Divide and recombine for large complex data: The subset likelihood modeling approach to recombination

Philip Gautier
Purdue University

Follow this and additional works at: http://docs.lib.purdue.edu/open_access_dissertations
Part of the Statistics and Probability Commons

Recommended Citation
http://docs.lib.purdue.edu/open_access_dissertations/458
This is to certify that the thesis/dissertation prepared

By Philip Gautier

Entitled
Divide and Recombine for Large Complex Data: The Subset Likelihood Modeling Approach to Recombination

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

William S. Cleveland
Chair

Chuanhai Liu

Bowie Xi

Lingsong Zhang

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University’s “Policy of Integrity in Research” and the use of copyright material.

Approved by Major Professor(s): William S. Cleveland

Approved by: Jun Xie 04/22/2015

Head of the Departmental Graduate Program Date
DIVIDE AND RECOMBINE FOR LARGE COMPLEX DATA:
THE SUBSET LIKELIHOOD MODELING APPROACH TO RECOMBINATION

A Dissertation
Submitted to the Faculty
of
Purdue University
by
Philip Gautier

In Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

May 2015
Purdue University
West Lafayette, Indiana
To Anne
ACKNOWLEDGMENTS

I’m very grateful to have had the opportunity to work with Bill Cleveland. The approach to data analysis he’s taught me is eminently practical, principled, and effective.

I am grateful to Chuanhai Liu for his suggestions on how to proceed with this work. I thank my committee members Bowei Xi and Lingsong Zhang.

I am grateful for the guidance and assistance of many others in the department, particularly Rebecca Doerge for her mentorship and leadership, and Doug Crabill for his excellent technical knowledge and generosity in sharing it. I’m grateful for the work of Diane Martin, Becca Miller, Marian Cannova, Jeffrey Klimes, Shaun Ponder, Stephanie Foster, Marshay Jolly, Mary Roe, and Holly Graef for running things so smoothly.

I’m grateful to Betty Malloy at American University for giving me a running start on my academic journey.

And I’m grateful to my friends and family for their support.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>vi</td>
</tr>
<tr>
<td>1 DIVIDE AND RECOMBINE</td>
<td>1</td>
</tr>
<tr>
<td>2 SUBSET LIKELIHOOD MODELING</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>4</td>
</tr>
<tr>
<td>2.2 Subset Likelihood Models</td>
<td>7</td>
</tr>
<tr>
<td>2.2.1 The All-Data Model</td>
<td>8</td>
</tr>
<tr>
<td>2.2.2 Quadratic Model</td>
<td>8</td>
</tr>
<tr>
<td>2.2.3 Cubic Model</td>
<td>9</td>
</tr>
<tr>
<td>2.2.4 Sparse Cubic Model</td>
<td>10</td>
</tr>
<tr>
<td>2.2.5 Skew-Normal Model</td>
<td>10</td>
</tr>
<tr>
<td>2.3 Model Fitting</td>
<td>11</td>
</tr>
<tr>
<td>2.3.1 Local Information of the Observed Subset Likelihood</td>
<td>12</td>
</tr>
<tr>
<td>2.3.2 Local Information of the Subset Likelihood Model</td>
<td>12</td>
</tr>
<tr>
<td>2.3.3 Solutions for the Subset Likelihood Model Parameters</td>
<td>14</td>
</tr>
<tr>
<td>2.4 Recombination</td>
<td>17</td>
</tr>
<tr>
<td>2.5 Subset Likelihood Model Diagnostics</td>
<td>19</td>
</tr>
<tr>
<td>3 WORKING WITH THE RECOMBINED LIKELIHOOD MODEL</td>
<td>25</td>
</tr>
<tr>
<td>3.1 Point Estimation</td>
<td>25</td>
</tr>
<tr>
<td>3.2 Interval Estimation</td>
<td>27</td>
</tr>
<tr>
<td>3.2.1 Credible Intervals</td>
<td>27</td>
</tr>
<tr>
<td>3.2.2 Normal Approximation</td>
<td>30</td>
</tr>
<tr>
<td>3.2.3 Diagnostics for Interval Estimation</td>
<td>32</td>
</tr>
<tr>
<td>4 SIMULATION RESULTS</td>
<td>36</td>
</tr>
<tr>
<td>4.1 Normality</td>
<td>37</td>
</tr>
<tr>
<td>4.2 Variance</td>
<td>44</td>
</tr>
<tr>
<td>4.3 Covariance</td>
<td>48</td>
</tr>
<tr>
<td>4.4 Bias</td>
<td>51</td>
</tr>
<tr>
<td>5 DISCUSSION</td>
<td>56</td>
</tr>
<tr>
<td>5.1 Comparison of Subset Likelihood Models</td>
<td>56</td>
</tr>
<tr>
<td>5.2 Related Work</td>
<td>59</td>
</tr>
<tr>
<td>5.2.1 Distributed Optimization</td>
<td>60</td>
</tr>
<tr>
<td>5.2.2 Streaming Optimization</td>
<td>61</td>
</tr>
<tr>
<td>5.2.3 Out-of-Core R Packages for Generalized Linear Models</td>
<td>62</td>
</tr>
</tbody>
</table>
ABSTRACT


Divide and recombine (D&R) is a statistical framework for the analysis of large complex data. The data are divided into subsets. Numeric and visualization methods, which collectively are analytic methods, are applied to each subset. For each analytic method, the outputs of the application of the method to the subsets are recombined. So each analytic method has associated with it a division method and a recombination method. Here we study D&R methods for likelihood-based model fitting. We introduce a notion of likelihood analysis and modeling. We divide the data and fit a likelihood model on each subset. The fitted model is characterized by a set of parameters much smaller than the subset data size, but retains as much information as possible about the true subset likelihood. Analysis of subset likelihoods and their fitted models consists of visualizations on an appropriate scale and region. These visualizations allow the analyst to verify the choice and fit of the model. The fitted models are recombined across subsets to form a model of the all-data likelihood, which we maximize to obtain a likelihood modeling estimate (LME). We present simulation results demonstrating the performance of our method compared with the all-data maximum likelihood estimate (MLE) for the case of logistic regression.
1. DIVIDE AND RECOMBINE

Divide and recombine (D&R) is a statistical framework for the analysis of large complex data [1]. The data are divided into subsets. Numeric and visualization methods, which collectively are analytic methods, are applied to each subset. For each analytic method, the outputs of the application of the method to the subsets are recombined. So each analytic method has associated with it a division method and a recombination method.

D&R computation for the application of an analytic method is embarrassingly parallel: the subset computations are independent and do not communicate with one another. D&R can exploit a distributed database and parallel compute engine like Hadoop [2]. Furthermore, an interactive language for data analysis such as R can be merged with Hadoop so that the data analyst or methodologist can program very efficiently in the analysis of large complex data, and not have to deal with the intricacies of database management and parallel computation. Such a merger has been accomplished by RHIPE [3], the R and Hadoop Programming Environment. RHIPE allows an R [4] user to apply D&R to large complex data wholly from within R. Almost any R function or package to be used in the analysis.

The final result of a D&R method is generally not the same as the result that would have been obtained, had it been feasible to apply the method directly to all of the data. Developing statistical theory and methods for D&R consists of developing division methods and recombination methods that do as well as possible given that the data must be divided as a matter of computational effectiveness. The direct all-data computation in many cases is not possible, and in other cases can take too long to make it practical. The accuracy of the D&R result depends on how division and recombination are done. D&R statistics research consists of finding “optimal” D&R methods that reduce statistical efficiency as little as possible.
While our focus will be finite sample results, asymptotics can be informative as well. Suppose there are \( n \) observations and \( r \) subsets each with \( m \) observations, so \( n = mr \). We let \( n \) get large. But the D&R approach is to develop theory and methods that work within the bounds of what is computationally possible. That means \( m \) needs to be limited, so we fix it, and let \( r \) increase.

D&R exploits the MapReduce computational system implemented in Hadoop. However nothing in this system brings a focus to how a data analyst divides the data, which translates to “chunks the data” in MapReduce language, and then recombines the output to provide best results.

The analysis of a large complex dataset often has several divisions that serve different analysis objectives that require different data structures. However, a division typically serves as the division method for many analytic methods. This is important because division, even when exploiting parallel computing, can be computationally expensive, but the up front cost is amortized across many analysis methods. D&R is a framework for an entire analysis: initial data exploration, model building, model checking, fitting models, applying procedures that serve goals like classification, and carrying out statistical inferences. A division needs to serve many analytic methods, both numeric and visual. This pushes the development of division methods in the direction of having good general performance across all analysis tasks. D&R makes analysis of large complex data a fundamentally statistical domain of basic research, but one that can readily exploit parallel computational environments.

Here we study D&R methods for likelihoods. We introduce a notion of likelihood analysis and modeling. We divide the data and fit a likelihood model on each subset. The fitted model is characterized by a set of parameters much smaller than the subset data size, but retains as much information as possible about the true subset likelihood. Analysis of subset likelihoods and their fitted models consists of visualizations on an appropriate scale and region. These visualizations allow the analyst to verify the choice and fit of the model. The fitted models are recombined across subsets to form
a model of the all-data likelihood, which we maximize to obtain a likelihood modeling estimate (LME).
2. SUBSET LIKELIHOOD MODELING

2.1 Introduction

Subset likelihood modeling is a method for working with statistical likelihoods within the divide and recombine (D&R) framework. When data are distributed into subsets, it is impractical to evaluate the exact likelihood, since doing so requires reading all of the data. Reading all of the data from all subsets is computationally intensive and slow. Subset likelihood modeling treats the subset likelihoods as data themselves. The subset likelihoods are statistics in the sense that they are functions of the data. We carry out modeling on these statistics - the subset likelihoods - much the same as data analysts typically carry out modeling of data.

The analyst applies a division method to the data, then finds a maximum likelihood estimate (MLE) on each subset. Using this subset MLE, along with the local information, the analyst fits a likelihood model on every subset. The fitted subset likelihood model replaces the observed subset likelihood with a much simpler function. Recombination consists of multiplying the fitted subset likelihood models (or on the log scale, summing them) to obtain a fitted model for the all-data likelihood.

The goal of subset likelihood modeling is much the same as the goal of data modeling. It is to summarize or describe the likelihood with a parsimonious model. Since the subset likelihood is, in general, a nontrivial function of all of the data in a given subset, it cannot be expressed without reading all of the data. The fitted subset likelihood model, however, can be expressed by reading only a small number of likelihood model parameters. The parameters of the subset likelihood model are estimated from the data. The subset likelihood model is a function of the parameter of interest, \( \theta \). So it is defined on the same space as the observed subset likelihood. The parameters of the subset likelihood model are chosen so that the fitted model is close
in shape to the observed subset likelihood. Though many methods could, in principle, be used to estimate the parameters of the subset likelihood model, we use derivative matching, akin to a Taylor expansion. This means that the mode of the fitted subset likelihood model occurs at the same point in $\theta$-space as the mode of the observed subset likelihood. If $\theta$ is a $p$-dimensional vector, then the $p$-dimensional gradient of the observed subset likelihood at the mode is by definition zero. This is also true for the fitted subset likelihood model. Further, the Hessian, which is the symmetric $p \times p$ matrix of second derivatives, also matches at the mode. Depending on the likelihood model, higher-order derivatives may also match. This method preserves the local information of the observed subset likelihood in the fitted subset likelihood model.

Many statistical modeling methods require the evaluation of a likelihood, or of its log and derivatives. For example, the calculation of an MLE often requires a numerical optimization of the log likelihood. Numerical optimization is a computationally intensive task. The gradient and the Hessian of the log likelihood must be evaluated many times before an iterative numerical optimization procedure such as Newton’s Method \cite{6} arrives at the MLE. Newton’s Method is a common choice for finding an MLE, and is the method implemented (under the equivalent formulation of iteratively re-weighted least squares) in the widely used R package \texttt{glm.fit} \cite{4}. Bayesian Markov Chain Monte Carlo (MCMC) methods also require the iterative evaluation of the likelihood \cite{7}. It may be possible to write the likelihood as a function only of some sufficient statistic, but in general, and often in practice, the likelihood, its log, and its first two derivatives, are functions of all the data, or at least all of the observations of the independent variables.

The benefit of summarizing the observed subset likelihood with a likelihood model is computational efficiency. This efficiency has three parts: central processing unit (CPU) time, size in memory, and cross-network transfer time. The CPU time is the time required for arithmetic operations involved in evaluating the likelihood. In order to perform these operations, the data needed to evaluate the likelihood must first be present in random access memory (RAM) in most computer architectures.
The size in memory is the storage size of the data in RAM. Data are typically stored as 8-byte floating point numbers, for example R’s `double` type. The cross-network transfer time is time needed to transfer data between nodes in a compute cluster. In D&R computation, the cross-network transfer time is often the bottleneck. This means that reducing cross-network transfer time has the greatest effect on reducing the overall time required to perform computations required for the analysis of large complex data.

Replacing the observed likelihood with a fitted likelihood model reduces the demand on all three of these resources. Because evaluation of the fitted likelihood model requires only the parameters of the fitted likelihood model, rather than all the data, far less CPU time is required. Only the parameters of the fitted likelihood model must be loaded into RAM, rather than the much larger data. And only the parameters of the fitted likelihood model must be transferred across the network.

The compelling reason to use D&R is that the analyst is faced with data that are too large to fit in RAM on a single machine. Likelihood modeling solves this problem, making D&R analysis of the likelihood possible. A fitted subset likelihood model is entirely described by its parameters. These are few enough to be transferred across the network in a reasonable amount of time. Recombination reduces the size of the fitted model even further, so that the fitted all-data likelihood model, which is the recombination of the fitted subset likelihood models, can fit in RAM on a single node. Evaluating the all-data likelihood model is then possible in cases where evaluation of the observed all-data likelihood would not have been. Then, the data analyst can proceed with a full analysis of the likelihood. The analyst might, for example, maximize the all-data likelihood model to find a Likelihood Modeling estimate (LME) as a replacement for the MLE which was previously unavailable due to computational constraints.

Likelihood modeling brings to bear all the tools of data modeling including visualization and the iterative process of model conjecture, fitting, and diagnostics. Visualizing the subset likelihoods is instrumental to the choice of likelihood model.
Visualization also serves as a diagnostic tool to check the appropriateness of a fitted likelihood model. In cases where the likelihood is asymptotically normal, we have a right to conjecture a normal model for the subset log likelihood. But we know that the fit will be imperfect in finite sample sizes. Therefore, even when the data size is large, we will not accept a normal model without verification.

This work is limited to discussion of logistic regression. However, the applications of subset likelihood modeling are by no means limited to logistic regression. The D&R concept of subset likelihood modeling applies to any analysis of large complex data which makes use of a likelihood function which is approximately log-concave. Such analyses encompass at least all exponential family linear models. The subset likelihood models presented below are not specific to logistic regression. They are motivated by the approximate normality of the likelihood, which is present in a very broad range of models used for data analysis. We focus on logistic regression for clarity, and because of its familiarity and widespread use. Since it is possible to obtain the exact all-data MLE for linear regression within the D&R framework, logistic regression is, in a sense, the simplest linear model which presents a serious problem. That makes it an excellent opportunity for expanding D&R methodology and providing new tools to those who wish to analyze large complex data.

2.2 Subset Likelihood Models

The purpose of subset likelihood modeling is to reduce the number of parameters necessary to describe the likelihood function. Computing the all-data likelihood requires all of the observations. In contrast, the model for the likelihood involves only a small number of parameters. Likelihood modelling is therefore a form of data reduction.

Any model for the subset likelihoods which achieves these goals can be used in the D&R paradigm. The analyst should carefully validate the choice of model using visual and numeric diagnostic tools, as discussed later. We present four possible
models here. Three are based on a Taylor expansion. The fourth is chosen to reflect what we observe in empirical study of the subset likelihoods.

A constant term may be added to the model and calculated for the purpose of visually comparing the subset likelihood to its fitted model. The constant will have no effect on the calculation of a maximum likelihood estimator. Nor will it have any effect on likelihood inference, since it does not change the local information. Distributional inference will also be unaffected, since normalizing the likelihood or any posterior derived from the likelihood will erase the estimated constant.

2.2.1 The All-Data Model

Recall that the data consist of \( n \) observations divided into \( r \) subsets, each with \( m \) observations. Each observation consists of \( x \in \mathbb{R}^p \) and \( y \in \{0, 1\} \). The model is indexed by \( \theta \in \mathbb{R}^p \). The subscript \( (s) \) indicates the \( s^{th} \) subset and the subscript \( (s)i \) indicates the \( i^{th} \) observation on the \( s^{th} \) subset. The data model is logistic regression with a binary response. The observations are independent, so the all-data log likelihood is the sum of the subset log likelihoods.

\[
\ell(\theta) = \sum_{s=1}^{r} \ell_{(s)}(\theta)
\]

\[
\ell_{(s)}(\theta) = \log \prod_{i=1}^{m} \left( \frac{e^{\theta^\top x_{(s)i}}}{1 + e^{\theta^\top x_{(s)i}}} \right)^{y_{(s)i}} \left( 1 - \frac{e^{\theta^\top x_{(s)i}}}{1 + e^{\theta^\top x_{(s)i}}} \right)^{1-y_{(s)i}}
\]

2.2.2 Quadratic Model

The quadratic subset log likelihood model is

\[
\omega_{(s)}^Q(\theta) = \sum_{t=1}^{p} \sum_{u=t}^{p} \alpha_{(s)tu} \theta_t \theta_u + \sum_{t=1}^{p} \beta_{(s)t} \theta_t
\]

with \( \alpha_{(s)tu} \) and \( \beta_{(s)t} \) being the real-valued parameters calculated on subset \( s \). The subscripts \( t \) and \( u \) are integers such that \( 1 \leq t \leq u \leq p \). The number of parameters estimated on each subset is \( \frac{p^2 + 3p}{2} \).
If fitted as described in section 2.3, the quadratic model is a second-order Taylor expansion of the subset log likelihood around the subset MLE.

There are two independent arguments for the appropriateness of the quadratic model. The first is Bayesian in nature. Since the likelihood is asymptotically normal by the Bernstein-von Mises theorem [8], then its log is approximately quadratic for sufficiently large sample sizes. So if the subset data size is reasonably large, the quadratic model will be a good fit to the observed log likelihood.

The second relies on the sampling distribution of the MLE, which is also asymptotically normal [9]. Its mean is the true value of $\theta$ and its variance is the inverse Fisher Information. Since the Fisher Information is unavailable in practice, a good estimate is the local information, which is sometimes called the Observed Fisher Information [10]. This local information consists of the Hessian of the log likelihood evaluated at the MLE. Consider the following scenario. We divide data into subsets and carry out logistic regression on each subset by calculating the MLE and local information for each subset. We now have $r$ draws from $r$ normal distributions with common mean $\theta$ and known variances. The best estimator of $\theta$ is then the covariance-weighted mean of the $r$ subset MLEs, which is identical to the LME obtained by recombining the quadratic subset likelihood models and maximizing.

### 2.2.3 Cubic Model

The cubic subset log likelihood model is

$$\omega^C_{(s)}(\theta) = \sum_{t=1}^{p} \sum_{u=t}^{p} \sum_{v=u}^{p} \alpha_{(s)tu} \theta_t \theta_u \theta_v + \sum_{t=1}^{p} \sum_{u=t}^{p} \beta_{(s)tu} \theta_t \theta_u + \sum_{t=1}^{p} \gamma_{(s)t} \theta_t$$

with $\alpha_{(s)tu}$, $\beta_{(s)tu}$, and $\gamma_{(s)t}$ being the real-valued parameters calculated on subset $s$. The subscripts $t$, $u$, and $v$ are integers such that $1 \leq t \leq u \leq v \leq p$. The number of parameters estimated on each subset is $\frac{p^3 + 6p^2 + 11p}{6}$.

If fitted as described in section 2.3, the cubic model is a third-order Taylor expansion around the MLE. In practical applications, the asymptotic results motivating the quadratic model may not hold. In particular, the likelihood for logistic regression
models will be skewed in finite samples. The cubic model accommodates this skew by allowing for an asymmetric fitted subset likelihood model. The $p \times p \times p$ array of third derivatives captures the local information on skewness, or departure from normality, in the observed subset likelihood and replicates this local information in the fitted subset likelihood model.

### 2.2.4 Sparse Cubic Model

The sparse cubic subset log likelihood model is

$$\omega_{SC}^s(\theta) = \sum_{t=1}^{p} \alpha_{(s)t}\theta_t^3 + \sum_{t=1}^{p} \sum_{u=t}^{p} \beta_{(s)tu}\theta_t\theta_u + \sum_{t=1}^{p} \gamma_{(s)t}\theta_t$$

with $\alpha_{(s)t}$, $\beta_{(s)tu}$, and $\gamma_{(s)t}$ being the real-valued parameters calculated on subset $s$. The subscripts $t$ and $u$ are integers such that $1 \leq t \leq u \leq p$. The number of parameters estimated on each subset is $\frac{p^2+5p}{2}$.

The sparse cubic model is a special case of the cubic model. It uses only the $p$ diagonal elements of the third derivative. The result is a far more parsimonious, and therefore less computationally demanding, subset likelihood model which still allows for skewness.

### 2.2.5 Skew-Normal Model

Azzalini and Dalla Valle [11] present the following multivariate skew-normal distribution:

$$f(\theta) = 2\phi(\theta)\Phi(\alpha^T\theta)$$

where $\phi$ is a multivariate normal density and $\Phi$ is a univariate normal cumulative density. The $p$-dimensional random vector $\theta$ follows the multivariate skew-normal distribution, and the marginal distribution of each element of $\theta$ has a univariate skew-normal distribution. The $p$-dimensional parameter $\alpha$ governs the level of skewness.
We refer to this density in three parts: the normalizing constant $2$, the normal kernel \( \phi(\theta) \) and the skewness factor \( \Phi(\alpha^T \theta) \).

The skew-normal subset log likelihood model is

$$
\omega_{SN}^{(s)}(\theta) = -\frac{1}{2} \left( (\theta - \mu^{(s)})^T \Sigma_{(s)}^{-1}(\theta - \mu^{(s)}) \right) + \log \Phi \left( \alpha^T_{(s)}(\theta - \eta^{(s)}) \right)
$$

with \( \mu^{(s)}, \alpha^{(s)}, \text{ and } \eta^{(s)} \) being \( p \)-dimensional parameter vectors and \( \Sigma_{(s)} \) being a symmetric \( p \times p \) matrix parameter calculated on subset \( s \). The number of real-valued parameters estimated on each subset is therefore \( \frac{p^2 + 7p}{2} \).

This is a variation on Azzalini and Dalla Valle’s formulation. The skewness factor has been centered at \( \eta \), the mode of the distribution, rather than \( \mu \). This change will allow us later to estimate the parameters of the skew-normal model in closed form, rather than with an iterative numerical procedure, greatly reducing the computational cost. As is the case with Azzalini and Dalla Valle’s original skew-normal, \( \mu \) and \( \Sigma \) are not equal to the first two moments of the distribution. Those two moments, as well as the mode, depend on \( \alpha \) and \( \eta \) as well as \( \mu \) and \( \Sigma \).

Like the cubic and sparse cubic models, the skew-normal model has an exact normal distribution as a special case (when \( \alpha = 0 \)) but allows for skewness in general.

### 2.3 Model Fitting

The quadratic and cubic likelihood models are Taylor expansions of the observed subset log likelihood at its maximum. The sparse cubic and skew-normal models are fit in much the same way. We calculate the subset MLE and local information from the observed subset likelihood and choose parameter values which result in a subset likelihood model with the same maximum and local information.
2.3.1 Local Information of the Observed Subset Likelihood

Model fitting is now described in detail. First find the usual MLE on each subset: \( \hat{\theta}_{(s)} = \arg \max_{\theta} \ell_{(s)}(\theta) \). Then, evaluate the first three derivatives of the log likelihood at the MLE:

\[
\left. \frac{\partial \ell_{(s)}}{\partial \theta_a} \right|_{\theta = \hat{\theta}_{(s)}} = 0
\]

\[
\left. \frac{\partial^2 \ell_{(s)}}{\partial \theta_a \partial \theta_b} \right|_{\theta = \hat{\theta}_{(s)}} = -\sum_{i=1}^{m} x_a x_b e^{x_i^T \hat{\theta}_{(s)}} \left( 1 + e^{x_i^T \hat{\theta}_{(s)}} \right)^2
\]

\[
\left. \frac{\partial^3 \ell_{(s)}}{\partial \theta_a \partial \theta_b \partial \theta_c} \right|_{\theta = \hat{\theta}_{(s)}} = -\sum_{i=1}^{m} x_a x_b x_c e^{x_i^T \hat{\theta}_{(s)}} \left( 1 - e^{x_i^T \hat{\theta}_{(s)}} \right) \left( 1 + e^{x_i^T \hat{\theta}_{(s)}} \right)^3
\]

2.3.2 Local Information of the Subset Likelihood Model

Next, find the corresponding derivatives of the model. To simplify notation, consider the parameters to be symmetric in the sense that \( \alpha_{(s)abc} = \alpha_{(s)a^*b^*c^*} \) where \( (a^*, b^*, c^*) \) is any permutation of \( (a, b, c) \). The notation \( 1 \{ x \} \) is the indicator function which takes the value 1 when \( x \) is true and 0 otherwise.

For the quadratic model, we need the \( p \)-dimensional gradient and the \( p \times p \) matrix of second partial derivatives, the Hessian. For the cubic model, we need these plus the \( p^3 \) third partial derivatives. For the sparse cubic and skew-normal models, we need the first and second partial derivatives, plus the \( p \) diagonal elements of the third partial derivative.

**Quadratic**

\[
\left. \frac{\partial}{\partial \theta_a} \omega_{(s)}^Q(\theta) \right|_{\theta = \hat{\theta}_{(s)}} = \sum_{t=1}^{p} \alpha_{(s)ta} \hat{\theta}_{(s)t} + \alpha_{(s)aa} \hat{\theta}_{(s)a} + \beta_{(s)a}
\]
\[ \frac{\partial^2}{\partial \theta_a \partial \theta_b} \omega_Q^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}_{(s)}} = \frac{21_{\{a=b\}}}{\alpha_{(s)ab}} \]

**Cubic**

\[ \frac{\partial}{\partial \theta_a} \omega_C^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}_{(s)}} = \sum_{t=1}^{p} \sum_{u=t}^{p} \alpha_{(s)tua} \hat{\theta}_{(s)t} + \alpha_{(s)aad} \hat{\theta}_{(s)a} + \sum_{t=1}^{p} \alpha_{(s)tad} \hat{\theta}_{(s)t} + \sum_{t=1}^{p} \beta_{(s)tad} \hat{\theta}_{(s)t} + \beta_{(s)aa} \hat{\theta}_{(s)a} + \gamma_{(s)a} \]

\[ \frac{\partial^2}{\partial \theta_a \partial \theta_b} \omega_C^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}_{(s)}} = \frac{21_{\{a=b\}}}{\alpha_{(s)ab}} \left( \sum_{t=1}^{p} \alpha_{(s)tub} \hat{\theta}_{(s)t} + \alpha_{(s)aub} \hat{\theta}_{(s)a} + \alpha_{(s)abu} \hat{\theta}_{(s)b} + \beta_{(s)ab} \right) \]

\[ \frac{\partial^3}{\partial \theta_a \partial \theta_b \partial \theta_c} \omega_C^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}_{(s)}} = \frac{21_{\{a=b \land a=c \land b=c\}}}{\alpha_{(s)abc}} \frac{31_{\{a=b=c\}}}{\alpha_{(s)abc}} \frac{\alpha_{(s)abc}}{3} \]

**Sparse Cubic**

\[ \frac{\partial}{\partial \theta_a} \omega_{SC}^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}_{(s)}} = 3\alpha_{(s)a} \hat{\theta}_{(s)a} + \sum_{t=1}^{p} \beta_{(s)tad} \hat{\theta}_{(s)t} + \beta_{(s)aa} \hat{\theta}_{(s)a} + \gamma_{(s)a} \]

\[ \frac{\partial^2}{\partial \theta_a \partial \theta_b} \omega_{SC}^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}_{(s)}} = 6\alpha_{(s)a} \hat{\theta}_{(s)a} \frac{1_{\{a=b\}}}{\alpha_{(s)a}} + \frac{21_{\{a=b\}}}{\beta_{(s)ab}} \beta_{(s)ab} \]

\[ \frac{\partial^3}{\partial \theta_a^3} \omega_{SC}^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}_{(s)}} = 6\alpha_{(s)a} \]

**Skew-Normal**

For the skew-normal, the mode of the likelihood model is made to match the subset MLE, so \( \hat{\eta}_{(s)} = \hat{\theta}_{(s)} \). Then the derivatives of the skew-normal model are:

\[ \frac{\partial}{\partial \theta_a} \omega_{SN}^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}_{(s)}} = -(\theta - \mu_{(s)})^T \Sigma_{(s)-1} a + \frac{2\alpha_{(s)a}}{\sqrt{2\pi}} \]

\[ \frac{\partial^2}{\partial \theta_a \partial \theta_b} \omega_{SN}^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}_{(s)}} = \Sigma_{(s)-1} a a^T + \Sigma_{(s)-1} b b^T \]

\[ \frac{\partial^3}{\partial \theta_a^3} \omega_{SN}^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}_{(s)}} = 2\Sigma_{(s)-1} a a a^T + 3\Sigma_{(s)-1} b b b^T \]
where $\Sigma_{(s)}^{-1}$ is the $a^{th}$ column of $\Sigma_{(s)}^{-1}$. Note that the third-order derivatives of the log skew-normal model do not depend on the mean and covariance parameters. This is because we centered the skewness term on the mode, which will result in closed-form solutions for the model parameters, a significant computational advantage.

### 2.3.3 Solutions for the Subset Likelihood Model Parameters

The final step in fitting the subset likelihood models is to equate the local information of the model with the local information of the observed subset likelihood and solve for the model parameters. The local information of the observed subset likelihood consists of the subset MLE and the derivatives of the log likelihood, given above for logistic regression. These are constants after the data has been observed. The local information of the subset likelihood model consists of the mode and the derivatives of the model, given above. These are functions of the model parameters. The solutions for the model parameters are:

**Quadratic**

\[
\frac{\partial^2}{\partial \theta_a \partial \theta_b} \omega_{SN}^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}(s)} = -\Sigma_{(s)}^{-1} - \alpha_{(s)a} \alpha_{(s)b} \frac{2}{\pi} \\
\frac{\partial^3}{\partial \theta_a^3} \omega_{SN}^{(s)}(\theta) \bigg|_{\theta = \hat{\theta}(s)} = \frac{\alpha_{(s)a} (8 - 2\pi)}{\pi \sqrt{2\pi}}
\]

where $\Sigma_{(s),a}$ is the $a^{th}$ column of $\Sigma_{(s)}^{-1}$. Note that the third-order derivatives of the log skew-normal model do not depend on the mean and covariance parameters. This is because we centered the skewness term on the mode, which will result in closed-form solutions for the model parameters, a significant computational advantage.
Cubic

\[
\frac{\partial^3}{\partial \theta_a \partial \theta_b \partial \theta_c} \hat{w}^C_{(s)}(\theta) \bigg|_{\theta = \hat{\theta}(s)} = \frac{\partial^3 \ell_{(s)}}{\partial \theta_a \partial \theta_b \partial \theta_c} \bigg|_{\theta = \hat{\theta}(s)} \quad \iff \quad \hat{\alpha}_{(s)abc} = 2^{-1}(a=b \lor a=c \lor b=c) \cdot 3^{-1}(a=b=c) \cdot \frac{\partial^3 \ell_{(s)}}{\partial \theta_a \partial \theta_b \partial \theta_c} \bigg|_{\theta = \hat{\theta}(s)}
\]

\[
\frac{\partial^2}{\partial \theta_a \partial \theta_b} \hat{w}^C_{(s)}(\theta) \bigg|_{\theta = \hat{\theta}(s)} = \frac{\partial^2 \ell_{(s)}}{\partial \theta_a \partial \theta_b} \bigg|_{\theta = \hat{\theta}(s)} \quad \iff \quad \hat{\beta}_{(s)ab} = 2^{-1}(a=b) \cdot \frac{\partial^2 \ell_{(s)}}{\partial \theta_a \partial \theta_b} \bigg|_{\theta = \hat{\theta}(s)} - \sum_{t=1}^p \hat{\alpha}_{(s)tab} \hat{\theta}_{(s)t} - \hat{\alpha}_{(s)aab} \hat{\theta}_{(s)a} - \hat{\alpha}_{(s)abb} \hat{\theta}_{(s)b}
\]

\[
\frac{\partial}{\partial \theta_a} \hat{w}^C_{(s)}(\theta) \bigg|_{\theta = \hat{\theta}(s)} = \frac{\partial \ell_{(s)}}{\partial \theta_a} \bigg|_{\theta = \hat{\theta}(s)} = 0
\]

\[
\iff \quad \hat{\gamma}_{(s)a} = - \sum_{t=1}^p \sum_{u=1}^p \hat{\alpha}_{(s)tau} \hat{\theta}_{(s)t} \hat{\theta}_{(s)u} - \hat{\alpha}_{(s)aaa} \hat{\theta}_{(s)a}^2 - \sum_{t=1}^p \hat{\alpha}_{(s)taa} \hat{\theta}_{(s)t} \hat{\theta}_{(s)a} - \sum_{t=1}^p \hat{\beta}_{(s)ta} \hat{\theta}_{(s)t} - \hat{\beta}_{(s)aa} \hat{\theta}_{(s)a}
\]

Sparse Cubic

\[
\frac{\partial^3}{\partial \theta_a^3} \hat{w}^{SC}_{(s)}(\theta) \bigg|_{\theta = \hat{\theta}(s)} = \frac{\partial^3 \ell_{(s)}}{\partial \theta_a^3} \bigg|_{\theta = \hat{\theta}(s)} \iff \quad \hat{\alpha}_{(s)a} = 6^{-1} \frac{\partial^3 \ell_{(s)}}{\partial \theta_a^3} \bigg|_{\theta = \hat{\theta}(s)}
\]

\[
\frac{\partial^2}{\partial \theta_a \partial \theta_b} \hat{w}^{SC}_{(s)}(\theta) \bigg|_{\theta = \hat{\theta}(s)} = \frac{\partial^2 \ell_{(s)}}{\partial \theta_a \partial \theta_b} \bigg|_{\theta = \hat{\theta}(s)} \quad \iff \quad \hat{\beta}_{(s)ab} = 2^{-1}(a=b) \left( \frac{\partial^2 \ell_{(s)}}{\partial \theta_a \partial \theta_b} \bigg|_{\theta = \hat{\theta}(s)} - 6 \hat{\alpha}_{(s)a} \hat{\theta}_{(s)a} 1\{a = b\} \right)
\]

\[
\frac{\partial}{\partial \theta_a} \hat{w}^{SC}_{(s)}(\theta) \bigg|_{\theta = \hat{\theta}(s)} = \frac{\partial \ell_{(s)}}{\partial \theta_a} \bigg|_{\theta = \hat{\theta}(s)} = 0
\]

\[
\iff \quad \hat{\gamma}_{(s)a} = -3 \hat{\alpha}_{(s)a} \hat{\theta}_{(s)a}^2 - \sum_{t=1}^p \hat{\beta}_{(s)ta} \hat{\theta}_{(s)t} - \hat{\beta}_{(s)aa} \hat{\theta}_{(s)a}
\]
Skew-Normal

\[
\frac{\partial^3 \hat{\omega}^{SN}_{(s)}}{\partial \theta_a^3} \bigg|_{\theta = \hat{\theta}(s)} = \frac{\partial^3 \ell(s)}{\partial \theta_a^3} \bigg|_{\theta = \hat{\theta}(s)} \iff \hat{\alpha}(s)_k = \left( \frac{\partial^3 \ell(s)}{\partial \theta^3_k} \bigg|_{\theta = \hat{\theta}(s)} \right)^{\frac{1}{3}} \frac{\pi \sqrt{2\pi}}{8 - 2\pi}
\]

\[
\frac{\partial^2 \hat{\omega}^{SN}_{(s)}}{\partial \theta_a \partial \theta_b} \bigg|_{\theta = \hat{\theta}(s)} = \frac{\partial^2 \ell(s)}{\partial \theta_a \partial \theta_b} \bigg|_{\theta = \hat{\theta}(s)} \iff \hat{\Sigma}^{-1}(s) = -H(s) - \frac{2}{\pi} \hat{\alpha}(s) \hat{\alpha}^T(s)
\]

\[
\frac{\partial \hat{\omega}^{SN}_{(s)}}{\partial \theta_a} \bigg|_{\theta = \hat{\theta}(s)} = \frac{\partial \ell(s)}{\partial \theta_k} \bigg|_{\theta = \hat{\theta}(s)} = 0 \iff \hat{\mu}(s) = \hat{\theta}(s) - \frac{2}{\sqrt{2\pi}} \hat{\Sigma}(s) \hat{\alpha}(s) \text{ and } \hat{\eta}(s) = \hat{\theta}(s)
\]

where $H$ is the Hessian, the matrix of second partial derivatives of the log likelihood at the subset MLE.

Note that $-H$ will be positive definite, since the subset log likelihood will be concave as long as the usual regularity conditions hold. But $\hat{\Sigma}^{-1}(s)$ as calculated above may not be positive definite, depending on $\alpha(s)$. If $\hat{\Sigma}^{-1}(s)$ is not positive definite, the normal kernel of the skew-normal model will be inverted, meaning it will go to infinity in the tails. This is problematic both from an inference perspective, since the fitted skew-normal model will not represent a probability density, and from a numerical perspective, since optimization will be unstable. An alternative parameter estimate is

\[
\hat{\Sigma}^{-1}(s) = -H(s)
\]

This alternative forces the inverse covariance matrix of the skew-normal model’s normal kernel to match the Hessian of the observed subset likelihood. The skew-normal model remains unchanged, but the estimator of $\Sigma^{-1}(s)$ is different. Therefore the fitted model is different. This variant is called the skew-normal model with forced second derivative. The advantage is numerical stability and a fitted subset likelihood model which is guaranteed to be a proper probability density function. The disadvantage is that the second-order local information of the fitted subset likelihood model will deviate slightly from the local information of the observed subset likelihood. The third partial derivatives will not change, and the gradient—and therefore the location
of the mode—will still match the observed subset likelihood. Simulation experiments show that this alternative estimator for $\Sigma_{(s)}^{-1}$ improves performance.

2.4 Recombination

To recombine the fitted subset likelihood models, simply sum them on the log scale:

$$\hat{\omega}(\theta) = \sum_{s=1}^{r} \hat{\omega}_{(s)}(\theta)$$

With independent observations, the all-data log likelihood is the sum of the subset log likelihoods. If the fitted subset log likelihood models are good approximations of the observed subset log likelihoods, then the sum of the fitted subset log likelihood models is a good approximation of the observed all-data log likelihood. The observed all-data likelihood is unavailable due to the large size of the data. The recombined likelihood model is its D&R estimator.

The three polynomial subset likelihood models—quadratic, cubic, and sparse cubic—are linear in their parameters. Therefore recombination consists of summing each parameter across the subsets.

**Quadratic**

The recombined quadratic likelihood model is

$$\hat{\omega}_Q(\theta) = \sum_{t=1}^{p} \sum_{u=t}^{p} \hat{\alpha}_{tu} \theta_t \theta_u + \sum_{t=1}^{p} \hat{\beta}_t \theta_t$$

where

$$\hat{\alpha}_{ab} = \sum_{s=1}^{r} \hat{\alpha}_{(s)ab}$$

$$\hat{\beta}_a = \sum_{s=1}^{r} \hat{\beta}_{(s)a}$$
The number of parameters in the recombined quadratic model is \( \frac{p^2 + 3p}{2} \) which is the same as the number of parameters in a single subset.

**Cubic**

The recombined cubic likelihood model is

\[
\hat{\omega}^C(\theta) = \sum_{t=1}^{p} \sum_{u=t}^{p} \sum_{v=u}^{p} \hat{\alpha}_{tvu} \theta_t \theta_u \theta_v + \sum_{t=1}^{p} \sum_{u=t}^{p} \hat{\beta}_{tu} \theta_t \theta_u + \sum_{t=1}^{p} \hat{\gamma}_t \theta_t
\]

where

\[
\hat{\alpha}_{abc} = \sum_{s=1}^{r} \hat{\alpha}_{(s)abc}
\]

\[
\hat{\beta}_{ab} = \sum_{s=1}^{r} \hat{\beta}_{(s)ab}
\]

\[
\hat{\gamma}_a = \sum_{s=1}^{r} \hat{\gamma}_{(s)a}
\]

The number of parameters in the recombined cubic model is \( \frac{p^3 + 6p^2 + 11p}{6} \) which is the same as the number of parameters in a single subset.

**Sparse Cubic**

The recombined sparse cubic likelihood model is

\[
\hat{\omega}^{SC}(\theta) = \sum_{t=1}^{p} \hat{\alpha}_t \theta_t^3 + \sum_{t=1}^{p} \sum_{u=t}^{p} \hat{\beta}_{tu} \theta_t \theta_u + \sum_{t=1}^{p} \hat{\gamma}_t \theta_t
\]

where

\[
\hat{\alpha}_a = \sum_{s=1}^{r} \hat{\alpha}_{(s)a}
\]

\[
\hat{\beta}_{ab} = \sum_{s=1}^{r} \hat{\beta}_{(s)ab}
\]

\[
\hat{\gamma}_a = \sum_{s=1}^{r} \hat{\gamma}_{(s)a}
\]
The number of parameters in the recombined sparse cubic model is \( p^3 + 6p^2 + 11p \) which is the same as the number of parameters in a single subset.

**Skew-Normal**

The recombined skew-normal likelihood model is

\[
\hat{\omega}^{SN}(\theta) = -\frac{1}{2} ((\theta - \hat{\mu})^\top \hat{\Sigma}^{-1}(\theta - \hat{\mu})) + \sum_{s=1}^{r} \log \Phi \left( \hat{\alpha}^\top_{(s)}(\theta - \hat{\eta}_{(s)}) \right)
\]

where

\[
\hat{\Sigma}^{-1} = \sum_{s=1}^{r} \hat{\Sigma}^{-1}_{(s)}
\]

\[
\hat{\mu} = \left( \sum_{s=1}^{r} \hat{\Sigma}^{-1}_{(s)} \right)^{-1} \sum_{s=1}^{r} \hat{\Sigma}^{-1}_{(s)} \hat{\mu}_{(s)}
\]

The number of parameters in the recombined skew-normal model grows with the number of subsets. It is \( p^2 + 3p + 4rp \).

### 2.5 Subset Likelihood Model Diagnostics

We approach the subset likelihood modeling as a single element in a larger set of tools which collectively make up an interactive data analysis. The tools of data analysis are well known to statisticians. They include tools for data visualization, data exploration, data description and summary, model building, and model diagnostics. All of these tools apply to subset likelihood modeling. Successful likelihood modeling begins when the analyst uses D&R methods to visualize the data at their finest granularity [12] [13]. After thoroughly exploring the data, the analyst may conjecture a data model. Likelihood modeling allows the analyst to fit the data model in the D&R framework. In this chapter, we have used the logistic regression data model as an example. Likelihood modeling continues the analytic process of describing the data.
We treat the likelihood as data. It is a statistic, in the sense that it is a function of the data. By modeling the likelihood, we refine our understanding of the data. Statistical model building is an iterative process of model conjecture, model fitting, and model diagnostics. We build a model for the likelihood using the same approach. In cases where the likelihood is asymptotically normal, or the sampling distribution of the MLE is asymptotically normal, the analyst has a right to conjecture a quadratic model. But even when \( m \) is large, we will not assume that a quadratic model works well without doing model diagnostics. In this section, we propose a visual diagnostic tool to verify the appropriateness of a fitted subset likelihood model.

Figure 2.1 shows an example diagnostic plot for three likelihood models fit on simulated data. The data are from a subset of size \( m = 2^8 \). The values of the independent variable \( x \) are drawn independently from a uniform distribution on \((-7, 7)\) and the response is Bernoulli with mean \((1 + \exp(-\theta x_i))^{-1}\) where \( \theta = 1 \). The top three panels show the observed subset likelihood in blue. A blue point highlights the subset MLE. The fitted subset likelihood model is shown in red. The bottom panels display the error of the subset likelihood model. This is the difference \( \hat{\omega}_s - \ell(s) \). The error will always be zero at the subset MLE if the model is fit using the methods described in section 2.3.
Figure 2.1. A diagnostic plot for three subset likelihood models fit to the same observed likelihood. The scale is determined by the parameter $\delta = 20$. The top row of panels shows the observed subset likelihood in blue and the fitted model in red. The subset MLE is highlighted with a blue point. The vertical scale is log base 2 and shifted so that the MLE is at 0. The lower row of panels shows the error of the fitted model $\hat{\omega}(s) - \ell(s)$. The plot uses a log base 2 scale so that the quantities are easily interpreted by the analyst. For example, in the bottom left plot, we see that the error of the quadratic model is approximately $2^{10}$ at the left edge of the plotting region and approximately $2^{-20}$ at the right edge. This means that the quadratic model overstates the likelihood by a factor of approximately 1000 at the left edge of the region and understates the likelihood by a factor of approximately 1000000 at the right edge of the region.

The plotting region has been chosen such that the likelihood has equal values at the left and right endpoints. This is so that any skewness in the likelihood is easily visible. A simple and efficient method for finding an interval with this property is described in algorithm 1. The analyst supplies the parameter $\delta$, which is the difference between the value of the log likelihood at its maximum and the value at the desired
endpoints. Figure 2.1 uses $\delta = 20$. This means that the plotted region includes all points where the likelihood is no less than $2^{-20} \approx 10^{-6}$ of its value at the subset MLE. This is a very large region. Therefore the diagnostic plots in figure 2.1 put a visual emphasis on the tail behavior of the likelihood and the fitted models. In this region, with this one example data set, we see that the skew-normal model is a good fit for the observed subset likelihood. It has the smallest maximum error by a large margin. We see that the quadratic model does not provide a good fit. The assumption of normality of the likelihood was not justified in this case. The cubic likelihood model demonstrates its instability on this region. In the right half of the plotted region, we see the cubic model turn towards positive infinity. By the edge of the region, the value of the fitted cubic model has already exceeded its value at the subset MLE. This is a gross misrepresentation of the observed subset likelihood. Inference on this region using the cubic likelihood model will give disastrously wrong results. Convex optimization of the fitted cubic model with a starting point near the right edge of this region will not converge.

Figures 2.2 and 2.3 demonstrate the importance of choosing an appropriate plotting region for the diagnostic plot. They show the same observed subset likelihood and the same three fitted likelihood models as in figure 2.1 but with smaller regions chosen by $\delta = 10$ and $\delta = 2$ respectively. A very different picture emerges of the appropriateness of the three likelihood models. In a small region around the subset MLE, as shown in figure 2.3, the cubic model provides a nearly perfect fit. The quadratic and skew-normal models each have maximum absolute error of approximately 0.5 on the log scale. This represents a deviation of about 40% on the scale of the likelihood. The error of the quadratic model is smaller than the error of the skew-normal model when $\theta$ is close to the subset MLE.
Figure 2.2. A diagnostic plot for three subset likelihood models fit to the same observed likelihood. The scale is determined by the parameter $\delta = 10$. The top row of panels shows the observed subset likelihood in blue and the fitted model in red. The subset MLE is highlighted with a blue point. The vertical scale is log base 2 and shifted so that the MLE is at 0. The lower row of panels shows the error of the fitted model $\hat{\omega}(s) - \ell(s)$. 
Figure 2.3. A diagnostic plot for three subset likelihood models fit to the same observed likelihood. The scale is determined by the parameter $\delta = 2$. The top row of panels shows the observed subset likelihood in blue and the fitted model in red. The subset MLE is highlighted with a blue point. The vertical scale is log base 2 and shifted so that the MLE is at 0. The lower row of panels shows the error of the fitted model $\hat{\omega}(s) - \ell(s)$.

These diagnostic plots, drawn for the same data set, demonstrate an important point about subset likelihood modeling. Which likelihood model is best depends on the goals of the analysis. In the example shown here, the skew-normal model performed best in the tails of the likelihood, while the cubic model performed best close to the subset MLE. If the goal of the analysis is to model tail behavior, the analyst might choose the skew-normal model based on these diagnostic plots. If the goal is to find a point estimate for $\theta$, the analyst might choose the cubic model based on the same set of diagnostic plots.
3. WORKING WITH THE RECOMBINED LIKELIHOOD MODEL

The recombined likelihood model is the divide and recombine (D&R) estimate for the all-data likelihood, which is computationally unavailable due to the large size of the data. In this chapter we describe in detail the D&R subset likelihood modeling approach to point estimation and interval estimation.

3.1 Point Estimation

The point estimate for $\theta$ is obtained by maximizing the recombined likelihood model. The maximizer is the Likelihood Modeling Estimate (LME). It is the D&R replacement for the maximum likelihood estimate (MLE). The LME obtained depends on the choice of subset likelihood model and the fitting method used to estimate the model parameters. The models and fitting methods described in the previous chapter—the quadratic, cubic, and sparse cubic likelihood models, plus the skew-normal model fit with or without a forced second derivative—result in five different estimators $\hat{\omega}$ for the all-data likelihood and a five different LMEs.

$$\hat{\theta}_{\text{LME}} = \arg\max_{\hat{\theta}} \hat{\omega}$$

Finding the quadratic LME from the recombined quadratic model can be done by directly maximizing the recombined log likelihood model with a convex optimization method, or by inverting the $p \times p$ matrix in the closed-form solution

$$\hat{\theta}_{\text{LME}}^Q = \left( \sum_{s=1}^{r} H_{(s)} \right)^{-1} \sum_{s=1}^{r} H_{(s)} \hat{\theta}_{(s)}$$

where $H_{(s)}$ is the Hessian from subset $s$ and $\hat{\theta}_{(s)}$ is the subset MLE from subset $s$. The multiplication $H_{(s)} \hat{\theta}_{(s)}$ is computed as a part of the analysis on each subset. The
two summations are computed as part of recombination. The matrix inversion is computed at the time the analyst requests an LME.

The cubic and sparse cubic LMEs are found by convex optimization, for example with the `optim` package \[4\] in R. Care must be taken here. The recombined cubic and sparse cubic models have only one local maximum—the LME—and are concave (“convex down”) in a neighborhood around the LME, but are infinite in the tails. Therefore the choice of starting point for numerical optimization should be reasonably close to the LME. We use the quadratic LME as a starting point. This choice was consistently appropriate in our simulations. Calculating the quadratic LME as a first step in calculation of the cubic or sparse cubic LME adds little computational overhead, since the quadratic is far less computationally intensive than the cubic. We also suggest using bounds during numerical optimization of the recombined cubic models. A reasonable choice for upper and lower bounds on the search region are the component-wise minimum and maximum of the subset MLEs. These are easily computed at recombination, and can be incorporated into, for example, the L-BFGS-B method \[14\] available via `method = "L-BFGS-B"` in `optim`. These bounds have been consistently successful in our experiments. It is immediately apparent to the analyst when incorrectly chosen bounds or a disadvantageous starting point has interfered with the calculation of an LME. In such cases, the calculated LME will be on the boundary of the search region, or the optimization will exhaust its maximum allowed number of iterations. In our experience, this was the case in fewer than 0.2% of 30000 simulation runs.

The skew-normal LME is found by convex optimization. The recombined log likelihood model is not guaranteed to be concave (“convex down”), but a unique maximum is guaranteed. Simulation experiments show that convex optimization methods perform consistently well for finding the LME. An arbitrary starting point is sufficient for convergence.
3.2 Interval Estimation

Any complete data analysis requires that we are able to make statistical inference. It is not enough to have point estimates. In the D&R framework, this means that the recombined data must bring back enough information from the subset analyses for inference after recombination. Likelihood modeling achieves this by providing an estimate for the full likelihood. The recombined likelihood model is our inference engine. We present here two methods for interval estimation using the recombined likelihood.

3.2.1 Credible Intervals

Consider a Bayesian analysis with a uniform prior. Then the recombined likelihood model is an estimate for the unnormalized posterior. We form a credible region for θ from this posterior. The p-dimensional credible region is the intersection of p marginal credible intervals.

To form a credible interval in the k\textsuperscript{th} dimension of θ, consider a slice of the log posterior running through the LME:

$$\hat{\omega}(\theta) \bigg|_{\theta_{-k} = \hat{\theta}_{-k}}$$

where $\theta_{-k}$ means all elements of the parameter vector $\theta$ except the $k$\textsuperscript{th} element. This is a one-dimensional function of $\theta_k$. For notational simplicity, let $f$ be this one-dimensional slice of the likelihood. We propose to find $a$ and $b$ such that $f(a) = f(b) = f(\hat{\theta}_k) - \delta$. The interval $(a, b)$ is the credible interval in the $k$\textsuperscript{th} dimension. The overall credible region is the intersection over $p$ such intervals. To find the endpoints of each interval, we provide algorithm 1.

The function root is any method to find the root of a monotone one-dimensional function on an interval, such as the \texttt{uniroot} function in the R stats package. The constant $\delta$ is the desired difference between the log likelihood at its maximum and
Algorithm 1 Credible Interval Calculation

1: lik_max := f(θ_k)
2: a := lik_max − σ
3: b := lik_max + σ
4: stop_a := FALSE
5: stop_b := FALSE
6: while ¬(stop_a ∧ stop_b) do
7:      stop_a := lik_max − f(a) < δ
8:      stop_b := lik_max − f(b) < δ
9:      if ¬ stop_a then
10:         a := 2a − θ_k
11:         end if
12:     if ¬ stop_b then
13:         b := 2b − θ_k
14:     end if
15: end while
16: a := root(f − lik_max + δ, (a, a+θ_k/2))
17: b := root(f − lik_max + δ, (b+θ_k/2), b))
18: return (a, b)
at the endpoints. The constant $\sigma$ is an initial step size. An appropriate choice is the square root of the $k^{th}$ diagonal element of the Hessian or simply the constant 1.

This method is quite simple, but provides a significant improvement over grid search in terms of computational efficiency. It begins at the maximum of the function $f$ and searches outward in both directions for the two points where the log likelihood has dropped by $\delta$ from its maximum. First, it overshoots by taking successive steps, each one doubling the distance from the mode until the change in $f$ is at least as large as desired. Then, the exact point desired is between the last and second-to-last step, and may be found efficiently by a method such as the one implemented in \texttt{uniroot}.

The interval found above is a $1 - \alpha$ credible interval where

$$\alpha = 1 - \left( \int_{-\infty}^{\infty} e^{f(\theta)} d\theta \right)^{-1} \int_{a}^{b} e^{f(\theta)} d\theta.$$ 

It is the smallest $1 - \alpha$ credible interval if the recombined likelihood is unimodal. But to find $\alpha$, we must calculate $\int_{-\infty}^{\infty} e^{f(\theta)} d\theta$ numerically. Numerical integration will be computationally expensive, and the result may not be of much practical use to the analyst. Instead, the analyst may choose to forgo the $1 - \alpha$ construction, and instead describe the credible interval as the region in which the posterior density is no less than $e^{-\delta}$ times the density at the posterior mode.

Finding a credible region is most appropriate for the skew-normal likelihood model. For the quadratic model, the normal approximation in the following section is simpler to compute, and corresponds to an exact credible region, since the quadratic model is already exactly normal. For the cubic and sparse cubic models, the credible region approach must be applied with care. Since the cubic and sparse cubic models are concave only in a neighborhood around the LME, and infinite beyond some point outside of that neighborhood, no credible region will exist if $\delta$ is chosen to be too large. The algorithm given above will not converge. The cubic and sparse cubic models provide good estimates of the likelihood only locally. They should not be used for inference outside of a small neighborhood around the LME. We advise the analyst to use the cubic and sparse cubic models for inference only in the region in which the recombined model is concave. To implement this strategy, check that the Hessian is
positive definite after each step away from the mode in the method described above. If a non-positive definite Hessian is found before convergence, decrease $\delta$ and begin again.

3.2.2 Normal Approximation

The second method for interval estimation is to use the local information of the recombined likelihood model to form a normal approximation to the all-data likelihood. This approach uses the LME and the Hessian of the recombined likelihood to form a confidence interval or credible interval. It is trivially simple to compute and justified by both Bayesian and frequentist asymptotic theory.

The normal approximation method for D&R confidence intervals can be thought of as fitting a quadratic likelihood model to the recombined likelihood. If the subsets were modeled as quadratic, the approximation is exact and nothing changes. If the subset models were cubic or skew-normal, we fit a normal approximation to the recombined likelihood to allow for simple estimation of intervals.

From a Bayesian point of view, the recombined likelihood model is an estimate of the posterior distribution under a uniform prior. Using a normal approximation to obtain credible intervals from the posterior is known as the Laplace approximation \[15\]. It uses two pieces of local information from the posterior, the mode and the Hessian. The normal distribution with the same mode and Hessian can be used to approximate desired posterior integrals, or quantiles of the normal can be used to form approximate credible intervals.

From a frequentist point of view, the LME is an estimate of the MLE. Our simulations show that the LME has an approximately normal sampling distribution with the same variance as the MLE. A common method for calculating confidence intervals is to rely on the asymptotic distribution of the MLE, which is normal with covariance matrix equal to the inverse Fisher information. A common estimate for the Fisher
information is the local information, also called the observed Fisher information, the negative Hessian of the log likelihood at the MLE. For a discussion, refer to Efron [10].

Confidence or credible intervals based on the normal approximation will be centered at the LME. The only piece of information required to determine the shape of the interval is the Hessian of the recombined likelihood model at its mode. If the analyst has already used a numerical optimization method to obtain the LME, it is very likely that the necessary information has already been computed. Many numerical optimization routines use either an exact or numerically differentiated Hessian in each step. If this is the case, the analyst may keep the Hessian from the last step before convergence and invert it for use in the normal approximation. Otherwise, the Hessian of the recombined likelihood model can be calculated using the formulae for local information given in Section 2.3.2.

The normal approximation is most appropriate for the polynomial likelihood models. It is a very natural choice indeed for the quadratic model, since that model is already exactly normal. Therefore applying this method to the recombined quadratic model involves no approximation, only direct inference on the recombined likelihood. For interval estimation with the skew-normal model, we advocate the method described in the previous section. Since the recombined skew-normal model is proportional to a proper density function, normal approximation is unnecessary for inference. For the cubic and sparse cubic models, the normal approximation has the disadvantage of discarding the third-order local information, but has the considerable advantage of producing a proper distribution with which to make inference. It solves the problems of non-concavity and infinite tails described in the previous section. The Bayesian perspective described in Section 2.2.2 gives reason to expect that the all-data likelihood is closer to a normal distribution than the subset likelihoods. So even if the analyst finds that a quadratic subset likelihood model is inappropriate due to skew in the observed subset likelihoods, and therefore chooses the cubic or sparse cubic model, it may yet be appropriate to apply a normal approximation to the recombined likelihood model.
3.2.3 Diagnostics for Interval Estimation

We propose a visual diagnostic tool to aid the data analyst in choosing between the two methods presented in this section for interval estimation. The credible interval method is appealing because it takes full advantage of the shape of the recombined likelihood to make more exact inference. The normal approximation is appealing because it is guaranteed to give a proper distribution and because inference is more computationally efficient, especially if \( p \) grows large. Another advantage of the normal approximation is that the analyst can attach probabilities to the estimated intervals without numerically integrating the recombined likelihood. If the analyst chooses the normal approximation for computational or other reasons, our visual diagnostic tool compares the recombined likelihood model to its normal approximation to reveal the error introduced by the approximation. If the analyst wishes to use the credible interval method, our tool verifies that the recombined likelihood model is well-behaved in the region of interest.

Figures 3.1 and 3.2 show two example diagnostic plots for a recombined cubic likelihood fit on simulated data. The simulated data size is \( n = 2^{11} \) with subset size \( m = 2^8 \) and the number of subsets is \( n/m = r = 2^3 \). A single independent variable \( x \) was drawn independently from a standard normal distribution, and the response is Bernoulli with mean \((1 + \exp(-\theta x_i))^{-1}\) where \( \theta = 1 \).

The region displayed in the diagnostic plot should be chosen to correspond at least roughly with the intervals to be estimated. One way to achieve this is to fit the normal approximation and use its quantiles. For example, if a 95% interval is desired, the region displayed in the diagnostic plot should extend 2\( \sigma \) left and right of the LME, where \( \sigma \) is the standard deviation of the fitted normal. This has been done in figure 3.1. Visual inspection of the recombined cubic likelihood model in figure 3.1 reveals that the model is well-behaved in the plotted region. It is concave, so credible interval estimation yielding an interval within this region is appropriate. If the analyst prefers estimation via the normal approximation, the maximum error
within this region will be approximately 0.35 on the log scale. This means that the normal approximation misrepresents the recombined likelihood model by a factor of at most $2^{0.35} \approx 1.4$ within this region. Either method is appropriate.

Figure 3.2 shows the same recombined likelihood and its normal approximation. This time the plotting region was chosen as if a credible interval was desired. The endpoints of the region are the output of algorithm 1 with $\delta = 22$. The dashed
red line shows the inflection point of the recombined cubic likelihood model. To the right of this line, the recombined model is no longer concave. Using the recombined likelihood model as an inference engine beyond this line is risky. The very heavy right tail seen in figure 3.2 may be an accurate representation of uncertainty about \( \theta \) in that region, or it may be an artifact of the improper posterior distribution given by the cubic likelihood model. The analyst may wish to conclude that no inference is possible so far away from the LME.
Figure 3.2. A visual diagnostic tool for interval estimation with a recombined likelihood model. The top panel shows the recombined cubic likelihood model in blue and its normal approximation in red. The range of the plot is given by algorithm 1 with $\delta = 22$. The LME is highlighted with a blue point. A dashed red line shows the inflection point of the recombined likelihood model. The bottom panel shows the error of the normal approximation.
4. SIMULATION RESULTS

In this section, we analyze the properties of likelihood modeling estimates (LMEs) using designed simulation experiments. Divide and recombine (D&R) results are substitutes for the results that would be obtained if it were computationally feasible to analyze the data without subset division. We are interested in comparing LMEs to the maximum likelihood estimate (MLE). But our interest is in the statistical properties of the estimators, not in how close they are numerically to the MLE. Purely numerical analysis ignores the fact that the LME is a new estimator with a sampling distribution different from the sampling distribution of the MLE. We describe the sampling distribution of our proposed LMEs by analyzing simulation results.

The factors in our designed experiment were the total data size \( n \), the subset size \( m \), and the number of subsets \( r \). The value of \( r \) is determined by the choice of \( m \) and \( n \), since \( mr = n \). We used integer powers of 2 for the subset and total data size to ease comprehension. Therefore the value of \( m \) can be at most \( n/2 \), since the smallest non-trivial number of subsets is \( n/m = r = 2 \).

We varied \( m \) from \( 2^{11} \) to \( 2^{19} \) and \( n \) from \( 2^{12} \) to \( 2^{20} \). The minimum subset size was chosen to preserve numerical stability of the likelihood. Too small subset sizes introduce a small probability of non-concavity in the likelihood, which poses problems over many runs of the simulation. The maximum total data size was chosen to allow a very large number of runs in a reasonable amount of time on the available hardware. D&R methods scale to accommodate very large data sets by taking advantage of embarrassingly parallel computation. But we wish to compare our LMEs to the familiar all-data MLE. This limited us to data sizes small enough that the MLE can be computed in a reasonable amount of time.

We held \( p \), the dimension of \( \theta \), fixed at 3. This allowed us to analyze the effects of multidimensionality, including covariance of the LME. Much higher values of \( p \)
introduce additional computational overhead and reduce the number of simulation runs possible in a reasonable amount of time.

The data were drawn from the generative model:

\[ \theta = (1, 1, 1)^T \]

\[ x_i \overset{iid}{\sim} \text{Normal}(0, 1) \]

\[ y_i | x_i \sim \text{Bernoulli}( (1 + e^{\theta^T x_i})^{-1} ) \]

So the model is logistic regression with three independent variables, binary response, and no intercept. The observations are independent. For each unique combination of \( m \) and \( n \), we simulated a single design matrix and 30000 runs, meaning 30000 response vectors. On each response vector, we computed the MLE and the five LMEs: the quadratic, cubic, sparse cubic, skew-normal, and skew normal with forced second derivative described in section 2.3.3.

We found that the LMEs were nearly identical to the MLE in variance, but had greater bias. Unlike the MLE, the bias of the LMEs does not go to zero as the total data size grows. Instead, the bias of the LMEs converges on some fixed constant. This constant is different for each of the five LMEs and depending on the value of \( m \). In general, the bias of the LMEs is very small in magnitude compared with the true value of the parameter.

4.1 Normality

The LMEs have an approximately normal distribution for all values of \( m \) and \( n \). This is encouraging for several reasons. The recombined likelihood model is an estimate of the all-data likelihood. The mode of the all-data likelihood, the MLE, is asymptotically normal. We would expect that the mode of a good estimate of the likelihood would also be approximately normal. Normality aids further analysis of the statistical properties of the LMEs’ sampling distributions, since we can fully characterize those distributions by their means and variances.
Figures 4.1, 4.2, 4.3, 4.4, and 4.5 show the marginal distributions of the LMEs. Each panel represents a unique combination of $m$ and $n$. In the top row of panels, the total data size is $n = 2^{14}$, in the second row it is $2^{16}$, in the third $2^{18}$ and in the bottom row it is $2^{20}$. The left column of panels has each data set of size $n$ divided into subsets of size $2^{17}$, the second column has subsets of size $2^{15}$, the third has subsets of size $2^{13}$, and the right column has subsets of size $2^{11}$. So for example, the bottom left panel of figure 4.1 shows the distribution of the quadratic LME when the total data size is $n = 2^{20}$ with $r = 2^3$ subsets each of size $2^{17}$. A single outlier is present in figure 4.4 where $\log n = 20$ and $\log m = 15$. This outlier was caused by a $\Sigma^{-1}$ matrix which was not positive definite, as described in section 2.3.3. The remedy proposed in that section succeeded, as we can see in figure 4.5.
Figure 4.1. Quantile plot of the first element of the Quadratic LME. The reference distribution is the standard normal. The sample mean and the true value $\theta_1 = 1$ are both shown in red. A reference line passing through the 0.25 and 0.75 quantiles is shown in black to aid visual assessment of normality.
Figure 4.2. Quantile plot of the first element of the Cubic LME. The reference distribution is the standard normal. The sample mean and the reference line passing through the 0.25 and 0.75 quantiles is shown in black to aid visual assessment of normality. The true values $\theta_1 = 1$ are both shown in red. A reference line passing through the sample mean and the reference line is also shown in black to aid visual assessment of normality.
Figure 4.3. Quantile plot of the first element of the Sparse Cubic LME. The reference distribution is the standard normal. The sample mean and the true value \( \theta_1 = 1 \) are both shown in red. A reference line passing through the 0.25 and 0.75 quantiles is shown in black to aid visual assessment of normality.
Figure 4.4. Quantile plot of the first element of the Skew-Normal LME. The reference distribution is the standard normal. The sample mean and the true value $\theta_1 = 1$ are both shown in red. A reference line passing through the 0.25 and 0.75 quantiles is shown in black to aid visual assessment of normality.
Figure 4.5. Quantile plot of the first element of the Skew-Normal with forced second derivative LME. The reference distribution is the standard normal. The sample mean and the true value $\theta_1 = 1$ are both shown in red. A reference line passing through the 0.25 and 0.75 quantiles is shown in black to aid visual assessment of normality.
Each LME is a vector of length 3, but only the first element is shown. The three elements are exchangeable random variables, so the marginal distribution of the first is the same as the marginal distributions of the others. The plots show the quantiles of the LMEs compared with quantiles of a standard normal distribution. The values of the LMEs lie very close to the reference lines drawn, indicating that their distributions are very close to normal distributions. A difference in mean with respect to the standard normal distribution shows up as a vertical shift. The means of the sampling distributions of the LMEs are close to 1, which is the true value of $\theta$ in our simulation. The means will be described in more detail below. A difference in variance with respect to the standard normal distribution shows up as a shift in slope. The variances of the sampling distributions of the LMEs vary with $n$. The variances will be described in more detail below. Three panels of each plot are blank because the subset size $m$ must be smaller than the total data size $n$.

4.2 Variance

The variances of the LMEs are nearly identical to the variance of the all-data MLE. This is a powerful result. It indicates that the significant advantages of D&R estimation come at little or no cost in terms of the variance of our estimators.

Figure 4.6 shows the log standard deviation of the estimates, grouped by $n$. There are two salient points in this plot. The first is that no difference exists between the six estimators—the MLE plus five LMEs. They have been plotted in different colors, as indicated, but the fact that they are obscured by overplotting demonstrates that their variances are nearly identical. The second point is that the variances of the estimates do not depend on the subset size. Each panel is dedicated to a unique total data size $n$. Within each panel, the subset size $m$ varies from left to right. The all-data MLE is by definition invariant to $m$, since it was calculated on the full data set without subset division. So it is no surprise that its variance does not change with $m$. But the fact that the standard deviation of the LMEs does not change with $m$ is of great
importance. It means that the variance of the LME depends only on the total data size, not on the analyst’s choice of subset size.

Figure 4.6. Log standard deviation of the estimates grouped by $n$, the total data size.

Figure 4.7 shows the log standard deviation again, this time grouped by the subset size $m$. Each panel is dedicated to a unique value of $m$. Within each panel, the total data size grows from left to right. Therefore the number of subsets grows from left to right. We see that the variance decreases as the total data size grows. Asymptotically, the variance of the MLE is proportional to $n^{-1/2}$. We can verify in the figure that the variance of our estimates, calculated on finite samples, is behaving as theory predicts. Since we have taken a log transformation of the standard deviation, the $n^{-1/2}$ rule
is a linear relationship. The next figure takes advantage of this rule to analyze more closely the differences between estimators.

![Figure 4.7. Log standard deviation of the estimates grouped by \(m\), the subset data size.](image)

We fit a line through the points in each panel of figure 4.7. The slope is the natural log of \(\sqrt{2}\). We plot the residuals in figure 4.8. The result is a plot of the log standard deviation of the estimates after correcting for sample size. We have seen that the variance of the estimates decreases proportional to \(n^{-1/2}\). Now we remove that effect and see what remains. All three elements of the parameter vector \(\theta\) are presented in the plot. Within each panel, \(n\) increases from left to right. So at a given value of \(m\), there are three points for each estimator, corresponding to the marginal
variances of the three elements of the parameter vector. In the bottom left panel, the subset size is $2^{11}$. We see that the skew-normal LME has the largest variance. The skew-normal LME with forced second derivative has a slightly lower variance. The variances of the quadratic, cubic, and sparse cubic LMEs are indistinguishable from the variance of the MLE.

Figure 4.8. Standard deviation by $m$, the subset data size, after correcting for $n$, the total data size. The plotted points are residuals of a regression performed on Figure 4.7 where the slope is fixed at $\log_e \sqrt{n}$.

The difference in the natural log of standard deviation is approximately equal to the percentage difference in standard deviation. This holds when the difference on the natural log scale is near zero. We can see in figure 4.8 that the standard deviations of the five LMEs are within 3% of the standard deviation of the MLE for the smallest
4.3 Covariance

We analyze the covariance matrices of the LMEs and compare them to the covariance of the MLE by examining the eigenvalues of the covariance matrix. There is one sample covariance matrix for each estimator at each combination of \( m \) and \( n \). This is the \( 3 \times 3 \) sample covariance matrix obtained from the 30000 runs of the simulation. Each run produced a length-3 realization of each of the 5 LMEs and a length-3 realization of the MLE. We calculated the maximum and minimum eigenvalues from each sample covariance matrix. The maximum and minimum eigenvalues of the covariance matrix give the maximum and minimum variance of any linear combination of \( \hat{\theta} \) if the linear combination has norm 1. So the square root of the eigenvalues gives the minimum and maximum standard deviation of such linear combinations.

Figure 4.9 shows the natural log of the square root of the eigenvalues, which is log standard deviation. Each panel corresponds to a fixed value of \( n \). The results are similar to those of the marginal variances. The maximum and minimum variances of linear combinations of the estimators does not depend on the analyst’s choice of subset size. The MLE and the five LMEs overlap entirely, indicating that differences between them are trivially small at the scale shown. The difference between estimators will be shown in detail shortly.
Figure 4.9. Log square root of the maximum and minimum eigenvalues of the covariance of the estimates by $n$, the total data size.

Figure 4.10 shows the natural log of the square root of the eigenvalues grouped by $m$. All five LMEs and the MLE show the same behavior. The standard deviation decreases proportional to $n^{-1/2}$ regardless of the subset size.

In figure 4.11 we view the residuals after correcting for $n$ as before. These are the residuals from fitting a line to each panel in figure 4.10 where the slope is the natural log of $\sqrt{2}$ to remove the effect of sample size. What remains is a clearer picture of the differences between the MLE and the five LMEs. As in the case of marginal variances, the maximum variance of unit vector linear combinations of the LMEs is very close to that of the MLE. The largest variance belongs to the skew-normal LME. The skew-normal LME with forced second derivative has a smaller variance,
but still slightly larger than the MLE or the polynomial LMEs. The quadratic, cubic, and sparse cubic LMEs are indistinguishable from the MLE at all subset sizes. The differences between the best and worst estimators is on the order of a few percentage points of standard deviation in the worst case, and shrinks as the subset size grows.
Figure 4.11. Log square root of the maximum and minimum eigenvalues of the covariance of the estimates by $m$, the subset data size, after correcting for $n$, the total data size. The plotted points are residuals of a regression performed on Figure 4.10 where the slope is fixed at $\sqrt{n}$.

4.4 Bias

The all-data MLE is unbiased asymptotically, but biased in any finite sample size. The LMEs have slightly larger bias than the MLE. The bias of the LMEs is where we pay the price for parallel computation, but it is a small price.

Figure 4.12 shows the bias of the estimates grouped by the total data size $n$. So each panel shows, for a given data size, how the performance of the estimators depends on the choice of subset size $m$. At each value of $m$, there are three points plotted for
each estimator. These show the bias for each of the three elements of the parameter vector. The skew-normal LME has the largest bias. The skew-normal LME with forced second derivative shows a slight improvement. The bias of the sparse cubic LME is smaller, and the bias of the quadratic LME is smaller still. The bias of the cubic LME is indistinguishable from the bias of the MLE at this scale. Overplotting obscures the MLE, so that only the cubic LME is visible. A later plot will reveal the differences between the MLE and the cubic LME.

Figure 4.12. Bias of the estimates grouped by $n$, the total data size.

For a fixed value of $n$, the bias of all the LMEs decreases as the subset size increases. The MLE is by definition invariant to subset size. It has the smallest bias
in general. As the subset size grows, the number of subsets shrinks, and the LMEs converge on the MLE.

Even in the worst case, which is the skew-normal with a subset size of $2^{11}$, the bias of the LMEs is very small in magnitude. Recall that the true value of $\theta$ is 1. The worst bias shown in figure 4.12 represents an error of less than 1%.

Figure 4.13 is the most important plot for understanding the behavior of the LMEs. It shows how the bias of the estimates behaves when the subset size is fixed and the number of subsets grows. This is the most relevant result for D&R analysis, since the computational environment puts a hard upper limit on the subset size, but the number of subsets is limited only by the amount of data available and the patience of the analyst.

![Figure 4.13](image_url)  
Figure 4.13. Bias of the estimates grouped by $m$, the subset data size.
For a fixed subset size, each LME has a persistent bias that does not go to zero even as the number of subsets grows. For \( m = 2^{11} \), the skew-normal LME has a persistent bias of about 0.006, the skew-normal with forced second derivative has a persistent bias of about 0.004, the quadratic has a persistent bias of about \(-0.0025\), the sparse cubic has a persistent bias of less than \(-0.001\), and the cubic has a persistent bias of about \(-0.00025\) (the cubic’s bias is discernible in figure 4.14). All of these represent less than 1% error with respect to the true parameter value of 1. The persistent bias associated with each LME shrinks towards zero as the subset size grows. The ordering between LMEs remains.

Because the bias of the cubic LME is so much smaller than the bias of the other LMEs, it must be plotted against the MLE on its own for visual perception. Figure 4.14 shows the bias of the cubic LME and the MLE. In the bottom left panel we see that as \( n \) grows, the bias of the MLE shrinks to zero as its asymptotic distribution promises. The bias of the cubic LME is persistent, though at a magnitude even smaller than the other LMEs.
Figure 4.14. Bias of the MLE and the cubic LME grouped by $m$, the subset data size.
5. DISCUSSION

5.1 Comparison of Subset Likelihood Models

We have presented four subset likelihood models—quadratic, cubic, sparse cubic, and skew-normal—and detailed parameter estimation methods resulting in five likelihood modeling estimates (LMEs)—quadratic, cubic, sparse cubic, skew-normal, and skew-normal with forced second derivative. We have presented point estimation and interval estimation procedures for making inference from the recombined likelihoods. We have provided the data analyst with visual diagnostic tools to evaluate the models and methods as applied to real data. And we have shown via designed experiments the statistical properties of our estimators. We attempt now to give some further guidance to the data analyst. In particular, we discuss how to choose between the models and how to apply them in practice.

The primary concerns in choosing a subset likelihood model are computation and statistical accuracy. We address computation first. The two computational bottlenecks associated with likelihood modeling are compute time on the subsets and cross-network transfer time. Compute time on the subsets is mostly determined by the time required to calculate the subset maximum likelihood estimate (MLE) and compute the necessary derivatives. It also includes time to calculate the parameters of the subset likelihood model. Though closed-form solutions were given for the parameters of all the likelihood models in section 2.3.3, the careful reader will have noticed that estimating the skew-normal model parameters requires inverting a $p \times p$ matrix, which must be done with an iterative numerical algorithm. Cross-network transfer time is determined by the number of parameters in the likelihood model.

The subset likelihood models have different requirements for computation of local information. The cubic is the most demanding. It requires the full array of third
derivatives, whose size is $\mathcal{O}(p^3)$. If $p$ is large, it will not be practical to compute the parameters of the cubic model or to transfer them across the network. The sparse cubic and skew-normal models require only the diagonal elements of the third derivative plus the Hessian. The quadratic is the least demanding, requiring only the Hessian. Since the size of the Hessian dominates the size of the diagonal of the third derivative, the sparse cubic, skew-normal, and quadratic all require the calculation of local information whose size is $\mathcal{O}(p^2)$.

If the only concern is getting an LME as quickly as reasonably possible, the quadratic is the right choice. But if possible, the analyst may wish to fit several models and compare them. Since the likelihood models all share local information from the observed subset likelihoods, it is possible to save time by fitting them simultaneously on each subset and recombinining only once.

If $p$ is not prohibitively large, the cubic model should provide the most accurate point estimate according to our simulations. If $p$ is too large, the sparse cubic model is a good compromise. But if the goal is to fit a model which is accurate in the tails of the likelihood, the cubic and sparse cubic models should not be used. Figure 5.1 was seen earlier in section 2.5 and is reproduced here. It shows an example of the tail behavior of the cubic model. Away from the mode, it behaves poorly. In general, the two cubic models provide an improper likelihood. We have not studied this extensively, but 5.1 suggests that the skew-normal model may be a better fit in the tails. Practitioners should apply the diagnostics tools presented in sections 2.5 and 3.2.3.
Figure 5.1. A diagnostic plot for three subset likelihood models fit to the same observed likelihood. The scale is determined by the parameter $\delta = 20$. The top row of panels shows the observed subset likelihood in blue and the fitted model in red. The subset MLE is highlighted with a blue dot. The vertical scale is log base 2 and shifted so that the MLE is at 0. The lower row of panels shows the error of the fitted model $\hat{\omega}(s) - \ell(s)$.

Lastly, a note on asymptotics. In the analysis of simulation results, a persistent bias was discovered. We discussed it in section 4.4. When the subset size was held fixed, and the number of subsets allowed to grow, the bias of the LMEs did not go to zero. This is in contrast to the all-data MLE, whose bias goes to zero as $n$ grows. However, simulation results also showed that the variance of the LMEs does indeed go to zero as $n$ grows, even when the subset size is fixed. In fact the variance appears not to depend on the subset size at all. The asymptotic behavior of the bias and variance of the LMEs suggests a reasonable course of action for their use in practice. The subset size should be made as large as possible, given the available computational environment. A practitioner may, for example, carry out timed likelihood modeling on a single subset with various subset sizes and extrapolate the time required to
perform the entire analysis at those sizes. A fixed subset size \( m \) has associated with it some level of persistent bias, regardless of the total data size. A total data size \( n \) has associated with it some level of variance, regardless of the subset size. Once the magnitude of the standard deviation is much smaller than the magnitude of the bias, there is little use in analyzing more subsets. Judged by the expected error of the LME, there may be nothing to lose by throwing away some of the data.

For example, in the discussion of simulation results, we found that the persistent bias of the skew-normal LME for a subset size of \( m = 2^{11} \) was approximately 0.006. And we found that the standard deviation of that estimator was smaller than \( e^{-5.8} \approx 0.003 \) by the time \( n \) grew to \( 2^{20} \). If the analyst has tens or hundreds of millions of observations, but is unable to compute on subsets larger than \( 2^{11} \approx 2000 \), a random sample of 1 million observations will do as well as the entire data set in terms of the expected error of the skew-normal LME. Similar reasoning can be applied to any LME and any subset size.

5.2 Related Work

The past few years have seen an explosion of interest in the analysis of large complex data from industry and academia. The importance of parallel computation to address this type of data is by now widely recognized. Two useful overviews of recent advances in statistical theory and methodology for large data are given by Wang [16] and Jordan [17]. In this section we briefly discuss some recent work relevant to divide and recombine (D&R) subset likelihood modeling.

The methods presented below fall broadly into two categories: numerical analysis and statistical analysis. For the numerical analyst, finding an MLE is an optimization problem. If the data are divided into subsets, the approach is called distributed optimization. The goal is to find a distributed optimization algorithm which produces a result approximately equal to the MLE within a reasonable time budget. This approach ignores the statistical properties of the data, the likelihood, and the MLE.
D&R likelihood modeling is a statistical approach. Its goal is to estimate $\theta$, not to approximate the MLE. Likelihood modeling takes advantage of the known statistical properties of the likelihood to improve estimation. It measures success in terms of the statistical properties of the estimates produced. And it provides the analyst with not just a point estimate, but also the recombined likelihood, which is an inference engine.

5.2.1 Distributed Optimization

Computation of an MLE usually requires numerical optimization. In the presence of large data, it may be advantageous to distribute the optimization over many nodes. One popular approach to distributed optimization is the Alternating Direction Method of Multipliers (ADMM).

ADMM is a distributed numerical algorithm for constrained optimization. A comprehensive overview of the ADMM algorithm is given by Boyd [18]. One form of ADMM called global consensus solves problems of the following form:

\[
\min_x \sum_i f_i(x_i) + g(z)
\]
\[s.t. \ x_i = z \ \forall i\]

For our purposes, $f_i$ is the negative log likelihood on the $i^{th}$ subset of data, $x_i$ is the vector of parameters estimated on that subset, and $g(z)$ is an (optional) regularizer such as a lasso [19] or ridge [20] penalty. For unconstrained logistic regression, let $g(z) = 0$. At first, separate parameter vectors are estimated on each subset. But the constraint $x_i = z \ \forall i$ ensures that they converge to a global consensus, which in our case is the MLE for the entire data set (or the lasso or Ridge estimate). The algorithm has three steps which iterate until convergence:

1. $x_i^{k+1} = \arg\max_x L(x, z_k, y_k)$
2. $z^{k+1} = \arg\max_z L(x_i^{k+1}, z, y^k)$
3. update $y$

where $L$ is the augmented Lagrangian and $y$ is the Lagrange parameter.

The algorithm works by finding an MLE on each subset, then averaging all subset MLEs together for a global consensus estimate. In the next iteration, each subset log likelihood is maximized again, but this time with a penalty that pushes toward the previous step’s global consensus estimate.

It is important to note that the first step is an optimization problem which can be carried out in parallel on each subset, but that the second and third steps require communication between subsets. For this reason, the authors say that “it is awkward to express ADMM in MapReduce.” The implementation they provide involves a complete MapReduce step within each iteration of the algorithm. They suggest that one should expect the number of iterations required for statistical modeling applications to be “in the tens.”

5.2.2 Streaming Optimization

Another approach to large-scale optimization is streaming optimization. Here “streaming” refers to the fact that the data are seen one observation at a time. A parameter estimate is held in memory and updated as each new observation arrives. Approximate optimization is complete after a single pass through the entire data set.

Stochastic gradient descent (SGD) \cite{21} is a popular method for streaming optimization. The update at step $t$ is

$$w^t = w^{t-1} - \alpha \nabla Q_i(w)$$

where $w$ is the parameter of interest, $Q_i$ is the negative log likelihood evaluated at the $i^{th}$ observation, and $\alpha$ is the step size, also known as the “learning rate.” Each time $t$ increments, a new observation $i$ is read from the stream.

The intuition behind SGD is this: at each step in a numerical optimization, we would like to take a step in the direction of the gradient. If the true gradient is
unavailable because the data are too large, then we use the gradient evaluated at a single observation as a good estimate.

5.2.3 Out-of-Core R Packages for Generalized Linear Models

A popular R package for fitting linear models and generalized linear models on data too large to fit in random access memory (RAM) is `biglm` [22]. It splits the data into chunks and loads one chunk at a time in RAM. It is up to the user to provide a back end for serving up chunks of data. The user's choice of back end will determine scalability and computational time.

Recall that each iteration in the numerical optimization for the logistic regression MLE can be represented as a weighted least squares problem. This is known as iteratively re-weighted least squares. Weighted least squares has an embarrassingly parallel solution which is exact. The method implemented in `biglm` is to iterate over the entire data set to find the MLE of a generalized linear model. At each iteration, there is a weighted least squares problem which is solved in parallel. That parallel solution is found by reading the data one chunk at a time.

5.2.4 Asymptotics for the Quadratic Subset Likelihood Model

Chen and Xie present an approach for the analysis of large data which they call Split and Conquer [23]. As applied to logistic regression, it is a covariance-weighted average of subset MLEs. This is equivalent to the LME produced by our quadratic likelihood model. They also consider the case of penalized regression. In this case, the penalty is applied on each subset, so that each subset estimate has some zero and some nonzero elements. The nonzero elements in the recombinmed estimator are chosen by majority vote among the subsets.

The authors show that their estimator is consistent when both the subset size and number of subsets grows to infinity. This is different from our notion of asymptotics. We consider the case where the subset size is fixed and the number of subsets grows.
Our choice is informed by practical concerns. In a D&R computational environment, the available hardware will limit the subset size. This is because subset computations are carried out by loading the entire subset into RAM on a single node. The number of subsets that can be analyzed has no such hard limit. It can continue to grow as long as the analyst can provide storage capacity, which is generally cheap, and time. Our simulations suggest that the quadratic LME, which is identical to Chen and Xie’s estimator, is inconsistent according to our notion of asymptotics, though still useful.

5.2.5 Timing Experiments for D&R Logistic Regression

Li presents a detailed analysis of the time needed to carry out a D&R logistic regression analysis [24]. The author presents simulation experiments varying the total data size, subset size, dimension of the parameter vector, and a variety of configuration parameters of the computational environment.

The results provide a guide to the optimal number of subsets as a function of total data size, dimension of the parameter vector, and configuration parameters. For example, when the total data size was $2^{30} \approx 1$ billion rows by 127 columns, the optimal subset size was $2^{12} \approx 4000$. Optimality here refers only to time. Li’s results also provide concrete examples of how long large D&R analysis might take. The fastest time for logistic regression with 1 billion rows and 127 columns was 18 minutes. The recombination method used was simple averaging of the subset MLEs. Subset likelihood modeling and recombination with the quadratic model should require the same amount of compute time on each subset, but a slightly larger amount of information will be transferred across the network compared with simple averaging. This is because estimation of the subset MLEs requires that the Hessian at the MLE be computed on each subset as part of numerical optimization, but recombination for the simple average requires only the $p$-dimensional subset MLE be transferred for recombination. The quadratic LME requires that the $p \times p$ Hessian be transferred as
well. The sparse cubic and skew-normal models require additional compute time and data transfer. The cubic model requires much more of both, if $p$ is large.

5.2.6 Distributed Markov Chain Monte Carlo Simulation

Bayesian data analysis often relies on numerical integration of the posterior via Markov Chain Monte Carlo (MCMC) simulation. Computational constraints in MCMC can be in the form of central processing unit (CPU) time or data size. If the bottleneck is CPU time, this problem may be addressed with multi-thread programming which leaves the statistical analysis unchanged. But if the data are too large to fit in RAM on a single machine, they must be divided into subsets. This requires a distributed approach to the analysis. For this case, Scott and coauthors present the Consensus Monte Carlo algorithm [25].

The method works by simulating $g$ MCMC draws of the parameter vector $\theta$ on each subset. For $i = 1, \ldots, g$ the recombined draw is a weighted combination of the $i^{th}$ draws from all the subsets:

$$\theta_i = \left( \sum_s W_s \right)^{-1} \sum_s W_s \theta_{si}$$

where $s$ indexes the subset, and $W_s$ are the weights. So we begin with $g$ draws on each subset, and end with $g$ recombined draws which are approximately from the all-data posterior. The authors suggest the sample covariance of $\theta_s$ on each subset as the weights $W_s$.

The Consensus Monte Carlo algorithm is optimal if $\theta$ is drawn from a normal distribution on each subset. The authors point out that the posterior distribution on each subset will be asymptotically normal for many models, so it should be approximately normal in practice.
5.2.7 Distributed Resampling

The bootstrap \cite{26} is a nonparametric method for assessing the uncertainty of an estimate. A random sample is taken from the data, with replacement, so that the sample size matches the size of the original data set. The estimator is applied to the sample, giving an estimate $\hat{\theta}_i$. This process is repeated many times to build up a distribution of $\hat{\theta}$. That distribution characterizes the uncertainty of the $\hat{\theta}$ computed on the original data set.

Suppose we wish to use the bootstrap to assess the uncertainty of an LME or another estimate calculated on distributed data. We could carefully collect samples from each subset to preserve the bootstrap’s intended resampling distribution, combine these samples into a new data set, divide it into subsets using the same division method applied to our original data, and calculate an estimate. This would be a time-consuming process. Repeating it enough times to build up a distribution would be impractical.

Kleiner and coauthors present the bag of little bootstraps (BLB) \cite{27} as an efficient variant on the bootstrap for distributed data. On each subset, a sample is taken with replacement. The size of this sample is equal to the total data size, much larger than the subset data size. An estimate is calculated on each subset’s sample. This process is repeated until enough bootstrap samples have been taken to provide a distribution. Amazingly, the authors are able to show that the BLB produces results asymptotically equivalent to the direct all-data bootstrap.

To understand why it is possible to compute with a bootstrap sample much larger than the subset size, consider the following simple example. Suppose we have a trivially small subset of size 2. It consists of data $x_1$ and $x_2$. Now suppose we have 1 billion of these subsets, so that the total data size is 2 billion. If we take a bootstrap sample of size 2 billion from $(x_1, x_2)$ and store each data point in memory, they will occupy approximately 16 terabytes of storage if each data point is a single 8-byte floating point number. But there will only be 2 unique values among them, since
every value is either $x_1$ or $x_2$. Instead, we could store only the number of times that $x_1$ appears in our sample and the number of times that $x_2$ appears in our sample. This requires only 16 bytes. So the size in memory, or on disk, of the bootstrap sample is proportional to the subset size, not the total data size. Adapting the estimation method to this representation of the data is trivial for many methods. For example, binary logistic regression with responses in \{0, 1\} becomes binomial logistic regression with count responses.

The bootstrap is a method for assessing the uncertainty of an estimate, not for producing a point estimate. So it does not provide an alternative to LMEs. It is presented here as an example of interesting new theory for analysis of distributed data. It could, for example, be used to establish an interval estimate associated with an LME.

5.3 Future Work

A promising direction for future work is to develop a Bayesian version of likelihood modeling. This could take two possible forms: either multiply the observed likelihood on each subset by a prior, sample from the resulting posterior, and summarize the posterior draws with an appropriate model, or multiply the recombined likelihood model by a prior and sample from the resulting posterior.

The first option is similar to the Consensus Monte Carlo algorithm \[25\] described above. After division, MCMC is carried out on each subset. At this point, the Consensus Monte Carlo algorithm combines posterior draws to recombine. We propose to model the posterior draws with an appropriate distribution. That is, we propose to carry out a data analysis on the posterior draws. Of the four likelihood models presented in this work, only the quadratic and skew-normal represent proper distributions. The cubic and sparse cubic models would not be appropriate for modeling posterior draws. Parameters of the quadratic could be estimated from posterior draws in the usual way with the sample mean and standard deviation. A maximum like-
likelihood approach to fitting the skew-normal is described in [1]. Appropriate data analysis will reveal further models for the posterior draws as warranted by the data.

The second option requires far less computation on each subset. It is also a much smaller deviation from our standard approach. The subset likelihood models are fit and recombined as in our standard approach, but point estimation differs. Rather than finding the mode of the recombined likelihood, we would multiply by a prior. The product is then a likelihood modeling posterior. A reasonable point estimate would be the posterior mean. Again, the cubic and sparse cubic models are not appropriate for this approach, unless approximated by a normal as in section 3.2.2. In the case of the quadratic and skew-normal models, a normal prior will be conjugate.
REFERENCES
REFERENCES


VITA
VITA

Philip Gautier was born in Washington, DC. He worked as an accounting assistant before earning a bachelor’s degree in economics from the University of Maryland, College Park, and a master’s degree in statistics from American University, Washington, DC. His interests include practical and scalable applications of data visualization, statistical modeling, and machine learning.