\Big\{1+\sum_{j=3}^{m}\frac{\gamma_j}{j!}H_j\Big(\frac{x-\mu}{\sigma}\Big)\Big\}, \end{aligned} \tag{4.1}$$

where $\gamma_j = \mathrm{E}_f H_j((X_i - \mu)/\sigma)$. Note that $\gamma_0 = 1$, $\gamma_1 = 0$, $\gamma_2 = 0$, while

$$\gamma_3 = \mathrm{E}_f Y^3, \quad \gamma_4 = \mathrm{E}_f(Y^4 - 3), \quad \gamma_5 = \mathrm{E}_f(Y^5 - 10Y^3)$$

are equal to the skewness, kurtosis, pentakosis and so on. The normal density is the one having each $\gamma_j = 0$ for $j \geq 3$. A natural Bayesian semiparametric scheme is to put priors on (μ, σ) and the $\gamma_3, \gamma_4, \ldots$ coefficients, and then calculate the posterior mean of (4.1). This amounts to Bayesian modelling of uncertainty around the normal density.

This quickly becomes quite cumbersome due to the prohibitively large number of terms involved, in principle $(m+1)^n$, when the likelihood product $f(x_1)\cdots f(x_n)$ is expanded as a sum. Some progress is nevertheless possible. Assume at this stage that μ and σ are known quantities, which we may then take to be 0 and 1. Consider a prior density of the form $\pi(\gamma_3, \ldots, \gamma_m)$. The key step to simplification of the problem is to note that the likelihood can be written as being proportional to

$$\prod_{i=1}^{n}\Big\{\sum_{j=0}^{m}\frac{\gamma_j}{j!}H_j(x_i)\Big\} = \sum_{C(n)}\Big(\frac{\gamma_0}{0!}\Big)^{k_0}\cdots\Big(\frac{\gamma_m}{m!}\Big)^{k_m}S_n(k_0, \cdots, k_m), \tag{4.2}$$

where $C(n)$ is the set of all nonnegative $(k_0, \ldots, k_m)$ with sum n, and $S_n(k_0, \ldots, k_m)$ is the sum of all products $H_{j_1}(x_1)\cdots H_{j_n}(x_n)$ for which exactly k_0 of the j_is are equal to 0, exactly k_1 of the j_is are equal to 1, and so on. A given $S_n(k_0, \ldots, k_m)$ is the sum of $n!/(k_0!\cdots k_m!)$ such terms, but they can be obtained recursively instead, via

$$S_n(k_0, \ldots, k_m) = \sum_{j=0}^{m} H_j(x_n)S_{n-1}(k_0, \ldots, k_j - 1, \ldots, k_m).$$

Here a S_n term is set to zero if one or more of its indices is -1, and the computations start out from $S_1(0, \ldots, 1, \ldots, 0) = H_j(x_1)$. The rearrangement (4.2) reduces the number of terms from the astronomical $(m+1)^n$ to the hopefully manageable $(n+1)\cdots(n+m+1)/(m!)$.

Let

$$M(k_0, \ldots, k_m) = \mathrm{E}\Big(\frac{\gamma_0}{0!}\Big)^{k_0}\cdots\Big(\frac{\gamma_m}{m!}\Big)^{k_m},$$

computed relative to the prior distribution. Then

$$\frac{\widehat{\gamma}_j}{j!} = \mathrm{E}\Big\{\frac{\gamma_j}{j!} \mid \text{data}\Big\} = \frac{\sum_{C(n)} M(k_0, \ldots, k_j + 1, \ldots, k_m)S_n(k_0, \ldots, k_m)}{\sum_{C(n)} M(k_0, \ldots, k_j, \ldots, k_m)S_n(k_0, \ldots, k_m)}, \tag{4.3}$$

and this defines the Bayesian density estimator $\widehat{f}_m(x) = \phi(x)\sum_{j=0}^m(\widehat{\gamma}_j/j!)H_j(x)$. All terms with $k_1 \geq 1$ and/or $k_2 \geq 1$ drop out since $\gamma_1 = \gamma_2 = 0$. Note also that further simplification is possible when $\gamma_3, \ldots, \gamma_m$ are taken independent in the prior, as would often be reasonable; see Remark 4.1 below. Observe finally that the full density curve $\widehat{f}(x)$ is available once the $m-2$ estimates $\widehat{\gamma}_3, \ldots, \widehat{\gamma}_m$ have been arrived at.

The scheme thickens further when uncertainty about (μ, σ) is taken into account. The Bayes solution is the posterior mean of (4.1). Formulae for the exact solution can be written down following the route above, but quickly become intractable. The simplest practical solution is perhaps as follows: The procedure above gives a way of computing $\widehat{f}_m(x \mid \mu, \sigma)$ for each fixed (μ, σ) (pre-transform to $(x_i - \mu)/\sigma$, use the above, and back-transform). Then average this over a suitable and separately obtained posterior distribution for these parameters. In practical terms, this would mean simulating perhaps 100 values of (μ, σ) from the posterior distribution, and then compute the average of the $\widehat{f}_m(x \mid \mu, \sigma)$ curves. See (4.8) below for one possibility.

REMARK 4.1. The method is quite general, of course, and does not rely on the specifics of the Hermite expansion. Another attractive framework, for example, this time for densities on $[0, 1]$, uses

$$f_m(x) = 1 + \sum_{j=1}^m c_j\sqrt{2}\cos(j\pi x) \quad \text{with } c_j = \mathrm{E}_f\sqrt{2}\cos(\pi j X), \tag{4.4}$$

with individual or simultaneous priors placed on the collection of c_js. It should also be realised that as long as the order m is fixed the underlying orthogonality structure does not matter much either, as far as the mathematical and computational details are concerned; the Hermite expansion (4.1) is for example nothing but an mth order polynomial in $(x - \mu)/\sigma$, and can be represented as such. Nevertheless the orthogonal structure is appealing, since we get explicit representations of the coefficients in terms of the underlying density; they would otherwise change in value and interpretation when going from order m to order $m+1$, say. This representation in terms of coordinates for orthogonal basis functions also invites the user to think in terms of independent prior distributions for each coefficient, or perhaps a joint prior with a simple dependence structure. This translation of prior knowledge into separate priors for coordinates would sometimes be too forced and not relevant enough, of course, but it greatly simplifies the task as well as the resulting mathematical structure.

REMARK 4.2. The recursive method outlined above leads to exact estimates (4.3). Of course even this method can be too cumbersome or too slow, if $(n + 1)\cdots(n + m + 1)/(m!)$ is very large. A different way of computing (4.3) is to stick to the left hand side of (4.2), multiply with the prior density, and carry out the necessary numerical integrations in the $(\gamma_3, \ldots, \gamma_m)$ space, perhaps via simulation. — Yet another numerical approach can be based on statistical methods from survey sampling theory. Both numerator and denominator consists of $(m + 1)^n$ terms (although many of them are zero), each of which is easy to compute once selected. Terms can be sampled from these large sets, leading to estimates of numerator and

denominator. Good sampling regimes are those that catch big terms with high probability.

These comments, and indeed the method that led to (4.3), are also relevant for the problem of computing Bayes estimates in parametric mixture models.

4.2. THE ROBUST HERMITE EXPANSION. There are some difficulties with the (4.1) expansion. The γ_j coefficients are not always finite, and estimates, whether frequentist or Bayesian, are quite variable. Another and perhaps more serious disadvantage surfaces when one rewrites the underlying loss function that led to the (4.1) approximation as $\int (f-f_m)^2\phi^{-2}\,dx$, that is, integrated weighted squared error with weight function proportional to $\exp\{(x-\mu)^2/\sigma^2\}$. This would mean caring too much about what happens outside the mainstream area.

Another Hermite expansion that in several senses gives more robust estimation is developed in Fenstad and Hjort (1994). This robust parallel to (4.1) is

$$f_m(x) = \phi\Big(\frac{x-\mu}{\sigma}\Big)\frac{1}{\sigma}\sum_{j=0}^{m}\delta_j H_j\Big(\sqrt{2}\frac{x-\mu}{\sigma}\Big)/\sqrt{j!}, \tag{4.5}$$

where

$$\delta_j = \sqrt{2}\mathrm{E}_f H_j(\sqrt{2}(X-\mu)/\sigma)\exp\{-\tfrac{1}{2}(X-\mu)^2/\sigma^2\}/\sqrt{j!}. \tag{4.6}$$

Note that the function being averaged is bounded, so all coefficients automatically exist, and robust estimation is unproblematic. The normal density has $\delta_0 = 1$ and $\delta_j = 0$ for $j \geq 1$. A Bayesian model for a random density is once more to place a prior distribution for $(\delta_0,\ldots,\delta_m)$, for example having independent δ_js, and perhaps having distributions concentrated around zero to model a density in the vicinity of the normal curve. The reasoning of the previous subsection can now be applied with the necessary modifications to give a method for computing the Bayes estimate, the posterior mean of (4.5). See below for some of the details.

4.3. SPECIAL CONSTRUCTION. Results and examples in Fenstad and Hjort (1994) provide more information on typical sizes of δ_js, and hence on how priors for them should be selected. To give a reasonably concrete example we focus on creating a nonparametric model 'around' the normal distribution. The functions that when averaged as in (4.6) produce δ_js are all bounded, so there is a maximal interval $[a_j,b_j]$ in which δ_j must lie. This interval is $[0,\sqrt{2}]$ for $j=0$ and is symmetric around zero for j odd, say $[-b_j,b_j]$, and since the average values are typically closer to zero anyway a natural simplification is to place a symmetric smooth prior on say $[-c_j,c_j]$ for each $j \geq 1$, where $c_j = \min\{|a_j|,b_j\}$. These are bounded (by 1.213, in fact) and go slowly towards zero as j increases. One reasonable scheme is to let $\delta_j = c_j(2B_j-1)$ for $j \geq 1$ where the B_js are independent symmetric Beta variables with decreasing variances, say $B_j \sim \mathrm{Beta}(\beta_j,\beta_j)$; see Remark 4.1 again. Choose $\beta_j = (kj^2-1)/2$, where k is a parameter determining the amount of prior uncertainty; the variance of δ_j is $c_j^2/(kj^2)$. In addition there should be separate priors for (μ,σ) and for δ_0 around 1 in $[0,\sqrt{2}]$. Altogether this models a random density around the normal, which is the limiting case of a large k. Symmetry means $\delta_j = 0$ for j odd, so approximate symmetry could be modelled by tighter B_js for j odd than for j even.

The computation of the Bayes estimate for f is as follows. For given (μ,σ), let

$$\widehat{f}_m(x\,|\,\mu,\sigma)=\phi\Big(\frac{x-\mu}{\sigma}\Big)\frac{1}{\sigma}\sum_{j=0}^{m}\widehat{\delta}_j(\mu,\sigma)H_j\Big(\sqrt{2}\frac{x-\mu}{\sigma}\Big)/\sqrt{j!}, \tag{4.7}$$

where

$$\widehat{\delta}_j(\mu,\sigma)=\frac{\sum_{C(n)}M(k_0,\ldots,k_j+1,\ldots,k_m)S_n(k_0,\ldots,k_m\,|\,\mu,\sigma)}{\sum_{C(n)}M(k_0,\ldots,k_j,\ldots,k_m)S_n(k_0,\ldots,k_m\,|\,\mu,\sigma)},$$

where $S_n(k_0,\ldots,k_m\,|\,\mu,\sigma)$ is the sum of all products

$$(j_1!)^{-1/2}H_{j_1}(\sqrt{2}(x_1-\mu)/\sigma)\cdots(j_n!)^{-1/2}H_{j_n}(\sqrt{2}(x_n-\mu)/\sigma)$$

with k_0 cases of $j_i=0$, k_1 cases of $j_i=1$, and so on, and where $M(k_0,\ldots,k_m)=\mathrm{E}\,\delta_0^{k_0}\cdots\delta_m^{k_m}$. The $M(\cdot)$s are easily found, and the $S_n(\cdot)$s are to be computed recursively as explained in Section 4.1. A high portion of the terms are zero, namely those where there is an odd j for which k_j is odd. In the end the (4.7) estimator is averaged over say 100 realisations of (μ,σ) drawn from a separately obtained posterior distribution for these.

One possibility for this particular ingredient of the scheme is

$$\begin{pmatrix}\mu/\sigma^*\\ \log\sigma\end{pmatrix}\approx\mathcal{N}_2\{\begin{pmatrix}\mu^*/\sigma^*\\ \log\sigma^*\end{pmatrix},\frac{1}{n}\begin{pmatrix}1 & \frac{1}{2}\gamma_3^*\\ \frac{1}{2}\gamma_3^* & \frac{1}{2}+\frac{1}{4}\gamma_4^*\end{pmatrix}\}, \tag{4.8}$$

in which μ^*, σ^*, γ_3^* and γ_4^* are the usual (frequentist) estimates for mean, standard deviation, skewness and kurtosis. This is based on the approximate binormal sampling distribution for (μ^*,σ^*) given (μ,σ), and (4.8) results from starting with a flat reference prior for $(\mu,\log\sigma)$. It is still valid as an approximation with any other continuous positive prior, if n is large. The traditionally obtained posterior in a normal model, based on a normal-Gamma start for $(\mu,1/\sigma^2)$, is like in (4.8) above, but with skewness and kurtosis parameters equal to zero, since then the normal-model likelihood has been used. Thus (4.8) is the robust modification of the traditional posterior.

5. Log-linear expansions. The Bayes methods of the previous section ended up being quite cumbersome computationally, due to the large number of possible combinations when the likelihood was expanded. There are also mild problems related to the fact that the resulting estimates occasionally may give negative values and may not integrate to precisely 1. The present section develops some theory for Bayesian estimation of multiplicatively expanded densities instead.

5.1. BASIC FRAMEWORK. Assume observations fall in a given bounded interval. For a fixed set of continuous basis functions $\psi_1,\psi_2,\psi_3,\ldots$ on this interval, consider densities of the form

$$f(x,\mathbf{c})=a(\mathbf{c})^{-1}\exp\Big\{\sum_{j=1}^{m}c_j\psi_j(x)\Big\}, \tag{5.1}$$

where $a(\mathbf{c}) = \int \exp\{\sum_{j=1}^m c_j \psi_j(x)\}\,dx$. If m is small this is nothing but an ordinary parametric model for f, but our intention is to let m be potentially large and even possibly infinite. The model corresponds to an additive expansion of the log-density in terms of the basis functions. Our Bayesian programme is to start out with prior distributions for all the c_j coefficients and compute the posterior distribution of $f(x, \mathbf{c})$, given a sample $X_1, \ldots, X_n$.

For a simple concrete example of such a setup, suppose the data are scaled to fall in $[0,1]$, and let $\psi_j(x) = x^j$ for $j \geq 1$. Then (5.1) amounts to expanding the log-density as a polynomial. To model uncertainty around a normal density, say scaled so as to have at least 0.999 of its probability mass on $[0,1]$, we would have suitable priors on c_1 and c_2 (and c_2 would have to be negative) and in addition have priors concentrated around zero for $c_3, c_4, \ldots$.

Note that any continuous density can be approximated uniformly well in this way, as a consequence of the Stone–Weierstrass theorem. This happens also in many other situations, with a suitably engineered system of basis functions. We may also assume without loss of generality that these are uniformly bounded (since scale factors can be moved from ψ_j to c_j). Thus (5.1) defines a bona fide density also in the infinite case provided $\sum_{j=1}^\infty |c_j|$ is finite. In the fully nonparametric case we should therefore make sure that this series is convergent with probability 1 under prior model circumstances.

The comments made in Remark 4.1 about representation and interpretation of the coefficients are valid in the present framework too. It aids our understanding of the parameters, and therefore of the structure of the prior distribution to be set, if the basis functions are made orthogonal. If they have been chosen to satisfy $\int \psi_j \psi_k\,dx = I\{j = k\}$, then $c_j = \int \psi_j(x) \log f(x)\,dx$, and any given density can be expanded as in (5.1) with coefficients determined from this. This also suggests using a prior for these where the c_js are either independent or obey some simple dependence structure.

5.2. THE POSTERIOR MODE. We now proceed to discuss general technical matters related to the calculation of the Bayes estimate. Let $\widehat{\mu}_j = n^{-1} \sum_{i=1}^n \psi_j(X_i)$ be the empirical ψ_j mean and let

$$\mu_j(c_1, \ldots, c_m) = \int \psi_j(x) f(x, \mathbf{c})\,dx = \mathrm{E}_{\mathbf{c}} \psi_j(X)$$

be the theoretical mean, under the assumption of the model. The log-likelihood of the data is

$$n \sum_{j=1}^m c_j \widehat{\mu}_j - n \log a(c_1, \ldots, c_m),$$

which is easily shown to be a concave function of the parameters. The maximum likelihood estimators $\widehat{c}_1, \ldots, \widehat{c}_m$ are the unique solutions to

$$\widehat{\mu}_j = \mu_j(c_1, \ldots, c_m), \quad j = 1, \ldots, m. \tag{5.2}$$

Now consider a prior distribution $\pi(c_1, \ldots, c_m)$ for the c_j-parameters. The posterior distribution is then proportional to

$$\pi(c_1, \ldots, c_m) \exp\Big\{ n \sum_{j=1}^m c_j \widehat{\mu}_j - n \log a(c_1, \ldots, c_m) \Big\}. \tag{5.3}$$

In several situations it would be possible to simulate from this posterior distribution, thus allowing us to compute the conditional mean of the (5.1) density for each x, for example. In general, rather than computing this squared-error loss Bayes estimate, it is simpler to go for the mode of (5.3), giving the Bayes estimate $\mathbf{c}^*$ of $\mathbf{c}$ under a sharp 0–1 loss function, and then use $f(x, \mathbf{c}^*)$ in the end.

Finding the mode of (5.3) posterior density can be cumbersome when m is large, but one is helped by the fact that its logarithm is often exactly or approximately concave, which means that even simple-minded numerical optimisation schemes, like Newton–Raphson, will work. The log of (5.3) is exactly concave when $\log \pi(\mathbf{c})$ is concave and otherwise approximately concave if n is larger than m.

5.3. A QUADRATIC APPROXIMATION. An approximation which also sheds light on the structure of the solution emerges by Taylor expanding the $\log a(\mathbf{c})$ term of (5.3) around the maximum likelihood estimator $\widehat{\mathbf{c}}$. One finds

$$\log a(\mathbf{c}) \simeq \log a(\widehat{\mathbf{c}}) + \sum_{j=1}^{m} \frac{\partial \log a(\widehat{\mathbf{c}})}{\partial c_j}(c_j - \widehat{c}_j) + \tfrac{1}{2} \sum_{j,k} \frac{\partial^2 \log a(\widehat{\mathbf{c}})}{\partial c_j \partial c_k}(c_j - \widehat{c}_j)(c_k - \widehat{c}_k) + \tfrac{1}{6} \sum_{j,k,l} \frac{\partial^3 \log a(\widehat{\mathbf{c}})}{\partial c_j \partial c_k \partial c_l}(c_j - \widehat{c}_j)(c_k - \widehat{c}_k)(c_l - \widehat{c}_l).$$

Furthermore, as a consequence of the exponential form of (5.1) one also has

$$\begin{aligned}
\frac{\partial \log a(\mathbf{c})}{\partial c_j} &= \mathrm{E}_{\mathbf{c}} \psi_j(X) = \mu_j(\mathbf{c}), \\
\frac{\partial^2 \log a(\mathbf{c})}{\partial c_j \partial c_k} &= \mathrm{cov}_{\mathbf{c}}\{\psi_j(X), \psi_k(X)\} = \omega_{j,k}(\mathbf{c}), \\
\frac{\partial^3 \log a(\mathbf{c})}{\partial c_j \partial c_k \partial c_l} &= \mathrm{E}_{\mathbf{c}}\{\psi_j(X) - \mu_j(\mathbf{c})\}\{\psi_k(X) - \mu_k(\mathbf{c})\}\{\psi_l(X) - \mu_l(\mathbf{c})\} = \gamma_{j,k,l}(\mathbf{c}).
\end{aligned}$$

Disregarding additive constants, the log-posterior density is therefore approximately equal to

$$\log \pi(\mathbf{c}) - \tfrac{1}{2} n (\mathbf{c} - \widehat{\mathbf{c}})' \Omega(\widehat{\mathbf{c}})(\mathbf{c} - \widehat{\mathbf{c}}) - \tfrac{1}{6} n \sum_{j,k,l} \gamma_{j,k,l}(\widehat{\mathbf{c}})(c_j - \widehat{c}_j)(c_k - \widehat{c}_k)(c_l - \widehat{c}_l),$$

where $\Omega(\mathbf{c})$ is the covariance matrix for the $\psi_j(X)$ variables. To illustrate further, suppose the prior distribution for $\mathbf{c}$ is multinormal with mean $\mathbf{c}_0$ and precision matrix Ω_0, i.e. covariance matrix Ω_0^{-1}. In cases where n is moderately large, compared perhaps to the number of c_js with significantly spread-out prior distributions, the basic quadratic approximation will be satisfactory and we can shave off the third order terms. The posterior mode $\mathbf{c}^*$ is then close to the minimiser of

$$\tfrac{1}{2}(\mathbf{c} - \mathbf{c}_0)' \Omega_0 (\mathbf{c} - \mathbf{c}_0) + \tfrac{1}{2} n (\mathbf{c} - \widehat{\mathbf{c}})' \Omega(\widehat{\mathbf{c}})(\mathbf{c} - \widehat{\mathbf{c}}),$$

that is,

$$\mathbf{c}^* \simeq \{\Omega_0 + n\Omega(\widehat{\mathbf{c}})\}^{-1}\{\mathbf{c}_0 + n\Omega(\widehat{\mathbf{c}})\widehat{\mathbf{c}}\}. \tag{5.4}$$

5.4. SPECIAL CONSTRUCTION. Suppose $f_0(x)$ is some prior guess density and that its log-expansion is $\sum_{j=1}^{m} c_{0,j}\psi_j(x) - \log a(\mathbf{c}_0)$. If m is allowed to be infinity

and the system of basis functions is rich enough this expansion would be the exact $\log f_0(x)$. Then write

$$f(x,\mathbf{c}) = \frac{\exp\{\sum_{j=1}^m c_{0,j}\psi_j(x)\}}{a(\mathbf{c}_0)} \frac{\exp\{\sum_{j=1}^m (c_j - c_{0,j})\psi_j(x)\}}{a(\mathbf{c})/a(\mathbf{c}_0)} = f_0(x) r(x,\mathbf{c}).$$

This rewriting in terms of a prior guess times a correction function is conceptually helpful but not of mathematical importance for the final estimation method. The idea is to take $c_1, c_2, \ldots$ independent and place a prior on each coefficient, as per comments made at the end of Section 5.1. A simple strategy is to have something like $c_j \sim \mathcal{N}\{c_{0,j}, \tau^2/j^2\}$, where τ is a fixed parameter determining the amount of prior uncertainty. The Bayesian density estimate is $f(x,\mathbf{c}^*)$, where $c_1^*, c_2^*, \ldots$ maximise

$$\sum_{j=1}^m \{-\tfrac{1}{2} j^2 (c_j - c_{0,j})^2/\tau^2\} + n\sum_{j=1}^m c_j \widehat{\mu}_j - n \log \int \exp\Big\{\sum_{j=1}^m c_j \psi_j(x)\Big\}\, \mathrm{d}x.$$

This is again a concave function with a unique maximum. An approximation is provided by (5.4). Note that c_j^* becomes close to $c_{0,j}$ for all large j because of the increased precision j^2/τ^2.

This scheme can with some efforts be generalised to a more semiparametrically inspired strategy, with a parametrical $f_0(x,\xi)$ as initial description, and a nonparametric correction function $r(x,\xi,\mathbf{c})$ to estimate. This is not pursued here.

We have not been very specific about the number m of terms to include. It is neat from a puristic nonparametric point of view that the apparatus can handle $m = \infty$, but then the priors used for c_js for large js need to have small variances. The observation made above, about closeness of c_j^* to $c_{0,j}$, suggests that even in cases with an infinity of priors, stopping at a moderate m could constitute a satisfactory numerical approximation. One natural scheme is to compute estimates for a numbers of ms, perhaps for m up to the sample size n, and then select one of them according to a suitable criterion. Such a criterion is the Bayesian information criterion of Schwarz (1978), choosing the model that maximises

$$\text{SIC}(m) = n\sum_{j=1}^m \widehat{c}_j \widehat{\mu}_j - n \log a(\widehat{c}_1, \ldots, \widehat{c}_m) - (\tfrac{1}{2}\log n)m.$$

Here $\widehat{c}_1, \ldots, \widehat{c}_m$ are the maximum likelihood estimates again, but the arguments used in Schwarz' proof of his main result also allow these to be replaced by the Bayesian mode estimates $c_1^*, \ldots, c_m^*$. The information criterion $\text{SIC}(m)$ is really based on large-sample approximations, and it should also be possible to work out useful finite-sample modifications in the present situation. The SIC criterion is similar to but more 'stingy' than the often used Akaike information criterion; Schwarz's criterion is less willing to allow many parameters.

The normal prior specification above gives an exponent function which is a Gaußian process. As such the method outlined here is related to the ones worked with in Lenk (1991, 1993), in spite of having a different starting point. See also Leonard (1978) and Thorburn (1986) for similarly spirited approaches. Lenk (1991) starts with such a Gaußian process for the exponent and uses Karhunen–Loéve representations to translate this into a specific form of (5.1), with a simple form for the

simultaneous prior for the coefficients. The present framework and methods are in several ways more general than in Lenk (1993), since non-normal priors easily can be used for the coefficients. The idea of building uncertainty around the normal density, for example, in the situation where the log-density is expanded like a polynomial, would typically require having non-normal priors for at least some of the coefficients.

6. Local priors on local parametric approximations. The basic idea of a semiparametric approach to (non-Bayesian) density estimation recently developed by Hjort and Jones (1994) is to work with parametric models, but only trusting them locally. That is, a parametric vehicle model $f(t,\theta)$ is used for t in a neighbourhood of a given x, and the final density estimate is of the form $f(x,\widehat{\theta}(x))$, where $\widehat{\theta}(x)$ is based only on data local to x. This aims at the best local parametric approximation to the true density. In this section semiparametric Bayesian analogues are developed, the idea being to use local priors on the local parameters.

6.1. LOCAL LIKELIHOOD FOR DENSITIES. Let $f(t,\theta)$ be a suitable parametric class of densities. The ordinary likelihood is of course $\prod_{i=1}^n f(x_i,\theta)$. But this conveys the information content of the data only when the model can be trusted fully. Suppose the density is modelled only locally, say $f(t)=f(t,\theta)$ for t in the window cell $C(x)=[x-\frac12 h, x+\frac12 h]$. The simplest modification of the likelihood would be to only include terms with $x_i\in C(x)$, but this is *not* correct; examples illustrate that this is an inadequate measure of information content, and that its maximiser is an unsatisfactory estimator of the parameter. A more appropriate modified likelihood is the likelihood that only uses information about X_is to the right of $x-\frac12 h$ and what happens to these during $[x-\frac12 h, x+\frac12 h]$. In other words, the appropriate distribution is the conditional one given $X_i\geq x-\frac12 h$, which is $f(t,\theta)/S(x-\frac12 h,\theta)$ for $t\in C(x)$, in terms of the survival function $S=1-F$, and with probability of further surviving $x+\frac12 h$ equal to $S(x+\frac12 h,\theta)/S(x-\frac12 h,\theta)$. Using $S(x,\theta)/S(a,\theta)=\exp[-\int_a^x\{f(t,\theta)/S(t,\theta)\}\,\mathrm{d}t]$, valid for $x\geq a$, this leads with further manipulations to

$$\begin{aligned}
L_{0,n}(x,\theta) &= \prod_{x_i\in C(x)}\frac{f(x_i,\theta)}{S(x-\frac12 h,\theta)}\prod_{x_i>x+\frac12 h}\frac{S(x+\frac12 h,\theta)}{S(x-\frac12 h,\theta)}\\
&= \prod_{x_i\in C(x)}\frac{f(x_i,\theta)}{S(x_i,\theta)}\prod_{x_i>x-\frac12 h}\frac{S(\min\{x_i,x+\frac12 h\},\theta)}{S(x_i-\frac12 h,\theta)}\\
&= \Big\{\prod_{x_i\in C(x)} f(x_i,\theta)\Big\}\exp\Big[-n\int_{C(x)}\Big\{\log S(t,\theta)\,\mathrm{d}F_n(t)+S_n(t)\frac{f(t,\theta)}{S(t,\theta)}\,\mathrm{d}t\Big\}\Big],
\end{aligned}$$

where $S_n=1-F_n$ is the empirical survival function. This is at the moment the exact likelihood of all data with $X_i\geq x-\frac12 h$, using information about what happens during $[x-\frac12 h, x+\frac12 h]$, and satisfactory Bayesian and non-Bayesian methods can be developed with $L_{0,n}$ as basis, see Hjort (1994a). This machinery tends to work better with hazard rates than for densities, however, and it turns out to be fruitful to work instead with a convenient approximation. This approximation emerges by

replacing the model-based $S(t,\theta)$ with the nonparametric $S_n(t)$, giving the *local likelihood* for the local data in $[x-\frac{1}{2}h, x+\frac{1}{2}h]$,

$$L_n(x,\theta) = \Big\{ \prod_{x_i \in C(x)} f(x_i,\theta) \Big\} \exp\Big\{ -n \int_{C(x)} f(t,\theta)\,\mathrm{d}t \Big\} \tag{6.1}$$

(ignoring a multiplicative factor not dependent on the parameter). This can also be written

$$L_n(x,\theta) = \Big\{ \prod_{i=1}^n f(x_i,\theta)^{\bar K(h^{-1}(x_i-x))} \Big\} \exp\Big\{ -n \int \bar K(h^{-1}(t-x)) f(t,\theta)\,\mathrm{d}t \Big\}, \tag{6.2}$$

where $\bar K(z) = 1$ on $[-\frac{1}{2},\frac{1}{2}]$ and zero elsewhere. Note that the scaled version $\bar K(h^{-1}(t-x))$ has support $[x-\frac{1}{2}h, x+\frac{1}{2}h]$. We shall also use (6.2) with more general kernel functions $\bar K(z)$, only requiring that they are smooth around zero and have 'correct level' $\bar K(0) = 1$. This also translates into $\bar K(h^{-1}(t-x))$ close to 1 for t near x. Typically $\bar K(z) = K(z)/K(0)$ for a symmetric unimodal probability density kernel function K.

These matters are further discussed in Hjort (1994a) and Hjort and Jones (1994), including other reasons favouring (6.1) and (6.2) for density estimation. See also Loader (1993). We call (6.2) the *kernel smoothed local likelihood* at x, and view it as carrying the local information on the local parameter θ. The argument is that it is a natural smoothing generalisation of (6.1) and that the local parametric model built around x is sometimes trusted a little less a little distance from x than at x itself. The requirement about 'correct level' for $\bar K$ is important. A scale factor in $\bar K$ does not matter for estimation theory based on maximisation of such local likelihoods, see Hjort and Jones (1994), but Bayesian consequences in what follows rest on the fact that if $\pi(\theta)$ is a prior for θ, then the simultaneous density for θ and local data is approximately proportional to $\pi(\theta)L_n(x,\theta)$. Note also that when h is large the local kernel smoothed likelihood becomes $\{\prod_{i=1}^n f(x_i,\theta)\}\exp(-n)$, which is the usual full likelihood (apart from the $\exp(-n)$ factor which is independent of the parameter). Thus ordinary parametric likelihood methods, whether frequentist or Bayesian, simply correspond to the large h case.

Before passing to the Bayesian consequences we note that, for the case of $\bar K(z) = K(z)/K(0)$,

$$\begin{aligned} f_n(x) &= n^{-1}\sum_{i=1}^n h^{-1}K(h^{-1}(x_i-x)) = K(0)(nh)^{-1}\sum_{i=1}^n \bar K(h^{-1}(x_i-x)), \\ g_n(x) &= n^{-1}\sum_{i=1}^n h^{-3}(x_i-x)K(h^{-1}(x_i-x)) \\ &= K(0)(nh^3)^{-1}\sum_{i=1}^n (x_i-x)\bar K(h^{-1}(x_i-x)). \end{aligned} \tag{6.3}$$

Here $f_n(x)$ is the classical kernel estimator already encountered in equation (2.3) and later on, while $g_n(x)$ similarly is a K-based kernel type estimator for the density derivative $f'(x)$ times the constant σ_K^2 (the variance of K). When the kernel used is $K = \phi$, the standard normal, $g_n(x)$ is simply the derivative of $f_n(x)$.

6.2. LOCAL BAYES ESTIMATES. Whereas Hjort and Jones (1994) maximise these local likelihoods and develop theory and special cases for the resulting $f(x, \widehat{\theta}(x))$ estimators, the present aim is to develop Bayesian estimators of the form

$$\widehat{f}(x) = \mathrm{E}\{f(x,\theta) \,|\, \text{local data}\}. \tag{6.4}$$

The posterior distribution in question is taken to be $\pi(\theta)L_n(x,\theta)/\int \pi(\theta)L_n(x,\theta)\,\mathrm{d}\theta$, where $\pi(\theta)$ is the prior density for θ. A fuller verbal description of (6.4) could be 'a locally parametric nonparametric Bayesian density estimator', and another informative mouthful is 'a nonparametric Bayesian estimator of the locally best parametric approximant to the true density'. The approximation in question here is in terms of a localised form of the Kullback–Leibler distance, see Hjort and Jones (1994). When h is large the local likelihood becomes the ordinary one and (6.4) is the familiar predictive density. When h is small the estimator is essentially nonparametric in nature.

Note that θ changes interpretation with x; for each new and temporarily fixed x there is a new parametric approximation to f near x and a new prior for the best fitting parameter $\theta = \theta_x$. See also Section 8.5.

It is also fruitful to generalise this method of local parameters to one with both global and local parameters present, in the spirit of two-stage priors or hierarchical Bayesian methods. One such framework is to model the density as $f(x,\xi,\theta)$, where ξ is a global 'background' parameter with prior $\pi_0(\xi)$ and θ is local to x. For each ξ the method above applies and gives $\widehat{f}(x,\xi) = \mathrm{E}\{f(x,\xi,\theta) \,|\, \text{local data}\}$. The final estimator is then to average this over the posterior density for ξ, say

$$\begin{aligned} \widehat{f}(x) &= \mathrm{E}\{\widehat{f}(x,\xi) \,|\, \text{all data}\} \\ &= \int \Big\{ \int f(x,\xi,\theta)\pi(\theta \,|\, \text{local data})\,\mathrm{d}\theta \Big\} \pi_0(\xi \,|\, \text{all data})\,\mathrm{d}\xi. \end{aligned} \tag{6.5}$$

Some of the examples below are of this sort.

7. Locally parametric Bayesian density estimators: Special cases. The development above, ending via the local likelihood (6.2) in the (6.4) and (6.5) estimators, is of course quite general, and there is a variety of different specialisations of the method. Below is a partial list of interesting special cases. We stress again that the estimators have nonparametric intentions, in spite of the fact that they use local parameterisations. When the smoothing parameter is large we are back to ordinary Bayesian fully parametric methods.

7.1. LOCAL CONSTANT WITH A GAMMA PRIOR. Let the local model simply be a constant, $f(t,\theta) = \theta$ for t in a neighbourhood around x. This is unrealistic as a fine description of the density, but makes sense locally; the main aim is to get hold of the 'local level' for f. Let furthermore θ have a Gamma prior $\{cf_0(x), c\}$, say, with prior mean $f_0(x)$ and prior variance $f_0(x)/c$. Via (6.3) the local likelihood (6.2) is seen to take the form

$$\theta^{nhf_n(x)/k_0} \exp\{-nh\theta/k_0\},$$

writing for simplicity $k_0 = K(0)$. It follows that θ given the local data is Gamma $\{cf_0(x) + nhf_n(x)/k_0, c + nh/k_0\}$, leading to

$$\widehat{f}(x) = \frac{cf_0(x) + nhf_n(x)/k_0}{c + nh/k_0}, \tag{7.1}$$

a weighted average of prior guess and kernel estimator. This is exactly as in (2.3), if the value for the prior strength parameter c used is ah/k_0.

7.2. LOCAL CONSTANT WITH A TWO-STAGE PRIOR. To generalise, let us keep the constant level model $f(t,\theta) = \theta$ for t near x, but let us employ a two-stage prior for θ: θ given a certain background parameter ξ is a Gamma $\{cf_0(x,\xi), c\}$, and ξ has a separate background prior $\pi_0(\xi)$. We think of $f_0(x,\xi)$ as the background model, which is next to be corrected on by the data. We have

$$\mathrm{E}\{f(x,\theta) \,|\, \text{local data}, \xi\} = \frac{c}{c + nh/k_0} f_0(x,\xi) + \frac{nh/k_0}{c + nh/k_0} f_n(x),$$

and the final Bayes estimator is

$$\widehat{f}(x) = \frac{c}{c + nh/k_0} \int f_0(x,\xi)\pi_0(\xi \,|\, \text{data})\, \mathrm{d}\xi + \frac{nh/k_0}{c + nh/k_0} f_n(x). \tag{7.2}$$

This linearly combines the predictive estimator of the parametric prior guess density and the nonparametric kernel estimator, and is quite similar to the (2.7) estimator that was derived in a quite different framework. It is not difficult to find the first term explicitly when the background model is Gaußian and the $(\mu, 1/\sigma^2)$ is given the conjugate normal-Gamma prior, for example.

7.3. LOCAL LEVEL AND LOCAL SLOPE. Another generalisation is to incorporate both local level and local slope in the vehicle model, say $f(x,\theta,\beta) = \theta \exp\{\beta(t-x)\}$ for t close to x. Let β have a prior $\pi(\beta)$ reflecting prior beliefs about the local slope at x, and let θ be a Gamma $\{cf_0(x), c\}$ (where the c parameter could depend on β). The local likelihood can be written

$$\theta^{nhf_n(x)/k_0} \exp\{-nh\theta\psi(\beta h)/k_0\} \exp\{\beta nh^3 g_n(x)/k_0\},$$

in view of (6.3) again, and where $\psi(\beta h) = \int K(z) \exp(\beta h z)\, \mathrm{d}z$. Thus

$$\theta \,|\, \text{local data}, \beta \sim \text{Gamma}\{cf_0(x) + nhf_n(x)/k_0, c + nh\psi(\beta h)/k_0\},$$

and

$$\mathrm{E}\{f(x,\theta,\beta) \,|\, \text{local data}, \beta\} = \frac{cf_0(x) + nhf_n(x)/k_0}{c + nh\psi(\beta h)/k_0}.$$

The Bayes solution $\widehat{f}(x)$ is to average this over the posterior distribution for β given the local data.

The local posterior for (β,θ) is proportional to

$$\pi(\beta)\, \theta^{cf_0(x) + nhf_n(x)/k_0 - 1} \exp[-\theta\{c + nh\psi(\beta h)/k_0\}] \exp\{nh^3 \beta g_n(x)/k_0\}.$$

Integrating out θ the result is proportional to

$$\pi(\beta)\frac{\exp\{nh^3\beta g_n(x)/k_0\}}{\{c+nh\psi(\beta h)/k_0\}^{cf_0(x)+nhf_n(x)/k_0}}.$$

If $K=\phi$, then $\psi(\beta h)=\exp(\frac{1}{2}\beta^2h^2)$, and an approximation gives

$$\frac{\exp\{cf_n(x)\exp(-\frac{1}{2}\beta^2h^2)\}}{\{\exp(\frac{1}{2}\beta^2h^2)+ck_0/(nh)\}^{cf_0(x)}}\pi(\beta)\exp\{-\tfrac{1}{2}nh^3f_n(x)\beta^2/k_0+nh^3\beta g_n(x)/k_0\}.$$

Supposing c to be small compared to nh, and letting the prior for β be a normal $(\beta_0,1/w_0^2)$, the resulting local posterior for β is approximately proportional to

$$\exp\big[-\tfrac{1}{2}w_0^2(\beta-\beta_0)^2-\tfrac{1}{2}nh^3f_n(x)\{\beta-g_n(x)/f_n(x)\}^2/k_0\big],$$

that is,

$$\beta\,|\,\text{local data}\approx\mathcal{N}\Big\{\frac{w_0^2\beta_0+nh^3f_n(x)\{g_n(x)/f_n(x)\}/k_0}{w_0^2+nh^3f_n(x)/k_0},\frac{1}{w_0^2+nh^3f_n(x)/k_0}\Big\}. \tag{7.3}$$

Note that the mean is a convex combination of β_0 and the natural nonparametric estimate of the log-derivative of the density. To a first order approximation the β given local data is a normal with mean $g_n(x)/f_n(x)$ and variance $\{nh^3f_n(x)/k_0\}^{-1}$. The final approximation for the density estimate itself becomes

$$\widehat{f}(x)\simeq f_n(x)\,\mathrm{E}\exp\{-\tfrac{1}{2}{}^2h^2\beta^2\,|\,\text{local data}\}\simeq f_n(x)\frac{\exp\{-\frac{1}{2}h^2\bar\mu^2/(1+h^2\bar\sigma^2)\}}{(1+h^2\bar\sigma^2)^{1/2}},$$

where $\bar\mu$ and $\bar\sigma^2$ are the newly found posterior parameters for β. This is similar to what Hjort and Jones (1994, Section 5.2) found for the maximum local likelihood density estimate for this local log-linear model.

7.4. LOCAL LEVEL, SLOPE AND CURVATURE. With efforts the previous example can be generalised to a model for local level, local slope and local curvature. That is, employ $\theta\exp\{\beta(t-x)+\frac{1}{2}\gamma(t-x)^2\}$ as vehicle model for t near x, give (β,γ) a normal prior, and let θ be a Gamma centred at some $f_0(x)$ again. The calculations are as above but become more complicated, and in addition to $f_n(x)$ and $g_n(x)$ they will involve an estimate of the second derivative of f. The end result is a Bayesian parallel to the maximum local likelihood density estimator found in Hjort and Jones (1994, Section 5.3) for the local log-quadratic model.

These constructions can also be generalised to the situation where the prior guess density involves a background parameter, just as estimator (7.2) generalised estimator (7.1).

7.5. PRIOR GUESS TIMES A LOCAL CONSTANT. This time write the density as a prior guess times a correction function which must then be locally estimated. With a local constant for this purpose this means using $f(t,\theta)=f_0(x)\theta$ for t near x. Let θ have a Gamma distribution around 1, say with parameters (c,c) (in particular

this means that $f(x,\theta)$ is seen as a Gamma with $\{c, c/f_0(x)\}$, and a differently structured variance than in 7.1 above). The local likelihood is

$$L_n(x,\theta) = \prod_{i=1}^n \{f_0(x_i)\theta\}^{\bar K(h^{-1}(x_i-x))} \exp\Big\{-n\theta \int \bar K(h^{-1}(t-x)) f_0(t)\,\mathrm{d}t\Big\}$$
$$= \prod_{i=1}^n \{f_0(x_i)\}^{\bar K(h^{-1}(x_i-x))}\, \theta^{nhf_n(x)/k_0} \exp\Big\{-nh\theta \int K(z) f_0(x+hz)\,\mathrm{d}z\Big\},$$

which leads to a different type of posterior for θ than in special case 7.1, namely

$$\theta \,|\, \text{local data} \sim \text{Gamma}\Big\{c + nhf_n(x)/k_0, c + nh\int K(z) f_0(x+hz)\,\mathrm{d}z/k_0\Big\}.$$

The density estimator becomes

$$\widehat f(x) = f_0(x) \frac{c + nhf_n(x)/k_0}{c + nh\int K(z) f_0(x+hz)\,\mathrm{d}z/k_0}. \tag{7.4}$$

This pushes the kernel estimator downwards in regions where f_0 is convex and upwards in regions where f_0 is concave.

Again extensions are possible, to two-stage priors with a global parameter in $f_0(x,\xi)$, and to the local log-linear model for f/f_0 with a local slope parameter in addition to θ.

7.6. A BETTER CORRECTION FACTOR FUNCTION. Again write $f(t,\theta) = f_0(t)\theta$ for a prior guess density f_0 and a local correcting constant, but this time consider using the alternative kernel function $\bar K(z) = k_0^{-1} K(z) f_0(x)/f_0(x+hz)$, that is, $\bar K(h^{-1}(t-x)) = k_0^{-1} h K_h(t-x) f_0(x)/f_0(t)$ for t near x. Then the local kernel smoothed likelihood becomes proportional to

$$\theta^{nhf_0(x)r_n(x)/k_0} \exp\Big\{-n\theta \int f_0(t) K(h^{-1}(t-x)) \frac{f_0(x)}{f_0(t)}\,\mathrm{d}t/k_0\Big\}$$
$$= \theta^{nhf_0(x)r_n(x)/k_0} \exp\{-nh\theta f_0(x)/k_0\},$$

where $r_n(x) = n^{-1}\sum_{i=1}^n K_h(x_i - x)/f_0(x_i)$ is the natural nonparametric kernel estimator of the correction factor function $r(x) = f(x)/f_0(x)$. The Bayesian density estimator becomes

$$\widehat f(x) = f_0(x) \frac{c + nhf_0(x) r_n(x)/k_0}{c + nhf_0(x)/k_0}.$$

This is close to $f_0(x) r_n(x)$, which is the simplest form of a class of density estimators recently developed by Hjort and Glad (1994), all of the form initial parametric start estimate times nonparametric correction factor.

An extension of the above is to write $f(t) = f(t,\xi)\theta$ for t near x, where θ is local to x and ξ is a global parameter, as with (6.5). The $f(t,\xi)$ could for example be a normal with a prior on (μ,σ). This leads to an estimator of the form

$$\widehat f(x) = \int f(x,\xi) \frac{c + nhf(x,\xi) r_n(x,\xi)/k_0}{c + nhf(x,\xi)/k_0} \pi_0(\xi \,|\, \text{data})\,\mathrm{d}\xi, \tag{7.5}$$

where $r_n(x,\xi) = n^{-1}\sum_{i=1}^n K_h(x_i - x)/f(x_i,\xi)$. If in particular c goes to zero, arguably corresponding to a noninformative prior for the local constant, then

$$\begin{aligned}\widehat{f}(x) &= \int f(x,\xi) r_n(x,\xi)\pi_0(\xi \mid \text{data})\,\mathrm{d}\xi \\ &= n^{-1}\sum_{i=1}^n K_h(x_i - x)\,\mathrm{E}\Big\{\frac{f(x,\xi)}{f(x_i,\xi)}\,\Big|\,\text{data}\Big\}.\end{aligned} \tag{7.6}$$

This is simply the classical kernel estimator when the prior model $f(x,\xi)$ is flat and noninformative, and otherwise aims to correct the kernel estimator so as to have smaller bias in a broad neighbourhood of the parametric model. The (7.5) and (7.6) estimators are Bayesian predictive versions of the $f(x,\widehat{\xi})r_n(x,\widehat{\xi})$ type estimator. In Hjort and Glad (1994) these semiparametric estimators have been shown to have frequentist properties generally comparable to and often better than those of the kernel estimator.

7.7. A RUNNING NORMAL DENSITY. This time let us try to estimate a 'running normal density'. The local model is now a normal (μ,σ^2), and the density estimate is

$$\widehat{f}(x) = \frac{1}{\sqrt{2\pi}}\int\int \frac{1}{\sigma}\exp\{-\tfrac{1}{2}(x-\mu)^2/\sigma^2\}\,\pi(\mu,\sigma \mid \text{local data})\,\mathrm{d}\mu\,\mathrm{d}\sigma. \tag{7.7}$$

To illustrate without too many technicalities we take σ known. Using the standard normal kernel the local likelihood is proportional to

$$\exp\Big\{-\tfrac{1}{2}\sum_{i=1}^n \frac{(x_i-\mu)^2}{\sigma^2}\frac{\phi(h^{-1}(x_i-x))}{\phi(0)}\Big\}\exp\Big\{-\frac{nh}{\phi(0)}\phi\Big(\frac{x-\mu}{\sqrt{\sigma^2+h^2}}\Big)\frac{1}{\sqrt{\sigma^2+h^2}}\Big\}.$$

Note again that a large h gives back the ordinary full likelihood, maximised for $\mu = \bar{x}$, the mean of the data. Using (6.3) again this is proportional to

$$\exp\Big[-\tfrac{1}{2}\frac{1}{\sigma^2}\frac{nhf_n(x)}{\phi(0)}\Big\{\delta - h^2\frac{g_n(x)}{f_n(x)}\Big\}^2\Big]\exp\Big\{-\frac{nh}{\sqrt{\sigma^2+h^2}}\exp\Big(-\tfrac{1}{2}\frac{\delta^2}{\sigma^2+h^2}\Big)\Big\},$$

in terms of $\delta = \mu - x$. If the local μ is given a normal (μ_0,τ^2) prior, for example, then this gives the posterior density. The (7.7) estimator is in the end found from numerical integration.

8. Supplementing remarks. This final section gives various additional results and remarks, several of which point to questions for further research.

8.1. FINE-TUNING OF PARAMETERS. Several of the estimators arrived at in the paper depend on one or more parameters, to be decided on for each application. Some of these parameters relate to specification of the prior distribution and others are 'smoothing parameters'. Parameters in the prior should ideally be set via the infamous 'prior considerations', perhaps explicitly involving previous data of similar nature, but can otherwise sometimes be estimated in an empirical Bayes manner. Using a hyper-prior and averaging over its posterior is another method, and this should usually be quite robust. It is also of interest to consider 'noninformative'

reference type priors. For the estimators of Sections 2 and 3 the natural version of this is to let the prior strength parameter a tend to zero. This leads to the traditional kernel estimator itself for cases (2.3) and (2.7), to a variable bandwidth kernel estimator using residuals in (2.9), to kernel estimators corrected for level in Section 3, and to a version of maximum penalised likelihood in (2.6). In particular this lends some Bayesian support to the kernel estimator and to the other limiting versions mentioned. Similarly, for the schemes in Section 7 that use Gamma priors with strength parameter c, the natural reference prior corresponds to letting c tend to zero. Again this is seen to lead to the kernel estimator for cases (7.1), (7.2), and to interesting competing versions for cases (7.4), (7.5), (7.6).

The h of Sections 2 and 3 is an example of an external smoothing parameter, perhaps more appropriately seen as an algorithmic parameter governing the amount of smoothing than a parameter of a statistical model. Bayesian versions of cross validation criteria can be invented for the purpose. The h of Sections 6 and 7 is also such a smoothing parameter, but it can also be interpreted in the context of a Bayesian model; if the kernel K in question is scaled so as to have support $[-\frac{1}{2}, \frac{1}{2}]$, for example, then $[x - \frac{1}{2}h, x + \frac{1}{2}h]$ is the interval around x in which the parametric approximation is believed to be adequate. As such the length of this interval of adequacy can be given a prior, and so on. If a data-driven method is wished for it is perhaps more natural, however, to use a suitable local goodness of fit criterion; the interval is stretched until the parametric model bursts, and this defines the h to be used for the given x. For an example, suppose a 'local constant' method is to be used, as in Sections 7.1, 7.2, 7.5 or 7.6. If the true density is constant over an interval $[a, b]$, then process convergence

$$\sqrt{n}\Big\{\frac{F_n[a,x]}{F_n[a,b]} - \frac{x-a}{b-a}\Big\} \to_d W^0\Big(\frac{x-a}{b-a}\Big)/F[a,b]^{1/2} \quad \text{for } a \le x \le b$$

can be demonstrated. Here $F_n[a,x]$ is the relative number of data points in $[a,x]$, and so on, and W^0 is the Brownian bridge. One can prove from this that the test which accepts $[a,b]$ as an interval of constancy of the underlying density, whenever

$$\frac{F_n[a,b]^{1/2}}{\sqrt{n}} \sum_{a \le x_i \le b} \Big|\frac{F_n[a,x_i]}{F_n[a,b]} - \frac{x_i - a}{b-a}\Big| \le 0.499,$$

has significance level about 10% (the 0.499 number is the upper 10% point of the distribution of $\int_0^1 |W^0(t)|\,\mathrm{d}t$). This can now be used for $x \pm \frac{1}{2}h$ intervals (with h above a suitable minimum). In the end the resulting h_x values, for the chosen xs at which the final estimator is to be computed, should be post-smoothed.

8.2. PERFORMANCE. With various starting constructions we have been able to give recipes for the computation of Bayes estimates. Non- and semiparametric Bayesian constructions are often so technically complicated that even this is sometimes an achievement. The question of performance analysis is typically even harder, and is presumably the reason why this aspect of the method is too often neglected. Different Bayesians might perhaps wish to stress different aspects of performance, but 'analysing performance' in the present context could for example mean studying exact or approximate risk functions under pointwise or integrated squared error

loss, that is, frequentist behaviour for given candidate densities. This might involve (a) assessing approximate biases and variances; (b) comparing these with those of standard estimators like the kernel method, for densities that are likely and not so likely under the prior; (c) proving or disproving large-sample consistency and normality; (d) assessing the reduction, if any, in terms of Bayes risk, compared to that of standard methods, both under the ideal prior that led to the estimator in question but also under other priors. These comments also point to simulation as a natural tool.

Most estimators derived in Section 2 can be analysed like this, at least for large n and small h. These estimators are typically asymptotically equivalent to appropriate kernel estimators, if the prior is held fixed, and the large sample theory for kernel estimators is very well developed. Estimators from Section 3 can also be analysed with suitable extensions of standard tools. Some but not all of the estimators from Section 6 and 7 can be analysed, for moderate to large n and small h, using methods of Hjort and Jones (1994). The estimators that rely on both global and local parameters would require more delicate tools for their analysis.

Estimators from Sections 4 and 5 are not so easily analysed. The task is not terribly hard if the expansion order m is fixed and small compared to n, but otherwise we find ourselves in need of more theory, particularly when m is allowed to be infinite. To illustrate, let us pose a 'typical question' about such estimators: For the (4.4) model mentioned in Remark 4.1, with m infinite, place independent priors $c_j = \sqrt{2}(2B_j - 1)$ on the c_js, where the B_js are symmetric Beta variables with decreasing variances, say $\mathrm{Var}\, c_j = k/j^2$. This assures convergence of the infinite expansion with probability 1. What is the behaviour of the resulting Bayes estimator $1 + \sum_{j=1}^{\infty} \widehat{c}_j \sqrt{2} \cos(j\pi x)$?

8.3. BAYESIAN HAZARD RATE AND REGRESSION CURVE ESTIMATION. Many of the methods and results presented here have parallels in the problem of estimating hazard rates in survival analysis and in more general counting process models for life history data. See Hjort (1991a, 'third general method') and Hjort (1991b, Section 8) for two frameworks involving nonparametric randomness around parametric models, the latter also extending to such Bayesian uncertainty around the semiparametric Cox regression model. In these settings, which are analogous to the present paper's Sections 2 and 3, Beta processes, a generalised class of hazard function relatives of the Dirichlet (Hjort, 1990), play the natural role. A Bayesian locally parametric approach for hazard rates estimation, analogous to Sections 6 and 7, should not be difficult to develop, with the local likelihood machinery already present in Hjort (1994a).

The local likelihood machinery of Sections 6 and 7 also has natural parallels in Bayesian nonparametric regression; model the unknown regression curve as being locally linear, place normal priors on the local line parameters, and compute the posterior given local data using local likelihood. The noninformative prior versions of this scheme correspond to recently developed local polynomial regression methods that have been shown to have very good performance properties, see for example Fan and Gijbels (1992). Empirical Bayesian and hierarchical Bayesian schemes can be developed as well, and these could easily perform better than standard methods in situations with several covariates. See Hjort (1994b).

8.4. FURTHER PROBLEMS. Standard density estimation methods work reasonably well in all reasonably populated areas. That is, they perform well in dimensions 1 and possibly 2, but often not at all in higher dimensions (the curse of dimensionality is precisely that no areas are well populated in higher dimensions), and often not well in the tails. These problems have not been focussed on in the present paper, but are presumably areas where the Bayesian method might improve significantly on standard methods. One needs a broader range of prior distributions that reflect various notions of what are likely and not so likely densities. Challenges include building Bayesian methods that are geared towards for example approximate unimodality or approximate bimodality (and for this mixtures approaches would be appropriate). Grander problems, where the Bayesian viewpoint is well worth exploring to a fuller extent than hitherto, include statistical pattern recognition, non- and semiparametric regression (particularly with many covariates), and neural networks.

8.5. FULL BAYESIAN MODELS FOR LOCAL PARAMETRIC ESTIMATION. Our locally parametric method requires a local model and a local prior around each x. A full global model for the complete density curve requires in addition a description of how these underlying local parameters change with x, but this was not necessary as far as the computation of the estimate is concerned. It is nevertheless of interest to build such a fuller stochastic process framework, which also would be valuable when it comes to estimation of prior parameters, and for performance evaluation, cf. comments made in 8.1 and 8.2 above. This is not an easy task. To illustrate, take the simplest of the special cases considered, that of Section 7.1. Take the local constant θ_x to be the result of a smoothed Gamma process with independent increments, say $\theta_x = T(x - \frac{1}{2}\,\mathrm{d}x, x + \frac{1}{2}\,\mathrm{d}x)/\mathrm{d}x$ with a certain small smoothing resolution $\mathrm{d}x$, where $T(a,b)$ is Gamma distributed with parameters $\{c_0 F_0(a,b), c_0\}$ for each given (a,b) interval. There is such a process by Kolmogorov's consistency theorem and the additive property of Gamma variables with equal shape parameter. Hence θ_x is approximately a Gamma with parameters $\{c_0 f_0(x)\,\mathrm{d}x, c_0\,\mathrm{d}x\}$, with mean $f_0(x)$ and variance $f_0(x)/(c_0\,\mathrm{d}x)$. This fits in with the treatment in Section 7.1. There are alternative constructions of interest and relevance, and the other special cases in Section 7 require various extensions.

8.6. THE LINEAR SEMIPARAMETRIC ESTIMATOR. Results (2.3), (2.7) and (7.1)–(7.2) give Bayesian support to estimators of similar form that were considered in non-Bayesian frameworks by both Schuster and Yakowitz (1985) and Olkin and Spiegelman (1987). Jones (1994) and others have pointed to certain difficulties with such estimators, primarily because the mixing parameter is somewhat imprecisely defined and difficult to estimate. In the present Bayesian framework the mixing parameter has a clear interpretation, however. For the case of (7.2), for example, a natural suggestion is to estimate c by empirical Bayesian techniques and choose h by a suitable local goodness of fit criterion. We leave the finer details for future work.

8.7. INCONSISTENT ESTIMATES OF RESIDUAL DENSITY IN HIGHER DIMENSIONS. The semiparametric framework in Section 2.6 for data that were parametric transformations of residuals can be generalised to situations with multidimensional data. One such situation of interest is $X_i = \mu + \Sigma^{1/2}\varepsilon_i$, say in dimension d, where μ is location vector and Σ is a symmetric and positive definite matrix, and where the

ε_is come from a residual distribution G in $\mathbb{R}^d$. Now give (μ, Σ) a prior and let G be a Dirichlet with parameter aG_0. Just as in Section 2.6 one finds the Bayes estimate

$$\widehat{G}(y) = w_n G_0(y) + (1 - w_n)\, n^{-1} \sum_{i=1}^{n} \Pr\{\Sigma^{-1/2}(x_i - \mu) \le y \,|\, \text{data}\},$$

where $t \le y$ means $t_1 \le y_1, \ldots, t_d \le y_d$. Again, using the analogue of (2.8), the exact or approximate posterior for (μ, Σ) is typically a multinormal centred close to the maximum likelihood estimators and with covariance matrix proportional to $1/n$ (assuming continuous data with no ties). The point to be made now is that this translates into a density estimate $\widehat{g}(y)$ for the residual density, the derivative of $\widehat{G}(y)$, which as its main term has a multivariate variable kernel density estimate that is not consistent. This is since the bandwidths h_i become proportional to $1/\sqrt{n}$, which is too small; the estimator smooths too little. In the case of a known Σ matrix and a standard d-dimensional normal for G_0, for example, the h_is are all equal to $1/\sqrt{n}$, and the variance of the kernel estimator is equal to $R(K)^d f(x)/(nh^d)$ plus smaller order terms, where again $R(K) = \int K^2 \, \mathrm{d}z$. The bias is proportional to h^2. This means that there is ordinary consistency for $d = 1$, a stable variance that does not go to zero for $d = 2$, and a variance going towards infinity for $d \ge 3$. Statisticians using the exact $\widehat{g}(\cdot)$ with no ties are using estimates with enormous variances, if $d \ge 3$.

The face of the Dirichlet is sort of saved, however. A more careful scrutiny shows that the posterior for the parameter is really approximately multinormal with covariance matrix proportional to $1/D_n$, the number of distinct data vectors among the n. And a somewhat artificial facet of the Dirichlet process is that it produces data vectors that have lots of ties; in fact, D_n increases quite slowly as $a \log n$, as shown by Korwar and Hollander (1973). Hence the bandwidths above are really of size const./$(\log n)^{1/2}$, theoretically speaking, that is, if the data vectors really are sampled from a Dirichlet process. And this is a large enough size for consistency of $\widehat{g}(y)$.

References

Antoniak, C.E. (1974). Mixtures of Dirichlet processes with application to Bayesian nonparametric problems. *Annals of Statistics* **2**, 1152–1174.

Bunke, O. (1987). Bayesian inference in semiparametric models. Proceedings of the First World Congress of the Bernoulli Society, Ташкент, USSR, 27–30. VNU Science Press.

Dalal, S.R. (1979). Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. *Stochastic Processes and Their Applications* **9**, 99–107.

Diaconis, P. and Freedman, D.A. (1986a). On the consistency of Bayes estimates (with discussion). *Annals of Statistics* **14**, 1–67.

Diaconis, P. and Freedman, D.A. (1986b). On inconsistent Bayes estimates of location. *Annals of Statistics* **14**, 68–87.

Doss, H. (1985a). Bayesian nonparametric estimation of the median. I: Computation of the estimates. *Annals of Statistics* **13**, 1432–1444.

Doss, H. (1985b). Bayesian nonparametric estimation of the median. II: Asymptotic properties of the estimates. *Annals of Statistics* **13**, 1445–1464.

Escobar, M.D. and West, M. (1994). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, to appear.

Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Annals of Statistics* **20**, 2008–2036.

Fenstad, G.U. and Hjort, N.L. (1994). Two Hermite expansion methods for density estimation, and a comparison with the kernel method. Submitted for publication.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics* **2**, 615–629.

Ferguson, T.S. (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics*, Chernoff Festschrift Volume (H. Rizvi, J.S. Rustagi and D.O. Siegmund, eds.), 287–302. Academic Press, New York.

Florens, J.-P., Mouchart, M., and Rolin, J.-M. (1992). Bayesian analysis of mixtures: some results on exact estimability and identification. In *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.), 503–524. Oxford University Press, Oxford.

Good, I.J. and Gaskins, R.A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58**, 255–277.

Good, I.J. and Gaskins, R.A. (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.

Hjort, N.L. (1986). Contribution to the discussion of Diaconis and Freedman's 'On the consistency of Bayes estimates'. *Annals of Statistics* **14**, 49–55.

Hjort, N.L. (1987). Semiparametric Bayes estimators. Proceedings of the First World Congress of the Bernoulli Society, Ташкент, USSR, 31–34. VNU Science Press.

Hjort, N.L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics* **18**, 1259–1294.

Hjort, N.L. (1991a). Semiparametric estimation of parametric hazard rates. In *Survival Analysis: State of the Art*, Proceedings of the *NATO Advanced Study Workshop on Survival Analysis and Related Topics*, Columbus, Ohio (P.S. Goel and J.P. Klein, eds.), 211–236. Kluwer, Dordrecht.

Hjort, N.L. (1991b). Bayesian and empirical Bayesian bootstrapping. Statistical Research Report, Department of Mathematics, University of Oslo.

Hjort, N.L. (1994a). Dynamic likelihood hazard rate estimation. *Biometrika*, to appear.

Hjort, N.L. (1994b). Local linear Bayesian regression. Submitted for publication.

Hjort, N.L. and Glad, I.K. (1994). Nonparametric density estimation with a parametric start. Submitted for publication.

Hjort, N.L. and Jones, M.C. (1994). Locally parametric nonparametric density estimation. Submitted for publication.

Hjort, N.L. and Pollard, D.B. (1994). Asymptotics for minimisers of convex processes. *Annals of Statistics*, to appear.

Jones, M.C. (1994). Kernel density estimation when the bandwidth is large. *Aus-*

tralian Journal of Statistics, to appear.

Korwar, R.M. and Hollander, M. (1973). Contributions to the theory of Dirichlet processes. *Annals of Probability* **1**, 705–711.

Kuo, L. (1986). Computations of mixtures of Dirichlet processes. *SIAM Journal of Scientific Statistical Computation* **7**, 60–71.

Lavine, M. (1992). Some aspects of Polya tree distributions for statistical modeling. *Annals of Statistics* **20**, 1222–1235.

Lenk, P.J. (1991). Towards a practicable Bayesian nonparametric density estimator. *Biometrika* **78**, 531–543.

Lenk, P.J. (1993). A Bayesian nonparametric density estimator. *Nonparametric Statistics* **3**, 53–69.

Leonard, T. (1978). Density estimation, stochastic processes, and prior information (with discussion). *Journal of the Royal Statistical Society B* **40**, 113–146.

Lindley, D.V. (1972). *Bayesian Statistics: A Review.* Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics* **12**, 351–357.

Loader, C.R. (1993). Local likelihood density estimation. Manuscript, AT&T Bell Laboratories.

Mauldin, R.D., Sudderth, W.D. and Williams, S.C. (1992). Polya trees and random distributions. *Annals of Statistics* **20**, 1203–1221.

Olkin, I. and Spiegelman, C.H. (1987). A semiparametric approach to density estimation. *Journal of the American Statistical Association* **82**, 858–865.

Rissanen, J., Speed, T.P., and Yu, B. (1992). Density estimation by stochastic complexity. *IEEE Transactions on Information Technology* **38**, 315–323.

Schuster, E. and Yakowitz, S. (1985). Parametric/nonparametric mixture density estimation with application to flood-frequency analysis. *Water Resources Bulletin* **21**, 797–804.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **2**, 461–464.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Simonoff, J.S. (1983). A penalty function approach to smoothing large sparse contingency tables. *Annals of Statistics* **11**, 208–218.

Thorburn, D. (1986). A Bayesian approach to density estimation. *Biometrika* **73**, 65–76.

Wand, M.P. and Jones, M.C. (1994). *Kernel Smoothing.* Chapman and Hall, London. To exist.

West, M. (1991). Kernel density estimation and marginalization consistency. *Biometrika* **78**, 421–425.

West, M. (1992). Modelling with mixtures. In *Bayesian Statistics 4* (J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds.), 503–524. Oxford University Press, Oxford.