Preface

Much of the work in Bayesian econometrics has focused on showing the value of Bayesian methods for parametric models (see, for example, Geweke (2005), Koop (2003), Li and Tobias (2011), and Rossi, Allenby, and McCulloch (2005)). This literature has documented the superior sampling properties of Bayesian parametric methods, particularly in situations with limited sample information. For example, Bayesian hierarchical models have become standard in the marketing literature because of the prevalence of panel data structures and consumer heterogeneity. While these models that have proved so useful in micro-econometrics and marketing applications involve very large numbers of parameters, both the panel-unit level model (typically a linear regression or multinomial logit specification) and the distribution of heterogeneity (the random "effects" distribution—commonly assumed to be multivariate normal) are parametric. The purpose of this book is to show how Bayesian non-parametric methods can be applied to econometric and marketing applications. The Bayesian non-parametric methods used here are finite and infinite mixture models and the central theme is to express the non-parametric problem as fundamentally a problem of density approximation/estimation.

A Bayesian approach to non-parametric problems is fundamentally what I will call a "full-information" approach. That is, all Bayesian approaches are likelihood-based and require an approximation to the joint distribution of all observables. This contrasts to the approach often favored in the non-Bayesian econometrics literature which I will call "partial-information." In the "partial-information" approach (often the basis for GMM methods), only certain assumptions (such as independence or orthogonality restrictions) are used to form the estimator of the model parameters and no attempt is made to model the entire distribution of the data. In the "partial-information" approach,

the sampling distribution of the estimator is derived (typically via asymptotic methods) without making explicit distributional assumptions. The resulting inference is often claimed to be "distribution-free" for this reason. However, it is well recognized that there can be considerable efficiency losses from this approach. The argument that is used to justify the approach is that it does not require arbitrary and unsubstantiated distributional assumptions. Bayesian parametric approaches are then criticized on the basis that they require arbitrary distribution assumptions which often are not examined. Concerns for model specification cause the "partial-information" advocate to be willing to trade-off efficiency for robustness of inference.

In principle, a fully non-parametric Bayesian approach should remove concerns regarding model mis-specification while retaining the desirable sampling properties of a "full-information" or likelihood-based procedure. As a practical matter, however, the full non-parametric approach may require the approximation of high dimensional distributions. Bayesian approaches will essentially involve very flexible or highly parameterized models that can be given a non-parametric interpretation. Non-Bayesian procedures for highly parameterized models often suffer from the over-fitting problem where small subsets of data drive the model fit and the model produces very non-smooth estimates. The advantage of a Bayesian method in a highly parameterized model is that, with proper priors, the tendency to "over-fit" is reduced substantially. This is because, in many models, proper priors impose what amounts to a penalty for highly parameterized models and implement a form of shrinkage.[1] For Bayesian procedures, the concerns regarding over-fitting are replaced with the problem of assessment of proper priors. There has been insufficient attention devoted to the assessment of priors in Bayesian non-parametric applications. The assumption often made is that quality of Bayesian density approximation is

[1]Here the term "shrinkage" refers to the property of Bayes estimators with proper priors to be somewhere between the location of the prior and the likelihood, thereby reducing the sensitivity of the estimators to sampling error.

not influenced a great deal by the prior settings. There is a sense in which this statement is obviously false—there certainly are prior settings which exert a great deal of influence and limit the flexibility of the Bayesian procedure. More importantly, some of the standard, proper but diffuse, settings used in practice can be highly informative in ways that, perhaps, are unintended. What is required is a procedure for prior assessment that retains flexibility and smoothing. I find that it is relatively straightforward to avoid overly and, perhaps, unintentionally informative priors and find a region of prior settings where prior sensitivity is limited and yet the procedure retains desirable smoothing properties.

In the non-Bayesian literature on non-parametric approaches, the style of research is to propose a "non-parametric" procedure and then prove that as the sample size approaches infinity this procedure "consistently" (based on some norm which measures the difference between the non-parametric approximation and the "true" model) recovers the true model. In some cases, further analysis is done to determine the optimal rate at which parameters must be added to the "non-parametric" approximation and to provide some sort of asymptotic distribution theory. Of course, any number of approaches are consistent in the sense that there are many different possible sets of basis functions which can be used to approximate an arbitrary distribution. In some cases, some sort of theory of "testing" or a penalty function is used to expand the model as the sample size increases at a rate sufficient to assure consistency but not too rapidly to avoid overfitting. For the user of the procedure, there is little guidance as to how to apply the approximation to one sample of a fixed size. Cross-validation methods are sometimes used to "tune" the approximation for a given sample, but have unknown finite sample properties.

In the approach taken here, I will use mixtures of normals as the basis for approximations. It is obvious that mixtures of normals have the desired "consistency" property required of a non-parametric approach. What is more important is to demonstrate that the Bayesian procedures provide reasonable estimates

for the demanding sorts of examples considered in applied work. With proper procedures for prior assessment, the "flexibility" vs "over-fitting" problem is largely avoided. Approximations involving literally thousands of parameters can be used without concern. The Bayes Factors[2] implicit in Bayesian inference impose such strong smoothing and shrinkage properties that I do not observe the over-fitting problem. This means that I don't have to couple my inference procedure with additional (often ad hoc) methods for avoiding the over-fitting problem. This provides a real advantage to a thoughtful Bayesian approach.

There is a sense, however, that all non-parametric methods, no matter how powerful, are very demanding of the data. Ultimately, inference regarding densities demands a uniform distribution or lack of sparseness in the data. It is not so much the number of observations that matters, but more that the observations are spread evenly across the, possibly high dimensional, space. For sparse data or data with "holes," non-parametric methods will not work as well. The great advantage of Bayesian procedures is that the inference procedure produces reliable information regarding the precision of inference regarding the density and any functional thereof. This inference comes essentially free as part of the Bayes computing procedure and without other delta method or bootstrap computations. The inference will show that in sparse areas of the space, the density is imprecisely estimated. Although it is reassuring to know that the Bayesian procedure will properly reflect lack of sample information, this does not remove the problem which is particularly acute with high dimensional distributions. For this reason, interest may focus on semi-parametric Bayesian methods. In semi-parametric methods, part of the problem is modeled with parametric assumptions and non-parametric

[2]The Bayes Factor is the quantity by which one transforms the prior probability of a model to its posterior probability. In any parametric model, the Bayes Factor can be expressed as $\int p(Data|\theta)\, p(\theta)\, d\theta$, where $p(Data|\theta)$ is the distribution of the observed data given the parameters (the "model") and $p(\theta)$ is the prior.

methods are employed for the remaining smaller dimensional part of the problem. Linear index models are a good example of a semi-parametric approach. For example, consider the problem of modeling the joint distribution of (y, x). If this is high dimensional, the semi-parametric approach would be to model the bivariate joint distribution of $(y, x'\gamma)$ using non-parametric density approximations. In this semi-parametric approach, the linear index serves as a dimension reduction device (see Chapter 4 for further discussion).

Panel data provide a good example of a natural setting where semi-parametric approaches are desirable. Invariably, panels involve a large number of cross-sectional units tracked over a limited duration of time. Here the time dimension is insufficient to allow for non-parametric approaches to the model at the cross-sectional level, but there are sufficient cross-sectional units to allow for a non-parametric determination of a random coefficient distribution. For example, with purchase data, we might specify a standard multinomial logit model for each unit indexed by a parameter vector which has an arbitrary distribution across units (see Chapter 5).

Ultimately, part of the output of a non-parametric or semi-parametric approach will be a density estimate or some functional of this object. For example, a non-parametric random coefficient model will yield a joint distribution of preference parameters across consumers. This joint distribution will be represented by a high dimensional density surface. A challenge will be to summarize this density in a meaningful way. The ability to compute univariate and bivariate marginals from this distribution will become important and this is another advantage of the mixture approach in that the implied marginal distributions are easy to calculate. Visualization techniques will play an important role in summarizing these distributions as moments lose much of their significance or interpretability for non-elliptical distributions.

I develop all of the methods here from first principles so that this work is accessible to anyone with a reasonable introductory knowledge of Bayesian statistics. Specifically, I assume

that readers are familiar with the Bayesian paradigm, Bayesian inference for multivariate regression models with normal errors, and the Gibbs sampler. This material is covered well in many texts including Rossi, Allenby, and McCulloch (2005) (which may be preferred due to notational compatibility).

In this book, all methods are illustrated with both simulated and actual data. In addition, the finite and infinite mixture approaches are implemented in my contributed R (R (2012)) package, `bayesm` (Rossi (2012)). Given the modular properties of MCMC methods, it will be a simple matter to use the `bayesm` routines to implement a computational method for the many possible models which can be crafted from a mixture of normals approach.

Summary

Most econometric models used in micro-economics and marketing applications involve arbitrary distributional assumptions. For example, the standard normal linear regression model assumes that the error terms in a regression are normally distributed and that the regression function is linear. Another important example is the use of the multivariate normal distribution as a model for heterogeneity or for the distribution of parameters across different units in a panel data structure. As more and less sparse data becomes available, there is a natural interest to provide methods which relax these distributional assumptions. In the Bayesian approach advocated here, specific distributional assumptions are replaced with more flexible distributions based on mixtures of normals. The Bayesian can use either a large but fixed number of normal components in the mixture or an infinite number bounded only by the sample size. By using flexible distributional approximations instead of fixed parametric models, the Bayesian can reap the advantages of an efficient method which models all of the structure in the data while retaining desirable smoothing properties. Non-Bayesian non-parametric methods often require additional ad

hoc rules to avoid "over-fitting" in which resulting density approximates are non-smooth. With proper priors, the Bayesian approach largely avoids over-fitting, while retaining flexibility. I provide methods for assessing informative priors that only require simple data normalizations. I apply the mixture of normals approximation method to a number of important models in micro-econometrics and marketing including the general regression model, instrumental variable problems, and models of heterogeneity.

Acknowledgments

I'd like to thank Herman van Dijk for inviting me to give the PUP lectures at Erasmus University. Thanks also to Sanjog Misra, Nick Polson, and Matt Taddy for many useful discussions. My co-authors, Jean-Pierre Dube, Guenter Hitsch, Chris Hansen, Tim Conley, and Rob McCulloch, are responsible for much of this material and I am very grateful for all of the insights I have obtained over the years by working with them. I am also very grateful to Renna Jiang for excellent research assistance. I am grateful for very detailed comments from Professor Dennis Fok and an anonymous reviewer that improved the readability of the manuscript.