FAST NONPARAMETRIC ESTIMATION OF A MIXING DISTRIBUTION

WITH APPLICATION TO HIGH-DIMENSIONAL INFERENC

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Ryan Martin

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2009

Purdue University

West Lafayette, Indiana
To my wife and parents, for their endless love and support.
ACKNOWLEDGMENTS

My time at Purdue has been very fruitful, both personally and professionally. For this, I must first thank my two wonderful advisors, Professors Jayanta K. Ghosh and Chuanhai Liu, who each, in their own way, had a positive influence on my development. Professor Ghosh is a scientist in the purest sense, whose knowledge of, and more importantly, passion for our subject is an inspiration to me (and others). Professor Liu has always encouraged me to build my own belief functions and not be bound by conventions. Perhaps the only way I can really repay my advisors for all they have done for me is to do the same for my own future advisees.

I have been lucky enough to have a third “advisor”—Professor Surya Tokdar, also a student of J. K. Ghosh—who has been a tremendous help to me during these last few years. Besides helpful technical pointers, his fresh perspective and advice on living the life of a young academic statistician has been invaluable. I cherish his friendship and look forward to many more years of fruitful collaborations.

Besides those directly involved with my research, there are other members of the Purdue faculty and staff that deserve some acknowledgement. I am grateful for the genuine support, guidance and encouragement Professor Rebecca Doerge has kindly given me over the past years, despite her terribly busy schedule. Ellen Gundlach, whom I have worked for in some capacity over the last three years, has given me numerous teaching opportunities, critical to my development as a teacher, for which I am very grateful. Lastly, the support and patience of the secretarial staff, in particular, Becca Miller and Darlene Wayman, is very much appreciated.

One of the best things about graduate school is the friendships made along the way. There are others, but Paul Livermore Auer, Paul Kidwell, and Andrea Rau have, from the very beginning, been there to support and challenge me, and I sincerely cherish their friendships. I am glad to have shared the Purdue experience with them.
Last, but definitely not least, I must thank my family for their unconditional love and support. My parents and my two sisters, Jenny and Cortney, have always been a positive influence in my life. While I was not blessed with any particularly outstanding abilities, my family taught me that I could be successful if I worked hard, and I have tried to apply that principle to all aspects of my life. But I am most grateful for my beautiful wife, Cheryl. She lovingly demands the very best from me and, quite literally, this work would not have been possible without her.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>x</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Mixture models</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Inference in nonparametric mixture problems</td>
<td>4</td>
</tr>
<tr>
<td>1.2.1 Nonparametric MLE</td>
<td>5</td>
</tr>
<tr>
<td>1.2.2 Dirichlet process mixtures</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Newton’s recursive algorithm</td>
<td>7</td>
</tr>
<tr>
<td>1.4 Empirical Bayes analysis</td>
<td>9</td>
</tr>
<tr>
<td>1.5 Summary of results</td>
<td>13</td>
</tr>
<tr>
<td>2 Newton’s Algorithm and Stochastic Approximation</td>
<td>15</td>
</tr>
<tr>
<td>2.1 Stochastic approximation</td>
<td>16</td>
</tr>
<tr>
<td>2.1.1 Algorithm and examples</td>
<td>16</td>
</tr>
<tr>
<td>2.1.2 ODE stability theory</td>
<td>20</td>
</tr>
<tr>
<td>2.1.3 SA Convergence Theorem</td>
<td>23</td>
</tr>
<tr>
<td>2.2 Newton’s estimate as SA</td>
<td>24</td>
</tr>
<tr>
<td>2.3 Estimating non-mixing parameters</td>
<td>27</td>
</tr>
<tr>
<td>2.4 Additional remarks</td>
<td>34</td>
</tr>
<tr>
<td>3 General Consistency Theorems</td>
<td>37</td>
</tr>
<tr>
<td>3.1 Kullback-Leibler projections</td>
<td>38</td>
</tr>
<tr>
<td>3.2 Robustness and rate of convergence</td>
<td>40</td>
</tr>
<tr>
<td>3.2.1 Preliminaries</td>
<td>41</td>
</tr>
<tr>
<td>3.2.2 Convergence of $m_n$</td>
<td>46</td>
</tr>
<tr>
<td>3.2.3 A quick numerical illustration</td>
<td>48</td>
</tr>
<tr>
<td>3.2.4 Convergence of $f_n$</td>
<td>50</td>
</tr>
<tr>
<td>3.3 Average over permutations</td>
<td>52</td>
</tr>
<tr>
<td>3.4 Additional remarks</td>
<td>54</td>
</tr>
<tr>
<td>4 Numerical Results</td>
<td>55</td>
</tr>
<tr>
<td>4.1 Regular mixtures</td>
<td>56</td>
</tr>
<tr>
<td>4.2 Irregular mixture</td>
<td>61</td>
</tr>
<tr>
<td>4.3 Massive data example</td>
<td>61</td>
</tr>
<tr>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>An Extension of the Recursive Algorithm</td>
</tr>
<tr>
<td></td>
<td>The RE+ algorithm</td>
</tr>
<tr>
<td></td>
<td>Convergence analysis</td>
</tr>
<tr>
<td>5.3</td>
<td>RE+ Simulations</td>
</tr>
<tr>
<td>5.4</td>
<td>Density estimation with RE+</td>
</tr>
<tr>
<td></td>
<td>5.4.1 Kernel-based density estimation</td>
</tr>
<tr>
<td></td>
<td>5.4.2 Finite mixture densities</td>
</tr>
<tr>
<td>5.5</td>
<td>Appendix: Simulated annealing details</td>
</tr>
<tr>
<td>6</td>
<td>A Nonparametric Empirical Bayes Framework for High-dimensional Inference</td>
</tr>
<tr>
<td>6.1</td>
<td>High-dimensional problems and sparsity</td>
</tr>
<tr>
<td>6.2</td>
<td>Two-groups model and simultaneous testing</td>
</tr>
<tr>
<td></td>
<td>6.2.1 Parametric and nonparametric Bayes</td>
</tr>
<tr>
<td></td>
<td>6.2.2 Efron’s nonparametric empirical Bayes</td>
</tr>
<tr>
<td></td>
<td>6.2.3 A new nonparametric empirical Bayes framework</td>
</tr>
<tr>
<td>6.3</td>
<td>Parameter estimation with RE+</td>
</tr>
<tr>
<td>6.4</td>
<td>Empirical Bayes decision rule</td>
</tr>
<tr>
<td>6.5</td>
<td>Microarray data examples</td>
</tr>
<tr>
<td>6.6</td>
<td>Simulation results</td>
</tr>
<tr>
<td>6.7</td>
<td>Additional remarks</td>
</tr>
<tr>
<td>6.8</td>
<td>Appendix: Recursive formula for pseudo-likelihood gradient</td>
</tr>
<tr>
<td>7</td>
<td>Future work</td>
</tr>
<tr>
<td>LIST OF REFERENCES</td>
<td>111</td>
</tr>
<tr>
<td>VITA</td>
<td>119</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>13</td>
</tr>
<tr>
<td>4.1</td>
<td>58</td>
</tr>
<tr>
<td>4.2</td>
<td>60</td>
</tr>
<tr>
<td>5.1</td>
<td>77</td>
</tr>
<tr>
<td>6.1</td>
<td>102</td>
</tr>
</tbody>
</table>

1.1 Mean (and standard deviation) of the loss $||\hat{\theta} - \theta||^2$ for the three estimates over 1000 iid samples of $(X, \theta)$, with $n = 250$, based on the Skew-Normal prior $F = SN(0, 1, \alpha)$, for $\alpha \in \{0, 10\}$.

4.1 Mean computation time (in seconds) for PARE, NPMLE and DPM over the $T = 100$ samples in the Beta-Normal and Gamma-Poisson models. RE (not displayed) is about 100 times faster than PARE.

4.2 Summary of KL divergences ($\times 100$) of the various estimates against (i) the true mixture, and (ii) the estimated DP mixture for each of the models BN and GP.

5.1 Mean (and standard deviation) of the indicated quantity over the 100 replications of the density estimation experiment in Example 5.3.

6.1 Estimates of $\pi$ for Models I and II based on RE+ and Jin-Cai procedures in the simultaneous significance testing simulation example.
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Sample path of the sequence ( {x_n} ) in Example 2.3. The dotted line is the value of ( \xi ) used for data generation.</td>
<td>19</td>
</tr>
<tr>
<td>3.1 Estimated mixing and mixture densities in the Exponential scale-mixture problem for weights ( w_i = (i + 1)^{-r}, r \in {0.67, 1} ), in the RE algorithm with ( n = 1000 ).</td>
<td>50</td>
</tr>
<tr>
<td>3.2 Observed ( L_1 ) errors for various ( n ) in the Exponential scale-mixture problem, along with the fitted curve ( \hat{L}_{1,n} \propto n^{-\hat{r}} ).</td>
<td>51</td>
</tr>
<tr>
<td>4.1 Plots of the mixing density estimates (top) and corresponding mixture density estimates (bottom) for model (BN). Each cell shows the true mixing or mixture density (black) with the ( T = 100 ) estimates (gray).</td>
<td>57</td>
</tr>
<tr>
<td>4.2 Plots of the mixing density estimates (top) and corresponding mixture density estimates (bottom) for model (GP). The cells show the true mixing or mixture densities (black) with the ( T = 100 ) estimates (gray).</td>
<td>58</td>
</tr>
<tr>
<td>4.3 Summary of the ( L_1 ) errors (left) and Bias-Spread tradeoff (right) for models (BN) (top) and (GP) (bottom).</td>
<td>59</td>
</tr>
<tr>
<td>4.4 Plots of the absolutely continuous parts of the mixing distributions (top) and corresponding mixture density estimates (bottom) for model (IN). The true ( \varphi ) or ( m ) are shown in black with the ( T = 100 ) estimates in gray.</td>
<td>62</td>
</tr>
<tr>
<td>4.5 Summary of the estimates of ( \Pr(\theta = 0) ) (left), summary of the ( L_1 ) errors (middle), and Bias-Spread tradeoff (right) for irregular mixture model (IN).</td>
<td>63</td>
</tr>
<tr>
<td>4.6 Plot of the absolutely continuous part of the mixing distribution (left) and corresponding mixture density estimates (right) for model (IN) in the massive data example.</td>
<td>63</td>
</tr>
<tr>
<td>5.1 Summary of the ( L_1 ) error of the estimates of ( m ) (top) and of the estimates of ( \sigma ) (bottom) for the RE+ algorithm over various ( n ) in the Beta-Normal example.</td>
<td>72</td>
</tr>
<tr>
<td>5.2 Scatterplot of ( \hat{\sigma}_n ) by ( \hat{K}_n(\hat{\sigma}_n) ) for various ( n ) in the Beta-Normal example. The points should concentrate around ((0.1, 0)).</td>
<td>73</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>5.3</td>
<td>Plots of density estimates based on the RE+, SJ and CV bandwidth selectors for the whale migration data in Example 5.2.</td>
</tr>
<tr>
<td>5.4</td>
<td>Plots of the RE+ estimates of the mixing (left) and mixture densities (right) for the galaxy data in Example 5.4.</td>
</tr>
<tr>
<td>5.5</td>
<td>Plots of the RE+ estimates of the mixing (left) and mixture densities (right) for the Thailand data in Example 5.5.</td>
</tr>
<tr>
<td>5.6</td>
<td>Plot of the simulated annealing sample path for the optimization in Example 5.4.</td>
</tr>
<tr>
<td>6.1</td>
<td>The HIV data set described in Example 6.3. Plots show estimates of $\pi m^{(0)}$ (thin lines), $(1-\pi)m^{(1)}$ (dashed lines) and $m$ (heavy lines) based on Efron and RE+. Cases falling between the two triangles are null.</td>
</tr>
<tr>
<td>6.2</td>
<td>Plots of the estimated local FDR based on the two empirical Bayes methods for the HIV data set described in Example 6.3. Tick marks on the z-axis show the $n = 7,680$ observed z-scores.</td>
</tr>
<tr>
<td>6.3</td>
<td>The leukemia data set described in Example 6.4. Plots show estimates of $\pi m^{(0)}$ (thin lines), $(1-\pi)m^{(1)}$ (dashed lines) and $m$ (heavy lines) based on the RE+ and regularized RE+ algorithms. Cases falling between the two triangles are null.</td>
</tr>
<tr>
<td>6.4</td>
<td>Alternative densities $m^{(1)}$ for the two models (black), along with the RE+ estimates (gray) when $\pi = 0.9$.</td>
</tr>
<tr>
<td>6.5</td>
<td>Plots of the power, MP and FDR of the RE+ (black, solid), JC (black, dashed) and Bayes Oracle (gray) testing procedures in the two models as functions of the true null proportion $\pi$.</td>
</tr>
</tbody>
</table>

Mixture distributions have, for many years, been used in a wide range of classical statistical problems, including cluster analysis and density estimation, but they are now finding new and interesting applications in the high-dimensional problems inspired by microarrays and other recent technological advances. Computational breakthroughs such as the EM and MCMC algorithms make fitting the mixture model relatively easy, but inference on the mixing distribution itself remains a challenging problem. Recently, M. A. Newton proposed a fast recursive algorithm for nonparametric estimation of the mixing distribution, motivated by heuristic Bayesian arguments, which has been shown to perform well in a host of applications. Theoretical investigations, on the other hand, have been rather limited. This thesis gives a thorough exploration of the theoretical properties of Newton’s recursive estimate (RE).

We begin with a rigorous justification for the recursive algorithm, showing that RE is just a special case of stochastic approximation. For finite mixtures, consistency of RE is established using classical stochastic approximation results; general mixtures, on the other hand, would require an infinite-dimensional stochastic approximation which is still not well studied in general. As an alternative approach in the general mixture problem, a martingale approximation is used to show, under mild conditions, that the estimated mixture density converges almost surely to the “best possible” mixture in a Kullback-Leibler sense, and a competitive bound on the rate of convergence is obtained. Under some extra conditions, including identifiability, we prove almost sure weak convergence of the estimated mixing distribution.
These general convergence results lead to an important extension of the recursive algorithm, which we call RE+, that is able to estimate a mixing distribution and additional non-mixing parameters simultaneously. Applications of the RE+ algorithm in density estimation are highlighted. Motivated by connections between RE and Dirichlet process mixtures, this new algorithm is further developed into a novel non-parametric empirical Bayes framework for high-dimensional inference. In particular, we propose an identifiable, nonparametric version of the popular two-groups model whose parameters are easily estimated by the RE+ algorithm. In the context of large-scale simultaneous testing, this new procedure gives believable answers in two very different microarray data examples, and is shown in simulations to be competitive with the Bayes Oracle test over a range of sparsity levels.
1. INTRODUCTION

For many years, mixture distributions have been used to successfully model data that show population heterogeneity or involve latent variables. Recent technological advances have rapidly brought high-dimensional problems to the forefront of statistical research, and mixture models are now finding new applications in these interesting and more difficult problems. With higher-dimensional data and the push for fast and flexible inference procedures comes new and difficult computational challenges. Well-established computational methods, namely the EM and various MCMC algorithms, are now available, making it possible to fit a mixture model to data with relative ease. But computation can be cumbersome, especially in high-dimensional problems, and inference on the mixing distribution itself remains a challenging problem.

Towards a potential solution to this simultaneous problem of computation and inference, M. A. Newton (and co-authors) recently proposed a fast recursive algorithm capable of “online” nonparametric estimation of the mixing distribution. While this algorithm performs quite well in applications, its motivation is heuristic and theoretical investigations have thus far been rather limited. The primary goal of this thesis is to carefully explore the theoretical properties of this recursive algorithm, giving a more rigorous motivation in terms of stochastic approximation, and proving very general convergence theorems. Along the way we will investigate the algorithm’s finite-sample performance and highlight some high-dimensional applications. An important extension of the recursive algorithm is also given which significantly expands the scope of applicability and leads quite naturally to a new nonparametric empirical Bayes framework for high-dimensional inference.

The remainder of this chapter will briefly introduce some basic concepts and notation that will appear throughout. A more detailed summary of the chapters to follow is given in Section 1.5.
1.1 Mixture models

Mixture distributions have been widely used in a variety of interesting and important statistical problems, such as cluster analysis, density estimation and empirical Bayes inference, which in turn have been used in all sorts of applications, such as astronomy, biology, economics, and genomics, to name a few. Here we give a brief overview of the mixture problem and set some notation which will be used throughout.

Let \((\mathcal{X}, \mathcal{A})\) and \((\Theta, \mathcal{B})\) be a pair of measurable spaces, called the sample and parameter spaces, respectively, and let \(\{p(\cdot|\theta) : \theta \in \Theta\}\) be a parametric family of probability densities on \(\mathcal{X}\) with respect to a dominating \(\sigma\)-finite measure \(\nu\) on \((\mathcal{X}, \mathcal{A})\). Then, for a probability measure \(F\) on \((\Theta, \mathcal{B})\), define the mixture density on \(\mathcal{X}\) as

\[
m_F(x) = \int_{\Theta} p(x|\theta) \, dF(\theta).
\] (1.1)

If \(F\) has a density \(f\) with respect to a dominating \(\sigma\)-finite measure \(\mu\) on \((\Theta, \mathcal{B})\), then we write \(m_f(x) = \int p(x|\theta)f(\theta) \, d\mu(\theta)\) for the mixture density. An important property of the general mixture model (1.1) is its flexibility: any probability density function \(m\) on \(\mathcal{X}\) can be approximated arbitrarily well by a mixture of the form (1.1) with a suitably chosen \(p(x|\theta)\). For example, Normal location-scale mixtures are dense in the space of probability distributions on \(\mathbb{R}\).

An alternative way to view the mixture (1.1) is as a hierarchical model

\[
\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} F \quad \text{and} \quad X_i|\theta_i \overset{\text{iid}}{\sim} p(\cdot|\theta_i), \quad i = 1, \ldots, n.
\] (1.2)

If we were to associate the mixing distribution \(F\) with a prior distribution on the parameter \(\theta\) in a Bayesian sense, then we see that all Bayesian problems can be viewed as a mixture problem. The situation where \(F\) is an unknown prior falls into the category of empirical Bayes where the data is used in some way to help choose an appropriate prior. The empirical Bayes approach, described in more detail in Section 1.4 and Chapter 6, will be a primary focus throughout.
An important special case of (1.1) is the so-called finite mixture model, with density given by

\[ m_f(x) = \sum_{s=1}^{S} p(x|\theta^s)f(\theta^s), \quad (1.3) \]

where \( p(x|\theta) \) is a parametric density, \( \Theta = \{\theta^1, \ldots, \theta^S\} \) is a finite set of cardinality \( S \), and \( f \) is a probability density with respect to counting measure \( \mu \). Note that \( \Theta \) could be known or unknown; in general, \( S \) too can be unknown.

Finite mixtures are an extremely useful tool in all sorts of statistical problems and, although they are just a special case of the general mixture (1.1), they are important enough to deserve their own discussion; McLachlan and Peel [74] is a particularly good one. A summary of the vast literature would be impractical so, instead, we make only one important remark. Finite mixtures provide a natural and mathematically simple way to encode population heterogeneity. This makes them ideally suited for density estimation, clustering and classification problems. In high-dimensional testing problems, we will see in Chapter 6 that a two-component mixture allows for a very simple rule for classifying cases as null or non-null. For further details on finite mixtures in classical statistics see, for example, McLachlan and Basford [75], Priebe [85], Cutler and Cordero-Braña [19], McLachlan, Bean and Peel [73], and Woo and Sriram [117]. For Bayesian treatments see, for example, West, Müller and Escobar [116], Escobar and West [38], and Roeder and Wasserman [97].

A critical issue in any inference problem is identifiability of the parameters involved. In classical parametric problems—like point estimation in the Exponential family, for example—identifiability of the parameters is trivial. Things are not as straightforward in mixture problems, however.

**Definition 1.1** A mixing distribution \( F \) is identifiable if and only if the mapping \( F \mapsto m_F \) is one-to-one or, equivalently, if and only if \( m_F(x) = m_G(x) \) for \( \nu \)-almost all \( x \) implies \( F = G \).

The mixture mapping \( F \mapsto m_F \) is linear and, in the case of finite mixtures (1.3), identifiability reduces to a matrix non-singularity condition; see, for example, Te-
ich [109] and Lindsay [65]. Yakowitz and Spragins [118] give also some interesting results on finite mixtures of multiparameter distributions. Identifiability results for general mixtures of the form (1.1), in particular those in Teicher [108], require that the mixands $p(x|\theta)$ have suitably non-vanishing Fourier transforms.

Finite mixture models have one peculiar feature related to identifiability, namely, the so-called “label-switching problem.” That is, simply permuting the mixture component labels—technically a different mixing distribution—yields an identical mixture. This can be particularly problematic in the Bayesian context where inference is based on posterior simulations. For these reasons, identifiability in finite mixture problems is usually defined modulo permutations of the mixture component labels. Here, when we consider finite mixtures, we avoid such concerns by assuming an ordering on the mixture components induced by the natural ordering on the parameter space. But see McLachlan and Peel [74, Secs. 1.14, 4.9] and Titterington, Smith and Makov [111, Sec. 3.1] for further discussion of the label-switching problem.

A further challenge in finite mixtures is a certain singularity problem. For example, in a simple two-component mixture ($S = 2$) setting $f(\theta_1) = 0$, $1$ reduces the dimension of the problem not just by 1 but, rather, $\dim(\Theta) + 1$. This can cause problems in both classical and Bayesian asymptotics. Indeed, the usual Laplace approximation may not be valid around such “boundary points.” See, for example, Ghosh and Sen [46], Hall and Stewart [50] and Drton [25]. These boundary cases must be handled with care in our approach as well; see Chapter 2.

1.2 Inference in nonparametric mixture problems

Nonparametric estimation of a mixing distribution $F$ based on iid data from the mixture $mF$ is a notoriously difficult problem. Here we briefly mention two popular nonparametric estimation methods. The relatively new procedure described in Section 1.3, which is the major focus of the remainder of this thesis, will often be compared to the following two procedures. We should mention that there are other
methods, such as deconvolution, but these will not be discussed here; see, instead, Zhang [119], Fan [39], Masry and Rice [72], and the references therein.

1.2.1 Nonparametric MLE

For iid observations $X_1, \ldots, X_n$ from the mixture distribution $m_F$ in (1.1) with unknown $F$, then the nonparametric likelihood function is defined as

$$
\ell(F) = \prod_{i=1}^{n} m_F(X_i) = \prod_{i=1}^{n} \int p(X_i|\theta) \, dF(\theta).
$$

The goal is then to estimate $F$ by maximizing $\ell(F)$ over all distributions on $\Theta$.

The simplest situation is when $F$ is restricted to discrete distributions with a known support size. But even in this relatively simple case, there are theoretical and practical challenges; see Chapter 33 of DasGupta [20]. The EM algorithm of Dempster, Laird and Rubin [23] has proved to be a very useful computational tool, and Redner and Walker [88] give a comprehensive overview of the EM in the finite mixture problem. See also McLachlan and Peel [74].

No surprise, the general case is even more challenging. Kiefer and Wolfowitz [55] were among the first to consider consistency of the MLE in the mixture problem; Laird [58] and Lindsay [64, 65] focus primarily on existence and characterizations of the MLE. It turns out that our approach of minimizing Kullback-Liebler numbers in Chapter 3 is related to the directional (Gâteaux) derivative characterization of the NPMLE given in Lindsay [64]; cf. Remark 3.5 below. Both Laird and Lindsay show, under various conditions, that the MLE $\hat{F}_n$ of $F$ is almost surely discrete, with at most $n$ support points. This fact makes the computations more manageable—see Wang [114]—but the discreteness of $\hat{F}_n$ also has obvious downsides.

1.2.2 Dirichlet process mixtures

For nonparametric Bayesian analysis, Ferguson [40] proposed the so-called Dirichlet process (DP) distribution, a distribution on distributions. A random probability
measure $F$ on the measurable space $(\Theta, \mathcal{B})$ has a DP distribution with parameters $\alpha$ and $F_0$, written $F \sim \mathcal{D}(\alpha, F_0)$, if for any integer $k$ and any measurable partition $B_1, \ldots, B_k$ of $\Theta$, the random probability vector $(F(B_1), \ldots, F(B_k))$ has a $k$-dimensional Dirichlet distribution $\text{Dir}_k(\alpha F_0(B_1), \ldots, \alpha F_0(B_k))$. Ferguson [40] shows that the DP process exists, its support is large in some sense, and also that it is a conjugate prior. We highlight two more specific properties below. For more details, see Ghosh and Ramamoorthi [45].

Blackwell and MacQueen [12] give a representation of the DP which has proven to be quite useful. For the model $\theta_1, \ldots, \theta_n | F \overset{\text{iid}}{\sim} F$ and $F \sim \mathcal{D}(\alpha, F_0)$, the marginal distribution of $\theta_1, \ldots, \theta_n$ is determined by the conditionals:

$$
\theta_i|\theta_1, \ldots, \theta_{i-1} \sim \frac{\alpha}{\alpha + i - 1} F_0 + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j}, \quad i = 2, \ldots, n.
$$

(1.4)

(Here $\delta_x$ denotes a degenerate distribution at $x$.) In (1.4) we have effectively integrated out the infinite-dimensional parameter $F$—this is particularly useful in MCMC computation in DP mixtures discussed below. This representation is called a Polya urn scheme due to its connection to the classical Polya urn problems in introductory probability texts, e.g., Hoel, Port and Stone [51].

It turns out that draws $F \sim \mathcal{D}(\alpha, F_0)$ are almost surely discrete! There are many proofs of this fact—see, for example, Ferguson [40], Blackwell [11], Blackwell and MacQueen [12] and Sethuraman [100]. In some applications, such as cluster analysis and point estimation, the discreteness is a good property since it allows for adaptive shrinkage toward random locations. But, in other applications, where the distribution $F$ is of primary interest, the discreteness can be inappropriate.

When modeling continuous data, the most popular way to dodge the discreteness of the DP is to use DP mixtures. That is, assume data $X_1, \ldots, X_n$ are observed from a distribution with density $m_F(x) = \int p(x|\theta) dF(\theta)$ where $p(\cdot|\theta)$ is a parametric density and $F$ is a realization from $\mathcal{D}(\alpha, F_0)$. This induces a prior on the density $m_F$; see Lo [67]. Antoniak [4] shows that the posterior distribution of $F$ is a mixture of DPs. Additional theoretical properties of DP mixtures, including posterior con-
sistency, are discussed in Barron, Schervish and Wasserman [6], Ghosal, Ghosh and Ramamoorthi [42], and Ghosal and van der Vaart [43].

Computation in DP mixtures is often facilitated by breaking the mixture, introducing the latent variables $\theta_i$ in (1.2). Typically the next step is to use a posterior version of the Polya urn representation (1.4) to “impute” the unobserved $\theta_i$'s. Once samples of the $\theta_i$'s are available from their marginal posterior ($F$ is integrated out), quantities such as the predictive distribution of a new $X^*$ can be estimated. See MacEachern [68], Escobar [37], Escobar and West [38], MacEachern and Müller [70] and Neal [78] for details. Sequential importance sampling is an alternative to the MCMC methods just described and has its own set of advantages; see Liu [66] and MacEachern, Clyde and Liu [69]. Two key points are that (i) a posterior draw of $F$ is still discrete, so the use of DP mixtures when $F$ itself of interest may be inappropriate and (ii) each of these procedures begins by integrating out $F$, so posterior inference on (functionals of) the mixing distribution is difficult.

1.3 Newton’s recursive algorithm

Two critical features of the NPML and DPM analyses described in Section 1.2 are (i) computations are non-trivial and can be slow and (ii) inference on the mixing distribution itself is challenging. These are particularly problematic in the high-dimensional applications we have in mind. As a fast and flexible alternative, we consider a relatively new algorithm, due to Newton [80], designed for recursive non-parametric estimation of the unknown mixing distribution $F$; see also Newton, Quintana and Zhang [81] and Newton and Zhang [82].

**RE Algorithm** Choose an initial estimate $F_0$ of $F$ and a deterministic sequence of weights $w_1, \ldots, w_n \in (0, 1)$. Then, for $i = 1, \ldots, n$ compute

$$F_i(B) = (1 - w_i)F_{i-1}(B) + w_iF_{i-1}(B | X_i),$$  (1.5)
where $B \in \mathcal{B}$ is measurable and $F_{i-1} (\cdot | X_i)$ is the posterior distribution measure of $\theta_i$, determined by Bayes Theorem, when $\theta_i \sim F_{i-1}$ a priori and $X_i | \theta_i \sim p(\cdot | \theta_i)$. If $F$ has a density $f$ with respect to $\mu$, then recursion (1.5) can be written as

$$f_i(\theta) = (1 - w_i)f_{i-1}(\theta) + w_i \frac{p(X_i|\theta)f_{i-1}(\theta)}{\int p(X_i|\theta')f_{i-1}(\theta') \, d\mu(\theta')} \quad \theta \in \Theta. \quad (1.6)$$

Ultimately produce $F_n$ and $m_n := m_{F_n}$ as the estimates of $F$ and $m_F$, respectively.

A few quick comments about the RE algorithm are in order. First, the weight sequence $w_n$ cannot be arbitrary; $w_n \to 0$ is a necessary condition for convergence and the rate determines the RE rate of convergence. More details are given in Chapter 3. Second, computation of $F_n$ is very fast. Indeed, computation is typically $O(n)$. Some applications of RE can be found in Tao, et al. [107] and Quintana and Newton [86].

The original motivation behind the algorithm is what we refer to as the “one-step correspondence” between the RE and DP mixture models, and is summarized in the following proposition.

**Proposition 1.2** (One-Step Correspondence). Suppose $F$ has a DP prior distribution $\mathcal{D}(1/w_1 - 1, F_0)$ and $X_1 | F \sim m_F$. Then, for any measurable $B \in \mathcal{B}$,

$$\mathbb{E}\{F(B) \mid X_1\} = (1 - w_1)F_0(B) + w_1 F_0(B | X_1).$$

Roughly speaking, the proposition states that the recursive update $F_1$ of $F_0$ corresponds to the posterior mean under the DP formulation. Of course, there is nothing special about the first step so the idea is to simply apply Proposition 1.2 iteratively. Thus, the motivation behind the RE algorithm is roughly as follows: if, at step $i$, we model $F$ with a DP distribution $F \sim \mathcal{D}(1/w_i - 1, F_{i-1})$, then the posterior mean of $F$, given the new observation $X_i$, is nothing but $F_i$ in (1.5). Therefore, the RE algorithm is, in a very rough sense, an approximation to the fully Bayes DP mixture model. This connection, which is described further in Chapter 5, is exactly what we need to justify our empirical Bayes approach in Chapter 6.

Despite the striking similarities between the RE and the DP prior Bayes procedures, it is important to note that $F_n$ cannot be a posterior quantity since it depends
on the order of the data and, therefore, is not a function of the sufficient statistics \( X_{(1)}, \ldots, X_{(n)} \). This dependence on the order is relatively weak, especially for \( n \) large, but still may be troubling in practice. A natural approach to reduce the dependence is to compute the estimates over several randomly chosen permutations of the data sequence and take a pointwise average. The computational efficiency of the RE algorithm makes this a feasible solution. We will revisit this idea in Chapter 3.

These heuristic properties and strong numerical performance of RE have been known for some time, but theoretical investigations have been rather limited. Important unanswered questions are:

- Under what conditions does RE converge?
- Are the estimates \( f_n \) and \( m_n \) consistent?
- If \( f_n \) and \( m_n \) converge, what are their rates?
- If the model (1.1) is formulated incorrectly, what happens to the RE?

In the upcoming chapters, we will give precise answers to the first, second, and fourth questions above; the third we give only a partial answer. Various extensions and applications of RE—particularly in empirical Bayes analysis of high-dimensional problems—will also be given along the way.

1.4 Empirical Bayes analysis

Empirical Bayes analysis, which has proven to be very useful in high-dimensional problems, will be a recurring theme throughout this thesis. Here we give a brief overview of the general empirical Bayes problem and some indications of what is to follow in subsequent chapters.

Consider the very general hierarchical Bayesian model (1.2) where the number of parameters \( \theta_1, \ldots, \theta_n \) is assumed to be rather large—a so-called high-dimensional problem—and \( F \) is a prior distribution, known only to fall within a specified class \( \mathcal{F} \).
of probability measures on the parameter space $\Theta$. To account for this uncertainty about $F$, a Bayesian might proceed by introducing a second stage prior distribution $\Pi$ on $F$. A key feature of this hierarchical Bayes approach is that inference on, say, $\theta_1$, which is based on $X_1$ alone in classical and non-hierarchical Bayes procedures, is improved by borrowing strength from the other $X_i$’s, through the exchangeability introduced by the common prior $\Pi$.

Empirical Bayes (EB) inference, the brainchild of Robbins [90–92] which Neyman [83] referred to as a “breakthrough” in statistical decision making, is based on the idea of using the observed data $X_1, \ldots, X_n$ to choose the prior $F \in \mathcal{F}$. Thus, EB gives an automatic, data-based procedure for choosing priors which is, in some sense, a middle ground between the classical and hierarchical Bayes approaches. Estimating the prior is a difficult proposal with small to moderate sample sizes. But, with recent technological advances that allow scientists to collect and process massive amounts of data, EB methods have become increasingly popular.

Considerable attention has been given to the special case of parametric empirical Bayes (PEB), developed in a series of papers by Efron and Morris [32–36] and also by Morris [77], in which $\mathcal{F}$ is indexed by a finite-dimensional parameter, say $\gamma$; i.e., $\mathcal{F} = \{F_\gamma : \gamma \in \Gamma\}$. The PEB approach shares the “strength borrowing” property of the hierarchical Bayes approach and also has the desirable property of reducing the high-dimensional problem about $\theta_1, \ldots, \theta_n$ to a relatively low-dimensional problem about the hyperparameter $\gamma$. Therefore, in the PEB framework, the high-dimensionality of the problem is actually a “blessing” rather than a “curse.” Moreover, unlike nonparametric empirical Bayes, PEB does not require a very large $n$.

**Remark 1.3** One drawback of the EB approach, in general, is that it is relatively difficult to account for uncertainty about the prior in the analysis. However, when the dimension $n$ is large—as in the modern applications we have in mind here—this is of less concern because estimates of the prior are usually of good quality. See, however, Scott and Berger [99] who show difference between hierarchical and empirical Bayes can be substantial in the case of sparse mixtures.
The hierarchical and empirical Bayes approaches have a common shortcoming, namely, that choosing the particular parametric model for $\mathbb{F}$ can be difficult. Moreover, this choice often has a dramatic effect on the conclusions reached. The following example illustrates this point and also anticipates our proposed solution based on the recursive algorithm and its RE+ extension to be described in Chapter 5.

**Example 1.4** Consider the case where $X_i$ are independent and $X_i \sim N(\theta_i, 1)$. This is the famous example of Stein [105, 106] where the standard estimate $X = (X_1, \ldots, X_n)$ of $\theta = (\theta_1, \ldots, \theta_n)$ is inadmissible, under $L_2$-norm loss, for $n \geq 3$. This was a very startling result, known for some time as Stein’s phenomenon; see Berger [8].

Stein [106] derives the James-Stein estimate which, as shown in Morris [77], coincides with the PEB estimate when $\mathbb{F} = \{N(0, \tau^2) : \tau^2 > 0\}$ is the location family of Normal distributions. In particular, the PEB estimate under the Normal prior is

$$\hat{\theta}_{PEB}^\ast(X) = \left(1 - \frac{n - 2}{\|X\|^2}\right) X,$$

and it shrinks the least-squares or maximum likelihood estimate $X$ towards the prior mean 0. This estimate performs well under the prior $N(0, \tau^2)$, uniformly dominating the least-squares estimate $X$ in terms of mean square error (MSE). However, the story changes when the true prior $\mathbb{F}$ is outside this relatively narrow class $\mathbb{F}$.

Consider the class of skew-Normal distributions generated by location-scale transformation of the following density function

$$2\varphi(x)\Phi(\alpha x), \quad x \in \mathbb{R},$$

where $\varphi$ and $\Phi$ are the PDF and CDF of the standard Normal distribution and $\alpha \in \mathbb{R}$ is a skewness parameter that controlling the symmetry. Note that $\alpha = 0$ corresponds to the usual $N(0, 1)$ distribution. We will write $SN(\mu, \sigma, \alpha)$ to denote a skew-Normal distribution with location $\mu$, scale $\sigma$ and skewness $\alpha$. See Azzalini [5] for further details. Here we will look at the performance of the PEB estimate (1.7) when $\mathbb{F}$ belongs to the class $\mathbb{F}$ of skew-Normal distributions.
As a robust alternative to the PEB estimate, we can easily estimate the prior \( F \) nonparametrically using the RE algorithm, and compute the Bayes estimate of \( \theta = (\theta_1, \ldots, \theta_n) \) based on the estimated prior \( F_n \):

\[
\hat{\theta}^{\text{RE}}_i(X) = \int \theta \, dF_n(\theta | X_i), \quad i = 1, \ldots, n
\]  

(1.8) where \( dF_n(\theta | X_i) \propto p(X_i | \theta) \, dF_n(\theta) \) is the “posterior” distribution of \( \theta_i \), given \( X_i \), assuming that the recursive estimate \( F_n \) is the true prior. For two choices of skewness parameter \( \alpha \), the \( L_2 \) error \( \| \hat{\theta}(X) - \theta \|_2 \) is summarized in Table 1.1. These calculations are based on 1000 iid samples of \((X, \theta)\) of dimension \( n = 250 \) from the true model: \( X | \theta \sim N_n(\theta, I_n) \) and \( \theta_1, \ldots, \theta_n \text{iid} \sim F \). The RE and PEB estimates are compared to the “oracle,” the Bayes rule when \( F \) is known:

\[
\hat{\theta}^{\text{oracle}}_i(X) = \int t \, dF(t | X_i), \quad i = 1, \ldots, n
\]

where \( dF(\cdot | X_i) \) is the true posterior distribution of \( \theta_i \), given \( X_i \). Here we see that, in the Normal case (\( \alpha = 0 \)), the PEB and RE estimates perform similarly, both being close to the oracle on average. However, when the skewness jumps from \( \alpha = 0 \) to \( \alpha = 10 \), the RE estimate remains relatively stable while the relative quality of the PEB estimate drops dramatically: for \( \alpha = 0 \), PEB’s estimated efficiency is \( 126/125 = 1.008 \) but, for \( \alpha = 10 \), it jumps to \( 126.7/118.4 = 1.07 \). Therefore, if there is any doubt about the particular parametric choice of \( F \), a nonparametric procedure such as the RE rule (1.8) might be more appropriate than PEB.

Avoiding specification of a parametric class \( F \) altogether, as suggested in Example 1.4, and considering, instead, a nonparametric class of priors results in a much more flexible analysis. However, this additional flexibility does not come free. Nonparametric methods typically require a large sample size to produce an adequate estimate and, although this condition is satisfied in the high-dimensional problems under consideration, computational issues can arise and, therefore, the problem remains unsolved. What we desire is a nonparametric empirical Bayes (NPEB) approach which is simultaneously accurate, computationally efficient, and justifiable from a Bayesian
Table 1.1
Mean (and standard deviation) of the loss $\|\hat{\theta} - \theta\|_2^2$ for the three estimates over 1000 iid samples of $(X, \theta)$, with $n = 250$, based on the Skew-Normal prior $F = SN(0, 1, \alpha)$, for $\alpha \in \{0, 10\}$.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>Oracle</th>
<th>PEB</th>
<th>RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>125.0 (6.612)</td>
<td>126.0 (6.560)</td>
<td>127.7 (6.830)</td>
</tr>
<tr>
<td>10</td>
<td>118.4 (5.808)</td>
<td>126.7 (7.125)</td>
<td>122.1 (6.133)</td>
</tr>
</tbody>
</table>

point of view. An NPEB approach based on the RE algorithm—and its extension, the RE+ algorithm, described in Chapter 5—might potentially fill this void.

1.5 Summary of results

Despite the promising performance of RE in simulations, theoretical investigations into the convergence properties have thus far been rather limited. The primary goal of this thesis is to explore the asymptotic properties of the RE algorithm.

We start in Chapter 2 by placing the algorithm on a firm mathematical foundation. The appropriate context, suggested by Newton himself, is stochastic approximation. There we show that the RE algorithm is just a special case of classical stochastic approximation and tools thereof are used to prove a consistency theorem for finite mixtures which is slightly stronger than that of Ghosh and Tokdar [47]. While the conditions are somewhat stronger than desirable, this form of analysis does shed some light on what the algorithm is actually doing.

In Chapter 3 we present a general analysis of the RE algorithm. There we show that, even when the mixture model (1.1) is mis-specified, the estimated mixture $m_n$ converges to the “best possible” mixture of the specified form, where “best” is measured in a Kullback-Leibler (KL) sense. From the proof of this result, we immediately obtain a bound on the rate of convergence, which turns out to be competitive with other convergence rates in the mixture density estimation context. Some numerical
calculations further demonstrate that this bound is not sharp. The proof of these results relies heavily on martingale theory and properties of the KL divergence which, incidentally, are closely related to the ODE and Lyapunov function theory used in Chapter 2. Under some additional conditions, including identifiability, convergence of $f_n$ is established. The dependence of RE on the order of the data is then considered, and it is shown that an average of RE over permutations (PARE) of the data sequence also leads to consistent estimates.

Some numerical investigations into the finite-sample performance of RE and PARE are presented in Chapter 4. There we see the accuracy and computational efficiency of the two estimates. Some comparisons with DP mixtures are also given.

Allowing unknown non-mixing parameters $\xi$ in the RE algorithm is important for many practical problems. In Chapter 5 we propose a “likelihood-based” extension, called the RE+ algorithm, that uses the RE algorithm to (asymptotically) find the best mixing distribution for given $\xi$ and then numerical optimization to find the “best” value of $\xi$. Again “best” is measured in terms of KL divergence. The RE+ algorithm is given a Bayesian perspective and then some applications in density estimation are highlighted. Motivated by the connection between RE+ and nonparametric Bayes methods, in Chapter 6 we then propose a general nonparametric empirical Bayes framework applicable in a host of interesting high-dimensional problems. Our version of the “two-groups model” has the flexibility and the structure necessary to produce reliable results under a wide range of models. Simulations and real-data examples demonstrate that our NPEB method using the RE+ algorithm performs well compared to other popular empirical Bayes methods in the literature.

Finally, in Chapter 7 we discuss several interesting unanswered questions and some directions for future work in this area.
2. NEWTON’S ALGORITHM AND STOCHASTIC APPROXIMATION

Stochastic approximation (SA) was introduced by Robbins and Monro [93] as an algorithmic method for finding the root of a function $h$ when only noisy observations on $h$ are available. It has since developed into an important area of systems control and optimization, with numerous applications in statistics. After a brief introduction, some intuition, and motivating examples, we turn to some of the theoretical aspects. The convergence analysis of a SA algorithm turns out to be closely related to martingales and ordinary differential equations (ODEs). In Section 2.1.2 we develop the necessary stability theory for ODEs, including Lyapunov functions, and then we state a theorem providing sufficient conditions for almost sure convergence of a SA algorithm in Section 2.1.3.

It turns out that Newton’s recursive algorithm is closely related to stochastic approximation. Consider the finite mixture model (1.3) of the general form

$$m_f(x) = \sum_{s=1}^{S} p(x|\theta^s)f(\theta^s),$$

where $\Theta = \{\theta^1, \ldots, \theta^S\}$ is a known finite set and $f$ is a density on $\Theta$ with respect to counting measure $\mu$. In Section 2.2 we show that, for finite mixtures, the recursive algorithm is nothing but a special case of SA. We then apply the general results for convergence of SA algorithm to prove consistency of the recursive estimates under relatively mild conditions. The critical part of the proof is showing that the Kullback-Leibler divergence is a Lyapunov function for the governing ODE.

In Section 2.3 we consider the case where additional non-mixing parameters $\xi$ are present in the mixture model (1.1). An important example is when the conditional distribution of $X_i$ is $N(\theta_i, \sigma^2)$, where the scale parameter $\sigma$ is unknown. As discussed
in Chapter 1, Newton’s original algorithm is not able to handle the additional parameter, thus limiting its scope of applications. Here we propose a modified algorithm, which we call the RE+R, capable of recursively estimating both the mixing density $f$ and the non-mixing parameter $\xi$. The critical assumption is that, in the hierarchical representation (1.2), replicates $X_{i1}, \ldots, X_{ir}$, for $r \geq 2$, are available from $p(\cdot | \theta_i, \xi)$ for each $i = 1, \ldots, n$. In microarray experiments, for example, where estimating the mixing distribution is crucial to the analysis, replicates are typically available. We express this algorithm as a general SA and prove consistency in the case of finite $\Theta$ under suitable conditions.

2.1 Stochastic approximation

2.1.1 Algorithm and examples

Consider the problem of finding the unique root $\xi$ of a function $h(x)$. If $h(x)$ can be evaluated exactly for each $x$ and if $h$ is sufficiently smooth, then various numerical methods can be employed to locate $\xi$. A majority of these numerical procedures, including the popular Newton-Raphson method, are iterative by nature, starting with an initial guess $x_0$ of $\xi$ and iteratively defining a sequence $\{x_n\}$ that converges to $\xi$ as $n \to \infty$. Now consider the situation where only noisy observations on $h(x)$ are available; that is, for any input $x$ one observes $y = h(x) + \varepsilon$, where $\varepsilon$ is a zero mean random error. This problem arises in situations where $h(x)$ denotes the expected value of the response when the experiment is run at setting $x$. Unfortunately, standard deterministic methods cannot be used in this problem.

In their seminal paper, Robbins and Monro [93] proposed a stochastic approximation algorithm for defining a sequence of design points $\{x_n\}$ targeting the root $\xi$ of $h$ in this noisy case. Start with an initial guess $x_0$. At stage $n \geq 1$, use the state $x_{n-1}$ as the input, observe $y_n = h(x_{n-1}) + \varepsilon_n$, and update the guess $(x_{n-1}, y_n) \mapsto x_n$. More precisely, the Robbins-Monro algorithm defines the sequence $\{x_n\}$ as follows.
Robbins-Monro Algorithm  Start with $x_0$ and, for $n \geq 1$, set

$$x_n = x_{n-1} + w_n y_n = x_{n-1} + w_n \{ h(x_{n-1}) + \varepsilon_n \}$$  \hspace{1cm} (2.1)$$

where $\{ \varepsilon_n \}$ is a sequence of iid random variables with mean zero, and the weight sequence $\{ w_n \}$ satisfies

$$w_n > 0, \quad \sum_n w_n = \infty, \quad \sum_n w_n^2 < \infty.$$  \hspace{1cm} (2.2)$$

While the SA algorithm above works in more general situations, we can develop our intuition by looking at the special case considered in Robbins and Monro [93], namely, when $h$ is bounded, continuous and monotone decreasing. If $x_n < \xi$ then $h(x_n) > 0$ and we have

$$E(x_{n+1} | x_n) = x_n + w_{n+1} \{ h(x_n) + E(\varepsilon_{n+1}) \} = x_n + w_{n+1} h(x_n) > x_n.$$  

Likewise, if $x_n > \xi$, then $E(x_{n+1} | x_n) < x_n$. This shows that the move $x_n \mapsto x_{n+1}$ will be in the correct direction on average.

Some remarks on the conditions in (2.2) are in order. While $\sum_n w_n^2 < \infty$ is necessary to prove convergence, an immediate consequence of this condition is that $w_n \to 0$. Clearly $w_n \to 0$ implies that the effect of the noise vanishes as $n \to \infty$. This, in turn, has an averaging effect on the iterates $y_n$. On the other hand, the condition $\sum_n w_n = \infty$ washes out the effect of the initial guess $x_0$. For further details, see Nevel’son and Has’minskii [79].

We conclude this section with three simple examples of SA to shed light on when and how the algorithm works. Example 2.1, taken from Kushner and Yin [57], is an important special case of the Robbins-Monro algorithm (2.1) which further motivates the algorithm as well as the conditions (2.2) on the sequence $\{ w_n \}$. Example 2.2 uses SA to find quantiles of a $t$-distribution, and Example 2.3 illustrates a connection between SA and empirical Bayes inference, two of Robbins’ greatest contributions.

**Example 2.1** Let $F_\xi$ be the cdf of a distribution with mean $\xi$. Then estimation of $\xi$ is equivalent to solving $h(x) = 0$ where $h(x) = \xi - x$. If $Z_1, \ldots, Z_n \sim F_\xi$, then
the average \( \overline{Z}_n \) is the least squares estimate of \( \xi \). To see that \( \{\overline{Z}_n\} \) is actually a SA sequence, recall the computationally efficient recursive expression for \( \overline{Z}_n \):

\[
\overline{Z}_n = \overline{Z}_{n-1} + n^{-1}(Z_n - \overline{Z}_{n-1}).
\]

(2.3)

If we let \( x_n = \overline{Z}_n \), \( w_n = n^{-1} \), and \( y_n = Z_n - \overline{Z}_{n-1} \), then (2.3) is exactly of the form of (2.1), with \( \{w_n\} \) satisfying (2.2). Moreover, if \( \varepsilon_n = Z_n - \xi \), then we can write \( y_n = h(x_{n-1}) + \varepsilon_n \). With this setup, we could study the asymptotic behavior of \( x_n \) using the SA analysis below (see Sections 2.1.2–2.1.3), although the SLLN already guarantees \( x_n \to \xi \) a.s.

**Example 2.2** Suppose we wish to find the \( \alpha \)th quantile of the \( t_\nu \) distribution; that is, we want to find the solution to the equation \( F_\nu(x) = \alpha \), where \( F_\nu \) is the cdf of the \( t_\nu \) distribution. While there are numerous numerical methods available (e.g. Newton-Raphson or bijection), we demonstrate below how SA can be used to solve this problem. Making use of the well known fact that the \( t_\nu \) distribution is a scale-mixture of normals, we can write

\[
F_\nu(x) = \mathbb{E} \left[ \Phi(x \mid \nu^{-1}Z) \right], \quad Z \sim \chi^2_\nu
\]

where \( \Phi(x\mid\sigma^2) \) is the cdf of the \( N(0,\sigma^2) \) distribution. Now, for \( Z_1, Z_2, \ldots \sim \chi^2_\nu \), the sequence \( \{y_n\} \) defined as \( y_n = \alpha - \Phi(x_{n-1} \mid \nu^{-1}Z_n) \) are noisy observations of \( h(x_{n-1}) = \alpha - F_\nu(x_{n-1}) \). This \( h \) is bounded, continuous and monotone decreasing so the Robbins-Monro theory says that the sequence \( \{x_n\} \) defined as (2.1) converges to the true quantile, for any initial condition \( x_0 \).

**Example 2.3** In this simple example, we demonstrate the connection between SA and empirical Bayes (EB) estimation, both of which are theories pioneered by Robbins. Consider the simple hierarchical model

\[
\lambda_1, \ldots, \lambda_n \sim \text{Exp}(\xi) \quad \text{and} \quad Z_i \mid \lambda_i \sim \text{Poi}(\lambda_i)
\]

for \( i = 1, \ldots, n \), where the exponential rate \( \xi > 0 \) is unknown. EB tries to estimate \( \xi \) based on the observed data \( Z_1, \ldots, Z_n \). Here we consider a recursive estimate of \( \xi \).
Fix an initial guess $x_0$ of $\xi$. Assuming $\xi$ is equal to $x_0$, the posterior mean of $\lambda_1$ is $(Z_1 + 1)/(x_0 + 1)$, which is a good estimate of $\xi^{-1}$ if $x_0$ is close to $\xi$. Iterating this procedure, we can generate a sequence

$$x_i = x_{i-1} + w_i \left[ \frac{1}{x_{i-1}} - \frac{Z_i + 1}{x_{i-1} + 1} \right],$$

(2.4)

where $\{w_i\}$ is assumed to satisfy (2.2). Let $y_i$ denote the quantity in brackets in (2.4) and take its expectation with respect to the distribution of $Z_i$:

$$h(x) = \mathbb{E}(y_i \mid x_{i-1} = x) = \frac{\xi - x}{\xi x(x + 1)}.$$ 

(2.5)

Then the sequence $\{x_n\}$ in (2.4) is a SA targeting a solution of $h(x) = 0$. Since $h$ is continuous, decreasing and $h(x) = 0$ iff $x = \xi$, it follows from the general theory that $x_n \to \xi$. Figure 2.1 shows the first 250 steps of such a sequence with $x_0 = 1.5$.

![Figure 2.1](image.png)

Figure 2.1. Sample path of the sequence $\{x_n\}$ in Example 2.3. The dotted line is the value of $\xi$ used for data generation.
The examples above emphasize one important property that \( h(x) \) must satisfy; namely, that it must be easy to “sample” in the sense that there is a function \( H(x, z) \) and a random variable \( Z \) such that \( h(x) = \mathbb{E}[H(x, Z)] \). Another thing, which is not obvious from the examples, is that \( h(x) \) must have certain stability properties. In general, a SA sequence need not have a unique limit point. However, conditions can be imposed which guarantee convergence to a particular solution \( \xi \) of \( h(x) = 0 \), provided that \( \xi \) is a stable solution to the ODE \( \dot{x} = h(x) \). This is discussed further in Section 2.1.2.

### 2.1.2 ODE stability theory

The asymptotic theory of ODEs plays an important role in the convergence analysis of a SA algorithm. After showing the connection between SA and ODEs, we briefly review some of the ODE theory that is necessary in the sequel.

Recall the general SA algorithm in (2.1) given by \( x_n = x_{n-1} + w_n y_n, \ n \geq 1 \). Assume there is a measurable regression function \( h \) such that \( h(x) = \mathbb{E}\{y_n|x\} \) and rewrite the algorithm as

\[
x_n = x_{n-1} + w_n h(x_{n-1}) + w_n \{y_n - h(x_{n-1})\}.
\]

Define \( M_n = y_n - h(x_{n-1}) \). Then \( \{M_n\} \) a zero mean martingale and, under suitable conditions, the martingale convergence theorem—see Billingsley [10]—guarantees that \( M_n \) becomes negligible as \( n \to \infty \), leaving us with

\[
x_n = x_{n-1} + w_n h(x_{n-1}) + w_n M_n \approx x_{n-1} + w_n h(x_{n-1}).
\]

But this latter “mean trajectory” is deterministic and essentially a finite difference equation with small step sizes. Rearranging the terms gives us

\[
\frac{x_n - x_{n-1}}{w_n} = h(x_{n-1}),
\]

which, for large \( n \), can be approximated by the ODE \( \dot{x} = h(x) \). It is now clear how SA algorithms are related to the asymptotic properties of solutions to ODEs.
Consider a general autonomous ODE \( \dot{x} = h(x) \), where \( h : \mathbb{R}^S \to \mathbb{R}^S \) is a bounded and continuous, possibly non-linear, function. A solution \( x(t) \) of the ODE is a trajectory in \( \mathbb{R}^S \) with a given initial condition \( x(0) \). Unfortunately, in many cases, a closed form expression for a solution \( x(t) \) is not available. For that reason, other methods are necessary for studying these solutions and, in particular, their properties as \( t \to \infty \).

Imagine a physical system, such as an orbiting celestial body, whose state is being governed by the ODE \( \dot{x} = h(x) \) with initial condition \( x(0) = x_0 \). Then, roughly speaking, the system is stable if choosing an alternative initial condition \( x(0) = \tilde{x}_0 \) in a neighborhood of \( x_0 \) has little effect on the asymptotic properties of the resulting solution \( x(t) \). The following definition makes this more precise.

**Definition 2.4** A point \( \xi \in \mathbb{R}^S \) is said to be locally stable for \( \dot{x} = h(x) \) if for each \( \varepsilon > 0 \) there is a \( \delta > 0 \) such that if \( \|x(0) - \xi\| < \delta \), then \( \|x(t) - \xi\| < \varepsilon \) for all \( t \geq 0 \). If \( \xi \) is locally stable and \( x(t) \to \xi \) as \( t \to \infty \), then \( \xi \) is locally asymptotically stable. If this convergence holds for all initial conditions \( x(0) \), then the asymptotic stability is said to be global.

Points \( \xi \) for which stability is of interest are equilibrium points of \( \dot{x} = h(x) \). Any point \( \xi \) such that \( h(\xi) = 0 \) is called an equilibrium point, since the constant solution \( x(t) \equiv \xi \) satisfies \( \dot{x} = h(x) \).

**Example 2.5** Let \( \dot{x} = Ax \), where \( A \) is a fixed \( S \times S \) matrix. For an initial condition \( x(0) = x_0 \), we can write an explicit formula for the particular solution: \( x(t) = e^{At}x_0 \) for \( t \geq 0 \). Suppose, for simplicity, that \( A \) has a spectral decomposition \( A = U\Lambda U' \), where \( U \) is orthogonal and \( \Lambda \) is a diagonal matrix of the eigenvalues \( \lambda_1, \ldots, \lambda_S \) of \( A \). Then the matrix exponential can be written as \( e^{At} = Ue^{\Lambda t}U' \), where \( e^{\Lambda t} \) is diagonal with \( s^{\text{th}} \) element \( e^{\lambda_s t} \). Clearly, if \( \lambda_s < 0 \), then \( e^{\lambda_s t} \to 0 \) as \( t \to \infty \). Therefore, if \( A \) is negative definite or, more generally, Hurwitz (each eigenvalue has negative real part), then the origin \( x = 0 \) is globally asymptotically stable.
When explicit solutions are not available, proving asymptotic stability for a given equilibrium point will require a so-called Lyapunov function; see, for example, LaSalle and Lefschetz [59].

**Definition 2.6** Let $\xi \in \mathbb{R}^S$ be an equilibrium point of the ODE $\dot{x} = h(x)$. A function $\ell : \mathbb{R}^S \rightarrow \mathbb{R}$ is called a Lyapunov function at $x = \xi$ if

L1. $\ell$ has continuous first partial derivatives in a neighborhood of $\xi$;

L2. $\ell(x) \geq 0$ with equality if and only if $x = \xi$;

L3. The time derivative $\dot{\ell}(x) = \nabla \ell(x)'h(x)$ of $\ell$ along the path $x = x(t)$ is $\leq 0$.

A Lyapunov function is said to be strong if $\dot{\ell}(x) = 0$ implies $x = \xi$.

Lyapunov functions are a generalization of the potential energy of a system, such as a swinging pendulum, and Lyapunov’s theory gives a formal extension of the stability principles of such a system. Theorem 2.7 is very powerful because it does not require an explicit formula for the solution. See LaSalle and Lefschetz [59] for a proof of this theorem and various extensions of the Lyapunov theory.

**Theorem 2.7** If there exists a (strong) Lyapunov function in a neighborhood of an equilibrium point $\xi$ of $\dot{x} = h(x)$, then $\xi$ is (asymptotically) stable.

There is no general recipe for constructing a Lyapunov function. In one important special case, however, a candidate Lyapunov function is easy to find. Suppose $h(x) = -\nabla g(x)$, for some positive-definite, sufficiently smooth function $g$. Then $\ell(x) = g(x)$ is a Lyapunov function since $\dot{\ell}(x) = -\|\nabla g(x)\|^2 \leq 0$.

**Example 2.8** Consider again the linear system $\dot{x} = Ax$ from Example 2.5, where $A$ is a $S \times S$ negative definite matrix. Here we will derive asymptotic stability by finding a Lyapunov function and applying Theorem 2.7. In light of the previous remark, we choose $\ell(x) = -\frac{1}{2}x'Ax$. Then $\dot{\ell}(x) = -\|Ax\|^2 \leq 0$ so $\ell$ is a strong Lyapunov function for $\dot{x} = Ax$ and the origin is asymptotically stable by Theorem 2.7.
Of interest is the stronger conclusion of global asymptotic stability, but Theorem 2.7 does not tell us how far the initial condition $x_0$ can be from the equilibrium in question. For the results to follow, we will prove the global part directly.

2.1.3 SA Convergence Theorem

Consider, for fixed $x_0$ and $\{w_n\}$ satisfying (2.2), the general SA algorithm

$$x_n = \text{Proj}_X \{x_{n-1} + w_n y_n\}, \quad n \geq 1. \quad (2.6)$$

where $X \subset \mathbb{R}^S$ is compact and $\text{Proj}_X(x)$ is a projection of $x$ onto $X$. The projection is necessary when boundedness of the iterates cannot be established by other means. The truncated or projected algorithm (2.6) is often written in the alternative form

$$x_n = x_{n-1} + w_n y_n + w_n z_n, \quad (2.7)$$

where $z_n$ is the “minimum” $z$ such that $x_{n-1} + w_n y_n + w_n z$ belongs to $X$; see Kushner and Yin [57] for the precise definition.

Next we state the main stochastic approximation result used in the sequel, a special case of Theorem 5.2.3 in Kushner and Yin [57]. First define the filtration sequence $\mathcal{A}_n = \sigma(y_1, \ldots, y_n)$.

**Theorem 2.9** For $\{x_n\}$ in (2.6) with $\{w_n\}$ satisfying (2.2), assume

SA1. $\sup_n E\|y_n\|^2 < \infty$;

SA2. There exists a continuous function $h(\cdot)$ and a random vector $\beta_n$ such that $E\{y_n|\mathcal{A}_{n-1}\} = h(x_{n-1}) + \beta_n$ a.s. for each $n$;

SA3. $\sum_n w_n\|\beta_n\|$ converges a.s.

*If $\xi$ is globally asymptotically stable for $\dot{x} = h(x)$, then $x_n \to \xi$ a.s.*
2.2 Newton’s estimate as SA

Let $\mu$ be counting measure and $S = \mu(\Theta)$. We can think of the density $f_n(\theta)$ in (1.6) as a vector $f_n = (f_n^1, \ldots, f_n^S)'$ in the probability simplex $\Delta^S$, defined as

$$\Delta^S = \left\{ (\varphi^1, \ldots, \varphi^S)' \in [0,1]^S : \sum_{s=1}^{S} \varphi^s = 1 \right\}.$$

Define $H : \mathcal{X} \times \Delta^S \to \mathbb{R}^S$ with $s^{th}$ component

$$H_s(x, \varphi) = p(x|\theta^s)\varphi^s - \varphi^s, \quad s = 1, \ldots, S,$$

(2.8)

where $m_\varphi(x) = \sum_s p(x|\theta^s)\varphi^s$ is the marginal density on $\mathcal{X}$ induced by $\varphi \in \Delta^S$. Then (1.6) becomes

$$f_n = f_{n-1} + w_n H(X_n, f_{n-1}).$$

(2.9)

Let $P_x = \text{diag} \{ p(x|\theta^s) : s = 1, \ldots, S \}$ be the diagonal matrix of the sampling density values at $x$, and define the mapping $h : \Delta^S \to \mathbb{R}^S$ to be the conditional expectation of $H(x, f_n)$, given $f_n = \varphi$:

$$h(\varphi) = \int_{\mathcal{X}} H(x, \varphi) m_f(x) d\nu(x) = \int_{\mathcal{X}} \frac{m_f(x)}{m_\varphi(x)} P_x \varphi d\nu(x) - \varphi,$$

(2.10)

where $f = (f^1, \ldots, f^S)'$ is the true mixing/prior distribution. From (2.10), it is clear that $f$ solves the equation $h(\varphi) = 0$ which implies (i) $f$ is an equilibrium point of the ODE $\dot{\varphi} = h(\varphi)$, and (ii) that $f$ is a fixed point of the map

$$T(\varphi) = h(\varphi) + \varphi = \int \frac{m_f(x)}{m_\varphi(x)} P_x \varphi d\nu(x).$$

Newton [80, p. 313] recognized the importance of this map in relation to the limit of $f_n$. Also, the use of $T$ in Shyamalkumar [103] for the $I_0$-projection problem is closely related to the SA approach taken here.

We have shown that (2.9) can be considered as a general SA algorithm, targeting the solution $\varphi = f$ of the equation $h(\varphi) = 0$ in $\Delta^S$. Therefore, the SA results of Section 2.1.3 can be used in the convergence analysis.

Theorem 2.10 below requires the following conditions:
A2.1. \( \Theta \) is finite and \( \mu \) is counting measure.

A2.2. \( \{w_n\} \) satisfies (2.2).

A2.3. \( f \) is identifiable; cf. Definition 1.1.

A2.4. \( p(\cdot|\theta) > 0 \) \( \nu\)-a.e. for each \( \theta \in \Theta \).

The condition A2.4 is considerably weaker than the assumption that \( p(x|\theta) \) is uniformly bounded away from 0 and \( \infty \) in the consistency result of Ghosh and Tokdar [47]. The following theorem, therefore, applies to many important cases, such as Normal or Gamma mixtures.

**Theorem 2.10** Assume A2.1–A2.4. Then \( f_n \to f \) a.s.

**Proof** To prove the theorem, we need only show that the algorithm (2.9) satisfies the conditions of Theorem 2.9. First note that \( f_n \) is, for each \( n \), a convex combination of points in the interior of \( \Delta^S \) so no projection as in (2.6) is necessary. Second, the random variables \( \beta_n \) in assumption SA2 are identically zero so SA3 is trivially satisfied. Let \( \{u_n\} \) be a convergent sequence in \( \Delta^S \), where \( u_n = (u_n^1, \ldots, u_n^S)' \). The limit \( u = (u^1, \ldots, u^S)' = \lim_{n \to \infty} u_n \) also belongs to \( \Delta^S \) so \( h(u) \) is well-defined. To prove that \( h = (h_1, \ldots, h_S)' \) is continuous, we show that \( h_s(u_n) \to h_s(u) \) for each \( s = 1, \ldots, S \) as \( n \to \infty \). Consider

\[
h_s(u_n) = \int \frac{p(x|\theta^s)u_n^s m_f(x)}{m_{an}(x)} d\nu(x) - u_n^s.
\]

The integrand \( p(\cdot|\theta^s)u_n^s / m_{an}(\cdot) \) is non-negative and \( \nu\)-a.e. bounded for each \( s \). Then by the bounded convergence theorem we get

\[
\lim_{n \to \infty} h_s(u_n) = h_s(u), \quad s = 1, \ldots, S.
\]

But \( \{u_n\} \subset \Delta^S \) was arbitrary so \( h \) is continuous.

Next, note that \( H(x, f_n) \) is the difference of two points in \( \Delta^S \) and is thus bounded independent of \( x \) and \( n \). Then SA1 holds trivially.
Finally, we show that $f$ is globally asymptotically stable for the ODE $\dot{\varphi} = h(\varphi)$ in $\Delta^S$. Note that $\sum_{s=1}^{S} \varphi_s = \sum_{s=1}^{S} h_s(\varphi) = 0$ so the trajectories lie on the connected and compact $\Delta^S$. Let $\ell(\varphi)$ be the KL divergence $K(f, \varphi)$ defined, in this case, by $K(f, \varphi) = \sum_{s=1}^{S} f_s \log(f_s / \varphi_s)$. We claim that $\ell$ is a strong Lyapunov function for $\dot{\varphi} = h(\varphi)$ at $f$. Certainly $\ell(\varphi)$ is positive definite. To check the differentiability condition L1, we must show that $\ell(\varphi)$ has a well-defined gradient around $f$, even when $f$ is on the boundary of $\Delta^S$. Suppose, without loss of generality, that $f^1, \ldots, f^d$ are positive, $1 \leq d \leq S$, and the remaining $f^{d+1}, \ldots, f^S$ are zero. By definition, $\ell(\varphi)$ is constant in $\varphi^{d+1}, \ldots, \varphi^S$ and, therefore, the partial derivatives with respect to those $\varphi$’s are zero. Thus, for any $1 \leq d \leq S$ and for any $\varphi$ such that $\ell(\varphi) < \infty$, the gradient can be written as

$$
\nabla \ell(\varphi) = -(r^1, r^2, \ldots, r^d)' + r^d 1_d',
$$

where $r^s = f^s / \varphi^s$ and $1_d$ is a vector whose first $d$ coordinates are one and last $S - d$ coordinates are zero. The key point here is that the gradient of $\ell(\varphi)$, for $\varphi$ restricted to the boundary which contains $f$, is exactly the same as (2.11). We can, therefore, extend the definition of $\nabla \ell(\varphi)$ continuously to the boundary if need be.

Given that $\nabla \ell(\varphi)$ exists on all of $\Delta^S$, the time derivative of $\ell$ along $\varphi$ is

$$
\dot{\ell}(\varphi) = \nabla \ell(\varphi)' h(\varphi)
= \int \frac{m_f(x) - m_{\varphi}(x)}{m_{\varphi}(x)} \nabla \ell(\varphi)' P_{x\varphi} d\nu(x)
= 1 - \int \frac{m_f}{m_{\varphi}} m_f d\nu
$$

(2.12)

It remains to show that $\dot{\ell}(\varphi) = 0$ if and only if $\varphi = f$. Applying Jensen’s inequality to $y \mapsto y^{-1}$ in (2.12) gives

$$
\dot{\ell}(\varphi) = 1 - \int_X \left( \frac{m_{\varphi}}{m_f} \right)^{-1} m_f d\nu \leq 1 - \left( \int_X \frac{m_{\varphi}}{m_f} m_f d\nu \right)^{-1} = 0
$$

(2.13)

where equality can hold in (2.13) iff $m_{\varphi} = m_f \nu$-a.e. We assume the mixtures are identifiable, so this implies $\varphi = f$. Therefore, $\dot{\ell}(\varphi) = 0$ if and only if $\varphi = f$, and
we have shown that \( \ell \) is a strong Lyapunov function on \( \Delta^S \). To prove that \( f \) is a globally asymptotically stable point for \( \dot{\varphi} = h(\varphi) \), suppose that \( \varphi(t) \) is a solution, with \( \varphi(0) = f_0 \), that does not converge to \( f \). Since \( \ell \) is a strong Lyapunov function, the sequence \( \ell(\varphi(t)) \), as \( t \to \infty \), is bounded, strictly decreasing and, thus, has a limit \( \lambda > 0 \). Then the trajectory \( \varphi(t) \) must fall in the set

\[
\Delta^* = \{ \varphi \in \Delta^S : \lambda \leq \ell(\varphi) \leq \ell(f_0) \}
\]

for all \( t \geq 0 \). In the case \( f \in \text{int}(\Delta^S) \), \( \ell(\varphi) \to \infty \) as \( \varphi \to \partial\Delta^S \), so the set \( \Delta^* \) is compact (in the relative topology). If \( f \in \partial\Delta^S \), then \( \Delta^* \) is not compact but, as shown above, \( \ell(\varphi) \) is well-defined and continuous there. In either case, \( \ell \) is continuous and bounded away from zero on \( \Delta^* \), so

\[
\sup_{\varphi \in \Delta^*} \ell(\varphi) = -L < 0.
\]

Then, for any \( \tau \geq 0 \), we have

\[
\ell(\varphi(\tau)) = \ell(f_0) + \int_0^\tau \dot{\ell}(\varphi(s)) \, ds \leq \ell(f_0) - L\tau.
\]

If \( \tau > \ell(f_0)/L \), then \( \ell(\varphi(\tau)) < 0 \), which is a contradiction. Therefore, \( \varphi(t) \to f \) for all initial conditions \( \varphi(0) = f_0 \), so \( f \) is globally asymptotically stable. Theorem 2.9 then implies \( f_n \to f \) a.s.

**Remark 2.11** One can show that the KL divergence \( \ell(\varphi) = K(m_f, m_\varphi) \) between the mixtures \( m_f \) and \( m_\varphi \) is also a Lyapunov function for the ODE \( \dot{\varphi} = h(\varphi) \) defined by (2.10); this fact is important in Chapter 3.

**Remark 2.12** Theorem 2.10 covers the interior case (when \( f \) is strictly positive) as well as the boundary case (when \( f^s = 0 \) for some \( s \)). The fact that \( f_0^s > 0 \) implies \( f_n^s > 0 \) for all \( n \) suggests that convergence may be slow in the boundary case.

### 2.3 Estimating non-mixing parameters

Suppose now that the condition distribution of \( X \) on \( \mathcal{X} \) is parametrized not only by \( \theta \) but by an additional parameter \( \xi \). An example of this is the Normal distribution
with mean $\theta$ and variance $\xi$. More specifically, we replace the sampling densities $p(x|\theta)$ of Section 2.2 with $p(x|\theta, \xi)$ where $\theta$ is the latent variable, and $\xi$ is also unknown. As discussed in Chapter 1, Newton's algorithm cannot be used in this situation since $\theta$ does not fully specify the sampling density.

In this section we introduce a modification of Newton's algorithm that, when replicates are available, can simultaneously and recursively estimate both a mixing distribution and an additional unknown parameter. This modification, called the RE$+R$ algorithm, is actually quite simple—at each step we use a plug-in estimate of $\xi$ calculated from the replicates in the update (1.6). We show that the RE$+R$ algorithm can be written as a general SA algorithm and, under certain conditions, prove its consistency.

Let $p(x|\theta, \xi)$ be a two-parameter family of densities on $\mathcal{X}$, and consider the model

$$
\begin{align*}
\theta_1, \ldots, \theta_n & \overset{iid}{\sim} f \\
X_{i1}, \ldots, X_{ir} & \overset{iid}{\sim} p(\cdot|\theta_i, \xi), \quad i = 1, \ldots, n
\end{align*}
$$

(2.14)

where $f$ is an unknown density on $\Theta$ and the parameter $\xi \in \Xi$ is also unknown. The number of replicates $r \geq 2$ is assumed fixed. In applications, such as DNA microarray experiments, the number of replicates $r$ might be around ten. Note that (2.14) is nothing but a traditional random effects model, except that we assume a completely nonparametric model for the random effects.

Assume, for simplicity, that $\Xi \subseteq \mathbb{R}$; the more general case $\Xi \subseteq \mathbb{R}^p$ is a natural extension of what follows. Let $\Theta = \{\theta^1, \ldots, \theta^S\}$ be a finite set and take $\mu$ to be counting measure on $\Theta$. Recall that $\Delta^S$ is the probability simplex. Assume:

A2.5. $\xi \in \text{int}(\Xi_0)$, where $\Xi_0$ is a compact and convex subset of $\Xi$.

A2.6. $f \in \text{int}(\Delta_0)$ where $\Delta_0 \subset \Delta^S$ is compact and, for each $\varphi \in \Delta_0$, the coordinates $\varphi^1, \ldots, \varphi^S$ are bounded away from zero.

The subset $\Xi_0$ can be arbitrarily large so assumption A2.5 causes no difficulty in practice. Assumption A2.6 is somewhat restrictive in that $f$ must be strictly posi-
tive. While A2.6 seems necessary to prove consistency, simulations suggest that this assumption can be weakened.

The RE+R algorithm uses an estimate of $\xi$ at each step in Newton’s algorithm. We assume here that an unbiased estimate of $\xi$ is available:

A2.7. $T(X)$, for $X \in X^r$, is an unbiased estimate of $\xi$, with variance $\tau^2 < \infty$.

Later we will replace the unbiased estimate with a Bayes estimate.

At time $i = 1, \ldots, n$, we observe an $r$-vector $X_i = (X_{i1}, \ldots, X_{ir})'$ and we compute $\hat{\xi}^{(i)} = T(X_i)$. An unbiased estimate of $\xi$ based on the entire data $X_1, \ldots, X_n$ would be the average $\hat{\xi}_n = n^{-1} \sum_{i=1}^n \hat{\xi}^{(i)}$, which has a convenient recursive expression

$$\hat{\xi}_i = i^{-1} [(i - 1)\hat{\xi}_{i-1} + \hat{\xi}^{(i)}], \quad i = 1, \ldots, n. \quad (2.15)$$

More importantly, by construction, $\hat{\xi}^{(1)}, \ldots, \hat{\xi}^{(n)}$ are iid random variables with mean $\xi$ and finite variance. It is, therefore, a consequence of the SLLN that $\hat{\xi}_n$, as defined in (2.15), converges a.s. to $\xi$. While this result holds for any unbiased estimate $T$, an unbiased estimate $T'$ with smaller variance is preferred.

Define the mapping $H: X^r \times \Delta_0 \times \Xi_0 \to \mathbb{R}^S$ with $s$th component

$$H_s(x, \varphi; \xi) = \frac{\overline{p}(x|\theta^s, \xi)\varphi^s}{\sum_t \overline{p}(x|\theta^t, \xi)\varphi^t} - \varphi^s, \quad (2.16)$$

for $s = 1, \ldots, S$, where $\overline{p}(\cdot|\theta, \xi)$ is the joint density of an iid sample of size $r$ from $p(\cdot|\theta, \xi)$. Note that we make no notational distinction here between a generic element $\xi \in \Xi_0$ and its true value.

**RE+R Algorithm** Choose an initial estimate $f_0 \in \Delta_0$, weights $w_1, \ldots, w_n \in (0, 1)$, and an arbitrary $\xi_0 \in \Xi_0$. Then for $i = 1, \ldots, n$ compute

$$\xi_i = \text{Proj}_{\Xi_0} \left\{ i^{-1} \left[ (i - 1)\xi_{i-1} + \hat{\xi}^{(i)} \right] \right\}$$

$$f_i = \text{Proj}_{\Delta_0} \{ f_{i-1} + w_i H(X_i; f_{i-1}, \xi_i) \}$$

and produce $(f_n, \xi_n)$ as the final estimate.
We claim that the RE+R algorithm for estimating $f$ can be written as a general SA involving the true but unknown $\xi$ plus an additional perturbation. Define the quantities
\begin{align}
h(f_{n-1}) &= E[H(X_n, f_{n-1}, \xi) | \mathcal{A}_{n-1}] \tag{2.17} \\
\beta_n &= E[H(X_n, f_{n-1}, \xi_n) | \mathcal{A}_{n-1}] - E[H(X_n, f_{n-1}, \xi) | \mathcal{A}_{n-1}] \tag{2.18}
\end{align}

where $\mathcal{A}_{n-1} = \sigma(X_1, \ldots, X_{n-1})$, so that

$E[H(X_n, f_{n-1}, \xi_n) | \mathcal{A}_{n-1}] = h(f_{n-1}) + \beta_n$.

Now the update $f_{n-1} \mapsto f_n$ can be written as

$f_n = f_{n-1} + w_n \{h(f_{n-1}) + M_n + \beta_n + z_n\}, \tag{2.19}$

where $z_n$ is the “minimum” $z$ keeping $f_n$ in $\Delta_0$, and

$M_n = H(X_n, f_{n-1}, \xi_n) - h(f_{n-1}) - \beta_n$

is a martingale adapted to $\mathcal{A}_{n-1}$. Notice that (2.19) is now in a form in which Theorem 2.9 can be applied. We will make use of the Law of Iterated Logarithm so define $u(t) = (2t \log \log t)^{1/2}$.

**Theorem 2.13** Assume A2.1–A2.7 and, in addition, assume

A2.8. $\frac{\partial}{\partial \xi} H(x; \varphi, \xi)$ is bounded on $X^r \times \Delta_0 \times \Xi_0$

A2.9. $\sum_n w_n n^{-1} u(n) < \infty$.

Then $(f_n, \xi_n) \rightarrow (f, \xi)$ a.s. as $n \rightarrow \infty$.

The proof of Theorem 2.13 requires the following lemma, establishing a Lipschitz-type bound on the error terms $\beta_n$ in (2.18). Its proof follows immediately from A2.8 and the Mean Value Theorem.

**Lemma 2.14** Under the assumptions of Theorem 2.13, there exists a finite positive number $A$ such that $\|\beta_n\| \leq AE\{ |\xi_n - \xi| \mid \mathcal{A}_{n-1} \}$. 
We are now ready to prove Theorem 2.13.

**Proof** (of Theorem 2.13) The map $h$ in (2.17) has $s^{th}$ component

$$h_s(\varphi) = \int H(x; \varphi, \xi) m_{f,\xi}(x) \, d\nu^r(x) = \int \frac{m_{f,\xi}(x)}{m_{\varphi,\xi}(x)} p(x|\theta^s, \xi) \varphi^s \, d\nu^r(x) - \varphi^s,$$

where $m_{f,\xi}(x) = \sum_s p(x|\theta^s, \xi)^s$ is the marginal density of $x$ and $\nu^r$ is the product measure on $X^r$. Notice that this $h$, which does not depend on the estimate $\xi_n$, is exactly the same as the $h$ in (2.10). Therefore, the continuity and stability properties derived in the proof of Theorem 2.10 are valid here as well. All that remains is to show that the $\beta_n$'s in (2.18) satisfy SA3 of Theorem 2.9.

By the SLLN, $\xi_n$ belongs to $\Xi_0$ for large enough $n$ so we can assume, without loss of generality, that no projection is necessary. Let $G_n = Z_1 + \cdots + Z_n$ where the $G_i = \frac{1}{\tau^2}(\hat{\xi}(i) - \xi)$ and $\tau^2$ is the variance of $\xi(i)$. Then $|\xi_n - \xi| = cn^{-1}|G_n|$, where $c > 0$ is a constant independent of $n$. Since $G_n$ is a sum of iid random variables with mean zero and unit variance, the Law of Iterated Logarithm states that

$$\limsup_{n \to \infty} \left\{ |G_n|/u(n) \right\} = 1 \quad a.s. \quad (2.20)$$

Now, by Lemma 2.14 and (2.20) we have

$$\|\beta_n\| \leq A cn^{-1} E\{|G_n| \mid \mathcal{A}_{n-1}\} = O(n^{-1}u(n))$$

and, therefore, $\sum_n w_n \|\beta_n\|$ converges a.s. by A2.9. We have shown that condition SA3 is satisfied, so the proof of Theorem 2.13 is complete.

We now remove the restriction to unbiased estimates of $\xi$, focusing primarily on the use of a Bayes estimate in place of the unbiased estimate. But first, let $\hat{\xi}_i = T(X_1, \ldots, X_i)$ be any suitable estimate of $\xi$ based on only $X_1, \ldots, X_i$. Then replace the RE+R update $f_{i-1} \mapsto f_i$ with

$$\tilde{f}_i = \text{Proj}_{\Delta_0} \left\{ \tilde{f}_{i-1} + w_i H(X_i, \tilde{f}_{i-1}, \hat{\xi}_i) \right\}.$$

While this adaptation is more flexible with regards to the choice of estimate, this additional flexibility does not come for free. Notice that the algorithm is no longer
recursive. That is, given a new data point \(x_{n+1}\), we need more information than just the pair \((\hat{f}_n, \hat{\xi}_n)\) to obtain \((\hat{f}_{n+1}, \hat{\xi}_{n+1})\).

**Corollary 2.15** If assumptions A2.7 and A2.9 in Theorem 2.13 are replaced by

A2.7'. \(|\hat{\xi}_n - \xi| = \mathcal{O}(\rho_n)\) a.s. as \(n \to \infty\)

A2.9'. \(\sum_n w_n \rho_n < \infty\)

then \((\hat{f}_n, \hat{\xi}_n) \to (f, \xi)\) a.s. as \(n \to \infty\).

Typically, for Bayes and maximum likelihood estimates, the rate is \(\rho_n = n^{-1/2}\). Then A2.9' holds if, for example, \(w_n \sim n^{-1}\).

To illustrate the RE+R and its modified version, consider the special case where \(p(\cdot|\theta, \xi)\) in (2.14) is a normal density with mean \(\theta\) and \(\xi = \sigma^2\) is the unknown variance parameter. That is, \(X_{i1}, \ldots, X_{ir} \iid N(\theta, \sigma^2)\) for each \(i = 1, \ldots, n\). The statistic \(U_i = \sum_i\) is sufficient for the mean and the density \(g(\cdot|\theta, \sigma^2)\) of \(U_i\) is known. Therefore, \(H\) in (2.16) can be written as

\[
H_s(u, \varphi, \sigma^2) = \frac{g(u|\theta^s, \sigma^2)\varphi^s}{\sum_t g(u|\theta^t, \sigma^2)\varphi^t} - \varphi^s
\]  

(2.21)

for \(s = 1, \ldots, S\), where \(g(u|\theta, \sigma^2)\) is the \(N(\theta, \sigma^2/r)\) density. Even in this simple example, it is not obvious that the function \(H\) in (2.21) satisfies A2.8.

**Proposition 2.16** A2.8 holds for \(H\) in (2.21).

**Proof** To prove the proposition we must show that the function \(H_s\), defined in (2.21), has bounded derivative. Clearly each component \(H_s\) of \(H\), defined in (2.21), is differentiable with respect to \(\sigma \in \Sigma_0\) and, after simplification,

\[
\frac{\partial}{\partial \sigma^2} H_s(u, \varphi, \sigma^2) = \frac{\varphi^s e^{-r(\theta^s)^2/2\sigma^2} e^{ru\theta^s/\sigma^2}}{2\sigma^4} \sum_t v_{st}(u) \varphi^t e^{-r(\theta^t)^2/2\sigma^2} e^{rs\theta^t/\sigma^2} \]  

where (as \(|u| \to \infty\))

\[
v_{st}(u) = (\theta^s)^2 - (\theta^t)^2 + 2u(\theta^t - \theta^s) = \mathcal{O}(|u|).
\]  

(2.22)
This derivative is continuous on \( u(\mathcal{X}^r) \times \Delta_0 \times \Sigma_0 \) and, since \( \Delta_0 \) and \( \Sigma_0 \) are compact, we know that

\[
A_s(u) := \sup_{\varphi \in \Delta_0} \sup_{\sigma^2 \in \Sigma_0} \left| \frac{\partial}{\partial \sigma^2} H_s(u; \varphi, \sigma^2) \right| \tag{2.23}
\]

is finite for all \( u \in u(\mathcal{X}^r) \) and for all \( s \). By the Mean Value Theorem,

\[
|H_s(u; \varphi, \sigma^2) - H_s(u; \varphi, \tilde{\sigma}^2)| \leq A_s(u)|\tilde{\sigma}^2 - \sigma^2|.
\]

It remains to show that \( A_s(u) \) is bounded in \( u \). For notational simplicity, assume that \( \varphi \) and \( \sigma \) are the values for which the suprema in (2.23) are attained. Making a change of variables \( y = ru/\sigma^2 \) we can, with a slight abuse of notation, write

\[
A_s(y) \leq \frac{C_s \varphi^s e^{y\theta^1} \sum_t |v_{st}(y)| \varphi^t e^{y\theta^t}}{\sum_t |\varphi^t e^{y\theta^t}|^2}.
\]

We must show that \( A_s(y) \) is bounded as \( |y| \to \infty \). Assume, without loss of generality, that the \( \theta^t \)'s are arranged in ascending order: \( \theta^1 < \theta^2 < \cdots < \theta^S \). Factoring out, respectively, \( e^{y\theta^1} \) and \( e^{y\theta^S} \), we can write

\[
A_s(y) \leq \frac{C_s \varphi^s e^{y(\theta^s - \theta^1)} \sum_t |v_{st}(y)| \varphi^t e^{y(\theta^t - \theta^1)}}{(\varphi^1)^2 + \sum_{t \neq 1} \sum_{r \neq 1} \varphi^t \varphi^r e^{y(\theta^r - \theta^1) + y(\theta^t - \theta^1)}}
\]

\[
A_s(y) \leq \frac{C_s \varphi^s e^{y(\theta^s - \theta^s)} \sum_t |v_{st}(y)| \varphi^t e^{y(\theta^t - \theta^s)}}{(\varphi^S)^2 + \sum_{t \neq S} \sum_{r \neq S} \varphi^t \varphi^r e^{y(\theta^r - \theta^S) + y(\theta^t - \theta^S)}}
\]

Note that since \( \varphi \in \Delta_0 \), each \( \varphi^t \) is bounded away from 0. If \( y \to -\infty \) then the term \( e^{y(\theta^s - \theta^1)} \to 0 \) dominates the numerator of the first inequality, while the denominator is bounded. Similarly, if \( y \to +\infty \), then the term \( e^{y(\theta^s - \theta^s)} \to 0 \) dominates the numerator in the second inequality, while the denominator is bounded. For the case \( s = 1 \) or \( s = S \), note that \( |v_{11}(y)| = |v_{SS}(y)| = 0 \), so the two inequalities can still be applied and a similar argument shows \( A_1 \) and \( A_S \) are also bounded. Therefore, \( A_s(y) \) is bounded for each \( s \) and the claim follows by taking \( A \) to be \( \max \{ \sup_y A_s(y) : 1 \leq s \leq S \} \). \[ \square \]

Let \( \Sigma_0 \) be the \( \Xi_0 \) defined in the general setup. For the RE+R, we choose \( T(x) \) to be the sample variance of \( x \), resulting in the recursive estimate

\[
\sigma_i^2 = \frac{1}{i(r-1)} \sum_{k=1}^{i} \sum_{j=1}^{r} (X_{kj} - \overline{X}_k)^2. \tag{2.24}
\]
For $\sigma^2$, take the standard non-informative prior $p(\sigma^2) = (\sigma^2)^{-1}$. Under squared-error loss, the Bayes estimate of $\sigma^2$ based on $X_1, \ldots, X_i$ is

$$\hat{\sigma}_i^2 = \mathbb{E}(\sigma^2 | \mathcal{A}_i) = \frac{1}{i(r-1)} - \frac{1}{2} \sum_{k=1}^{i} \sum_{j=1}^{r} (X_{kj} - \overline{X}_k)^2.$$  \hspace{1cm} (2.25)

Since $|\hat{\sigma}_n^2 - \sigma^2| = \mathcal{O}(n^{-1/2})$ a.s., the conclusion of Corollary 2.15 obtains if $w_n \sim n^{-1}$. In numerical examples not presented here, we see very little difference in the performance of the unbiased RE+R compared to that of the biased RE+R with the Bayes estimate $\hat{\sigma}_i^2$ in (2.25).

### 2.4 Additional remarks

In this chapter we have used general stochastic approximation results to prove a consistency theorem for RE in the case of a finite parameter space $\Theta$. It is natural to wonder if this theorem can be extended to the case where $\Theta$ is, say, a compact metric space. This problem is the subject of Chapter 3. The argument there is based on the approximate martingale representation used in Ghosh and Tokdar [47] but applied to the KL divergence between the mixtures; cf. Remark 2.11. The proof of Theorem 2.10 would seem to go through for the general-$f$ case if only there was an infinite-dimensional version of Theorem 2.9. There are Hilbert space versions—see, for example, Robbins and Siegmund [94]—but, to our knowledge, there is no result of this kind currently available that suits our needs.

It should also be pointed out that our approach in this chapter to the finite mixture problem is somewhat less general than would be desirable. In particular, we are assuming that the support of $f$ is within a known finite set of points. In general, however, what is known is that the support of $f$ is contained in, say, a bounded interval. In this case, a set of grid points $\Theta = \{\theta^1, \ldots, \theta^S\}$ are chosen to approximate the unknown support $\Theta^*$ of $f$. Newton’s algorithm will produce an estimate $f_n$ on $\Theta$ in this case, but it is impossible to directly compare $f_n$ to $f$ since their supports $\Theta$ and $\Theta^*$ may be entirely different. There is no problem comparing the mixture, however.
This observation leads us to the following important conjecture, closely related to the so-called $I_0$-projections in Liese and Vadja [63] and Shyamalkumar [103].

**Conjecture 2.17** Let $m_{f_n}$ and $m_f$ be the mixture densities corresponding to $f_n$ on $\Theta$ and $f$ on $\Theta^*$, respectively. Then, as $n \to \infty$, $$K(m_f, m_{f_n}) \to \inf_{\varphi} K(m_f, m_{\varphi}) \quad \text{a.s.,}$$

where $\varphi$ ranges over all densities on $\Theta$.

We show in Chapter 3 that a similar conjecture happens to be true in the general mixture problem, and these results are used in the subsequent chapters to suitably modify the recursive algorithm to apply to a very broad class of statistical problems.
3. GENERAL CONSISTENCY THEOREMS

In the present chapter we investigate the asymptotic properties of the recursive estimates $f_n$ and $m_n := m f_n$ defined on page 7 under very general conditions on the parameter space $\Theta$. In practice, $\Theta$ would typically be a subset of Euclidean space, but our theory is general enough to apply even when $\Theta$ is infinite-dimensional.

Let $m$ be a probability density with respect to a dominating $\sigma$-finite measure $\nu$ on $\mathcal{X}$, and suppose independent observations $X_1, \ldots, X_n$ are available from $m$. We model $m$ with a mixture density $m_f$ of the form

$$m_f(x) = \int_{\Theta} p(x|\theta) f(\theta) d\mu(\theta), \quad f \in F,$$

where $F = F(\Theta, \mu)$ is a class of densities with respect to a dominating $\sigma$-finite measure $\mu$ on $\Theta$, and $p(\cdot|\theta)$, for $\theta \in \Theta$, is a parametric family of densities on $\mathcal{X}$. In keeping with the spirit of this thesis, we interpret the mixture model (3.1) as a Bayesian model with known likelihood $p(x|\theta)$ and unknown prior $f \in F$.

Our ultimate goal is to ascertain the asymptotic properties of $f_n$, the estimate of the mixing/prior distribution. It turns out, however, that it is best to do so indirectly by studying the performance of the mixture $m_n$ as an estimate of the unknown density $m$—roughly, if $m_n$ is close to $m$, then the Bayes model (1.2) with estimated prior $f_n$ accurately represents the observed data. Therefore, the problem can be formulated as one of density estimation, so the remainder of this chapter will be noticeably non-Bayesian. To interpret our results, it is important to keep in mind that estimation of $m$ is not our primary goal: the mixture $m_n$ is nothing but a tool to measure the performance of $f_n$. Of course, one can dispense with the Bayesian interpretation altogether and treat $m_n$ simply as a mixture-based nonparametric density estimate. See Remark 3.10 for a comparison of our convergence results with others in the mixture density estimation context.
In the general case, when the mixture/Bayesian model (3.1) does not contain the true \( m \), there is no hope of \( m_n \) being a consistent estimate. But perhaps \( m_n \) will be “close” to \( m \) in some sense, suggesting that our estimated prior \( f_n \) is robust to model mis-specification. Indeed, we show in Section 3.2 that the recursive estimate \( m_n \) converges almost surely in the \( L_1 \) topology to the “best possible” mixture of the prescribed form (1.1) in a Kullback-Leibler sense. This “best possible” mixture is called the KL projection and defined in Section 3.1. From this, we can deduce convergence of \( f_n \) under relatively mild conditions. Moreover, we establish a bound on the rate of convergence of \( m_n \) to the KL projection.

In Section 3.3 we turn our attention to the problem that the mixing density estimate \( f_n \) depends on the order in which the data is run through the recursion (1.6). We propose a permutation invariant version of RE obtained by averaging over all permutations of the data sequence and establish the convergence properties of this permutation-averaged RE, or PARE. This result theoretically justifies Newton’s [80] suggestion to weaken the order-dependence by taking a pointwise average of the estimates over several random permutations of the data sequence.

### 3.1 Kullback-Leibler projections

For probability measures \( P \) and \( Q \), with \( P \ll Q \), the Kullback-Leibler (KL) divergence of \( Q \) from \( P \) is

\[
K(P, Q) = \int \log(dP/dQ) \, dP,
\]

where \( dP/dQ \) is the Radon-Nikodym derivative of \( P \) with respect to \( Q \). If both \( P \) and \( Q \) have densities \( p \) and \( q \) with respect to a common dominating measure \( \nu \), then write \( K(p, q) = \int p \log(p/q) \, d\nu \) instead of \( K(P, Q) \). Some authors write \( I_0(P, Q) \) instead of \( K(P, Q) \) and call (3.2) the \( I_0 \)-divergence. Incidentally, the \( I_1 \)-divergence interchanges the role of \( P \) and \( Q \) in (3.2); see Liese and Vajda [63].

KL divergences arise quite frequently in both classical and Bayesian statistics, primarily due to the ubiquity of likelihood ratios. In our case, the occurrence of
the KL divergence is quite natural. In Chapter 2, it was shown that the asymptotic
behavior of the iterates \( f_n \) is determined by that of a particular differential equation,
and that the KL divergence is a Lyapunov function for this differential equation.
Thus, the KL divergence controls the dynamics of the RE algorithm in some sense,
forcing the estimates to converge to a stable equilibrium.

Let \( \mathbb{P} \) and \( \mathbb{Q} \) be sets of probability measures, where \( \mathbb{Q} \) is a convex subset of \( \mathbb{P} \). We
will be interested in identifying \( Q^* \in \mathbb{Q} \) such that \( Q^* \) is closest to \( P \) in a Kullback-
Leibler sense; that is,

\[
K(P, Q^*) = K(P, Q) := \inf \{ K(P, Q) : Q \in \mathbb{Q} \}
\]

We call \( Q^* \) the KL projection of \( P \) onto \( \mathbb{Q} \). See Leroux [61], Shyamalkumar [103]
and Kleijn and van der Vaart [56] for fairly recent work involving KL (or \( I_0 \)-
projections. \( I_1 \)-projections, which minimize \( K(P, Q) \) over \( P \), also appear in the liter-
ature. Csiszar [18] mentions applications in large deviation theory and statistical
physics while Dykstra [28] and Bhattacharya [9] focus on the relationship between
\( I_1 \)-projections and the popular iterative proportional fitting procedure.

In the study of KL projections, typically the most important issue is existence;
uniqueness follows automatically from convexity of \( \mathbb{Q} \) and convexity of the mapping
\( Q \mapsto K(P, Q) \) for fixed \( P \). A well known general result is that if \( \mathbb{Q} \) is convex and
closed with respect to the total variation distance topology, then there exists a KL
projection \( Q^* \) of \( P \) onto \( \mathbb{Q} \); see Proposition 8.5 of Liese and Vadja [63]. In this section
we prove a simple theorem which gives sufficient conditions for the existence of a KL
projection in our special case of mixtures.

First we need some notation. Let \( \mathbb{M} = \mathbb{M}(\mathcal{X}, \nu) \) be the space of all probability
densities on \( \mathcal{X} \) with respect to \( \nu \). Now, for a parametric family of probability densities
\( \mathcal{P} = \{ p(\cdot|\theta) : \theta \in \Theta \} \subset \mathbb{M} \), let \( \mathbb{M}_\Theta \) consist of all mixtures of the form (3.1); that
is, \( \mathbb{M}_\Theta = \{ m_f : f \in \mathcal{F} \} \). Here \( \mathbb{M} \) plays the of \( \mathbb{P} \) and \( \mathbb{M}_\Theta \) the role of \( \mathbb{Q} \); notice that,
by definition, \( \mathbb{M}_\Theta \) is convex. The following theorem gives conditions which imply the
existence of a mixing/prior density \( f \in \mathcal{F} \) such that \( K(m, m_f) = K(m, \mathbb{M}_\Theta) \).
Theorem 3.1 If $\mathbb{F}$ is compact with respect to the weak topology and $\theta \mapsto p(x|\theta)$ is bounded and continuous for $\nu$-a.e. $x$, then there exists a mixing density $f \in \mathbb{F}$ such that $K(m, m_f) = K(m, \mathbb{M}_\Theta)$.

Proof Choose any $\varphi \in \mathbb{F}$ and a sequence $\{\varphi_s\} \subset \mathbb{F}$ such that $\varphi_s \to \varphi$ weakly. Let $m_\varphi(x) = \int p(x|\theta)\varphi(\theta) d\mu(\theta)$ define the corresponding sequence $m_s = m_{\varphi_s}$ of mixture densities in $\mathbb{M}_\Theta$. Then weak convergence of $\varphi_s$ to $\varphi$ and the assumption that $\theta \mapsto p(x|\theta)$ is $\nu$-a.e. $x$ bounded and continuous implies, by Schefl’s theorem, that $m_s \to m_\varphi$ in $L_1(\nu)$. Let $\kappa(\varphi) = K(m, m_\varphi)$. Since $m_s \to m_\varphi$ in $\nu$-measure, it follows from Fatou’s lemma that

$$\kappa(\varphi) = \int \lim_{s \to \infty} \log(m/m_s) m d\nu \leq \liminf_{s \to \infty} \int \log(m/m_s) m d\nu = \liminf_{s \to \infty} \kappa(\varphi_s)$$

Thus $\kappa$ is lower semi-continuous with respect to the weak topology and, therefore, must attain its infimum on the compact $\mathbb{F}$.

Remark 3.2 Convexity of $\mathbb{M}_\Theta$ and of $K(m, \cdot)$ together imply that the KL projection $m_f$ in Theorem 3.1 is unique. However, in general, there could be many mixing densities $f \in \mathbb{F}$ that generate the KL projection $m_f$. Identifiability (cf. Definition 1.1) is needed to guarantee the mixing density $f$ is unique.

3.2 Robustness and rate of convergence

In this section we prove the claim that $m_n$ is asymptotically robust to mixture model mis-specification, and we also establish a lower bound on the rate of convergence. We begin with some assumptions and preliminary lemmas.
3.2.1 Preliminaries

For a parametric family of probability densities $\mathcal{P} = \{p(\cdot|\theta) : \theta \in \Theta\}$ on $\mathcal{X}$, recall that $\mathcal{M}_\Theta$ is the convex hull of $\mathcal{P}$. For the weight sequence $\{w_n\}$ in the RE algorithm, define the sequence of partial sums $a_n = \sum_{i=1}^{n} w_i$.

Our main results will require the following conditions:

A3.1. $w_n > 0$, $w_n \downarrow 0$, $\sum_n w_n = \infty$, $\sum_n w_n^2 < \infty$ and $\sum_n a_n w_n^2 < \infty$.

A3.2. $\mathcal{F}$ is compact with respect to the weak topology.

A3.3. $\theta \mapsto p(x|\theta)$ bounded and continuous for $\nu$-a.e. $x \in \mathcal{X}$.

A3.4. There exists $B < \infty$ such that

$$\sup_{\theta_1, \theta_2 \in \Theta} \int_{\mathcal{X}} \left[ \frac{p(x|\theta_1)}{p(x|\theta_2)} \right]^2 m(x) d\nu(x) < B.$$ 

The condition A3.1 on the weight sequence $\{w_n\}$ is natural, given the stochastic approximation representation presented in Chapter 2—in fact, A3.1 is just a slightly stronger version of condition (2.2). These conditions are satisfied if $w_n = O(n^{-\alpha})$, for $\alpha \in (2/3, 1]$. Conditions A3.2 and A3.3 together imply the existence of a density $f \in \mathcal{F}$ such that $K(m,m_f) = K(m,\mathcal{M}_\Theta)$ but this $f$ need not be unique without identifiability. A sufficient condition for A3.2 is that $\Theta$ be a compact metric space.

Condition A3.4 is by far our strongest assumption. Since $m(x)$ need not be of the form (3.1), A3.4 cannot be written as a condition on the parametric family $\mathcal{P}$. However, A3.4 does hold for many common families $\mathcal{P}$, such as

- Normal with mean $\theta$ and any fixed variance $\sigma^2 > 0$
- Gamma with rate $\theta$ and any fixed shape $\alpha > 0$
- Poisson with mean $\theta$

provided that $\Theta$ is compact and $m$ admits a moment-generating function. If we are willing to assume that $m \in \mathcal{M}_\Theta$, then A3.4 can be replaced by the following less restrictive condition:
A3.4’. There exists $B < \infty$ such that

$$\sup_{\theta_1, \theta_2, \theta_3 \in \Theta} \int_X \left[ \frac{p(x|\theta_1)}{p(x|\theta_2)} \right]^2 p(x|\theta_3) \, d\nu(x) < B.$$ 

Next is an important implication of A3.4 (or A3.4’), whose proof follows immediately from Jensen’s inequality.

**Lemma 3.3** For any mixing densities $\varphi, \psi \in \mathcal{F}$, if either condition A3.4 or A3.4’ holds, then $\int (m_\varphi/m_\psi)^2 \, m \, d\nu < B$.

The basic idea behind our analysis is to approximate the Kullbach-Leibler (KL) divergence $K(m, m_n)$ by a suitable martingale and take advantage of the classical, fully developed theory on martingale convergence. The starting point is a first-order Taylor expansion of $\log(1 + x)$ at $x = 0$: specifically

$$\log(1 + x) = x - x^2 R(x), \quad x > -1, \quad (3.4)$$

where the remainder term $R(x)$ satisfies

$$0 \leq R(x) \leq \max\{1, (1 + x)^{-2}\}, \quad x > -1. \quad (3.5)$$

Next express the estimated mixture density $m_n \in \mathcal{M}_\Theta$ recursively:

$$m_n(x) = (1 - w_n)m_{n-1}(x) + w_n h_{n,x_n}(x),$$

where

$$h_{n,x'}(x) = \int_{\Theta} \frac{p(x|\theta)p(x'|\theta)f_{n-1}(\theta)}{m_{n-1}(x')} \, d\mu(\theta), \quad x, x' \in \mathcal{X}. \quad \text{For notational convenience, also define the function}$$

$$H_{n,x'}(x) = \frac{h_{n,x}(x)}{m_{n-1}(x)} - 1, \quad x, x' \in \mathcal{X}. \quad \text{Then the KL divergence} \quad K_n = K(m, m_n) \quad \text{satisfies} \quad$$

$$K_n - K_{n-1} = \int_X m \log(m_{n-1}/m_n) \, d\nu$$

$$= - \int_X m \log(1 + w_n H_{n,x_n}) \, d\nu$$

$$= -w_n \int_X m H_{n,x_n} \, d\nu + w_n^2 \int_X m H_{n,x_n}^2 R(w_n H_{n,x_n}) \, d\nu$$
where \( R(x) \) is the remainder term in (3.4). Let \( \mathcal{A}_{n-1} \) be the \( \sigma \)-algebra generated by the data sequence \( X_1, \ldots, X_{n-1} \). Since \( K_{n-1} \) is \( \mathcal{A}_{n-1} \)-measurable, upon taking conditional expectation with respect to \( \mathcal{A}_{n-1} \) we get

\[
E(K_n|\mathcal{A}_{n-1}) - K_{n-1} = -w_n T(f_{n-1}) + w_n^2 E(Z_n|\mathcal{A}_{n-1}) \tag{3.6}
\]

where \( T(\cdot) \) and \( Z_n \) are defined as

\[
T(\varphi) = \int_\Theta \left\{ \int_{X \setminus \varphi(x)} m(x) p(x|\theta) d\nu(x) \right\}^2 \varphi(\theta) d\mu(\theta) - 1, \quad \varphi \in \mathcal{F} \tag{3.7}
\]

\[
Z_n = \int_\mathcal{X} m R^2(w_n H_{n,X_n}) d\nu \tag{3.8}
\]

The following property of \( T(\cdot) \) will be critical in the proof of our main result.

**Lemma 3.4** \( T(\varphi) \geq 0 \) with equality if and only if \( K(m, m_{\varphi}) = K(m, \mathcal{M}_\Theta) \).

**Proof** Treat \( \theta \) as a random element in \( \Theta \), whose distribution is determined by the density \( \varphi \), and define the random variable

\[
g_{\varphi}(\theta) = \int_\mathcal{X} \frac{m(x)}{m_\varphi(x)} p(x|\theta) d\nu(x), \quad \varphi \in \mathcal{F}. \tag{3.9}
\]

Then \( E_{\varphi}[g_{\varphi}(\theta)] = \int g_{\varphi} \varphi d\mu = 1 \) and \( T(\varphi) = V_{\varphi}[g_{\varphi}(\theta)] \geq 0 \), with equality if and only if \( g_{\varphi} = 1 \ \mu\text{-a.e.} \). To show that \( T(\varphi) = 0 \) implies \( K(m, m_{\varphi}) = K(m, \mathcal{M}_\Theta) \), we follow Shyamalkumar [103]. Define

\[
G(\varphi) = \log \left\{ \int_\Theta g_{\varphi}(\theta) f(\theta) d\mu(\theta) \right\},
\]

where \( f \in \mathcal{F} \) is such that \( K(m, m_f) = K(m, \mathcal{M}_\Theta) \). Note that \( T(\varphi) = 0 \) implies \( G(\varphi) = 0 \). By Jensen’s inequality

\[
G(\varphi) = \log \left\{ \int_\Theta \left[ \int_{X \setminus \varphi(x)} m(x) p(x|\theta) d\nu(x) \right] f(\theta) d\mu(\theta) \right\}
\]

\[
= \log \left\{ \int_{X \setminus \varphi(x)} \frac{m_f(x)}{m_\varphi(x)} m(x) d\nu(x) \right\}
\]

\[
\geq \int_{X \setminus \varphi(x)} \log \left( \frac{m_f(x)}{m_\varphi(x)} \right) m(x) d\nu(x)
\]

\[
= K(m, m_{\varphi}) - K(m, m_f)
\]

\[
\geq 0
\]

so that \( G(\varphi) = 0 \) implies \( K(m, m_{\varphi}) = K(m, m_f) \). \( \square \)
Remark 3.5 For some further insight into the relationship between the functional $T(\varphi)$ and the KL divergence $K(m, m_\varphi)$, define

$$D(\theta; \varphi) := g_\varphi(\theta) - 1 = \int \left[ \frac{p(x|\theta)}{m_\varphi(x)} - 1 \right] m(x) d\nu(x).$$

Simple analysis shows that $D(\theta; \varphi)$ is the Gâteaux derivative of $K(m, \eta)$ at $\eta = m_\varphi$ in the direction of $p(\cdot|\theta)$. Now, if $T(\varphi) = 0$, then $D(\theta; \varphi) = 0$ for $\mu$-a.e. $\theta$ and, hence, $D(\psi; \varphi) := \int D(\theta; \varphi) \psi(\theta) d\mu(\theta)$, the Gâteaux derivative of $K(m, \eta)$ at $\eta = m_\varphi$ in the direction of $m_\psi$, is zero for all $\psi \in F$. But the fact that the Gâteaux derivative of $K(m, \eta)$ at $\eta = m_\varphi$ vanishes in all directions suggests that $m_\varphi$ is a point at which the infimum $K(m, \mathcal{M}_\Theta)$ is attained. This is exactly the conclusion of Lemma 3.4.

From (3.6) we see the makings of a supermartingale in $K_n$. Indeed, were it not for the term involving $Z_n$ in (3.6), we would have a non-negative supermartingale and convergence of $K_n$ would follow immediately from the martingale convergence theorem [10]. Fortunately, the presence of $Z_n$ causes only minor difficulties.

Lemma 3.6 Under A3.1 and A3.4, $\sum_{n=1}^{\infty} w_n^2 E(Z_n) < \infty$.

Proof Note that for $a > 0$ and $b \in (0, 1)$, we have

$$(a - 1)^2 \max\{1, (1 + b(a - 1))^{-2}\} \leq \max\{(a - 1)^2, (1/a - 1)^2\}. \quad (3.10)$$

Combining inequalities (3.5) and (3.10) we see that

$$H_{n,X_n}^2 R(w_n H_{n,X_n}) \leq \max \left\{ \left( \frac{h_{n,X_n}}{m_{n-1} - 1} - 1 \right)^2, \left( \frac{m_{n-1}}{h_{n,X_n} - 1} - 1 \right)^2 \right\}$$

and, since $h_{n,X_n}, m_{n-1} \in \mathcal{M}_\Theta$ for each $n$, we conclude from Lemma 3.3 that $Z_n < 1 + B$ a.s. The claim now follows from A3.1.

Our last preliminary result of this section, Lemma 3.7 below, gets us closer to our end result. Define

$$K^*_n = K(m, m_n) - K(m, \mathcal{M}_\Theta).$$
Lemma 3.7 Under A3.1 and A3.4, \( \sum_n w_n K^*_n < \infty \) a.s.

Proof Let \( f \in \mathcal{F} \) be such that \( K(m, m_f) = K(m, M_\Theta) \). Upon considering the KL divergence between mixing densities, application of a telescoping sum reveals

\[
K(f, f_n) - K(f, f_0) = -\sum_{i=1}^n \int \log(f_i/f_{i-1}) \, d\mu
= -\sum_{i=1}^n \int \log \left[ 1 + w_i \left( \frac{p(X_i|\theta)}{m_{i-1}(X_i)} - 1 \right) \right] f(\theta) \, d\mu(\theta)
\]

Now apply the Taylor approximation in (3.4), which gives

\[
K(f, f_n) - K(f, f_0) = \sum_{i=1}^n w_i \left( 1 - \frac{m(X_i)}{m_{i-1}(X_i)} \right) + \int \Theta R_i(X_i, \theta) f(\theta) \, d\mu(\theta)
= \sum_{i=1}^n w_i V_i - \sum_{i=1}^n w_i M_i + \sum_{i=1}^n E_i
\]

where

\[
R_i(x, \theta) = w_i^2 \left( \frac{p(x|\theta)}{m_{i-1}(x)} - 1 \right)^2 R \left( w_i \left( \frac{p(x|\theta)}{m_{i-1}(x)} - 1 \right) \right)
M_i = -E \left[ 1 - \frac{m_f(X_i)}{m_{i-1}(X_i)} \right] = \int \frac{m_f(x)}{m_{i-1}(x)} - 1 \, m(x) \, d\nu(x)
V_i = 1 - \frac{m_f(X_i)}{m_{i-1}(X_i)} + M_i
E_i = \int \Theta R_i(X_i, \theta) f(\theta) \, d\mu(\theta)
\]

From the inequality \( \log(x) \leq x - 1 \) and the proof of Lemma 3.4, we get

\[
M_i \geq \int \log(m_f/m_{i-1}) m \, d\nu = K^*_{i-1} \geq 0.
\]

Therefore, it suffices to show that \( \sum_{i=1}^n w_i M_i \) converges almost surely to a finite random variable. Non-negativity of the KL divergence implies that

\[
\sum_{i=1}^n w_i M_i \leq K(f, f_0) + S_\infty + T_\infty \quad \forall n,
\]

where \( S_\infty := \sum_{i=1}^\infty w_i V_i \) and \( T_\infty := \sum_{i=1}^n E_i \). Next we show that \( S_\infty \) and \( T_\infty \) are each almost surely finite.
Start with $V_\infty$. By definition of $V_i$, $S_n = \sum_{i=1}^n w_i V_i$ form a zero mean martingale with respect to the filtration $\mathcal{A}_n$. Furthermore, by A3.4,

$$E(S_n^2) = E\left\{ \sum_{i=1}^n w_i^2 E[V_i^2 | \mathcal{A}_{i-1}] \right\} \leq E\left\{ \sum_{i=1}^n w_i^2 \left[ \left( 1 - \frac{m_f(X_i)}{m_{i-1}(X_i)} \right)^2 \right] | \mathcal{A}_{i-1} \right\} \leq 2(1 + B) \sum_{i=1}^n w_i^2$$

Since $\sum_{i=1}^\infty w_i^2 < \infty$ by A3.1, $E(S_n^2)$ is uniformly bounded in $n$. Therefore, by the martingale convergence theorem [10], $S_n \to S_\infty$ almost surely and $E(S_\infty) < \infty$.

Next, let $T_n = \sum_{i=1}^n E_i$ so that $T_\infty = \lim T_n$; $T_\infty$ always exists (since $E_i$’s are nonnegative) but may equal infinity. In light of inequality (3.10) we have

$$R_i(x, \theta) \leq \frac{w_i^2}{2} \left\{ \left( \frac{p(x|\theta)}{m_{i-1}(x)} - 1 \right)^2 \max \left\{ 1, \left( 1 + w_i \left( \frac{p(x|\theta)}{m_{i-1}(x)} - 1 \right) \right)^{-2} \right\} \right\} \leq \frac{w_i^2}{2} \max \left\{ \left( 1 - \frac{p(x|\theta)}{m_{i-1}(x)} \right)^2, \left( 1 - \frac{m_{i-1}(x)}{p(x|\theta)} \right)^2 \right\}$$

and hence, by assumption A3.5,

$$E[E_i | \mathcal{A}_{i-1}] = \int_\Theta E[R_i(X_i, \theta) | \mathcal{A}_{i-1}] f(\theta) \mu(d\theta) \leq w_i^2(1 + B).$$

Finally, by Fatou’s lemma and assumption A3.1,

$$E(T_\infty) \leq \liminf_{n \to \infty} E(T_n) = \liminf_{n \to \infty} E\left( \sum_{i=1}^n E_i | \mathcal{F}_{i-1} \right) \leq \liminf_{n \to \infty} (1 + B) \sum_{i=1}^n w_i^2 < \infty$$

which proves $T_\infty$ is finite with probability 1.

We have shown that the right-hand side of (3.11) is finite so $\sum_{i=1}^n w_i M_i$ is bounded above uniformly in $n$. Therefore, since the $M_i$’s are non-negative, $\sum_{i=1}^n w_i M_i$ must converge to a finite random variable.

3.2.2 Convergence of $m_n$

We now have the notation and machinery to precisely state and prove our main convergence result. Recall that $a_n = \sum_{i=1}^n w_i$. 


Theorem 3.8 Under assumptions A3.1–A3.4, there exists a non-negative, integrable random variable $K_\infty$ such that $a_nK_n^* \to K_\infty$ a.s.

Proof Define two sequences of random variables:

$$\delta_n = \sum_{i=n+1}^{\infty} a_{i-1}w_i^2E(Z_i|\mathcal{A}_n) \quad \text{and} \quad \varepsilon_n = \sum_{i=n+1}^{\infty} w_iE(K_i^*|\mathcal{A}_n).$$

Note that $\varepsilon_n$ is finite for each $n$ by Lemma 3.7. Next, define $\tilde{K_n} = a_nK_n^* + \delta_n + \varepsilon_n$.

Using (3.6) we find that

$$E(\tilde{K}_n|\mathcal{A}_{n-1}) - \tilde{K}_{n-1} = E(a_nK_n + \delta_n + \varepsilon_n | \mathcal{A}_{n-1})$$

$$- a_{n-1}K_{n-1} - \delta_{n-1} - \varepsilon_{n-1} - w_nK(m, \mathcal{M}_\Theta)$$

$$\leq a_nE(K_n|\mathcal{A}_{n-1}) + \sum_{i=n+1}^{\infty} a_{i-1}w_i^2E(E(Z_i|\mathcal{A}_n) | \mathcal{A}_{n-1})$$

$$+ \sum_{i=n+1}^{\infty} E(E(K_i|\mathcal{A}_n) | \mathcal{A}_{n-1}) - a_{n-1}K_{n-1}$$

$$- \sum_{i=n}^{\infty} a_{i-1}w_i^2E(Z_i|\mathcal{A}_{n-1}) - \sum_{i=n}^{\infty} w_iE(K_i|\mathcal{A}_{n-1}).$$

After some simplification, using the fact that $\mathcal{A}_{n-1} \subset \mathcal{A}_n$, we have

$$E(\tilde{K}_n|\mathcal{A}_{n-1}) - \tilde{K}_{n-1} \leq a_nE(K_n|\mathcal{A}_{n-1}) - a_{n-1}K_{n-1}$$

$$- a_{n-1}w_n^2E(Z_n|\mathcal{A}_{n-1}) - w_nE(K_n|\mathcal{A}_{n-1})$$

$$= a_{n-1} \left[ E(K_n|\mathcal{A}_{n-1}) - K_{n-1} - w_n^2E(Z_n|\mathcal{A}_{n-1}) \right]$$

$$= -a_{n-1}w_nT(f_{n-1})$$

$$\leq 0$$

Thus, $\tilde{K}_n$ forms a non-negative supermartingale and, by the martingale convergence theorem, has a limit $\tilde{K}_\infty \geq 0$, which has finite expectation. But assumption A3.1, Lemma 3.6, and the dominated convergence theorem ensure that $\delta_n$ and $\varepsilon_n$ both converge a.s. to zero; therefore, $a_nK_n^* \to \tilde{K}_\infty$ a.s. \qed

The following corollary summarizes our main result on robustness and rate of convergence of the mixture density estimate $m_n$. 

Corollary 3.9 Choose $w_n = O(n^{-2/3-\alpha})$ for some $\alpha \in (0, 1/3]$. Then under assumptions A3.1–A3.4, $d(m_n, m_f) = O(n^{-1/6+\alpha/2})$ a.s., where $d(\cdot, \cdot)$ is either $L_1$ or Hellinger distance and $m_f$ is the KL-projection of $m$ onto $\mathbb{M}_\Theta$. Moreover, if $m$ is a mixture in $\mathbb{M}_\Theta$, then $d(m_n, m) = O(n^{-1/6+\alpha/2})$ a.s.

Remark 3.10 Despite the fact that mixture density estimation is not our primary goal, it is worth mentioning that the results of Corollary 3.9 compare favorably to other recent work in mixture density estimation. In particular, for the important case where $m$ is a mixture (1.1) and the corresponding mixing density is compactly supported, our nearly $n^{1/6}$ convergence rate in the Hellinger metric requires only mild conditions on the mixands $p(x|\theta)$ and is essentially the rate obtained in Genovese and Wasserman [41] which requires stronger assumptions. Li and Barron [62] and Ghosal and van der Vaart [43] obtain better (nearly parametric) rates but their assumptions are even stronger—Li and Barron’s bound [62] involves constants which are infinite in many common problems and Ghosal and van der Vaart’s result [43] seems to rely heavily on the smoothness of the Gaussian mixands.

3.2.3 A quick numerical illustration

Here we give a simple numerical investigation into the theoretical bound on the convergence rate given in Corollary 3.9 whose purpose is two-fold:

1. We show a practical improvement of the RE when the “nominal” weights $w_i = (i + 1)^{-1}$ are replaced by the “optimal” weights $w_i = (i + 1)^{-0.67}$, and

2. We estimate the rate of convergence empirically, providing evidence to suggest that the theoretical $n^{-1/6+\varepsilon}$ bound on the $L_1$ error is not sharp.

Let $p(x|\theta) = \theta e^{-\theta x}$, $x \geq 0$, be an Exponential density with rate $\theta \geq 0$. Since $\theta$ is a scale parameter, the mixture $m(x) = \int p(x|\theta)f(\theta)\,d\theta$ is called a scale-mixture. Scale mixtures such as these are known to be particularly challenging. Here the true mixing
distribution is taken to be a log-Normal, $LN(\mu, \sigma^2)$, suitably truncated; that is, the density $f$ has formula
\[ f(\theta) = \kappa \cdot \frac{1}{\theta \sqrt{2\pi \sigma^2}} \exp\{-\frac{(\log(\theta) - \mu)^2}{2\sigma^2}\}, \quad 0 \leq \theta \leq 30, \]
where $\kappa$ is the normalizing constant. (In this example, we take $\mu = 1.5$ and $\sigma = 0.5$.) Figure 3.1 shows estimates of $f$ and $m$ from the RE algorithm, based on 25 samples of size $n = 1000$, using the two “extreme” weights $w_i = (i + 1)^{-r}$ where $r \in \{0.67, 1\}$. In each case, the initial guess $f_0 = \text{Unif}([0, 30])$ was used. Here we see that RE with $r = 1$ does a rather poor job. However, the estimates with $r = 0.67$ are substantially better on both the mixing and mixture density scales. Indeed, the $L_1$ error for the mixtures is, on average, about 4 times larger for $r = 1$ than for $r = 0.67$.

Using the same scale-mixture of Exponentials model, we compute the $L_1$ error $L_{1,n} = \int |m - m_n|$ for various (large) values of $n$ and, to estimate the rate of convergence, we use least squares to fit the log-linear model
\[ \log L_{1,n} = \log C - r \log n + \varepsilon_n, \quad (3.12) \]
where $C$ is a constant and $\varepsilon_n$ is zero-mean noise. We are particularly interested in the slope $r$ in the log-linear model since it characterizes the rate of convergence. Figure 3.2 shows the observed $L_1$ errors along with the fitted curve, obtained by fitting (3.12) via least-squares. The estimated rate $\hat{r} = 0.34$ is considerably larger than the $\approx 1/6$ bound given in Corollary 3.9. Therefore, empirical evidence suggests that the theoretical bound on the rate of convergence can be significantly improved.

Finally, we point out that the convergence results of Genovese and Wasserman [41] and Ghosal and van der Vaart [43] do not apply in this problem since the mixands $p(x|\theta)$ are not Gaussian, but the convergence theory of, say, histograms and kernel density estimates would apply. See DasGupta [20] and Wasserman [115].

**Remark 3.11** In some examples, the RE algorithm performs better with smaller weights $w_i = (i + 1)^{-1}$ in the sense of finite-sample accuracy. When estimating non-mixing parameters (see Chapter 5), larger weights like $w_i = (i + 1)^{-0.67}$ tend to yield severely biased estimates.
Figure 3.1. Estimated mixing and mixture densities in the Exponential scale-mixture problem for weights $w_i = (i + 1)^{-r}$, $r \in \{0.67, 1\}$, in the RE algorithm with $n = 1000$.

### 3.2.4 Convergence of $f_n$

Theorem 3.8 seems to suggest that the recursive estimates $f_n$ converge, in some sense, to some $f \in \mathcal{F}$ at which the infimum $K(m, M_\Theta)$ is attained. However, to conclude weak convergence of $f_n$ from the $L_1$ convergence of $m_n$ in Corollary 3.9, we need two additional conditions:

A3.5. Identifiability: $m_\varphi = m_\psi \nu$-a.e. implies $\varphi = \psi \mu$-a.e.
A3.6. For any $\varepsilon > 0$ and any compact $X_0 \subset \mathcal{X}$, there exists a compact $\Theta_0 \subset \Theta$ such that $\int_{X_0} p(x|\theta) \, \, d\nu(x) < \varepsilon$ for all $\theta \notin \Theta_0$.

The following theorem, which establishes conditions under which $L_1$ convergence of mixture distributions implies weak convergence of mixing distribution, will be needed to prove convergence of the recursive estimate $f_n$. Since this result requires no assumption on the construction of the sequence of mixing distribution estimates, it is presented in more generality than required.

**Theorem 3.12** Let $\tilde{F}$ and $\tilde{F}_n$ be probability measures on $\Theta$ with respective mixture densities $\tilde{m}(x) = \int p(x|\theta) \, d\tilde{F}($ and $\tilde{m}_n(x) = \int p(x|\theta) \, d\tilde{F}_n(\theta)$. Suppose $\tilde{m}_n \to \tilde{m}$ in $L_1$. Then, under A3.2–A3.3 and A3.5–A3.6, $\tilde{F}_n \to \tilde{F}$ in the weak topology.

**Proof** We first show that $\tilde{F}_n$ forms a tight sequence. Fix any $\varepsilon > 0$. It suffices to show existence of a compact $\Theta_0 \subset \Theta$ such that $\tilde{F}_n(\Theta_0) > 1 - \varepsilon$ for sufficiently large $n$. Take any compact $X_0 \subset \mathcal{X}$ such that $\int_{X_0} \tilde{m} \, d\nu > 1 - \varepsilon/2$. By A3.6, there exists

![Figure 3.2. Observed $L_1$ errors for various $n$ in the Exponential scale-mixture problem, along with the fitted curve $\hat{L}_{1,n} \propto n^{-r}$.](image)
a compact $\Theta_0$ such that $\int_{\Theta_0} p(x|\theta) \, d\nu(x) < \varepsilon/2$ for all $\theta \notin \Theta_0$. Now apply the $L_1$ convergence of $\tilde{m}_n$ to $\tilde{m}$ to conclude

$$1 - \frac{\varepsilon}{2} < \int_{\Theta_0} \tilde{m} \, d\nu \leq \lim_{n \to \infty} \int_{\Theta_0} \tilde{m}_n \, d\nu \leq \liminf_{n \to \infty} \left\{ \tilde{F}_n(\Theta_0) + \frac{\varepsilon}{2} \tilde{F}_n(\Theta_0^c) \right\}.$$  

Thus, $\tilde{F}_n$ is tight and the final assertion will follow once we show every weakly convergent subsequence $\tilde{F}_{n(k)}$ converges to $\tilde{F}$. Now, if $\tilde{F}_{n(k)} \rightarrow \tilde{F}^*$ for some probability measure $\tilde{F}^*$ on $\Theta$, then, by assumption A3.3, $\tilde{m}_{n(k)} \rightarrow \tilde{m}^*$ pointwise and hence in the $L_1$ topology (via Scheffé’s theorem) where $\tilde{m}^*(x) = \int p(x|\theta) \tilde{F}^*(d\theta)$. Therefore $\tilde{m}^* = \tilde{m}$, which, under A3.5, implies $\tilde{F}^* = \tilde{F}$.

In light of conditions A3.5–A3.6 and Theorem 3.12, weak convergence of the sequence of mixing densities $f_n$ follows immediately from the $L_1$ convergence of the mixtures implied by Theorem 3.8.

**Corollary 3.13** If conditions A3.1–A3.6 hold, then $f_n \rightarrow f$ a.s. in the weak topology, where $f \in F$ is the unique mixing density that satisfies $K(m,m_f) = K(m,M_{\Theta})$. Moreover, if $m \in M_{\Theta}$, then $f_n$ is a consistent estimate of the true mixing density $f$, in the weak topology.

### 3.3 Average over permutations

Despite the good convergence results for RE proved in Section 3.2, the dependence of $f_n$ on the order of the observations could be a cause of concern in applications, especially when $n$ is not very large. In some cases a particular ordering can be justified by problem specific considerations—see Bogdan, Ghosh and Tokdar [14]. In the absence of such justification a permutation invariant version of $f_n$ may be desirable. Newton [80] recommends calculating the average over a large number of random permutations. This can be seen as a Monte Carlo approximation to

$$\bar{f}_n = \frac{1}{n!} \sum_{\pi \in \Pi_n} f_{n,\pi} \quad (3.13)$$
where $\Pi_n$ is the permutation group on $\{1, \ldots, n\}$ and $f_{n,\pi}$, for $\pi \in \Pi_n$, represents the estimate $f_n$ with the observations arranged as $X_{\pi(1)}, \ldots, X_{\pi(n)}$. The estimate $\tilde{f}_n$ in (3.13), and the corresponding mixture $\tilde{m}_n := m_{f_n}$, will be called permutation-averaged RE, or PARE for short. The goal of the present section is to establish convergence properties of the PARE.

Begin by noticing that the permutation-averaged estimates satisfy

$$\tilde{f}_n = \mathbb{E}\{f_n \mid X^{(1)}, \ldots, X^{(n)}\} \quad \text{and} \quad \tilde{m}_n = \mathbb{E}\{m_n \mid X^{(1)}, \ldots, X^{(n)}\}.$$ 

Therefore, $\tilde{f}_n$ and $\tilde{m}_n$ produce a Rao-Blackwellization of $f_n$ and $m_n$, respectively, by making these functions of the sufficient statistic—the order statistics. As one might guess, this results in a smaller expected error in estimation, when error is measured by a divergence $d$ that is convex in the estimate. Indeed,

$$\mathbb{E}[d(f, \tilde{f}_n)] = \mathbb{E}[d(f, \mathbb{E}\{f_n \mid X^{(1)}, \ldots, X^{(n)}\})] \leq \mathbb{E}\mathbb{E}[d(f, f_n) \mid X^{(1)}, \ldots, X^{(n)}] = \mathbb{E}[d(f, f_n)]$$

and likewise,

$$\mathbb{E}[d(m, \tilde{m}_n)] \leq \mathbb{E}[d(m, m_n)]. \quad (3.14)$$

Examples of such divergence measures $d$ include the KL divergence and the $L_1$ distance. The following theorem is a consequence of Theorem 3.8 and (3.14).

**Theorem 3.14** Under A3.1–A3.4, $\tilde{m}_n \rightarrow m_f$ a.s. in $L_1$ where $m_f$ is the KL projection of $m$ onto $\mathbb{M}_\Theta$.

We next show that Theorem 3.14 leads to weak convergence of $\tilde{f}_n$. However, we prove convergence only in probability and not almost surely. Recall that $Y_n \rightarrow Y$ in probability if and only if every subsequence $n_k$ contains a further subsequence $n_{k(l)}$ such that $Y_{n_{k(l)}} \rightarrow Y$ a.s., whenever the underlying topology is metrizable.

**Corollary 3.15** Under A3.1–A3.6, $\tilde{f}_n$ converges weakly in probability to the unique mixing density $f \in \mathbb{F}$ such that $K(m, m_f) = K(m, \mathbb{M}_\Theta)$. 


Proof From Theorem 3.14 it follows that \( d(\bar{\mu}_n, m_f) \to 0 \) a.s., where \( d \) is the \( L_1 \) distance. Since the \( L_1 \) distance is bounded by 2, it follows by the dominated convergence theorem that \( E[d(\bar{\mu}_n, m_f)] \to 0 \) and, hence, \( d(\bar{\mu}, m_f) \to 0 \) in probability. Take an arbitrary subsequence \( n_k \). It must contain a further subsequence \( n_{k(l)} \) such that \( d(\bar{\mu}_{n_{k(l)}}, m_f) \to 0 \) a.s. Then Theorem 3.12 implies that \( \bar{f}_{n_{k(l)}} \to f \) a.s. in the weak topology. The assertion now follows since the weak topology is metrizable.

Remark 3.16 Even for moderate \( n \), there are too many permutations to compute \( \bar{m}_n \) exactly, so a Monte Carlo estimate \( \hat{m}_n \) based on a large number of random permutations is used as an approximation. Therefore, what we can conclude from Corollary 3.15 is a sort of practical consistency of \( \hat{m}_n \)—roughly, for large \( n \) and sufficiently many random permutations, \( \hat{m}_n \approx \bar{m}_n \) and \( \bar{m}_n \approx m \), which implies \( \hat{m}_n \approx m \).

3.4 Additional remarks

The theoretical work here parallels that done in Chapter 2 using results from stochastic approximation. Indeed, the KL divergence \( K(m, m_\varphi) \) plays the role of a Lyapunov function, and the functional \( T(\varphi) \), as pointed out in Remark 3.5, is related to the derivative of \( K(m, \cdot) \) in the direction of \( m_\varphi \). The non-negativity in Lemma 3.4 establishes a sort of “stochastic” Lyapunov property of the KL divergence.

In addition to finding a sharper bound on the rate in Corollary 3.9, there are other open problems. First, the rate of convergence for \( f_n \) is important, particularly in empirical Bayes problems, and it is unclear how it might be related to the rate of convergence for \( m_n \). Second, it is obvious from its definition that the rate of convergence for the PARE \( \bar{m}_n \) cannot be any slower than that of the RE \( m_n \). In fact, intuition suggests that the averaging in the PARE might yield a significantly better rate but, at this time, it is unclear how to incorporate the averaging over permutations into the rate of convergence analysis.
4. NUMERICAL RESULTS

In the following subsections we compare the performance of the recursive estimate (RE) and the recursive estimate averaged over permutations (PARE), starting with initial guess $f_0$, with that of the Dirichlet process mixture (DPM) estimate, where the mixing distribution $f$ is given a Dirichlet process prior with base measure $f_0$ and precision constant set to 1. While RE and PARE are easy to compute, computation of DPM is non-trivial. Here we employ a new importance sampling method, based on a collapsing of the Pólya Urn scheme; for details, see Appendix A of Tokdar, Martin and Ghosh [112]. We set the following simulation parameters:

- $T = 100$ samples of size $n = 200$ are taken from the model.
- For PARE, 100 random permutations of the data are selected.
- For RE and PARE, the weights satisfy $w_i = (i + 1)^{-1}$.
- For DPM, $R = 10,000$ importance samples are used. The efficiency is measured by the effective sample size (ESS); see, for example, Liu [66].

Remark 4.1 Estimation of a mixing distribution in the Dirichlet process mixture formulation is an extremely difficult problem. The current Monte Carlo approaches for DPM models, including the one used here, are based on some sort of exploration of the space of clustering configurations of the observations. Unfortunately, the conditional expectation of the mixing distribution, given the clustering, is highly variable—much more so than the conditional expectation of the mixture density. Consequently, one needs an extremely thorough exploration of the clustering space to obtain a reliable estimate of the mixing distribution. This is nearly impossible to achieve in finite time as this space grows exponentially with the number of observations.
4.1 Regular mixtures

In this subsection we will consider two regular mixture models, “regular” in the sense that $f$ is a density with respect to Lebesgue measure and smooth on its interval of support. Define the Beta-Normal (BN) and the Gamma-Poisson (GP) mixtures:

$$
\theta_i \sim \frac{1}{3} \text{Beta}(3, 30) + \frac{2}{3} \text{Beta}(4, 4) \quad X_i|\theta_i \sim \text{Normal}(\theta_i, \sigma^2) \quad (\text{BN})
$$

$$
\theta_i \sim \text{Gamma}_{[0,50]}(2, 0.4) \quad X_i|\theta_i \sim \text{Poisson}(\theta_i) \quad (\text{GP})
$$

In each case, the samples are independent across $i = 1, \ldots, n$. In (GP), the usual Gamma$(2, 0.4)$ distribution is truncated to $\Theta = [0, 50]$. Also, the mixture resulting from (GP) has a density with respect to counting measure. One can easily check that conditions A3.2–A3.4 are verified for these models; in particular, A3.4 follows immediately from the compactness of $\Theta$. For (BN) we choose $\sigma = 0.1$ but our conclusions hold for a range of $\sigma$ containing 0.1. We also choose $f_0$ to be a Unif$(\Theta)$ density in each case.

Figures 4.1 and 4.2 display the estimates for model (BN) and (GP), respectively. The DPM estimate of $f$ is overall quite poor, thanks to the difficulties described in Remark 4.1—the average ESS over the 100 datasets for the (BN) and (GP) models are only 538 and 324, respectively, compared to the $R = 10,000$ total importance samples. On the other hand, RE and PARE are much more stable across samples. Moreover, as expected from the Rao-Blackwellization, we see less variability in the PARE than in the RE, in the sense that the PARE estimate $\hat{f}_n$ hugs the true $f$ closer than does $f_n$.

That the sampling distribution is discrete in model (GP) has an interesting implication. In Figure 4.2 there is a (false) peak at zero for the mixture density $m_n$. This is due to the fact that the data $X_1, \ldots, X_n$ were generated by replicating each value according to its count. That is, the data sequence consists of all the 0s first, followed by all the 1s, etc. Therefore, permutation is necessary for count data stored in a particular deterministic order. Table 4.1 summarizes the computation time for PARE and DPM in the two models (the results for the NPMLE, described in Section 1.2.1,
based on Wang’s [114] algorithm are also shown). Clearly PARE (and RE) have a distinct advantage—DPM took over 300 times longer than PARE for (BN) and over 250 times longer for (GP).

Figure 4.3 summarizes the $L_1$ distances (left) as well as what we call a bias-spread summary (right) for the 100 estimates in the two regular examples. This bias-spread summary is similar to the traditional bias-variance decomposition of mean-square error: if $\hat{m}_{nt}$ is an estimate of the mixture density $m$ based on the $t^{\text{th}}$ sample ($t = 1, \ldots , T$) of size $n$, then

$$\text{Bias} = \int_X |\hat{m}_{nt} - m| dv \quad \text{and} \quad \text{Spread} = \frac{1}{T} \sum_{t=1}^{T} \int_X |\hat{m}_{nt} - \hat{m}_n| dv,$$

(4.1)
Figure 4.2. Plots of the mixing density estimates (top) and corresponding mixture density estimates (bottom) for model (GP). The cells show the true mixing or mixture densities (black) with the $T = 100$ estimates (gray).

Table 4.1
Mean computation time (in seconds) for PARE, NPMLE and DPM over the $T = 100$ samples in the Beta-Normal and Gamma-Poisson models. RE (not displayed) is about 100 times faster than PARE.

<table>
<thead>
<tr>
<th>Model</th>
<th>PARE</th>
<th>NPMLE</th>
<th>DPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN</td>
<td>0.14</td>
<td>1.11</td>
<td>43.77</td>
</tr>
<tr>
<td>GP</td>
<td>0.12</td>
<td>0.20</td>
<td>31.41</td>
</tr>
</tbody>
</table>
where \( \hat{m}_n(x) = T^{-1} \sum_{t=1}^{T} \hat{m}_n(x) \) is the point-wise average of the \( T \) estimates of \( m(x) \). We consider the sum of the bias and spread as a measure of overall variability and look at how the two components contribute to the sum. In both examples, PARE performs better in terms of overall variability, spread and, most importantly, \( L_1 \) loss. Compared to the other estimates, it appears that PARE does a better job of simultaneously controlling bias and spread. In the Beta-Normal example, RE also performs well. Due to the deterministic ordering issue mentioned above, RE performs quite poorly for (GP) and is not displayed. Note that these relative comparisons remain the same when the \( L_1 \) distance is replaced by the KL divergence.

Figure 4.3. Summary of the \( L_1 \) errors (left) and Bias-Spread tradeoff (right) for models (BN) (top) and (GP) (bottom).
The original motivation for RE was to approximate the DP mixture, so an interesting question is whether RE or PARE is close to the DP mixture in some sense. Table 4.2 summarizes the KL divergence of RE, PARE, and DPM from the true mixture, as well as the RE and PARE from the DPM for each of the two models being considered. Of course, since RE is not suitable for the GP model, these data do not appear in the table. That PARE is closer to the true mixture than DPM on average was apparent from Figures 4.1–4.2 and this is confirmed in the table. What is particularly interesting is that PARE is also fairly close to DPM on average. However, there is a striking pattern: on average, we see the relationship

\[ \text{KL(Truth, DPM)} \approx \text{KL(Truth, PARE)} + \text{KL(DPM, PARE)}, \]

suggesting that PARE falls somewhere in between the True mixture and the Bayes estimate, in some sense. Other simulations, not presented here, with larger sample sizes like \( n = 500 \) also show a similar pattern.

### Table 4.2
Summary of KL divergences (×100) of the various estimates against (i) the true mixture, and (ii) the estimated DP mixture for each of the models BN and GP.

<table>
<thead>
<tr>
<th>Model</th>
<th>Pair</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Mean</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>BN</td>
<td>RE vs. True</td>
<td>0.14</td>
<td>0.82</td>
<td>1.30</td>
<td>1.40</td>
<td>1.9</td>
<td>5.4</td>
</tr>
<tr>
<td></td>
<td>PARE vs. True</td>
<td>0.11</td>
<td>0.44</td>
<td>0.79</td>
<td>0.92</td>
<td>1.2</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>DPM vs. True</td>
<td>0.20</td>
<td>0.78</td>
<td>1.20</td>
<td>1.40</td>
<td>1.7</td>
<td>4.3</td>
</tr>
<tr>
<td>RE</td>
<td>DPM</td>
<td>0.120</td>
<td>0.64</td>
<td>0.90</td>
<td>0.96</td>
<td>1.20</td>
<td>2.1</td>
</tr>
<tr>
<td>PARE</td>
<td>DPM</td>
<td>0.094</td>
<td>0.34</td>
<td>0.49</td>
<td>0.54</td>
<td>0.73</td>
<td>1.4</td>
</tr>
<tr>
<td>GP</td>
<td>PARE vs. True</td>
<td>0.068</td>
<td>0.32</td>
<td>0.56</td>
<td>0.63</td>
<td>0.8</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>DPM vs. True</td>
<td>0.200</td>
<td>0.78</td>
<td>1.10</td>
<td>1.20</td>
<td>1.6</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>PARE vs. DPM</td>
<td>0.12</td>
<td>0.36</td>
<td>0.47</td>
<td>0.5</td>
<td>0.62</td>
<td>1.4</td>
</tr>
</tbody>
</table>
4.2 Irregular mixture

For an irregular mixture, we take $f$ to have both a discrete and an absolutely continuous component. In particular, consider the Irregular-Normal (IN) mixture

$$
\theta_i \sim \frac{2}{3} \delta_{\{0\}} + \frac{1}{3} \text{Normal}_{[-10,10]}(0, 4), \quad X_i | \theta_i \sim \text{Normal}(\theta_i, 1) \quad \text{(IN)}
$$

where the samples are independent across $i = 1, \ldots, n$, $\delta_{\{0\}}$ denotes a point-mass at zero, and the usual $N(0, 4)$ distribution is truncated to $\Theta = [-10, 10]$. Note that the choice of dominating measure $\mu$ is Lebesgue measure on $\Theta$ plus a unit mass at zero. The initial guess/hyperparameter $f_0$ is taken to be $\frac{1}{2} \delta_{\{0\}} + \frac{1}{2} \text{Unif}(\Theta)$ density. In this section we focus on just the PARE and DPM estimates.

Figure 4.4 shows the 100 estimates of the absolutely continuous part $\varphi$ of the mixing distribution as well as the corresponding estimates of the mixture. Just as in Section 4.1, we see PARE has considerably less variability than DPM (which has an average ESS of about 330) on the $\theta$-scale, while both perform comparably on the $x$-scale. The left-most plot in Figure 4.5 summarizes the 100 estimates of $\pi = \Pr(\theta = 0)$. Both procedures tend to overestimate $\pi = 0.667$ (horizontal line). Most likely, this is because $\varphi$ is also fairly concentrated around $\theta = 0$. The right two plots in Figure 4.5 summarize the $L_1$ errors and the bias-spread over the 100 samples, respectively. PARE, again, tends to be much more accurate under $L_1$ loss: on average, the $L_1$ loss is about 34% larger for DPM than it is for PARE. Also, PARE seems to handle the twin bias-spread problems better than DPM.

4.3 Massive data example

The irregular mixture (IN) in Section 4.2 arises in many important applications. In microarray analysis [98] or quantitative trait loci (QTL) mapping [14], each $\theta$ represents the expression level of a single gene or the association level of a single genetic marker, respectively. For the nonparametric regression problem [17], the $\theta$'s represent coefficients of, say, a wavelet basis expansion of the regression function. In
each example, the \( \theta \)-vector is assumed to be “sparse” in the sense that most of the \( \theta \)'s are zero. To account for sparseness, a Bayesian formulation assumes that the \( \theta \)'s are independent observations from a common prior distribution

\[
F(d\theta) = \pi \delta_{\{0\}}(d\theta) + (1 - \pi) \phi(\theta) \, d\theta.
\]

The dimension \( n \) of the \( \theta \)-vector is quite large in applications so an empirical Bayes approach is an attractive alternative to a fully Bayes procedure. It was shown in Section 4.2 that both the PARE and DPM can be used to estimate \((\pi, \phi)\), but when \( n \) is extremely large, computation becomes very expensive, particularly for DPM.
Figure 4.5. Summary of the estimates of Pr(θ = 0) (left), summary of the L_1 errors (middle), and Bias-Spread tradeoff (right) for irregular mixture model (IN).

Figure 4.6. Plot of the absolutely continuous part of the mixing distribution (left) and corresponding mixture density estimates (right) for model (IN) in the massive data example.
We take a simulated dataset of size $n = 50,000$ from the model (IN) in Section 4.2. Figure 4.6 shows the PARE and DPM estimates of $(\pi, \varphi)$ in (4.2). Although the PARE has two modes, it is a much closer approximation to the true $\varphi$ compared to the spiky DPM estimate. An important point is that, even with $R = 10,000$ importance samples, the ESS is only 1—see Remark 4.1. The estimates $\hat{\pi}$ of $\pi$ are 0.733 for PARE and 0.772 for DPM; both are slightly larger than the target $\pi = 0.667$. Figure 4.6 also shows the estimates of the mixture density $m$. With $n = 50,000$, $L_1$-consistency of the DPM posterior mean (see, for example, Barron, Schervish and Wasserman [6] and Ghosal, Ghosh and Ramamoorthi [42]) and of the RE estimate $\bar{m}_n$ (see Theorem 3.14) has kicked in, and that of DPM and PARE follows by Remark 3.16. What is perhaps most important in massive data problems—where almost any estimate will perform well—is computational efficiency. Here, the PARE was obtained in 45 seconds, while the DPM estimate took nearly 6 hours.

Evidence suggests that PARE is a much better procedure than DPM in this problem for an empirical Bayes analysis. Compared to DPM, the PARE algorithm is easier to implement, the computation is significantly faster, and the resulting estimates of $(\pi, \varphi)$ are much more accurate. We should point out, however, that DP mixtures are not meant to handle problems where the mixing distribution has a density. Perhaps a nonparametric prior for $f$ that sits on densities, such as Polya tree or Gaussian process priors, would fare better in our comparisons.
5. AN EXTENSION OF THE RECURSIVE ALGORITHM

In this chapter we introduce an important extension of Newton’s original recursive algorithm. As discussed in Chapter 1, the recursive algorithm has several limitations that are difficult to overcome. Here we tackle the mixture model problem with additional unknown non-mixing parameters. Motivated by the large-sample robustness of RE (cf. Theorem 3.8), we propose an extended RE algorithm, called the RE+, that features a “likelihood-based” estimate of the non-mixing parameters. A Bayesian interpretation of the RE+ algorithm can be given based on the relationship between RE and DP mixtures and general marginal likelihood methods in Bayesian statistics. Some convergence theory for the RE+ is presented in Section 5.2 and some simulation results are given in Section 5.3. Finally, the RE+ is applied to two density estimation problems in Section 5.4, highlighting the flexibility of the RE+ algorithm.

5.1 The RE+ algorithm

As discussed in Chapter 1, a serious limitation of the RE algorithm is that the sampling density $p(x|\theta)$ must be fully specified by the latent/mixing variable $\theta$. This restriction seems to severely limit the algorithm’s scope of application. We considered this problem in Chapter 2, but there it was assumed that replicates were available. Our approach here is entirely different. The robustness of the recursive algorithm under mis-specification (Theorem 3.8) will prove to be critical in our development.

In this section we consider a more general mixture model

$$m_{f,\xi}(x) := \int_\Theta p(x|\theta, \xi)f(\theta)\,d\mu(\theta), \quad x \in \mathcal{X},$$

(5.1)

where $p(x|\theta, \xi)$ is a two-parameter family of densities with respect to a dominating measure $\nu$ on $\mathcal{X}$ and, as before, $f$ is a density on $\Theta$ with respect to the dominating
measure $\mu$. Note that the measure $\mu$ might also depend on the parameter $\xi$; see Chapter 6. In general, we will suppress this dependence in the notation. We will assume that independent data $X_1, \ldots, X_n$ arise from a distribution $md\nu$ on $\mathcal{X}$.

Inclusion of an additional non-mixing parameter $\xi \in \Xi$ clearly makes the mixture model (5.1) more flexible than (1.1). Unfortunately, the RE algorithm cannot be used for models of this generality; see Bogdan, Ghosh and Tokdar [14]. But given the good theoretical properties and computational efficiency of the RE, one might be tempted to use guess of $\xi$ as a plug-in in the RE algorithm. Let $m_{n,\xi} := m_{f_n,\xi}$ be the estimated mixture based on using the value $\xi$ in the RE algorithm. Then, under suitable conditions, Theorem 3.8 states that, for each fixed $\xi$ we have

$$K(m, m_{n,\xi}) \to \inf \{K(m, m_{f,\xi}) : f \in \mathbb{F}\} \quad \text{a.s.} \quad n \to \infty. \quad (5.2)$$

Our proposed extension of the RE algorithm, which we call the RE+ algorithm, chooses $\xi$ to minimize a “sample version” of the KL divergence $K(m, m_{n,\xi})$ that appears on the left-hand side of (5.2). Minimizing this sample version of the KL divergence is equivalent to maximizing a suitable pseudo-likelihood function. The estimation procedure based on maximization of this pseudo-likelihood is akin to marginal likelihood methods in Bayesian problems.

Our jumping off point is that, for fixed $\xi \in \Xi$, the mixture density $m_{n,\xi}$ asymptotically minimizes the KL divergence $K(m, m_{f,\xi})$ over all mixtures of $p(\cdot | \theta, \xi)$ densities. This observation suggests that we fit the general mixture model (5.1) by minimizing $K(m, m_{n,\xi})$ over $\xi \in \Xi$. Unfortunately, the density $m$ is unknown so this quantity cannot be calculated. Instead, we propose to minimize a “sample version” of the KL divergence, namely,

$$\tilde{K}_n(\xi) = \frac{1}{n} \sum_{i=1}^{n} \log \{m(X_i)/m_{i-1,\xi}(X_i)\}. \quad (5.3)$$

Note that minimizing $\tilde{K}_n(\xi)$ is equivalent to maximizing a pseudo-likelihood function $L_n(\xi)$, given by

$$L_n(\xi) = \sum_{i=1}^{n} \log m_{i-1,\xi}(X_i), \quad (5.4)$$
so evaluation of \( m \) is not required. This leads to the following extension of the RE algorithm, which we call the RE+ algorithm.

**RE+ Algorithm**

Maximize the pseudo-likelihood \( L_n(\xi) \) in (5.4) over \( \Xi \), to produce the estimate \( \hat{\xi}_n = \arg\max L_n(\xi) \). Then, run the RE algorithm with the plug-in \( \xi = \hat{\xi}_n \) to produce the estimate \( \hat{f}_n := f_{n,\hat{\xi}_n} \) as an estimate of \( f \).

Evaluation of the pseudo-likelihood \( L_n(\xi) \) in (5.4) is performed by passing through the recursive algorithm, either with or without averaging over permutations of the data, with the specified \( \xi \) in the sampling density \( p(x|\theta,\xi) \). Maximization can then be performed using any available optimization procedure. In our experience, maximization of \( L_n(\xi) \) is relatively fast, usually requiring only a few, relatively inexpensive function evaluations.

In many cases, when additional unknown parameters are introduced, the model may become unidentifiable. For example, a location mixture of Normal densities with unknown variance \( \sigma^2 \) is unidentifiable. This can be problematic when the parameters of interest are “real-world” quantities. However, the mixture model structure is often artificial—for modeling simplicity and flexibility—so any set of parameters that provide an adequate fit to the data would suffice.

For some additional insight, we next show that the RE+ algorithm can be expressed an approximate filtering algorithm under a suitable DP prior Bayes formulation. First, consider the very general Bayesian model with independent observations

\[
X_1, \ldots, X_n \sim m(\cdot|\psi, \xi),
\]

where \( \psi \)—possibly a mixing density—has a prior distribution \( \psi \sim \Pi \) and \( \xi \) is a hyperparameter. The log marginal likelihood of \( \xi \) equals

\[
\sum_{i=1}^{n} \log m_{i-1}(X_i|\xi),
\]

where \( m_{i-1}(\cdot|\xi) \) is the posterior predictive density of \( X_i \) given \( X_1, \ldots, X_{i-1} \), obtained by integrating out \( \psi \) with respect to the posterior \( \Pi(\cdot|X_1, \ldots, X_{i-1}) \); i.e.,

\[
m_{i-1}(x|\xi) = \int m(x|\psi, \xi) d\Pi(\psi|X_1, \ldots, X_{i-1}). \tag{5.5}
\]
When \( \xi \) is unspecified, the analysis might proceed by maximizing \( \sum_{i=1}^{n} m_{i-1}(X_i|\xi) \) over \( \xi \). A filtering algorithm is a modification, or approximation, of the approach outlined above. The main difference being that, at the \( i \)-th step, \( m_{i-1}(\cdot|\xi) \) is replaced by \( m(\cdot|\hat{\psi}_{i-1},\xi) \), for some estimate \( \hat{\psi}_{i-1} \) of \( \psi \) based on \( X_1, \ldots, X_{i-1} \), such as the posterior mean or mode. This approximation effectively eliminates the need for the (possibly difficult, high-dimensional) integration in (5.5).

Now think back to our RE+ formulation. In light of the one-step correspondence (Proposition 1.2) between RE and DP mixtures, the pseudo-likelihood \( L_n(\xi) \) in (5.4) can be explained as the log marginal likelihood of \( \xi \) under the Bayesian formulation \( f \sim \mathcal{D}(1/w_1 - 1, f_0) \) obtained by an approximate filtering algorithm in which, at each step \( i = 1, \ldots, n \), the conditional posterior distribution of \( f \), given \( X_1, \ldots, X_{i-1} \), is approximated by \( \mathcal{D}(1/w_i - 1, f_{i-1,\xi}) \).

Further Bayesian considerations can be incorporated into the RE+ formulation. For example, from a Bayesian point of view, it would be natural to incorporate a prior distribution \( \Pi(\xi) \) for the non-mixing parameter \( \xi \)—a hierarchical structure is very flexible and leads to a more “data-driven” analysis. This is easy to do in the present framework since inclusion of a prior \( \Pi(\xi) \) changes the RE+ algorithm in only the most obvious way: rather than maximizing the pseudo-likelihood (5.4), this regularized RE+ algorithm would instead maximize a pseudo-posterior of the form

\[
L_n(\xi) + \log \Pi(\xi).
\]

This approach will be described further in Chapter 6 for the special case of simultaneous significance testing. There a default prior \( \Pi(\xi) \) is proposed, based on both subjective and objective considerations.

### 5.2 Convergence analysis

Towards a rigorous justification of the RE+ algorithm, we show that the convergence in (5.2) continues to hold when \( K(m, m_n, \xi) \) is replaced by \( \tilde{K}_n(\xi) \). Assume the following conditions, similar to those in Chapter 3:
A5.1. \( w_n > 0, w_n \downarrow 0, \sum_n w_n = \infty, \) and \( \sum_n w_n^2 < \infty. \)

A5.2. \( \Theta \) and \( \Xi \) are compact metric spaces.

A5.3. \( (\theta, \xi) \mapsto p(x|\theta, \xi) \) bounded and continuous for \( \nu \)-a.e. \( x \in \mathcal{X}. \)

A5.4. There exists \( B < \infty \) such that
\[
\int_{\mathcal{X}} \left[ \frac{p(x|\theta_1, \xi_1)}{p(x|\theta_2, \xi_2)} \right]^2 m(x) \, d\nu(x) < B \quad \forall (\theta_1, \xi_1), (\theta_2, \xi_2) \in \Theta \times \Xi.
\]

**Theorem 5.1** Under conditions A5.1–A5.4,
\[
\tilde{K}_n(\xi) \to \inf \{ K(m, m_{f, \xi}) : f \in F \} \quad \text{a.s.}
\]
as \( n \to \infty \) for each fixed \( \xi \in \Xi. \)

**Proof** Define the random variables
\[
U_i = \log \left[ \frac{m(X_i)}{m_{i-1, \xi}(X_i)} \right] - K(m, m_{i-1, \xi}), \quad i \geq 1
\]
and note that \( \mathbb{E}[U_i | \mathcal{A}_{i-1}] = 0, \) where \( \mathcal{A}_{i-1} = \sigma(X_i, \ldots, X_{i-1}). \) Therefore, \( \{(U_n, \mathcal{A}_n) : n \geq 1\} \) forms a zero mean martingale. Next, if we let \( \mathcal{E} = \{ m < m_{i-1, \xi} \} \subset \mathcal{X}, \) then by A5.4 and several applications of Jensen’s inequality we get
\[
\mathbb{E}\{U_i^2 | \mathcal{A}_{i-1}\} \leq \int_{\mathcal{X}} \left( \log \frac{m}{m_{i-1, \xi}} \right)^2 m \, d\nu
\]
\[
= \int_{\mathcal{E}} \left( \log \frac{m_{i-1, \xi}}{m} \right)^2 m \, d\nu + \int_{\mathcal{E}^c} \left( \log \frac{m}{m_{i-1, \xi}} \right)^2 m \, d\nu
\]
\[
\leq \int_{\mathcal{E}} \left( \frac{m_{i-1, \xi}}{m} - 1 \right)^2 m \, d\nu + \int_{\mathcal{E}^c} \left( \frac{m}{m_{i-1, \xi}} - 1 \right)^2 m \, d\nu
\]
\[
< C := 2(B + 1)
\]

Then the so-called *conditional variance* of the process \( \{U_n\}, \) defined as
\[
V_n^2 = \sum_{i=1}^n \mathbb{E}\{U_i^2 | \mathcal{A}_{i-1}\},
\]
satisfies \( r_n \sqrt{V_n} \leq n^{1/2} r_n \sqrt{C} \to 0 \), where \( r_n = n^{-1} \log \log n \). Also, by Chebyshev’s inequality,
\[
\sum_{n=1}^{\infty} P\{|U_n| > r_n^{-1} \alpha_{n-1}\} \leq C \sum_{n=1}^{\infty} r_n^2 < \infty \quad \text{a.s.}
\]
and it follows from Corollary 2 of Teicher [110] (with \( b_n = n \) and \( \beta = 1 \)) that \( n^{-1} \sum_{i=1}^{n} U_i \to 0 \) a.s. Therefore, we can conclude that
\[
\left| L_n(\xi) - \frac{1}{n} \sum_{i=1}^{n} K(m, m_{i-1, \xi}) \right| \to 0 \quad \text{a.s.} \quad (5.6)
\]
However, we know from Theorem 3.8 that \( K(m, m_{i-1, \xi}) \) and, hence, the average \( n^{-1} \sum_{i=1}^{n} K(m, m_{i-1, \xi}) \) converges to \( \inf \{ K(m, m_f, \xi) : f \in \mathfrak{F} \} \). The claim now follows immediately from (5.6).

The pointwise convergence in Theorem 5.1 implies that RE+ adapts well to the particular choice of \( \xi \). It does not imply convergence of \( \hat{\xi}_n \) unless \( \Xi \) is a finite set—for general \( \Xi \), uniform convergence in Theorem 5.1 would be needed. Unfortunately, the standard tools for proving uniform laws of large numbers, such as empirical processes (see, for example, Chapter 2 of Pollard [84]), seem inappropriate here because the \( i \)th summand in \( \tilde{K}_n(\xi) \) depends on the full data sequence \( X_1, \ldots, X_i \). Other approaches are available—for example, Andrews [3] and Datta and Hannan [21]—but the required stochastic equicontinuity conditions seem difficult to verify in the RE+ context.

### 5.3 RE+ Simulations

Perhaps the simplest problem in which the RE+ could be applied might be a location-mixture of Normals with a common unknown scale parameter \( \sigma \). In this section we study the performance of the RE+ algorithm in such a problem, paying close attention to the behavior of \( \tilde{K}_n(\hat{\sigma}_n) \) in (5.3), the “empirical version” of the KL divergence evaluated at the maximizer \( \hat{\sigma}_n \) of the pseudo-likelihood \( L_n \) in (5.4).

Suppose \( m(x) \) is a location-mixture of Normals. Specifically, take \( \Theta = [0, 1] \) and
\[
\theta_i \sim \frac{1}{3} \text{Beta}(5, 30) + \frac{2}{3} \text{Beta}(5, 4), \quad X_i | \theta_i \sim N(\theta_i, \sigma^2), \quad i = 1, \ldots, n,
\]
where sampling is independent and the Normal scale parameter is \( \sigma = 0.1 \). The simulations in Chapter 4 show that, when \( \sigma \) is known, the RE performs as well as the Bayes DP mixture estimates in terms of computational efficiency and estimation accuracy. Here we investigate the performance of the RE+ algorithm, as described above, that treats \( \sigma \) as unknown and estimates it from the observed data.

In this case, we want choose \( \hat{\sigma}_n \) to maximize the pseudo-likelihood \( L_n(\sigma) \). Evaluation of \( L_n(\sigma) \) is done via a single pass through the recursive algorithm with weight sequence \( w_i = (i+1)^{-1} \); see Remark 3.11. Once the maximizing \( \hat{\sigma}_n \) is found, we estimate the mixing density with the plug-in recursive estimate. For \( N = 100 \) samples of size \( n \in \{125, 250, 500, 1000\} \), Figure 5.1 summarizes the \( L_1 \) error \( \int |m - m_{n,\hat{\sigma}_n}| d\nu \), as well as the RE+ estimates of \( \sigma \). In each case, we see the estimates becoming increasingly more accurate as \( n \) increases—specifically, as \( n \) increases the average is draws nearer to zero and the spread shrinks. Computation time for the RE+ is relatively short, with significant improvement when implementation includes a formula for the derivative; see Section 6.8 where a (more complicated) derivative formula is given.

In Figure 5.1 we see the performance of the RE+ estimates of \( m \) and \( \sigma \) independent of one another. However, it is also of interest (partly from a theoretical point of view) to see the joint performance of \( \hat{\sigma}_n \) and \( m_{n,\hat{\sigma}_n} \). Figure 5.2 shows scatterplots of \( \hat{\sigma}_n \) versus \( \tilde{K}_n(\hat{\sigma}_n) \), the empirical KL divergence, for the 100 repetitions at each value of \( n \). When \( n \) is relatively small, there is a lot of scatter, but as \( n \) increases the points become more and more concentrated around the point \((0.1, 0)\). This observation has two important consequences: (i) there seems to be no identifiability problems in this example and (ii) it suggests that \( \tilde{K}_n(\hat{\sigma}_n) \) indeed converges to zero.

### 5.4 Density estimation with RE+

Nonparametric density estimation is a fundamental problem in statistics and mixture distributions, thanks to their built-in flexibility, play a major role in many popular methods, both frequentist and Bayesian. Here we describe two of the most popular
methods—kernel density estimation and finite mixtures—and demonstrate how the RE+ algorithm presented in Section 5.1 can be applied to each. For each problem below, Θ will be either a compact set or a finite set, μ will be either Lebesgue or counting measure, and the initial guess $f_0$ will be a Uniform density on Θ with respect to μ.
Figure 5.2. Scatterplot of $\hat{\sigma}_n$ by $\tilde{K}_n(\hat{\sigma}_n)$ for various $n$ in the Beta-Normal example. The points should concentrate around (0.1, 0).

5.4.1 Kernel-based density estimation

Kernel-based methods are among the most popular of density estimation procedures. Silverman [104] gives a complete treatment and Sheather [101] gives a recent survey of the various methods available as well as their software implementations.
The basic idea is to smooth out the empirical distribution—which is discrete, supported on the observed data points. That is, for a kernel function $\kappa(\cdot)$, and for an independent sample $X_1, \ldots, X_n$ from $m(x)$, the kernel estimate is

$$m_n^\kappa(x) = \frac{1}{nh} \sum_{i=1}^n \kappa\left(\frac{x - X_i}{h}\right),$$

where $h$ is a suitable tuning parameter, called a bandwidth, chosen by the user. There is substantial literature on the choice of the bandwidth $h$; a nice review and collection of references is given in Chapter 32 of DasGupta [20]. These choices are typically based on an asymptotic minimization of some error measurement, such as integrated mean square error (IMSE), and will depend on the unknown $m$. Therefore, the “optimal” $h$ must be estimated. There are various ways to do this and software is readily available; see Sheather [101].

In this subsection we will apply the RE+ algorithm which treats the unknown $m$ as a continuous mixture of Normal densities and $h$, the common scale of these Normal densities, as an unknown non-mixing parameter. Here we will be using RE+ as a “bandwidth selector” but one could also think of this as an “nonparametric empirical Bayes density estimate” in the spirit of Chapter 6.

For the unknown density $m$, consider a model of the form

$$m_{f,h}(x) = \int p(x|\theta, h)f(\theta) \, d\theta,$$

where $p(\cdot|\theta, h)$ is a Normal density with mean $\theta$ and scale $h$. In this context, the “bandwidth” $h$ plays the role of the non-mixing parameter $\xi$. The RE+ algorithm can now be applied in an obvious way; in our implementation, optimization of the pseudo-likelihood (5.4), defined in this case as

$$L_n(h) = \sum_{i=1}^n \log m_{i-1,h}(X_i),$$

is performed using the nlm routine in the R statistical software package [87], with a formula included for the derivative. It is worth pointing out here that the RE+ choice of bandwidth is solely data-based—no asymptotic considerations are involved.
In what follows we compare the performance of the RE+ kernel density estimate with that of several popular alternatives, namely, (biased) cross-validation (CV) and the Sheather-Jones [102] (SJ) plug-in rule. Givens and Hoeting [48], Chapter 10, give an introduction to these two methods. These two methods, as pointed out above, choose the bandwidth to minimize the asymptotic IMSE so, contrary to the RE+ selection rule, here the observed data is only indirectly involved.

First we give one real-data example showing that the RE+ estimate can, in fact, be preferable to these more classical kernel-based estimates.

**Example 5.2** Data on the times (in hours after midnight April 5th, 2001) of \( n = 121 \) bowhead whale calf sightings off the coast of Alaska are shown in Figure 5.3. These animals are endangered so characterizing the migration pattern of these whales is of scientific interest. Here we fit a three nonparametric density estimates to this whale data: Gaussian kernel density estimates based on the SJ and CV bandwidth selection rules and the RE+ estimate obtained by minimizing the pseudo-likelihood \( L_n(h) \). In this example, we see that the CV kernel estimate is much too smooth, failing to capture some of the important aspects of the histogram, while the SJ estimate is perhaps too wiggly, giving an unrealistically good fit to the observed data. On the other hand, we see that the RE+ falls somewhere in the middle, being smoother than the SJ estimate and rougher than the CV estimate.

Next we give some simulation results comparing the performance of RE+ with SJ and CV for a moderately large sample size \( n \).

**Example 5.3** Let \( m(x) \) be the density of the following two-component mixture distribution: \( 0.5 \mathcal{N}(4, 1^2) + 0.5 \mathcal{N}(9, 2^2) \). Here we take 100 samples of size \( n = 1000 \) from the above mixture and, for each sample, the three estimates—RE+, SJ and CV—are computed. The RE+ algorithm uses the model

\[
m_{f,h}(x) = \int_{\Theta} p(x|\theta, h) f(\theta) \, d\theta,
\]

(5.7)
where \( \Theta = [1, 15] \) and \( p(x|\theta, h) \) is a \( N(\theta, h^2) \) density. As before, RE+ is used to estimate the “bandwidth” \( h \) and then this \( h \) is used as a plug-in to compute the mixing density \( f(\theta) \) using RE, averaged over 25 random permutations. The \( L_1 \) error and the estimates of the bandwidth \( h \) are summarized in Table 5.1. We see very little difference between the SJ and CV estimates in this case, but the RE+ estimates are noticeably different. First, the RE+ bandwidths are dramatically different from the SJ and CV bandwidths; this is not surprising since RE+ is doing something entirely different. In particular, the quantity \( h \) has a meaningful interpretation under the RE+ framework—it is the data’s choice of the “best” scale parameter in the Gaussian location-mixture model—whereas, in kernel density estimation, \( h \) is a meaningless tuning parameter. Second, we see that the \( L_1 \) error of the RE+ estimates is about 7% less than that of the SJ and CV estimates, on average. This, by itself, is an interesting observation but note also that, in this problem, the RE+ model (5.7)
does not contain the true density \( m \). The conclusions of Corollary 3.9, along with these two observations, suggest that the RE+ estimate may converge faster than the kernel estimates but, in general, to the wrong thing, namely the KL projection. If the empirical rate \( \hat{r} = 0.34 \) presented in Figure 3.2 of Chapter 3 is accurate, then RE+ convergence rate might be close to that of classical kernel estimates. The \( L_1 \) error of the kernel estimates, which are consistent, would eventually surpass that of the inconsistent RE+, but perhaps the additional flexibility of a continuous mixture leads to a practical improvement in finite samples.

<table>
<thead>
<tr>
<th>Table 5.1</th>
<th>Mean (and standard deviation) of the indicated quantity over the 100 replications of the density estimation experiment in Example 5.3.</th>
</tr>
</thead>
<tbody>
<tr>
<td>RE+</td>
<td>SJ</td>
</tr>
<tr>
<td>( L_1 ) error</td>
<td>0.0772 (0.0212)</td>
</tr>
<tr>
<td>“Bandwidth”</td>
<td>0.981 (0.0489)</td>
</tr>
</tbody>
</table>

5.4.2 Finite mixture densities

Kernel-based density estimation is more or less standard in that the “bandwidth” \( h \) is a parameter in the usual sense. However, the RE+ setup allows for a much more general choice of non-mixing parameter and this next application—density estimation with finite mixtures—illustrates just how flexible the RE+ algorithm is.

Suppose \( m(x) \) is a finite mixture of the form

\[
m(x) = \sum_{s=1}^{S} p(x|\theta^s)f(\theta^s),
\]

where \( f \) has an unknown finite support \( \{\theta^1, \ldots, \theta^S\} \) of unknown size \( S \) within some known bounded set \( \Theta \). A naive approach, described briefly in Chapter 2, is as follows: choose a suitably fine grid of points \( \Theta = \{\theta_1, \ldots, \theta_G\} \) from \( \Theta \) as a candidate support
and estimate $f$ with Newton’s estimate $f_n$ on $\Theta$. Conjecture 2.17 claims that, for large $n$, the recursive estimate would produce the KL-best mixture over the candidate support $\Theta$; Theorem 3.8 above confirms this conjecture. This approach will produce an answer very quickly but, unfortunately, there are drawbacks. Most importantly, $f_n$ is too smooth in the sense that too many points in $\Theta$ are given positive mass; see, for example, Figure 3.1 in [71] or Figure 19.1 in [47]. To improve upon this naive approach, we propose a solution based on the RE+ method described in Section 5.1. The idea is to treat the support itself as an unknown non-mixing parameter $\xi$.

Let $\Theta = \{\theta_1, \ldots, \theta_G\}$ be the candidate support grid and let $\xi$ be a binary $G$-vector, with $\xi_g \in \{0, 1\}$ indicating whether or not $\theta_g$ receives positive mass. That is, $\xi$ controls which points of $\Theta$ are included in the mixture. More precisely, we consider mixtures of the form

$$m_{f,\xi}(x) = \frac{\sum_{g=1}^G p(x|\theta_g) f(\theta_g) \xi_g}{\sum_{g=1}^G f(\theta_g) \xi_g}.$$ 

It follows from Theorem 5.1 that $\tilde{K}_n(\xi) \to \inf_f K(m, m_{f,\xi})$ for fixed $\xi$ as $n \to \infty$. Moreover, since $\Xi = \{0, 1\}^G$ is a finite set, the estimated support indicator vector $\hat{\xi}_n = \arg \max L_n(\xi)$ converges a.s. to the true support indicator and, consequently, $\sum_{g=1}^G \hat{\xi}_{n,g}$ is a consistent estimate of the mixture complexity.

In this context, maximizing $L_n(\xi)$ over $\Xi$ is a combinatorial optimization problem. The solution space has $2^G$ elements, which is too large for an exhaustive search, even for moderate $G$. Instead, we will use a simulated annealing procedure to maximize $L_n(\xi)$ over $\Xi$. Simulated annealing is a numerical procedure designed to optimize functions over finite state spaces; see, for example, Robert and Casella [95, Sec. 5.2.3] and Givens and Hoeting [48, Sec. 3.4] for an introduction. Like MCMC, simulated annealing constructs a stochastic process designed to explore the state space, looking for modes in the objective function. Again, like MCMC, simulated annealing has implementation parameters—namely, the proposal distribution and cooling schedule—that have a significant effect on the performance of the algorithm. The details of our implementation can be found in Section 5.5.
Example 5.4 Evidence shows that the universe is in a never-ending process of expansion. Under the Big Bang model, galaxies should form clusters and the relative velocities of the galaxies should be similar within clusters. Roeder [96] considers velocity data for \( n = 82 \) galaxies. She models this data as a finite mixture of Normal distributions, with the number and location of mixture components unknown. The assumption is that each galactic cluster is a single component of the normal mixture. Multiple mixture components is consistent with the hypothesis of galaxy clustering.

We apply the methodology outlined above to estimate the mixing distribution itself, which immediately gives an estimate of the mixture complexity. Other authors, including Escobar and West [38] and Richardson and Green [89], have fit Bayesian hierarchical models that require a fairly complex Monte Carlo sampling scheme for posterior inference on the number of mixture components. To illustrate our approach, we will consider a simple Normal mixture model in which each component has variance \( \sigma^2 = 1 \). This choice is based on the \( a \) priori considerations of Escobar and West [38]: their common prior for the variance of each Normal component has unit mean.

From the observed velocities, it is apparent that the mixture components must be centered somewhere in the interval \( \Theta = [5, 40] \), so we choose a grid of candidate support points \( \Theta = \{5.0, 5.5, 6.0, \ldots, 39.5, 40.0\} \); here \( G = 71 \). Figure 5.4 shows the estimates of the mixing and mixture densities based on the RE+ algorithm using a SA optimization procedure. The RE+ estimate of the mixing distribution clearly identifies six mixture components, closely matching the conclusions in [38, 89, 96] and the RE+ mixture density provides a very good fit to the observed velocities.

In the next example, we fit a Poisson mixture to a set of count data used in Example 1.2 of Böhning [15] and in Wang [114].

Example 5.5 The data, given in Table 1 of Wang [114], come from a cohort study in north-east Thailand where the number of illness spells for \( n = 602 \) pre-school children is monitored between June 1982 and September 1985. The number of illness spells ranges from 0 to 24; a histogram of the data is given in the right-hand panel of Fig-
Böhning [15] points out that an ordinary Poisson model is inadequate for this data due to overdispersion. A Poisson mixture, therefore, seems more appropriate.

Upon investigation of the observed data, it is apparent that the support of the finite mixture must be within $\Theta = [0, 20]$. Therefore, we take our candidate support $\Theta$ to be a grid of $G = 100$ equi-spaced points from 0 to 20. The RE+ algorithm is run and the resulting mixing distribution is displayed in the left-hand panel of Figure 5.5. There we see four clearly identified support points. The overlay on the right-hand panel is the corresponding mixing density (the dots being the tips of the probability mass function). Here we see that the RE+ solution provides a very good fit to the observed counts. Moreover, our mixing density estimate closely matches the nonparametric MLE found by Wang [114], both in support values and in weights.

While this version of the RE+ algorithm seems to work well, there are drawbacks to the use of simulated annealing, or any other stochastic optimization procedure for that matter. For one, stochastic optimization algorithms are notoriously slow to converge, so the computations in Examples 5.4–5.5 are relatively expensive. The second drawback comes from the fact that there are control parameters—proposal
function, cooling schedule, etc—that determine how fast the process explores the solution space. Choosing these control parameters is an iterative process that takes time. However, with well-chosen control parameters, the performance of the algorithm can be significantly improved.

5.5 Appendix: Simulated annealing details

There are two important features of a simulated annealing algorithm that have a dramatic effect on the performance: the cooling schedule and the proposal distribution. The proposal distribution is familiar to users of MCMC, but the cooling schedule is unique to simulated annealing-type procedures. Unfortunately, there are no satisfactory “default choices” for these important implementation parameters that work in general. Some guidelines are available, however; see, e.g., Givens and Hoeting [48]. In this appendix, we present the our choices of cooling schedule and proposal distribution for the class of problems presented in Subsection 5.4.2.
The cooling schedule is just a decreasing sequence of temperatures \( \{ T_t : t \geq 0 \} \).

For good performance, \( T_t \) must approach 0 as \( t \to \infty \), but not too fast. Here we choose \( T_t = a / \log(1 + t) \), \( t \geq 0 \), for a suitably chosen value of \( a \) (found by trial-and-error).

Next we describe the proposal distribution \( \pi(t) \) that tells us how to generate a candidate point \( \xi_{new} \) at the \( t^{th} \) stage of the algorithm. In our examples, a draw \( \xi_{new} \) from \( \pi(t) \) is a binary \( G \)-vector differing from \( \xi(t) \) exactly one position. In other words, one of the \( G \) components of \( \xi(t) \) is chosen and then flipped from 0 to 1 or from 1 to 0. The choice of component is not made uniformly, however. To encourage a “sparse” solution, we want \( \pi(t) \) to assign greater mass to those components \( \xi_g(t) \) in \( \xi(t) \) such that \( \xi_g(t) = 1, \ g = 1, \ldots, G \). The particular choice of weights is

\[
\omega_g(t) \propto 1 + \left[ G / \sum_{g=1}^G \xi_g(t) \right] \xi_g(t), \quad g = 1, \ldots, G.
\]

Here we see that when most of the components of \( \xi(t) \) are 1, the sampling is near uniform, whereas, when \( \xi(t) \) is “sparse,” those components with value 1 have a greater chance of being selected. The majority of the early stages of the algorithm will eliminate candidate support points, causing the algorithm to favor smaller supports. At later stages, support points can be added in the move \( \xi(t) \rightarrow \xi(t+1) \), but this would typically require an increase in the pseudo-likelihood function.

Finally, we present the version of simulated annealing used in Subsection 5.4.2. Let \( \xi(0) \) be a specified starting point—here we take \( \xi(0) \) to be the full vector of 1s, corresponding to the support of Newton’s original estimate \( f_n \). Choose a large fixed stopping time \( t_0 \) and set \( \xi_{opt} = \xi(0) \). Then we generate the stochastic sequence \( \{ \xi(t) : 0 \leq t \leq t_0 \} \) as follows.

1. At iteration \( t + 1 \), simulate \( \xi_{new} \) from a probability distribution \( \pi(t) \) on \( \Xi \) depending on the current iterate \( \xi(t) \).

2. Define the acceptance probability

\[
\rho = \exp \{ (L_n(\xi_{new}) - L_n(\xi(t))) / T_t \} \wedge 1,
\]
where $L_n$ is the pseudo-likelihood defined in (5.4), and set

$$
\xi^{(t+1)} = \begin{cases} 
\xi_{\text{new}} & \text{with probability } \rho \\
\xi^{(t)} & \text{with probability } 1 - \rho
\end{cases}
$$

If $L_n(\xi^{(t+1)}) > L_n(\xi_{\text{opt}})$, set $\xi_{\text{opt}} = \xi^{(t+1)}$.

3. If $t = t_0$, return $\xi_{\text{opt}}$; otherwise, set $t \leftarrow t + 1$ and return to Step 1.

There are certainly more advanced stochastic optimization procedures that could potentially be more efficient; this is the focus of current research. However, our simulated annealing algorithm performs quite well in the two examples presented in Subsection 5.4.2. Figure 5.6 shows a sample path of the sequence $\{L_n(\xi^{(t)}) : t \geq 0\}$ from Example 5.4 and we see good mixing as well as the desirable property that some downhill moves are accepted.

Figure 5.6. Plot of the simulated annealing sample path for the optimization in Example 5.4.
6. A NONPARAMETRIC EMPIRICAL BAYES FRAMEWORK
FOR HIGH-DIMENSIONAL INFERENCE

In Chapter 5, we saw how the RE+ algorithm could be used in a variety of interesting statistical problems. In the present chapter we take a closer look at an important high-dimensional inference problem, namely, large-scale simultaneous hypothesis testing. Such large-scale inference problems arise in many applied fields such as genomics, proteomics, astrophysics, education sciences, and medical imaging, to name a few. Here we propose a unified framework for nonparametric empirical Bayes (NPEB) inference based on the RE+ algorithm.

6.1 High-dimensional problems and sparsity

Let $\theta = (\theta_1, \ldots, \theta_n)$, for $n$ large, be the parameter vector about which inference is to be made. Data $X = (X_1, \ldots, X_n)$ are observed, with $X_i$ related to $\theta_i$. Here, with a slight abuse of the usual notation, $n$ will represent the dimension of the problem and “large $n$” means high-dimensional. Compared to the traditional low-dimensional problems, such as inference about a common unknown mean based on iid data, these high-dimensional problems have some peculiar properties which stem from the fact that the $n$ experiments are only similar and not identical as is common in low-dimensional problems. One of the most common high-dimensional problem of this form is Stein’s example—see Example 1.4—where the $X_i$’s are independent $N(\theta_i, 1)$ random variables. There it is well known in that the natural estimate $\hat{\theta} = X$, which corresponds to the least-squares and maximum likelihood estimate, is inadmissible under sum-of-squared-error loss. Therefore, we cannot simply rely on classical statistical theory in such problems so something else is needed.
These high-dimensional problems have been a major focus of research efforts in recent years and many new procedures have emerged. The fundamental idea, which is common to both frequentist and Bayesian approaches, is that inference can be improved by sharing information between cases, despite the fact that a separate decision is required for each case. It is well known that both the hierarchical and empirical Bayes frameworks have this "information sharing" property built in; see Section 1.4. However, many frequentists methods, such as the false discovery rate (FDR) controlling procedure of Benjamini and Hochberg [7], also share this property.

High-dimensional problems typically have another interesting feature: it is often assumed that $\theta$ is "sparse" in the sense that many of the $\theta_i$'s are equal to zero (or some other fixed value). For example, in DNA microarray analysis, $\theta_i$ is the expression level of the $i$th gene and biological assumptions dictate that the majority of the genes have no effect on the characteristic under investigation. Nonparametric regression via an orthogonal basis expansion, such as wavelets, is another similar problem. When the degree of sparseness is known, asymptotically optimal procedures exist; see, for example, Donoho and Johnstone [24]. Rarely, however, is the degree of sparseness known so, recently, emphasis has shifted to estimates that adapt to unknown sparsity. For example, Abramovich, Benjamini, Donoho and Johnstone [1] propose a thresholding procedure based on FDR considerations and prove asymptotic optimality uniformly over various degrees of sparseness. Empirical Bayes methods are quite popular in such problems, particularly in the wavelet thresholding literature; see, for example, Johnstone and Silverman [53, 54] and the reference therein.

Our primary interest is in Bayesian methods where sparseness assumptions are incorporated by considering a mixture prior like that in (4.2), which corresponds to an important special case of the so-called two-groups model championed by Efron [31] and others. Specifically, we consider prior distributions $F$ of the form

$$F(d\theta) = \pi \delta_{\{\vartheta\}}(d\theta) + (1 - \pi)\varphi(\theta)\,d\theta,$$

where $\pi \in [0, 1]$ denotes the prior probability that a $\theta$ equals $\vartheta$, $\delta_{\{x\}}$ denotes a degenerate distribution at $x$, and $\varphi$, a continuous density with respect to Lebesgue measure
on $\Theta \subset \mathbb{R}$, represents the prior distribution when $\theta \neq \vartheta$. Note that $F$ is absolutely continuous, with density $f$, with respect to the dominating measure $\mu = \delta(\vartheta) + \lambda_{\text{Leb}}$, a point mass at $\vartheta$ plus Lebesgue measure on $\mathbb{R}$.

A fully Bayes analysis would proceed by introducing a second-stage prior on the parameters $(\vartheta, \sigma, \pi, \varphi)$ in the two-groups model (6.1). However, as demonstrated by Clyde and George [16], Scott and Berger [98] and others, elicitation of this second-stage prior is a formidable task. Moreover, the choice of second-stage prior can have a dramatic effect on the final conclusions. Therefore, recent attention has been focused on an empirical Bayes framework which, like the fully Bayes approach, allows for information sharing between cases and, in addition, provides an automatic, data-based choice of second-stage hyperparameters. Various parametric empirical Bayes analyses under the two-groups model (6.1) have been proposed; see, for example, Clyde and George [16, 17] and Johnstone and Silverman [53, 54] and the references therein. Bogdan, Ghosh and Tokdar [14] give a thorough comparison of various testing procedures, including some PEB tests. But, as indicated in Section 1.4, specification of the parametric form of the prior (in this case, for $\varphi$) can itself be a difficult task, and the final conclusions would likely depend on the particular choice. To avoid such dependencies, a robust nonparametric analysis that determines the entire prior distribution automatically would be potentially useful. Analyses of this form are few in number, the most notable being the very clever approach of Efron [29–31] which we describe in Section 6.2.2. Here we propose an alternative NPEB framework, based on the RE+ analysis described in Chapter 5, that estimates the parameters $(\vartheta, \sigma, \pi, \varphi)$ from the observed data. Three important and distinguishing features of our framework are as follows:

1. Our model is both flexible and mathematically rigorous;

2. The parameters are identifiable (see Theorem 6.1); and
3. In some sense, our approach can be viewed as an empirical Bayes counterpart to the hierarchical Bayes DP mixture model—recall the *one-step correspondence* between RE and DPMs, Proposition 1.2.

We will elaborate on each of these points in the sections to follow.

There are two (roughly equivalent) inference goals in these sparse high-dimensional problems, namely estimation/thresholding and testing; Abramovich, Grinshtein and Pensky [2] refer to this general problem as *testimation*. In this chapter will focus on simultaneous testing, but the estimation/thresholding problem can be handled in a completely analogous way.

Thanks to recent technological advances, there has been dramatic increase in the amount of data that scientists are able to collect, store and analyze. This analysis often requires testing hundreds or even thousands of hypotheses simultaneously. An abstract representation of the problem is testing a large set of hypotheses

\[ H_{0i} : \text{the } i^{th} \text{ case manifests a "null" behavior, } \quad i = 1, \ldots, n \]  

based on summary test statistics \( Z_1, \ldots, Z_n \), which will be called \( z \)-scores. In the applications mentioned above, \( n \) can range from a few hundred to several thousand; for example, in the HIV study described in Efron [30] and also in Example 6.3 below, there are \( n = 7,680 \) genes under investigation. Clearly, the scale of modern “large-scale” testing problems dwarfs that considered in classical multiple testing problems described in Miller [76] or Lehmann and Romano [60]. Therefore, classical techniques, such as the Bonferroni correction, applied in today’s large-scale problems lead to testing procedures which are too conservative. The idea of controlling FDR has lead to many new developments in this area. A review of this vast literature would be impractical, plus our main focus here is on Bayesian methods, so we refer the reader to Chapter 34 of DasGupta [20]. It is worth mentioning, however, that a lot of current research is devoted to finding connections between frequentists and Bayesian rules in the sparse high-dimensional inference problem; see for example Abramovich, et al. [1], Bogdan, Ghosh and Tokdar [14] and Bogdan, Chakrabarti and Ghosh [13].
6.2 Two-groups model and simultaneous testing

The general two-groups model assumes that the z-scores $Z_1, \ldots, Z_n$ arise from a mixture density

$$m(z) = \pi m^{(0)}(z) + (1 - \pi)m^{(1)}(z), \quad (6.3)$$

with $m^{(0)}$ and $m^{(1)}$, respectively, describing the null and alternative behavior of the z-scores $Z_1, \ldots, Z_n$. In most applications the model is “sparse” in the sense that the prior probability $\pi$ of a case being null is high, say $\pi \approx 0.9$, reflecting the fact that most of the $n$ cases are null or, as Efron [29] says, “uninteresting.” Due to the very construction of the z-scores, typically $m^{(0)}$ is taken to be a Normal density with possibly unspecified location/scale, but there is no standard choice of $m^{(1)}$ in the literature. In what follows we describe several popular choices for modeling $m^{(1)}$ and also introduce our new NPEB setup.

6.2.1 Parametric and nonparametric Bayes

Scott and Berger [98] propose a very clean parametric Bayes analysis based on the following hierarchical model:

$$Z_i \mid (\theta_i, \pi, \sigma, \tau) \sim N(\theta_i, \sigma^2)$$

$$\theta_i \mid (\pi, \sigma, \tau) \sim \pi \delta_{\theta_i} + (1 - \pi)N(0, \tau^2)$$

$$(\pi, \sigma, \tau) \sim \Pi(\pi) \times \Pi(\sigma, \tau) \quad (6.4)$$

where $\Pi(\pi)$ and $\Pi(\sigma, \tau)$ are prior distributions determined by both subjective and objective considerations. This results in a two-groups model (6.3) with $m^{(1)}$ being a $N(0, \sigma^2 + \tau^2)$ density. Their fully Bayes approach proceeds by computing

$$p_i := P\{\text{case } i \text{ is null } \mid Z_1, \ldots, Z_n\}, \quad i = 1, \ldots, n,$$

(using importance sampling or MCMC) and rejecting $H_{0i}$ if $p_i$ is less than some threshold $r$. Choosing $r = 0.5$ corresponds to 0-1 loss, but $r$ can vary for different loss functions—see Section 6.4.
It is important to point out that in these high-dimensional problems, the effect of the priors are not washed away as the dimension $n$ increases. Therefore, careful elicitation of hyperparameters for the highest stage of the hierarchy is important and, as Scott and Berger [98] point out, the final results greatly depend on the choices made. To reduce/eliminate this dependence, one might consider instead an empirical Bayes approach which automatically chooses hyperparameters that result in a good fit to the observed data. In the high-dimensional parametric problem, empirical Bayes methods are generally acceptable to Bayesians because the answers often match those of a full Bayes analysis. However, in certain problems, some might still feel too confined by the parametric restriction, so a nonparametric model might be more suitable.

Bogdan, Ghosh and Tokdar [14] propose to extend model (6.4) by removing the parametric restrictions on $m^{(1)}$. They do so by adding an additional nonparametric layer to the hierarchy (6.4) using a Dirichlet process (DP) prior. Specifically,

$$\begin{aligned}
Z_i \mid (\theta_i, F, \alpha, \pi_0, \sigma, \tau) &\sim N(\theta_i, \sigma^2) \\
\theta_i \mid (F, \alpha, \pi_0, \sigma, \tau) &\sim F \\
F \mid (\alpha, \pi_0, \sigma, \tau) &\sim \mathcal{D}(\alpha, \pi_0 \delta\{0\} + (1 - \pi_0)N(0, \tau^2)) \\
(\alpha, \pi_0, \sigma, \tau) &\sim \Pi(\alpha) \times \Pi(\pi_0) \times \Pi(\sigma, \tau)
\end{aligned}$$

(6.5)

where $\mathcal{D}(\alpha, F_0)$ is the DP distribution with base measure $F_0$ and precision constant $\alpha > 0$. With this choice $F_0 = \pi_0 \delta\{0\} + (1 - \pi_0)N(0, \tau^2)$ of base measure, a draw $F \sim \mathcal{D}(\alpha, \Phi_0)$ will, with probability 1, put positive mass on 0. The DP prior in (6.5) results in a model for $m^{(1)}$ which is, in some sense, centered around $N(0, \sigma^2 + \tau^2)$. In principle, the Bayesian decision analysis proceeds just as in the parametric case above: compute $p_i$ using MCMC and reject $H_{0i}$ if $p_i \leq r$. The additional flexibility is clear but computation becomes much more challenging, particularly when $n$ is large.

Currently, there is no widely accepted empirical Bayes version of the nonparametric Bayes model (6.5): the proposals of Robbins are “too frequentist”—see Remark 6.2—and standard procedures for estimating $F$ nonparametrically, such as NPMLEs or DPMs, are typically unsatisfactory. In the next section we describe
a relatively new nonparametric empirical Bayes proposal due to Efron that seems to bypass most of these difficulties.

### 6.2.2 Efron’s nonparametric empirical Bayes

Recently, Efron has written extensively on an empirical Bayes framework for the multiple testing problem that requires minimal mathematical modeling compared to the fully Bayes approach described in Section 6.2.1; see Efron [31] for a review of this work. His jumping off point is the realization that the frequentist false discovery rate (FDR) of Benjamini and Hochberg [7] has a nice Bayesian interpretation the two-groups model. Following Efron [29], define the “local false discovery rate”

\[
\text{fdr}(z) = \frac{\pi m^0(z)}{m(z)}. \tag{6.6}
\]

From Bayes’ theorem, fdr(z) is roughly the posterior probability that a case is null given its z-score is z. Therefore, the use of fdr(z) to classify cases as null or non-null based on the observed z-scores is quite natural. Efron’s key observation is that fdr does not depend on the alternative; therefore, \(m^{(1)}\) may be left more or less unspecified in the analysis and focus can be devoted entirely to the null component \(\pi m^{(0)}\).

In “classical” problems, by design, the null distribution \(m^{(0)}\) of the z-scores is supposed to be \(N(0, 1)\), the so-called theoretical null. But, as Efron [31] argues, \(m^{(0)}\) often appears quite different from the theoretical null in practice. A number of factors can contribute to this phenomenon, inter-case correlation being one of them. This necessitates estimating \(m^{(0)}\) from the observed z-scores—the estimate of \(m^{(0)}\) is called the empirical null.

Estimating both \(m^{(0)}\) and \(m^{(1)}\) (and maybe \(\pi\)) from the data, however, is fraught with many dangers. The most severe of these is a lack of identifiability. To counter this, strong assumptions on the components of \(m\) are required. For example, the zero-assumption in Efron [31] states that

most of the z-scores near zero come from null cases. \tag{6.7}
In terms of the mixture components, (6.7) implies that the supports of \( m^{(0)} \) and \( m^{(1)} \) are, in some sense, nearly disjoint—see Figure 6.1.

With the segregation of cases determined by (6.7), Efron [29, 31] then estimates \( \pi m^{(0)} \) with a scaled Normal distribution fit to the central peak of the z-score empirical distribution. Unfortunately, despite the strong conditions on \( m^{(0)} \) and \( m^{(1)} \) in (6.7), there can be major difficulties. For example, in some problems, like one presented in Section 6.5, this procedure can produce absurd estimates, such as \( \hat{\pi} > 1 \).

Certain aspects of Efron’s approach are nice. However, in light of the strong assumptions made and the possibility of very strange conclusions, it is apparent that improvements can be made. In the next section we describe our new nonparametric empirical Bayes approach that combines some of Efron’s ideas with the mathematically rigorous DP mixture model (6.5).

### 6.2.3 A new nonparametric empirical Bayes framework

Our version of the two-groups model is very similar to the DP mixture model (6.5) presented in Section 6.2.1. We model \( m^{(1)} \) as a location mixture of Normals with an unspecified mixing density \( \varphi \) which, incidentally, is equivalent to the model

\[
m(z) = \int_{\Theta} p(z|\theta, \sigma) F(d\theta)
\]

for the mixture \( m(z) \), where \( F \) is given by (6.1). A justification for our model and the resulting analysis can be given based solely on this similarity to the DP mixture model. But there is something to be said about Efron’s approach of starting with a fundamental assumption about the problem at hand and building an analysis directly from this assumption. Here we propose a similar motivation for our model based on a weaker version of the zero-assumption.

Consider a different scenario where the null/non-null segregation between the observables in (6.7) is deemed unlikely to occur, but the basic essence of Efron’s
zero-assumption prevails. Instead of focusing directly on the $z$-scores, we make a zero assumption about $m^{(0)}$ and $m^{(1)}$ directly as follows:

$$m^{(1)} \text{ has strictly heavier tails than } m^{(0)}.$$  \hfill (6.9)

This is weaker than Efron’s zero-assumption in the sense that the $z$-scores near zero are only more likely to have come from $m^{(0)}$ than from $m^{(1)}$.

A simple model that encodes our zero assumption (6.9) into the two-groups model (6.3) is the following:

$$m(z) = \pi p(z|\vartheta, \sigma) + (1 - \pi) \int_{\Theta} p(z|\theta, \sigma) \varphi(\theta) d\theta,$$  \hfill (6.10)

which is nothing but a particular realization of the DP mixture-version of the two-groups model described in Section 6.2.1 except the location $\vartheta$ is also unknown. While other models satisfying (6.9) could be considered, the simple model (6.10) is both sufficiently flexible and, as we see below, mathematically tractable. Also notice that since $\vartheta$ and $\sigma$ are to estimated, $\hat{m}^{(0)}(z) = p(z|\hat{\vartheta}, \hat{\sigma})$ is nothing but a version of Efron’s empirical null [29].

The following theorem answers the important practical question of identifiability of the parameters $(\vartheta, \sigma, \pi, \varphi)$ in model (6.10).

**Theorem 6.1** Let $\mathcal{F}$ denote the space of probability densities with respect to Lebesgue measure on $\Theta$; i.e., $\mathcal{F} = \{ \varphi \in L_1(\Theta) : \varphi \geq 0, \int \varphi = 1 \}$. Then the map $M : \mathbb{R} \times \mathbb{R}^+ \times (0, 1) \times \mathcal{F} \to \mathcal{F}$, given by

$$M(\vartheta, \sigma, \pi, \varphi)(z) = \pi p(z|\vartheta, \sigma) + (1 - \pi) \int_{\Theta} p(z|\theta, \sigma) \varphi(\theta) d\theta$$  \hfill (6.11)

is one-to-one.

**Proof** Assume $M(\vartheta_1, \sigma_1, \pi_1, \varphi_1) = M(\vartheta_2, \sigma_2, \pi_2, \varphi_2)$. Then, in terms of characteristic functions, we must have

$$e^{-\sigma_1^2 t^2/2} \left[ \pi_1 e^{it\vartheta_1} + (1 - \pi_1) \psi_1(t) \right] = e^{-\sigma_2^2 t^2/2} \left[ \pi_2 e^{it\vartheta_2} + (1 - \pi_2) \psi_2(t) \right]$$  \hfill (6.12)
for every \( t \in \mathbb{R} \), where \( \psi_1 \) and \( \psi_2 \) are the characteristic functions of \( \varphi_1 \) and \( \varphi_2 \), respectively. Since \( \varphi_1, \varphi_2 \in \mathcal{F} \), we know that, for \( k = 1, 2 \),

\[
\psi_k(t) \to 0 \quad \text{as} \quad t \to \pm \infty.
\]

(6.13)

Now, suppose \( \sigma_1 > \sigma_2 \). Choose a sequence \( \{t_s\} \subset \mathbb{R} \) such that \( t_s \to \infty \) and \( e^{it_s \vartheta_2} = 1 \) for all \( s \). Then, for large enough \( s \), (6.13) would imply that \( \pi_2 + (1 - \pi_2)\psi_2(t_s) \neq 0 \). On rearranging the terms in (6.12) we get

\[
e^{\frac{s}{2}(\sigma_1^2 - \sigma_2^2)} = \frac{\pi_1 e^{it_s \vartheta_1} + (1 - \pi_1)\psi_1(t_s)}{\pi_2 + (1 - \pi_2)\psi_2(t_s)}.
\]

(6.14)

As \( s \to \infty \), the left-hand side of (6.14) blows up to infinity while the right-hand side is bounded. Therefore, to avoid contradiction, we need \( \sigma_1 \leq \sigma_2 \). By symmetry, it follows that \( \sigma_1 = \sigma_2 \). With this equality, relation (6.12) easily leads to the equalities \( \vartheta_1 = \vartheta_2 \), \( \pi_1 = \pi_2 \) and \( \varphi_1 = \varphi_2 \).

Since our version (6.10) of the two-groups model is identifiable, an empirical Bayes analysis would be sensible. But there remains the difficult question of how to estimate the model parameters \((\vartheta, \sigma, \pi, \varphi)\). It turns out that the RE+ algorithm described in Chapter 5 is perfectly suited for this task.

### 6.3 Parameter estimation with RE+

To implement the NPEB procedure outlined in the previous section, we first need to estimate the model parameters \((\vartheta, \sigma, \pi, \varphi)\). It turns out the the RE+ algorithm described in Chapter 5 is ideally suited for this purpose. In the present section we will describe the implementation details for this two-groups problem.

In implementing the PR+ in the two-groups problem (6.10), the choice \( \xi = (\theta_0, \sigma) \) is a natural choice for the non-mixing parameter to be used in the pseudo-likelihood (5.4). However, one needs to take care in specifying the initial estimate \( F_0 = \pi_0 \delta_{\{\theta_1\}} + (1 - \pi_0)\varphi_0 \) to be used in the recursive algorithm. In spite of having large \( n \), the initial guess \( \pi_0 \) can have a substantial effect on the final estimate of \( F \) when one of the two
groups is scarce—which is the case in most modern applications. This motivates us to include \( \pi_0 \) as one more (tuning) parameter and carry out the RE+ maximization over the augmented parameter \( \xi = (\vartheta, \sigma, \pi_0) \).

It was remarked in Chapter 5 that, in high-dimensional problems, maximizing the pseudo-likelihood \( L_n(\xi) \) in (5.4) could be time-consuming. In specific problems, however, one might be able to improve the computations by developing problem-specific algorithms or deriving gradient and/or hessian formulae analytically. In Section 6.8 a recursive formula for the gradient of \( L_n(\vartheta, \sigma, \pi_0) \) is provided which can be combined with a standard optimization routine to significantly accelerate computation.

From a Bayesian perspective, the regularized version of the RE+ algorithm, described in Chapter 5, might be preferred since it can easily incorporate available prior information about the non-mixing parameters. In the absence of subjective prior information on \( \xi = (\vartheta, \sigma, \pi_0) \), we recommend the following default choice of prior:

\[
\Pi(\vartheta, \sigma, \pi_0) = \text{Normal}(0, 0.33^2) \times \text{InvChi}(4, 1) \times \text{Beta}(22.7, 1). \tag{6.15}
\]

Efron [30] restricts the null observables to the range \([-1, 1]\) and the choice of standard deviation 0.33 puts almost all the prior mass for \( \vartheta \) on that range. The inverse Chi prior for \( \sigma \) is equivalent to a Gamma prior for \( 1/\sigma^2 \) centered at 1 with shape parameter 2—the small shape parameter keeps prior mass for \( \sigma \) away from 0. Finally, the 22.7 in the Beta prior for \( \pi_0 \) sets the prior median equal to 0.9, reflecting the belief that the null proportion \( \pi \) should be large. Prior (6.15) is used in Example 6.4.

**Remark 6.2** This new NPEB approach, based on model (6.10) and the RE+, should appeal to Bayesians because the motivation comes from minimizing a KL divergence, a popular tool in Bayesian analysis; see, e.g., Ghosh, Delampady and Samanta [44, Sec. 3.8]. In the large-\( n \) problems being considered here, one might argue that other more established procedures could also be used; for example, NPML estimates or perhaps a Bayes empirical Bayes analysis using DP mixtures, akin to Deely and Lindley [22]. But in the two-groups testing problem, where distinguishing between \( m^{(0)} \) and \( m^{(1)} \) is critical, identifiability is an issue that must be handled very care-
fully. Here, it turns out that ML and DP estimates cannot satisfactorily identify the continuous mixing density $\varphi$, due to their a.s. discreteness (see Section 1.2). Therefore, the NPEB approach based on the model (6.10) and the RE+ algorithm has the advantage of simultaneously being extremely flexible and accurate, computationally efficient, and justifiable from a Bayesian point of view.

### 6.4 Empirical Bayes decision rule

The hypothesis testing problem can be approached from a decision theoretic point of view. This requires specification of a loss function and, in the testing context, the most common choice is the following:

$$\begin{align*}
\kappa_1 &= \text{loss for Type I error} \quad \text{and} \quad \kappa_2 = \text{loss for Type II error}.
\end{align*}$$

(6.16)

For testing the “point-null” in (6.2), it can be shown that the Bayes Oracle rule is

\[
\text{Reject } H_0 i \text{ if } \text{fdr}(Z_i) \leq r, \quad i = 1, \ldots, n
\]

where fdr is defined in (6.6) and $r$ depends on the losses $\kappa_1$ and $\kappa_2$.

Naturally, our NPEB approach is to mimic the Bayes Oracle rule. That is, the RE+ estimates of $\pi$, $m^{(0)}$ and $m$ are inserted into the definition of fdr, giving

\[
\hat{\text{fdr}}(z) = \hat{\pi} \hat{m}^{(0)}(z)/\hat{m}(z)
\]

and the test is performed based on the “estimated” decision rule

\[
\text{Reject } H_0 i \text{ if } \hat{\text{fdr}}(Z_i) \leq r, \quad i = 1, \ldots, n.
\]

But how do we choose the threshold $r$? As shown in Berger [8, pp. 163–164], the Oracle test uses the threshold $r = \kappa_2/(\kappa_1 + \kappa_2)$, which is the relative cost of a Type II error. Efron makes a case for choosing $r = 0.2$, which corresponds to $\kappa_1 = 4\kappa_2$; that is, a Type I error is four times as costly as a Type II error. This makes sense in the high-dimensional problem because the goal is to identify interesting cases; in such problems, concluding that too many cases are interesting is bound to waste researchers’ resources.
6.5 Microarray data examples

In this section we give two real-data examples demonstrating the performance of our proposed NPEB testing procedure based on the RE+ algorithm. The two data sets we use are taken from the microarray analysis literature but, certainly, our methods can be applied in other contexts. For a recent review of multiple testing in the microarray context, see Dudoit, Shaffer and Boldrick [27].

Example 6.3 Gene expressions for four HIV+ males are compared to the same in four normal males in van’t Wout, et al. [113]. The histogram in Figure 6.1 shows the z-scores for 7680 genes under investigation. The z-score $Z_i$ was calculated by suitably transforming a two-sample $t$-statistic that compares the expression levels for gene $i$ in the HIV+ patients against the normal subjects. The genes which had similar expression levels in the two groups were likely to produce $z$-scores close to zero, while the differentially expressed ones were likely the produce $z$-scores away from zero. The goal is to identify genes which are differentially expressed.

On applying the RE+ procedure described above, with $\xi = (\vartheta, \sigma, \pi_0)$, to the HIV data set, we estimated the empirical null $\hat{m}^{(0)}$ to be a $\mathcal{N}(-0.11, 0.74^2)$ density. The optimization was carried out numerically using the optim routine in R [87], each evaluation of $L_n(\xi)$ was made based on a RE derived from a fixed set of 25 permutations of the data. The optimum $\hat{\xi} = (-0.11, 0.74, 0.57)$ was then used to estimate $\hat{\pi} = 0.86$ and $\hat{\varphi}$ through a longer run of RE based on 100 permutations. Estimates of $\pi m^{(0)}$, $(1 - \pi)m^{(1)}$ and $m$ are shown in the right panel of Figure 6.1. Efron’s analysis was also carried out for this problem using the locfdr package in R; see the left panel of Figure 6.1. We clearly see here that the supports of Efron’s estimated $m^{(0)}$ and $m^{(1)}$ are, in some sense, nearly disjoint. Our RE+ estimate of the empirical null density closely matches the one reported in Efron [31], namely $\mathcal{N}(-0.11, 0.75^2)$. But our estimate $\hat{\pi} = 0.86$ is substantially lower than Efron’s $\tilde{\pi} = 0.93$. This is due to the difference in the zero-assumptions (6.7) and (6.9) underlying the two methods. Our method allows a small fraction of non-null $z$-scores to be
close to zero, while Efron rules out this possibility at the outset. Consequently, the estimated FDR values for the central $z$-scores are quite different for the two methods; see Figure 6.2. Note, however, the strikingly similar treatment of the non-central $z$-scores. In fact, with cut-off $\text{FDR}(z) \leq 0.2$, our method identifies 173 differentially expressed genes, closely matching the 160 genes identified by Efron.

Efron’s empirical Bayes analysis of the HIV data in Example 6.3 appears to be right on the mark. Therefore, it is comforting that our new framework gives similar results. In our next example, however, Efron’s analysis seems to break down since the number of non-null cases is relatively large, making the zero-assumption (6.7) somewhat questionable. On the other hand, thanks to the additional flexibility built into our version of the zero-assumption (6.9), the analysis based on our approach is much more believable.

Figure 6.1. The HIV data set described in Example 6.3. Plots show estimates of $\pi m^{(0)}$ (thin lines), $(1-\pi)m^{(1)}$ (dashed lines) and $m$ (heavy lines) based on Efron and RE+. Cases falling between the two triangles are null.
Figure 6.2. Plots of the estimated local fdr based on the two empirical Bayes methods for the HIV data set described in Example 6.3. Tick marks on the z-axis show the $n = 7,680$ observed z-scores.

**Example 6.4** Golub, et al. [49] report on a DNA microarray experiment to compare genetic profiles of patients with acute lymphoblastic leukemia (AML) against those with acute myeloid leukemia (AML). Originally, there were 6,817 genes under investigation but only $n = 3,051$ genes have been analyzed. This reduction is due to some pre-processing—see Dudoit, Fridlyand and Speed [26] for details—that eliminates those genes with little fluctuation in expression levels across the cases. Thus, a number of the null genes are removed at the outset, causing the proportion of null genes to fall dangerously low. In light of this pre-processing, our version (6.9) of the zero-assumption seems much more believable than Efron’s.

The reduced data set, available in the multtest package in R [87], comes in the form of a $n \times p$ matrix that contains expression level measurements for $n = 3,051$ genes on $p = 38$ patients—27 ALL patients and 11 AML patients. The z-scores $Z_1, \ldots, Z_n$
are obtained by suitably transforming the two-sample \( t \)-statistics comparing ALL and AML expression levels. Golub, et al. [49, p. 533] point out that a fairly large proportion of the genes under investigation are highly correlated with the ALL-AML class distinction; therefore, we might expect our estimates of \( \pi \) to be relatively small.

Efron’s estimation procedure outlined in Section 6.2.2 produced an absurd estimate of the null proportion (\( \hat{\pi} > 1.0 \)) so we will not consider this analysis further. Our method, on the other hand, produced a good fit to the histogram—see the left panel of Figure 6.3—but estimated the null proportion to be quite small, i.e., \( \hat{\pi} = 0.33 \). Over a thousand genes were classified as significant based on the 0.2 fdr threshold. This is an unsatisfactory reduction of the number of interesting genes, so we take our analysis one step further, using the regularized RE+ algorithm described in Section 6.3 with the prior (6.15). The fit of the regularized PR+ is shown in the right panel of Figure 6.3. Like in the RE+ analysis, the estimated mixture fits the histogram very well. The key difference is the relative weight of the null. Indeed, the regularized RE+ estimates the null proportion as \( \tilde{\pi} = 0.55 \), which is still small but favorable to \( \hat{\pi} = 0.33 \). From these estimates, about 760 genes were classified as significant.

### 6.6 Simulation results

In the previous section, RE+ was shown to perform well in two microarray data analyses. Here we investigate its performance in some relatively large scale simulations where we can actually check if it is doing the right thing.

As our model for \( z \)-scores \( Z_1, \ldots, Z_n \), we take the null component \( m^{(0)} \) to be a \( N(0, \sigma^2) \) density where \( \sigma^2 = 1 \) and consider two choices for the alternative \( m^{(1)} \):

I. \( m^{(1)} \) is a \( N(0, \sigma^2 + \tau^2) \) density, where \( \tau^2 = 2 \log(n) \);

II. \( m^{(1)} \) is a skew-Normal density—see Chapter 1—with variance \( 1.3^2(\sigma^2 + \tau^2) \) and skewness parameter \( \alpha = 3 \).
Model I is nothing but the model considered in Bogdan, Ghosh and Tokdar [14], and Model II is a skewed version of that, whose peak is off-center and its right-tail is thicker than the Normal. Figure 6.4 shows plots of these two alternative densities.

In each case, we want to test the sequence of $n$ hypotheses (6.2). Here we take $n = 1000$, so our tests are of relatively large scale compared to those considered in Bogdan, Ghosh and Tokdar [14]. For $N = 100$ replications, and at each of a range of $\pi$ values, the two-groups mixture model (6.10) is fit using the RE+ (under the same configurations as in Section 6.5) and various summary data is collected. It turns out that the R function `locfdr` is prone to return errors when the proportion of nulls $\pi$ is relatively low and/or the alternative is non-Normal. Therefore, we compare RE+ to the local fdr testing procedure outlined in Jin and Cai [52]. Their estimates are based Fourier methods and are shown to perform well, both theoretically and in practice, compared to Efron’s estimation method. Some limited calculations were done with the `locfdr` procedure and the results were similar to those of JC.
Table 6.1 shows the estimated means and standard deviations of the RE+ and Jin-Cai (JC) estimates of $\pi$ for various values of $\pi$. It is clear that, compared to the JC estimates which are always large, RE+ is much more adaptive to the sparsity level in the sense that when the true $\pi$ is small, the RE+ estimate is also small. This adaptivity is further illustrated by the monotonicity in the RE+ estimates of $\pi$ compared to the JC estimates that reach their minimum at $\pi = 0.7$ for each model. The corresponding estimates of $\sigma$ and $\vartheta$ are quite accurate for both RE+ and JC, so are not shown.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\pi$</th>
<th>RE+ Mean</th>
<th>Std. Dev.</th>
<th>JC Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.500</td>
<td>0.68</td>
<td>0.140</td>
<td>0.87</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>0.600</td>
<td>0.74</td>
<td>0.032</td>
<td>0.81</td>
<td>0.033</td>
</tr>
<tr>
<td></td>
<td>0.700</td>
<td>0.78</td>
<td>0.027</td>
<td>0.80</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>0.800</td>
<td>0.83</td>
<td>0.029</td>
<td>0.84</td>
<td>0.041</td>
</tr>
<tr>
<td></td>
<td>0.900</td>
<td>0.88</td>
<td>0.019</td>
<td>0.92</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td>0.950</td>
<td>0.90</td>
<td>0.018</td>
<td>0.96</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>0.990</td>
<td>0.95</td>
<td>0.031</td>
<td>0.99</td>
<td>0.008</td>
</tr>
<tr>
<td>II</td>
<td>0.500</td>
<td>0.65</td>
<td>0.065</td>
<td>0.84</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>0.600</td>
<td>0.70</td>
<td>0.075</td>
<td>0.75</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>0.700</td>
<td>0.76</td>
<td>0.023</td>
<td>0.72</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>0.800</td>
<td>0.82</td>
<td>0.023</td>
<td>0.82</td>
<td>0.044</td>
</tr>
<tr>
<td></td>
<td>0.900</td>
<td>0.88</td>
<td>0.020</td>
<td>0.91</td>
<td>0.028</td>
</tr>
<tr>
<td></td>
<td>0.950</td>
<td>0.91</td>
<td>0.018</td>
<td>0.95</td>
<td>0.023</td>
</tr>
<tr>
<td></td>
<td>0.990</td>
<td>0.93</td>
<td>0.026</td>
<td>0.99</td>
<td>0.008</td>
</tr>
</tbody>
</table>
Figure 6.4 shows the RE+ estimates of the alternative density \( m^{(1)} \) under the two models over the \( N = 100 \) replications at the \( \pi = 0.9 \) level. The estimates have an unexpected bump near \( z = 0 \), but we see the symmetry in the Model I estimates and skewness in the Model II estimates. The \( L_1 \) error is also shown and we see a bit more variability in the Model II estimates compared to those under Model I. It is important to keep in mind that, despite the fact that we have a sample of \( n = 1000 \) \( z \)-scores to work with, technically only 100 of those come from the estimand \( m^{(1)} \). Therefore, more variability is to be expected.

Finally, we compare the RE+ and JC testing procedures, not on the quality of their estimates, but in terms of their performance in the testing problem itself. We consider three performance criteria: power, mis-classification probability (MP) and false discovery rate (FDR). Figure 6.5 shows the values of these performance measures for each of the two models as a function of the true null proportion \( \pi \). The Bayes Oracle test is also performed, using the fdr-based decision rule with the true parameters, and the results are displayed. The three tests perform comparably as the sparseness increases (as \( \pi \to 1 \)). From a Bayesian point of view, the mis-classification probability—Bayes risk—is most important; here we see that as \( \pi \to 1 \), the risks of
both RE+ and JC converge to that of the Oracle. This same result was observed for
the Bonferroni, Benjamini-Hochberg, and various PEB procedures in Bogdan, Ghosh
and Tokdar [14]. Bogdan, Chakrabarti and Ghosh [13] have since rigorously proved
similar optimality results, so it is our conjecture that some sort of optimality can be
shown for RE+ (and JC for that matter). Some simple results (not presented here)
along these lines are promising, but more work is needed.

Figure 6.5. Plots of the power, MP and FDR of the RE+ (black, solid), JC (black, dashed) and Bayes Oracle (gray) testing procedures
in the two models as functions of the true null proportion $\pi$. 
6.7 Additional remarks

Here we have presented a nonparametric empirical Bayes framework for the large-scale simultaneous testing problem. The formulation of our approach is, on the surface at least, obviously similar to the nonparametric Bayes approach using DP mixtures in Section 6.2.1—the similarity comes from the “one-step correspondence,” the relationship between the RE and DP mixture updates. We also pointed out that, like the now widely accepted PEB, our NPEB solution based on the RE+ algorithm and minimizing a KL number might also be acceptable to Bayesians. Everything else aside, the strongest justification for PEB is that, when \( n \) is large, the PEB solutions typically match those of a full hierarchical Bayes analysis (Berger [8, p. 195]). Such evidence suggesting our NPEB answers would be similar to those of a full nonparametric Bayes analysis for large \( n \) has yet to be found—this is difficult both theoretically and numerically. However, we suspect that asymptotic optimality results (as \( n \to \infty \) and \( \pi = \pi_n \to 1 \)) along the lines of Bogdan, Chakrabarti and Ghosh [13] can be established here as well.

An important issue in high-dimensional problems—particularly in large-scale simultaneous testing—is dependence among the data \( Z_1, \ldots, Z_n \). Our theoretical results for RE/RE+ do not cover the case of dependent data but the fact that our NPEB approach works well on real data which are not independent (Efron [31] points out that genes in the HIV data in Example 6.3 are highly correlated) suggests that independence may not be crucial.

Finally, despite the fact that our discussion of NPEB was limited to the testing problem, the framework can be applied in an obvious way to other important high-dimensional problems, such as thresholding and variable-selection in orthogonal regression. Moreover, if one is willing to accept that the RE+ provides a general framework for approximating a full nonparametric Bayes analysis based on DP mixtures, then the kernel-based RE+ density estimation procedure shown in Section 5.4.1 can be thought of as a NPEB density estimation procedure.
6.8 Appendix: Recursive formula for pseudo-likelihood gradient

To speed up computation of the RE+ implementation in the simultaneous testing problem, it is helpful to have a formula for evaluating the gradient of \( \log m_{i-1}(X_i) \). It turns out that this can be done recursively, and including this gradient formula in the optimization routine accelerates the computations dramatically.

Recall that \( \xi = (\vartheta, \sigma, \pi_0) \). Start by defining an unconstrained version of \( \xi \), namely, \( \zeta = (\vartheta, \log \sigma, \text{logit} \, \pi_0) \), where \( \text{logit} \, x = \log \left( \frac{x}{1-x} \right) \). In what follows, \( \nabla \) will denote a gradient with respect to \( \zeta \) and if \( g \) is a function of a variable \( \theta \), then define \( \nabla g(\theta) \) to be the gradient with respect to \( \zeta \), pointwise in \( \theta \).

Using the notation in the earlier sections, the following algorithm shows how to compute \( u_i := m_{i-1}(X_i) \) and \( \nabla \log u_i \) for \( i = 1, \ldots, n \).

1. Start with user specified \( \pi_0 \) and \( \varphi_0 \), and set
   \[
   \nabla \pi_0 = (0, 0, \pi_0(1 - \pi_0)) \quad \text{and} \quad \nabla \varphi_0(\theta) \equiv (0, 0, 0).
   \]

2. For \( i = 1, \ldots, n \), repeat the following three steps:

   (a) Define \( z_0 = (X_i - \vartheta)/\sigma \) and \( z_1(\theta) = (X_i - \theta)/\sigma \). Then construct
   \[
   L_0 = p(X_i|\vartheta, \sigma) \\
   \nabla L_0 = (z_0 L_0/\sigma, (z_0^2 - 1) L_0, 0) \\
   L_1(\theta) = p(X_i|\theta, \sigma) \\
   \nabla L_1(\theta) = (0, (z_1^2(\theta) - 1) L_1(\theta), 0)
   \]
   where \( p(x|\theta, \sigma) \) is a \( N(\theta, \sigma^2) \) density.
(b) Compute
\[
\begin{align*}
    h_i &= \int L_1(\theta) \varphi_{i-1}(\theta) \, d\theta \\
    u_i &= \pi_{i-1} L_0 + (1 - \pi_{i-1}) h_i
\end{align*}
\]
\[
\nabla \log h_i = \frac{1}{h_i} \int \{ L_1(\theta) \nabla \varphi_{i-1}(\theta) + \nabla L_1(\theta) \varphi_{i-1}(\theta) \} \, d\theta
\]
\[
\nabla \log u_i = \frac{\nabla \pi_{i-1} L_0 + \pi_{i-1} \nabla L_0 + h_i \{ (1 - \pi_{i-1}) \nabla \log h_i - \nabla \pi_{i-1} \}}{u_i}
\]

(c) Update
\[
\begin{align*}
    \pi_i &= A_0 \pi_{i-1} \\
    \nabla \pi_i &= A_0 \nabla \pi_{i-1} + \nabla A_0 \pi_{i-1} \\
    \varphi_i(\theta) &= BA_1(\theta) \varphi_{i-1}(\theta) \\
    \nabla \varphi_i(\theta) &= \{ \nabla B A_1(\theta) + B \nabla A_1(\theta) \} \varphi_{i-1}(\theta) + BA_1(\theta) \nabla \varphi_{i-1}(\theta)
\end{align*}
\]
where
\[
\begin{align*}
    A_0 &= 1 + w_i (L_0/u_i - 1) \\
    A_1(\theta) &= 1 + w_i (L_1(\theta)/u_i - 1) \\
    B &= (1 - \pi_{i-1})/(1 - A_0 \pi_{i-1})
\end{align*}
\]
and
\[
\begin{align*}
    \nabla A_0 &= w_i \{ \nabla L_0 - L_0 \nabla \log u_i \} / u_i \\
    \nabla A_1(\theta) &= w_i \{ \nabla L_1(\theta) - L_1(\theta) \nabla \log u_i \} / u_i \\
    \nabla B &= (BA_0 - 1) \nabla \pi_{i-1} + w_i B \pi_{i-1} \{ \nabla L_0 - L_0 \nabla \log u_i \} / u_i / (1 - A_0 \pi_{i-1})
\end{align*}
\]

3. Return \( \sum_{i=1}^{n} \nabla \log u_i \).
7. FUTURE WORK

Here we conclude by discussing a few important unanswered questions and some directions for future work along these lines.

In Chapter 3 we presented some very general theorems for convergence of the recursive estimates, but a fundamental assumption underlying all these results is that the observed data sequence is iid. In light of our interest in high-dimensional problems, simultaneous testing in particular, dependence among the observations is a very important issue, and two practically relevant questions arise. First, if the original algorithm is applied to a dependent sequence of data, will the estimates converge and, if so, to what? The fact that the NPEB methods in Chapter 6 work well for data known to be dependent (see Example 6.3) is promising. Our jumping off point in our analysis of RE in the independent case was stochastic approximation, and it turns out that there are modifications and corresponding convergence theorems for stochastic approximation with correlated errors. These results could potentially shed light on how one should handle RE in the dependent case. Second, if the data has a known dependence structure, how can the original algorithm be modified to incorporate this additional information? Important applications include time series analysis and empirical Bayes variable selection in regression.

Numerical evidence like that presented in Section 3.2.3 suggest that the actual convergence rate for RE might be significantly faster than the theoretical bound. A natural question is whether this theoretical bound can be sharpened. This is a challenging problem but we are optimistic. An important follow-up question is how does the convergence rate for the RE mixture distribution estimate impact that of the corresponding mixing distribution. Although RE is difficult analytically, we believe that the actual estimates are nicer than the NPMLE, say, in practice, so a good rate for the RE mixing distribution estimate would have a significant impact.
From the preceding chapters it should be clear that stochastic approximation is a powerful tool for studying recursive algorithms such as RE. But, as mentioned in Chapter 2, the theory for infinite-dimensional stochastic approximation is still yet to be fully developed. Of course, should such theory become available, it would most likely be useful in studying the performance of RE. Alternatively, perhaps the convergence analysis here on the special case of RE could be extended to more general stochastic approximation procedures.

The RE+ algorithm in Chapter 5 is extremely flexible and can potentially be used in all sorts of problems. With a versatile methodology in hand, the natural next step is to look for applications. For example, preliminary work has shown that a version of the RE+ that incorporates data censoring performs provides a smooth alternative to the usual step-function estimates of classical nonparametric survival analysis.

But despite its strong performance in applications, the convergence theory for RE+ remains incomplete. The complexity of the pseudo-likelihood function seems to cause serious problems when trying to apply standard theorems on convergence of stochastic processes. Perhaps a new theory for processes with a pointwise martingale property would be needed to prove the desired convergence. If such a theory were available, it would be applicable in various other Bayesian filtering problems.

Lastly, there is a more philosophical question left unanswered, related to the Bayesian interpretation of RE. It is clearly not a Bayesian posterior quantity, but its similarity to the posterior mean under the DP mixture model suggests that it may be approximately Bayes. However, simulations presented in Chapter 4 would seem to indicate that RE is in fact not an approximation of the DP mixtures. The question, therefore, is if RE is not an approximation to the posterior mean of a DP mixture, is there some other prior such that RE approximates the corresponding posterior mean? This is a difficult question, but a satisfactory answer would help further justify the use of RE (and RE+) in Bayesian applications.
LIST OF REFERENCES
LIST OF REFERENCES


VITA
Ryan Glenn Martin was born on December 12, 1979 in Elkhart, Indiana. He first studied at Franklin College, receiving a B.A. in mathematics in 2003, and he later earned an M.S. in Statistics in 2005 from Indiana University-Purdue University Indianapolis. RM entered the Purdue Statistics Department Ph.D. program in the Fall of 2005 and began work with Professors Jayanta K. Ghosh and Chuanhai Liu early in the Spring of 2007.