EFFICIENT DESIGNS OF MULTIPLE SCLEROSIS CLINICAL TRIALS

by

Dean Emile Vrecko
B.Sc., University of British Columbia, 2004

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the School
of
Statistics and Actuarial Science

SIMON FRASER UNIVERSITY
2007

APPROVAL

Name: Dean Emile Vrecko

Degree: Master of Science

Title of project: Efficient Designs of Multiple Sclerosis Clinical Trials

Examining Committee: Dr. Brad McNeney
Chair

Dr. Rachel MacKay Altman
Senior Supervisor
Simon Fraser University

Dr. Joan Hu
Simon Fraser University

Dr. Boxin Tang
Simon Fraser University
External Examiner

Date Approved: _____

Abstract

Multiple sclerosis (MS) is a debilitating disease that attacks the central nervous system. Much research has been conducted to investigate the efficacy of various treatments in reducing the number of active brain lesions in patients, an indicator of disease activity. However, there has been little research regarding the time series nature of these lesion counts.

This project focuses on sample size recommendations for Phase II MS/MRI clinical trials using a longitudinal model. We explore design recommendations based on two estimators. One is based on summary statistics, while the other, $\widehat{T}_{ML}$, uses the time series nature of lesion counts. $\widehat{T}_{ML}$ was found to provide robust sample size recommendations and, over sample size ranges found commonly in current Phase II MS/MRI clinical trials, was a substantial improvement over $\widehat{T}_{POST}$ in terms of sensitivity We further demonstrated that hypothesis tests based on $\widehat{T}_{ML}$ are very powerful even for modest sample sizes.

To my pops. Thanks for everything you have given of yourself for me.

"Ninety percent of all statistics are made up on the spot."

— The ignorant or unbelieving masses

Acknowledgments

First off, I would like to thank God for his love and support throughout my life. I'm sure I don't even realise all you've done for me, but I will try to appreciate it nonetheless.

I would also like to thank the people that make up the Department of Statistics and Actuarial Science here at SFU. Thanks to both the faculty and staff for all the effort in making the time at SFU an enjoyable, yet growing experience. In particular, thanks for the support given to me by Rick Routledge in my first year at SFU, and to Richard Lockhart for his role as graduate chair, a position I like to think I put to the test.

Thanks to my family for being supportive and giving me some much needed grace over the last few years. In particular, thanks, mom and dad, for always being big supporters of education and my happiness - I'm sure I'll never know how much that support has affected who I am and will become.

Thanks also to John Petkau for his involvement with me as a temporary supervisor and as someone who, despite his busy schedule, welcomed me to seek his guidance.

Finally, I would like to thank Rachel. I can't imagine having a more supportive, more understanding, more kind supervisor than Rachel. Her supervision has been a highlight of my time at SFU. She has been relentless in her support of me, both financially and academically. I have come to appreciate her welcoming personality, her sharp mind, and her frankness. I am thankful to have had her as a supervisor and consider her a friend.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Multiple Sclerosis (MS) is a highly unpredictable and often debilitating disease affecting the central nervous system, that is, the brain and spinal cord. The disease attacks the myelin covering that protects the nerve fibers present in the central nervous system causing myelin loss and scarring. The presence of such damaged areas, referred to as *plaques* or *lesions*, disrupts the ability of nerve fibers to conduct electrical impulses to and from the brain. Depending on where these lesions occur, and which nerve impulses are disrupted, a large variety of symptoms are associated with MS, including numbness, double or blurred vision, loss of cognitive functions, loss of balance, disabling fatigue, inability to swallow or to control breathing, pain, or the partial or complete loss of any other function that uses the brain or spinal cord.

There are four main varieties of MS, each characterized by the distinct manner in which clinical symptoms manifest themselves: relapsing remitting (RRMS), secondary progressive (SPMS), progressive relapsing (PRMS), and primary progressive (PPMS). Of these four categories, RRMS is by far the most common variety, and will be the focus of this project. It is characterized by clearly defined periods of exacerbations of symptoms, or relapses, followed by periods of remission where the symptoms experienced during relapse are absent.

Because of the highly heterogeneous nature of the disease, the design of efficient clinical trials can be challenging. One major difficulty when designing such trials is the determination of sample size parameters such as the number of patients and duration of the study. A related issue is determining the scope for which sample size recommendations are valid. Given the importance of such challenges, both with respect to ethical considerations (such

as minimizing the time a placebo group receives an inert treatment when an effective treatment is available) and monetary considerations associated with the costs per patient during the trial, for this project we have chosen to focus on recommending sample size parameters over a broad range of potential patients.

In the past, researchers have generally used clinical responses to diagnose and track MS in patients. However, routine neurological MS examinations tend to be highly subjective and depend heavily on clinical measures such as characterization of symptoms or exacerbation rate; as a result, the ability to accurately diagnose and assess the status of the disease fluctuates with the severity of the symptoms and other subjective outcomes. More recently, non-clinical outcome measures of disease activity, such as magnetic resonance imaging (MRI) outcomes, which are able to detect the lesions associated with MS, have been introduced and are increasingly used to supplement standard clinical examinations.

Since research indicates that lesion severity is correlated with the activity of clinical outcomes ([3], [5]), and that even during periods of remission where no symptoms may be present, lesion activity is consistently active over time, the use of MRI has the potential to provide a more sensitive indicator of disease activity ([5]). It is this added sensitivity that has made MRI a common practice not only in diagnosing MS patients, but also in MS clinical trials, where increased sensitivity could result in shorter trials with fewer patients. Though most studies use several different clinical and MRI measures, for this project we use only one MRI outcome measure, that of *combined unique activity per scan*, which is a count that summarizes the number of active lesions per scan.

MRI outcomes are currently used as primary outcome measures in most Phase II MS clinical trials (Petkau, Personal Communication), but have yet to be accepted in Phase III trials. Thus, this project restricts itself to the use of MRI in designing Phase II MS clinical trials. Phase II trials are designed to assess clinical efficacy of the therapy being investigated and usually consist of patients being randomly assigned to one of two arms: the first arm being the placebo group where patients receive an inert substance and the second arm consisting of those patients who will receive a treatment that is thought to be an effective therapy.

Over the last ten years there has been an increasing interest in developing parametric models for MS/MRI data ([7]). Significant research has been done in parametric modelling of independent MS/MRI lesion counts ([7], [8]) which has provided methods for sample size

recommendations that would not be possible in the absence of such models ([7]). Longitudinal models for lesions counted over time also exist ([1], [2]). However, little is known about designing clinical trials based on a longitudinal model.

In an attempt to design an efficient RRMS/MRI Phase II clinical trial, that is, to design trials that minimize costs by optimizing the number of patients and length of time patients remain in the trial, our project uses the (MRI) outcome of combined unique activity per scan to make sample size recommendations over a reasonable range of patients and treatments. While previous work has focused on using summary outcomes to design such trials, our work focuses on making use of the longitudinal nature of the MRI data. It is hoped that the additional longitudinal information from our chosen estimator will yield additional power for detecting treatment efficacy, while at the same time generating insights on issues concerning the design of MS/MRI trials in general.

For this project we use data from the PRISMS (Prevention of Relapses and Disability by Interferon $\beta - 1a$ Subcutaneously in Multiple Sclerosis) study, which included 560 patients from 22 centres across nine countries. The patients, who were selected for high disease activity based on scans taken prior to treatment, were randomly assigned to three different groups, with one group receiving a placebo and the other groups receiving two different doses of interferon $\beta - 1a$. Of these patients, all 560 received biannual post-treatment scans as well as two pre-treatment scans (scans prior to beginning treatment) as a reference for subsequent scans. Of the 560 patients, 205 underwent additional monthly MRI scanning. This project focuses on the cohort receiving the monthly MRI scans. Given the considerable scope of the PRISMS study, and the rigorous way in which it was conducted, we feel that the design recommendations yielded from this project will be applicable to many future RRMS/MRI Phase II clinical trials.

In chapter 2 we begin by specifying a model for longitudinal count data developed by Altman and Petkau (manuscript in progress) which, over the course of this project, we use to explore two different estimators. We also introduce the two estimators we use extensively in chapters 3 and 4.

Chapter 3 deals with the POST estimator, the first of the two estimators used in this project. This estimator is based on independent summary statistics. We examine the standard deviation of the POST estimator, particularly in reference to its usefulness in designing MS/MRI clinical trials. We also consider how the estimator performs over a broad range of values for important clinical design parameters such as the number of patients

and the number of post-treatment scans (scans taken after treatment has begun). After discussing the POST estmator's robustness of its performance with respect to our model parameters, we draw upon our findings to make some optimal design recommendations for MS/MRI clinical trials.

Similarly, chapter 4 examines the maximum likelihood (ML) estimator. This estimator differs from the POST estimator in that it uses the longitudinal information from the data. This chapter's structure and organization mirrors that of chapter 3, but given the added complexity of a longitudinal estimator, includes some additional discussion regarding subsequent complications.

Chapter 5 compares the two sets of recommendations as well as the power of the two estimators. We also compare our work with the work of Smith ([6]), that is, the work wich this project is building off of. Finally, chapter 6 summarizes the results of this project and suggests future work using the Altman-Petkau model and ML estimator.

Chapter 2

The Setup

The difficulty of modeling an outcome of a disease such as MS is mainly due to the extreme heterogeneity that characterizes the disease. This is apparent when considering the range of clinical symptoms associated with MS as well as their degree of severity. In only considering one very specific and relatively objective outcome of the disease, we have restricted the variability to a single, relatively objective number.

This chapter reviews an intuitive yet statistically sophisticated model due to Altman and Petkau (manuscript in progress) that captures the main features of the PRISMS data. The model is motivated by the idea that the relapsing-remitting nature of RRMS is controlled by some latent process in the body and that such a process can be modeled using a Markov chain. At this time, the model incorporates only post-treatment scans; work is in progress to make use of all available information on the lesion counts.

Our key assumption in this project is that this model not only adequately describes the PRISMS data, but also provides a satisfactory characterization of lesion counts under other (future) treatments. This assumption seems reasonable at least for treatments in the interferon class as these might all be expected to have a similar mechanism or action.

2.1 Basic model description

Let Y_{hit} be the lesion count for patient i $(i = 1, ..., m_h)$ at month t $(t = 1, ..., n_{ih})$ in group h $(h = 1, 2, 3)$, noting that with $t \geq 1$, we are considering only post-treatment scans. Here, the treatments $1, 2, 3$ correspond to placebo (PL), low dose (LD), and high dose (HD) respectively. Let ϵ_{hit} be a latent variable associated with patient i in group h at time t,

where for each h and i, the process $\{\epsilon_{hit}\}_{t=1}^{n_{hi}}$ is assumed to be stationary with $\mathrm{E}[\epsilon_{hit}] = 1$. It is convenient to write $\epsilon_{hit} = e^{a_{Z_{hit}}}$ where Z_{hit} takes on the values of 1 or 2. Altman and Petkau use the notation $\sigma_h^2 \equiv \mathrm{Var}[\epsilon_{hit}]$ and $\gamma_h(|t-s|) \equiv \mathrm{Corr}[\epsilon_{his}, \epsilon_{hit}]$. In addition, let $u_{hi} \sim lognormal(-\frac{1}{2}\lambda_h^2, \lambda_h^2)$ be a patient specific random effect with $\{u_{hi}\}$ being independent and identically distributed while also being independent of $\{\epsilon_{hit}\}$, denoting $\mathrm{Var}[u_{hi}]$ by λ_h^2. Altman and Petkau assume $\mathrm{E}[u_{hi}] = 1$. They also assume that $Y_{hit}|u_{hi}, \epsilon_{hi}$ is Poisson distributed with mean $\mu_{hit}^* = g_h(t)u_{hi}\epsilon_{hit}$, where $g_h(t)$ is a known function of time and treatment. Finally, they assume that $g_h(t)$, u_{hi}, and ϵ_{hit} are non-negative so that μ_{hit}^* is also non-negative. Note that the assumption $\mathrm{E}[\epsilon_{hit}] = \mathrm{E}[u_{hi}] = 1$ was made for both model identifiability and convenience in computing $\mathrm{E}[Y_{hit}]$.

The marginal first and second moments are calculated as follows:

$$\begin{aligned} E[Y_{hit}] &= E[E(Y_{hit}|u_{hi}, \epsilon_{hit})] \\ &= g_h(t) \equiv \mu_{hit} \end{aligned} \tag{2.1}$$

and

$$\begin{aligned} \mathrm{Var}[Y_{hit}] &= E[\mathrm{Var}(Y_{hit}|u_{hi}, \epsilon_{hit})] + \mathrm{Var}[E(Y_{hit}|u_{hi}, \epsilon_{hit})] \\ &= g_h(t) + g_h^2(t)[\sigma_h^2\lambda_h^2 + \sigma_h^2 + \lambda_h^2]. \end{aligned} \tag{2.2}$$

Therefore, the model assumes that the data are overdispersed relative to the Poisson distribution, and that the variance is quadratic in the mean. Also, for $t > s$, we have

$$\begin{aligned} \mathrm{Cov}[Y_{his}, Y_{hit}] &= \mathrm{Cov}[g_h(s)u_{hi}\epsilon_{his}, g_h(t)u_{hi}\epsilon_{hit}] \\ &= g_h(s)g_h(t)\{E[u_{hi}^2\epsilon_{his}\epsilon_{hit}] - 1\} \\ &= g_h(s)g_h(t)[\gamma_h(|t-s|)\sigma_h^2(1+\lambda_h^2) + \lambda_h^2]. \end{aligned} \tag{2.3}$$

In choosing the function $g_h(t)$, we need to consider the nature of our data. Given that the patients for the PRISMS study were selected for high disease activity, we reasoned that the initial scans would yield high counts and that we would expect the subsequent counts of all treatment groups, including the placebo group, to decrease significantly after the study began. For this reason, and for convenience, Altman and Petkau suggest $g_h(t) = e^{\beta_0 + \beta_{h1}t + \beta_{h2}t^2}$. Note that this choice for the mean structure also reflects the idea that an

effective treatment for RRMS should correspond to a decrease in counts over the course of the study.

The parameters of interest in this project are those that depend on treatment. We can see from the specification of $g_h(t)$ the introduction of two model parameters per treatment group: β_{h1} and β_{h2}. While β_{h1} and β_{h2} have no direct clinical interpretation, they do have the practical interpretation as representing the shape of the mean structure of the model, $g_h(t)$. The model parameters associated with u_{hi} are simply the λ_h for each treatment group, which, from its definition above, can easily be seen as representing the inter-patient variability in treatment group h. Finally, the processes $\{\epsilon_{hit}\}$ are characterized by two parameters per treatment group: P_{h1} and P_{h2}. These parameters are two of the transition probabilities for the Markov chain described by ϵ_{hit} and can be thought of as the probability of remaining in relapse or remission respectively. Note that since the rows of the transition probability matrix for a given treatment group sum to 1, we need only to consider these two transition probabilities to specify all four transition probabilities.

To design a clinical trial where we compare a treated group to a placebo group, we need to specify a total of 14 model parameters. However, it seems reasonable to assume that MRI data from the placebo group of the PRISMS trial are fairly representative of typical MRI placebo data from other trials with the same protocol, entry criteria, etc.. With that in mind, we take the maximum likelihood estimates (MLEs) from our model, given in table 2.1, of P_{11}, P_{12}, β_{11}, β_{12}, and λ_1 to be representative of those parameters for the placebo groups of future trials.

h	$\hat{P}_{h1}$	$\hat{P}_{h2}$	$\hat{\beta}_{h1}$	$\hat{\beta}_{h2}$	$\hat{\lambda}_h$
1	0.839585	0.887490	-0.043937	-0.000033	1.513662
2	0.903200	0.904831	-0.262744	0.007714	2.261470
3	0.894576	0.900605	-0.300234	0.012645	5.857623

Table 2.1: Maximum likelihood estimates of model parameters for PRISMS study

Likewise, we assume that the PRISMS trial MLEs of the parameters which do not depend on treatment ($\hat{\beta}_0 = 0.844476$ and $\hat{a}_1 = 0.639148$) are representative of future trials. Thus, for the designs considered in the remainder of this work, we treat only 5 parameters (β_{h1}, β_{h2}, λ_h, P_{h1}, and P_{h2} for the treated group) as variables.

2.2 The POST estimator and its variance

The first estimator we consider is the POST estimator used by Smith (1999) to study the PRISMS data. Using independent summary statistics it provides an estimate of the difference between the mean combined unique activity per scan over the post-treatment scans of the placebo group and a group undergoing treatment, $\mu_{1\cdot\cdot} - \mu_{h\cdot\cdot}$, where h denotes a treated group. Essentially, the POST estimator measures the average difference between the treated group and placebo group in terms of lesion activity.

Assuming, as is the case in most clinical trials of this type, that m_h and n_{hi} are constant across patients and treatment groups ($m_h = m$ and n_{hi} for all h and i), the POST estimator, $\widehat{T}_{POST}$, is easily calculated. Note that

$$\bar{Y}_{hi\cdot} = \frac{1}{n}\sum_{t=1}^{n} Y_{hit}$$

is the ith patient's summary statistic whose expected value is $\mu_{hi\cdot}$. We define

$$\begin{aligned} \widehat{T}_{POST} &= \bar{Y}_{1\cdot\cdot} - \bar{Y}_{h\cdot\cdot} \\ &= \frac{1}{m}\sum_{i=1}^{m} \bar{Y}_{1i\cdot} - \frac{1}{m}\sum_{i=1}^{m} \bar{Y}_{hi\cdot} \end{aligned} \tag{2.4}$$

and therefore $E[\widehat{T}_{POST}] = \mu_{1\cdot\cdot} - \mu_{h\cdot\cdot}$, that is, $\widehat{T}_{POST}$ is an unbiased estimator for the true difference in mean combined unique activity per scan between the placebo and treated group.

We can now easily calculate the variance of $\widehat{T}_{POST}$ by denoting the variance-covariance matrix for a given treated group h as described by equations (2.2) and (2.3) by $\mathbf{V}_h$, and denoting the average of all its entries by $\bar{V}_h$ and then noticing that

$$\begin{aligned} \mathrm{Var}\left(\bar{Y}_{hi\cdot}\right) &= \mathrm{Var}\left(\frac{1}{n}\sum_{t=1}^{n} Y_{hit}\right) \\ &= \frac{1}{n^2}\left[\sum_{t=1}^{n} \mathrm{Var}\left(Y_{hit}\right) + \sum_{r\neq s}^{n} \mathrm{Cov}\left(Y_{hir}, Y_{his}\right)\right] \\ &= \bar{V}_h, \end{aligned}$$

and hence,

$$\begin{aligned}
\text{Var}\left(\widehat{T}_{POST}\right) &= \text{Var}\left[\frac{1}{m}\sum_{i=1}^{m}\bar{Y}_{1i\cdot} - \frac{1}{m}\sum_{i=1}^{m}\bar{Y}_{hi\cdot}\right] \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\text{Var}\left(\bar{Y}_{1i\cdot}\right) + \frac{1}{m^2}\sum_{i=1}^{m}\text{Var}\left(\bar{Y}_{hi\cdot}\right) \\
&= \frac{1}{m}\left[\bar{V}_1 + \bar{V}_h\right]. \qquad (2.5)
\end{aligned}$$

2.3 The ML estimator and its variance

Like the POST estimator, the ML estimator, $\widehat{T}_{ML}$, provides an estimate of $\mu_{1\cdot\cdot} - \mu_{h\cdot\cdot}$, the difference between the mean combined unique activity per scan over the post-treatment scans of the placebo group and a group undergoing treatment. But instead of relying on summary statistics in estimating $\mu_{1\cdot\cdot} - \mu_{h\cdot\cdot}$, as is the case with the POST estimator, $\widehat{T}_{ML}$ is a function of the MLEs of our model and hence uses the longitudinal information in the data.

From equation 2.1 it is easy to see that

$$\bar{\mu}_{h\cdot\cdot} = \frac{1}{n}\sum_{t=1}^{n} g_h(t)$$

and so

$$\begin{aligned}
\widehat{T}_{ML} &= \widehat{\bar{\mu}}_{1\cdot\cdot} - \widehat{\bar{\mu}}_{h\cdot\cdot} \\
&= \frac{1}{n}\sum_{t=1}^{n}\left(\hat{g}_1(t) - \hat{g}_h(t)\right) \equiv f(\hat{\theta}), \qquad (2.6)
\end{aligned}$$

where $\hat{\theta}$ represents the MLEs of the parameters of the model.

Since we cannot compute the variance of $\widehat{T}_{ML}$ analytically, we must settle for an estimate. We will consider two methods for obtaining an estimate of the variance of the ML estimator. The first method will be to estimate the variance using parametric bootstrapping with the model described in section 2.1. We first use the model to produce the sampling distribution of $\widehat{T}_{ML}$ by simulating from the specified model, and then we simply calculate the empirical standard deviation to estimate the SD of $\widehat{T}_{ML}$, denoted as $\text{SD}(\widehat{T}_{ML})$[1].

[1]Note that $\text{SD}(\widehat{T}_{ML})$ is actually an estimated standard deviation. We make this choice for notational simplicity.

In the second method, we appeal to the usual consistency and asymptotic normality properties associated with MLEs. We then apply the delta method to obtain an approximation for the distribution of $\widehat{T}_{ML}$ in order to get an estimate for the asymptotic variance of $\widehat{T}_{ML}$, that is, asymptotically,

$$f(\hat{\theta}) \sim N\left[f(\theta), \left(\frac{\partial f(\theta)}{\partial \theta}\right)^T \mathrm{Var}(\hat{\theta}) \left(\frac{\partial f(\theta)}{\partial \theta}\right)\right].$$

From equation 2.6 we see that

$$\frac{\partial f(\theta)}{\partial \theta} = \begin{pmatrix} \frac{f(\theta)}{\partial \beta_0} \\ \frac{f(\theta)}{\partial \beta_{11}} \\ \frac{f(\theta)}{\partial \beta_{12}} \\ \frac{f(\theta)}{\partial \beta_{h1}} \\ \frac{f(\theta)}{\partial \beta_{h2}} \end{pmatrix} = \frac{1}{n} \sum_{t=1}^{n} \begin{pmatrix} g_1(t) - g_h(t) \\ t g_1(t) \\ t^2 g_1(t) \\ t g_h(t) \\ t^2 g_h(t) \end{pmatrix}.$$

So, the estimated variance of the MLE is

$$\widehat{\mathrm{Var}}\left(\widehat{T}_{ML}\right) = \left(\frac{\partial f(\hat{\theta})}{\partial \theta}\right)^T \widehat{\mathrm{Var}}(\hat{\theta}) \left(\frac{\partial f(\hat{\theta})}{\partial \theta}\right) \tag{2.7}$$

where $\widehat{\mathrm{Var}}(\hat{\theta})$ is the estimated variance-covariance matrix of the model parameter estimates. We denote the estimated standard deviation of $\widehat{T}_{ML}$ obtained through the delta method by $\widehat{\mathrm{SD}}(\widehat{T}_{ML})$. Note that this calculation involves only those model parameters present in $f(\theta)$, namely, β_0, β_{11}, β_{12}, β_{h1}, and β_{h2}.

Chapter 3

The POST Estimator

This chapter looks at the POST estimator, an estimator that relies on independent summary statistics, and the relationship between its standard deviation and the sample size and model parameters. It also makes optimal sample size recommendations for clinical trials where we assume that the trial costs are fixed.

For the duration of this project, we will make the assumptions that, as in the PRISMS design, the numbers of patients in each arm of the study are equal, that is, $m_h = m$ for all h, and the number of scans for each patient are equal, that is, $n_{hi} = n$ for all h and all i. These assumptions are reasonable since they are commonly made in clinical trials.

In section 2.1 we noted that, given our model, we have a total of five model parameters to deal with. Thus there remains a considerable amount of complexity in determining any potential relationships between the standard deviation of our estimator and the model parameters and sample size parameters. However, by applying elementary experimental design theory and taking advantage of some clinical knowledge of the disease, some simplifications become available.

3.1 Designing an experiment

In an effort to reduce the number of model parameters involved in examining the trade-off between n and m in a design, we ran a simple screening experiment in hopes that one or more of the model parameters would be found not to be important. However, given that the model parameters have limited clinical interpretations, designing such an experiment presents a major challenge in determining clinically appropriate ranges for them. Given

the difference between the function of the parameters of the mean structure and that of the parameters of the hidden process, it is not surprising that finding appropriate levels requires a different method for each of the different types of parameters. Note that in the experimental design context, we treat the estimated standard deviation of the POST estimator as the *response*, with the parameters as *factors* and their values as their respective *levels*.

Before proceeding to a discussion of the levels of the model parameters, it is useful, for the sake of clarity, to give a brief description of how the sample size levels were chosen. The levels of the sample size parameters were among the easiest to choose because, unlike the model parameters, time and monetary limitations of clinical trials provide natural constraints. The number of scans, n, has a lower limit that is given by its definition: the smallest number of post-treatment scans is a single scan. We chose our high level for n to be 18 since the duration of typical Phase II MS/MRI clinical trials is generally less than a year and a half. Similarly, the lower bound for the number of patients, m, is also 1, but since most MS/MRI Phase II clinical trials consist of cohorts that are in excess of 10 patients, we chose 10 as a more plausible low level. We took our high level to be $m = 70$ because most such trials have treatment groups that are smaller than those of the PRISMS study which had roughly 70 patients apiece.

It was also relatively easy to specify ranges for the transition probabilities, P_{h1} and P_{h2}, of the latent processes, $\{\epsilon_{hit}\}$. Intuitively, since RRMS is defined by clear periods of relapse and remission, one would expect the transition probabilities corresponding to staying in either relapse or remission to be fairly high, that is, the tendency would be for the process to stay in its current state. This intuition is supported by the ML estimates for P_{h1} and P_{h2}, given in table 2.1. A reasonable low and high, then would be 0.825 and 0.91 respectively, given that the values of the ML parameter estimates did not change much under treatment and these levels cover the entire range of values attained by the ML parameter estimates. Note that a potential difficulty arises if we treat the two transition probabilities separately: an elementary factorial design would produce the pairing of P_{h1+} with P_{h2-}, whose interpretation would be a drug that increases the chances of patients staying in relapse while simultaneously decreasing the chances of staying in remission, that is, a drug that actually worsens the disease it is supposed to help fight. Since this possibility is highly unlikely in any Phase II clinical trial, we eliminate it from the design by treating the transition probabilities as a single factor, $\boldsymbol{P}_h = (P_{h1}, P_{h2})$, with three levels.

Despite the fact that the interpretation of the parameters associated with the mean structure, β_{h1} and β_{h2}, is not straightforward, it nevertheless provides a convenient means to determine reasonable ranges from a clinical standpoint. Regarding β_{h1} and β_{h2} as a single factor, $\boldsymbol{\beta}_h$, as opposed to two separate parameters, lends itself to easily conceiving of a low level: since we would expect any treatment to be at least as effective as a treatment with no efficacy, i.e. a placebo, using $\widehat{\boldsymbol{\beta}}_{PL}$ from the PRISMS study seems like a reasonable choice. Specifically, $g_-(t) \equiv \widehat{g}_{PL}(t)$, that is, $\boldsymbol{\beta}_- \equiv \widehat{\boldsymbol{\beta}}_{PL}$. A conceptually simple means of choosing a high level of $\boldsymbol{\beta}_h$ is to choose a mean structure, a $g_+(t)$, that lies far enough below $g_-(t)$ so that most mean structures of future treatments lie between $g_-(t)$ and $g_+(t)$. It is likely that new treatments being considered for a clinical trials will be on average as effective as the high dose treatment of the PRISMS study (Petkau, Personal Communication), so for our purposes, we take $\boldsymbol{\beta}_0 \equiv \widehat{\boldsymbol{\beta}}_{HD}$ to be a moderately effective treatment, where HD in a subscript denotes the MLE of the HD group from the PRISMS study. Therefore we choose a high level for $\boldsymbol{\beta}_h$ by finding $a \in \mathbb{R}$ such that $g_0(t)$ lies roughly halfway between $g_-(t)$ and $g_+(t)$, where $g_+(t) \approx a g_0(t)$. Ultimately, $a = 0.4$ was used because it produced such a graph, giving $\beta_{h1+} = -0.57$ and $\beta_{h2+} = 0.02$ where $\boldsymbol{\beta}_+ = (\beta_{h1+}, \beta_{h2+})$. It is also important to notice that in using the plot of $\widehat{g}_h(t)$ to choose the levels of $\boldsymbol{\beta}_h$, we have illustrated a simple method for a clinician to graphically select approximate values of β_{h1} and β_{h2}.

The most difficult levels to choose were those of λ_h. Given that λ_h can be thought of as measuring the inter-patient variability in treatment group h and its high degree of variability as demonstrated by the PRISMS study (see table 2.1), it is difficult to determine what range is reasonable for MS/MRI clinical trials in general. However, using the same reasoning as for choosing the levels of the $\boldsymbol{\beta}_h$, we can choose what seem to be appropriate levels for λ_h. We choose the low level of λ_h to be equal to the ML estimate of λ_{PL} for the PRISMS trial, that is, $\lambda_- \equiv \widehat{\lambda}_{PL}$. Similarly, for the high level of λ_h, we treat $\widehat{\lambda}_{HD}$ from the PRISMS study to represent the inter-patient variability for a moderately effective treatment. Taking $\lambda_0 \equiv \widehat{\lambda}_{HD}$, we therefore want to determine a $b \in \mathbb{R}$ such that $\lambda_+ = b\lambda_0$ where $b > 1$. Analyzing the PRISMS study, we noticed that $\widehat{\lambda}_{LD} \approx 2\widehat{\lambda}_{PL}$ and $\widehat{\lambda}_{HD} \approx 2\widehat{\lambda}_{LD}$. Ultimately, we chose the same multiplicative factor when picking λ_+, that is, we chose $b = 2$, therefore $\lambda_+ = 11.8$.

Applying an elementary experimental design to our parameters described above results in a $3 \cdot 2^4$ full factorial experiment. Notice that the response in this experiment, $\mathrm{SD}(\widehat{T}_{POST})$, is deterministic and thus we cannot use a standard ANOVA to determine the effects of the

parameters. Instead, we can only look at the sum squares (SS) of the effects of the factors and their interactions and, from their relative sizes, gauge which factors are important or not. Using the `aov()` function in *R* (see table A.1), we can see that only the transition probabilities were not important, that is, they have little to no effect on the standard deviation of $\widehat{T}_{POST}$. This reduces the number of factors we need to consider when evaluating $\widehat{T}_{POST}$ as an estimator from five factors to four.

3.2 Examining the effects of the parameters on $\mathrm{SD}(\widehat{T}_{POST})$

Now that we have reduced the parameters that we need to examine when considering $\mathrm{SD}(\widehat{T}_{POST})$ to just n, m, λ_h, and $\boldsymbol{\beta}_h$, we can begin investigating their effect on the standard deviation of the POST estimator under different clinical trial design considerations. To do so we use plots of $\mathrm{SD}(\widehat{T}_{POST})$ for various parameter combinations. Notice that both λ_h and $\boldsymbol{\beta}_h$ are model parameters which respectively represent the inter-patient variability and mean structure of combined unique activity per scan, and also depend on the characteristics of the treatment groups of a particular trial. So, in designing a clinical trial we do not have the opportunity to choose and fix their levels explicitly. The other two parameters, n and m, however, are sample size parameters; we have the freedom to specify their levels as we see fit. Thus, the goal of this section is to investigate how the choice of n and m impact $\mathrm{SD}(\widehat{T}_{POST})$ for various values of λ_h and $\boldsymbol{\beta}_h$.

We first investigate the effect of inter-patient variability on $\mathrm{SD}(\widehat{T}_{POST})$. From the screening experiment whose results are tabulated in table A.1, we can see that out of the parameters we are considering, inter-patient variability, or λ_h, has the largest effect on $\mathrm{SD}(\widehat{T}_{POST})$. Its effect on $\mathrm{SD}(\widehat{T}_{POST})$ can be seen graphically by comparing figures 3.1(a-c). Looking at figures 3.1(a-c), and noting that each curve on a plot is for the specified number of patients, we see that as λ_h gets larger, the standard deviation of $\widehat{T}_{POST}$ increases dramatically. This effect is robust over the ranges we have chosen for mean structure, number of scans, and number patients as can be seen in appendix B.1, and in fact, is increasingly evident as $\boldsymbol{\beta}_h$ decreases (which is consistent with the high SS associated with the lambda:beta interaction effect in table A.1.

The screening experiment also gives us some insight into the effect of the mean structure on our response. From section 3.1, recall that we define $\boldsymbol{\beta}_h$ as "low" if it corresponds to a less effective treatment, and $\boldsymbol{\beta}_h$ as "high" if it corresponds to a more effective treatment.

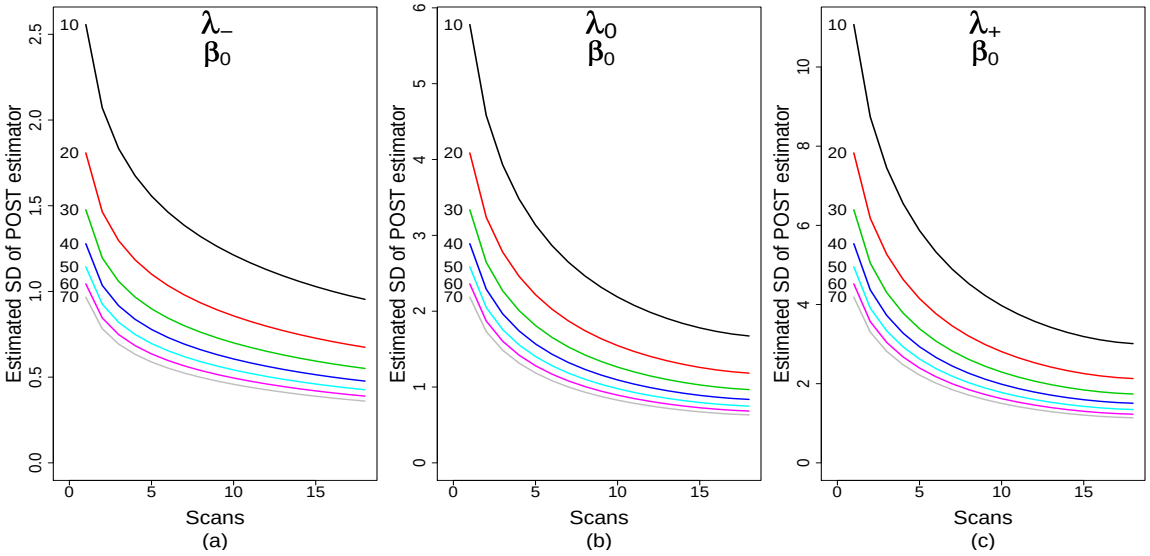

Figure 3.1: SD($\widehat{T}_{POST}$) vs. n for varying inter-patient variabilities: each curve represents the estimated standard deviation curve for the specified number of patients

As table A.1 makes clear, $\boldsymbol{\beta}_h$ has a substantial effect on SD($\widehat{T}_{POST}$). This can also be seen from figures 3.2(a,b) where an increase in $\boldsymbol{\beta}_h$ results in a smaller SD($\widehat{T}_{POST}$) (1-5 correspond to five different levels of $\boldsymbol{\beta}_h$, 1 and 5 representing the least and most effective treatments, respectively). Clinically, we would hope that a more effective treatment would result in a lowered and more homogeneous lesion count, and hence a lowered SD($\widehat{T}_{POST}$). The model does indeed capture this behavior, as can be seen on any plot with varying levels of $\boldsymbol{\beta}_h$ (e.g. Fig 3.2). Furthermore, as is evidenced in the figure in appendix B.1, the effect is robust over the ranges considered for the three remaining parameters. However, from the plots in the appendix we can also note that this effect is more pronounced as λ_h gets larger.

We now consider the effect that the number of scans has on the standard deviation of $\widehat{T}_{POST}$. From figure 3.3(a) we see that the curves for the various number of patients all decrease as n increases. This effect holds true across m, λ_h, and $\boldsymbol{\beta}_h$. Figure 3.3(a) also shows that the curves for the various number of patients all begin with a steep negative slope starting at $n = 1$ and level off substantially at around $n = 5$. This demonstrates that there is little gain in terms of a smaller SD($\widehat{T}_{POST}$) by increasing the number of scans taken in an RRMS clinical trial past five or six when using the POST estimator. The same effect

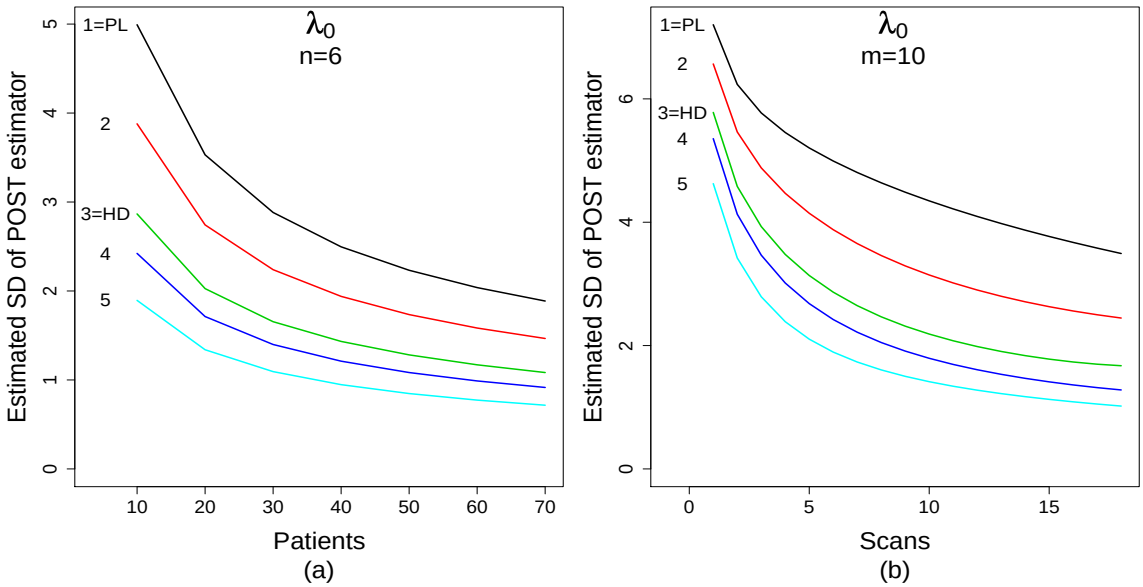

Figure 3.2: SD($\widehat{T}_{POST}$) for varying mean structures: each curve represents the estimated standard deviation curve for a given treatment group with 1 to 5 denoting the mean structures that represent the weakest to strongest treatment effect respectively

can also be seen from figure 3.4(a), which shows that increasing n by more than 5 yields only a relatively small decrease in the standard deviation of $\widehat{T}_{POST}$. From figure 3.3(b) we can see that this result holds true under the ranges considered regardless of which mean structure we are dealing with. Figures 3.1(a-c) demonstrate that this result is robust over the range of λ_h considered as well.

Finally, we examine the effect that the number of patients has on the standard deviation of $\widehat{T}_{POST}$. As with examining the effect that the number of scans has on SD($\widehat{T}_{POST}$), we use figures to determine the relationship between the number of patients and SD($\widehat{T}_{POST}$). From figures 3.4(a,b) it is immediately apparent that increasing the number of patients results in a corresponding decrease in SD($\widehat{T}_{POST}$). In particular, figures 3.4(a,b) show a substantial decrease in the slopes of curves at around $m = 30$, with 3.4(b) indicating that this result is robust over the mean structures we are considering. Similarly, 3.3(a) illustrates this result by examining the ever diminishing decrease in SD($\widehat{T}_{POST}$) that is gained by increasing the number of patients by increments of 10. Referring to figures B.1-B.4, it is easily seen that this result is robust over the range of inter-patient variabilities that we are considering.

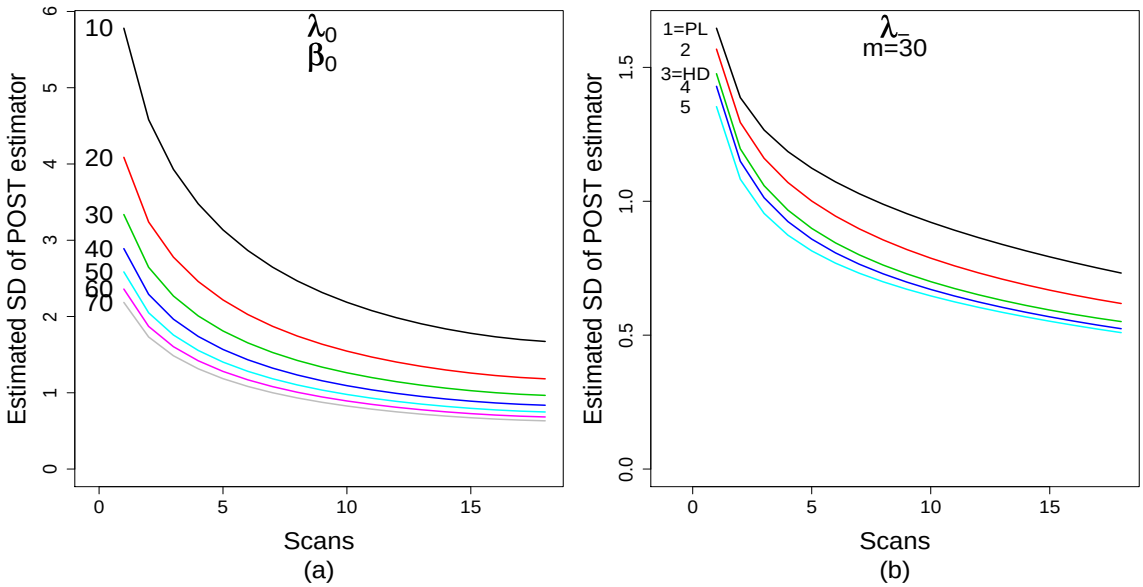

Figure 3.3: SD($\widehat{T}_{POST}$) vs. n for varying numbers of patients, (a), or mean structures, (b)

Notice that table A.1 indicates a clear patient:scan interaction effect. From figures B.2-B.4 we see that as m increases, a change in n produces an ever decreasing change in SD($\widehat{T}_{ML}$). Likewise, as n increases, increasing m has less impact on SD. These effects are robust across $\boldsymbol{\beta}_h$ and λ_h.

Assuming the monetary restriction of a fixed cost for a clinical trial, we briefly consider the trade-off between the number of scans and the number of patients. Since most Phase II MS/MRI clinical trials last longer than 6 months (Petkau, Personal Communication) with at least 20 patients per treatment group ([6]), we will examine the trade-off between increasing scans versus patients beyond $n = 6$ and $m = 20$. An increase of 3 scans per patient with m fixed at 20 is cost equivalent to an increase in 10 patients with n fixed at 6 (i.e. both choices require an additional 60 scans in total). From figures B.13 and B.14 we can see that when $m < 40$ and $n \geq 12$, for nearly every combination of $\boldsymbol{\beta}_h$ and λ_h, choosing to have an additional 10 patients almost always results in greater sensitivity than choosing to have an additional 3 scans per patient. This is not quite the case when we are considering a smaller number of scans per patient, say $n < 12$, with stronger treatment and stronger inter-patient variability combinations. In such cases, the choice between scans and patients reverses, with increasing the number of scans per patient being preferable. However, when $m > 40$, the

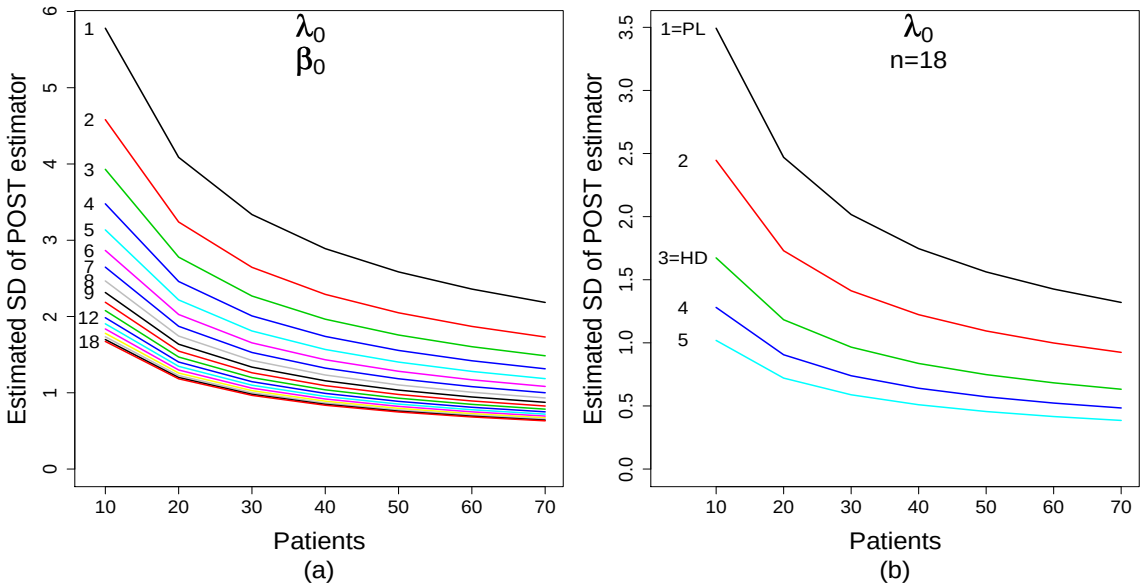

Figure 3.4: SD($\widehat{T}_{POST}$) vs. m for varying levels of scans, (a), or mean structures, (b)

advantageous choice seems to be to increase the number of scans per patient over increasing the number of patients, though the difference can often be small or, at times, negligible.

Thus, based on our model and the POST estimator, we now have some guidelines for appropriately choosing sample size parameters as well as an idea of the robustness to different treatment effects of those choices in order to optimize designs of Phase II MS/MRI clinical trials. To summarize, increasing the number of scans past 5 produces meager gain in terms of sensitivity, as does increasing the number of patients past 30. We also showed that these choices of sample size parameters are robust across the entire range of each parameter we considered, making them ideal candidates in terms of a balance between minimizing the number of scans and patients while maximizing sensitivity, regardless of the expected mean structure and inter-patient variability within the ranges of those parameters we explored. We also saw that the sensitivity of SD($\widehat{T}_{POST}$) is susceptible to changes in the model parameters, which represent population and treatment characteristics. Finally, we have commented on the case where there is a fixed budget, and a choice must be made as to whether to recruit more patients or to have more scans per patient.

Chapter 4

The Maximum Likelihood Estimator

This chapter examines the maximum likelihood estimator, $\widehat{T}_{ML}$, which, unlike the POST estimator, is an estimator that depends on longitudinal information. As in chapter 3, we investigate any potential relationships that exist between the estimated standard deviation of the estimator and both the sample size and model parameters, but also summarize some of the difficulties encountered using $\widehat{T}_{ML}$ with our choice of model. Though the manner in which we examine the ML estimator in this chapter will have many similarities to the approach used to explore the POST estimator in chapter 3, there are some minor differences in how we analyze the estimator.

For this chapter we will use a similar approach to investigate the same seven[1] parameters that we investigated in chapter 3, namely, $\mathbf{P}_h$, $\boldsymbol{\beta}_h$, λ_h, m, and n. Recall from chapter 3 that we have made the assumptions of equal numbers of patients in each arm ($m_h = m$ for all h) and equal numbers of scans for each patient ($n_{hi} = n$ for all h and i). We compute $\widehat{T}_{ML}$ and its standard error using a quasi-Newton routine (e.g., Nash 1979). This method has the added benefit of producing, as a by-product, the estimated variance-covariance matrix of the parameter estimates (as given by the inverse of the observed Fisher information matrix).

[1]Recall that both $\mathbf{P}_h$ and $\boldsymbol{\beta}_h$ are vectors.

4.1 Computational details

Due to the complexity added from considering an estimator based on a longitudinal model, there arise some computational problems. In particular, the complexity of the model along with the number of parameters to be estimated sometimes causes convergence problems when attempting to determine maximum likelihood estimates of the parameters.

One of the reasons for convergence problems seems to be due to the chosen mean structure. The mean structure was chosen to accurately reflect how lesion counts behave under treatment, but such behavior is quite difficult to deal with computationally. For example, under an effective treatment, lesion counts tend to fall sharply to zero after the very first post-treatment scan. Capturing this effect in the estimate of $g_h(t)$ is challenging for small sample sizes. Another cause of convergence issues seems to be the large number of zero lesion counts produced when a group is under treatment: as with the PRISMS data, even a moderately effective treatment, as defined in section 3.1, produces mostly zeros for the vast majority of patients. With so little data, estimation of a relatively large number of parameters quickly becomes difficult. It thus appears that the complexity of the model is both an advantage and a disadvantage: it captures the difficult structure of the data reasonably well, but moderate sample sizes are required in order to compute the MLEs of its parameters.

A simple solution is available for the difficulties identified above, namely, to decrease the range of the sample size parameters. By considering larger values of m and n, we were able to compute the MLEs without problems.

Unfortunately, even though the modified ranges of n and m produced reasonable estimates for the model parameters, they still did not always produce acceptable estimates for the variance-covariance matrix of the model parameters (which would be required to determine an estimate for the standard deviation of $\widehat{T}_{ML}$ using the delta method). Although increasing the lower bounds on n and m further would result in adequate estimates of the standard deviation of $\widehat{T}_{ML}$, we chose instead to use the bootstrapped estimate of the SD, $\mathrm{SD}(\widehat{T}_{ML})$, and hence to maintain a greater scope for recommendations.

4.2 Designing another experiment

Given the aforementioned challenge presented by our choice of model and estimator, any appropriate reduction of complexity is both useful and welcome. In this section we attempt such a simplification by using a screening experiment similar to chapter 3 in the hopes that one or more parameters will be found to have little effect on our response of interest, $\mathrm{SD}(\widehat{T}_{ML})$. This screening experiment has the same number of parameters and corresponding number of levels to consider as the experiment used to examine the POST estimator. Likewise, the interpretations for these parameters, the setup, and notation used in chapter 3 are unchanged.

While the previous chapter considered an estimator based solely on independent summary statistics, this chapter deals with an estimator that takes advantage of the additional information provided by the time series nature of the counts. As mentioned in section 4.1, our choices of estimator and model require some additional care including reducing the range of our model parameters. Ultimately, we found that changing from low levels $n = 1$ and $m = 10$ of the POST estimator to $n = 6$ and $m = 20$ for the low levels of the ML estimator was sufficient to produce acceptable parameter estimates. While it may seem that the ranges of sample size parameters for the ML estimator are restrictive, as mentioned in chapter 3, current Phase II MS/MRI clinical trials are commonly longer than 6 months in duration, with monthly scans usually taken for up to a year and, typically, have at least 20 patients per treatment arm.

With the levels of our parameters fixed, we are again left with a $3 \cdot 2^4$ full factorial experiment. Although we could have performed the standard ANOVA by dividing the total number of simulations to form groups of replicates, the number of simulations[2] required to have reliable estimates for the standard deviation of $\widehat{T}_{ML}$ for each replicate was prohibitive. Using the methods of chapter 3 and the ANOVA table A.2, we can see that the transition probabilities are not important. So, again, we exclude them from our consideration of parameter effects. Table A.2 also suggests that $\boldsymbol{\beta}_h$ may not be important. However, as a conservative measure, we chose to include the mean structure in our analysis, since some of the interactions involving $\boldsymbol{\beta}_h$ are substantial.

[2]Each simulation provides an estimate of T_{ML}, whereas for each SD estimate, $\mathrm{SD}(\widehat{T}_{ML})$, we used between 750 and 1500 simulations.

4.3 Examining the effects of the parameters on $\mathrm{SD}(\widehat{T}_{ML})$

As done previously, we first examine the effects of the model parameters on the estimated standard deviation of the ML estimator and consider the robustness of these effects. We subsequently do the same with sample size parameters, before making some recommendations for future trials based on our findings.

We begin by looking at the effect of inter-patient variability, or λ_h, on the estimated standard deviation of the ML estimator. By comparing figures 4.1(a-c), we first notice that the general shape of the plots remain the same across the different levels of λ_h. It is also clear that changing the level of λ_h has little to no impact of the value of $\mathrm{SD}(\widehat{T}_{ML})$. Examining the plots in section B.2 reveals that this also appears to be the case for any $\boldsymbol{\beta}_h$. The fact that λ_h does not have much of an influence on the magnitude of $\mathrm{SD}(\widehat{T}_{ML})$ is not surprising if we consider its associated sum of squares in table A.2. This value is relatively small when compared to those of the number of scans or patients.

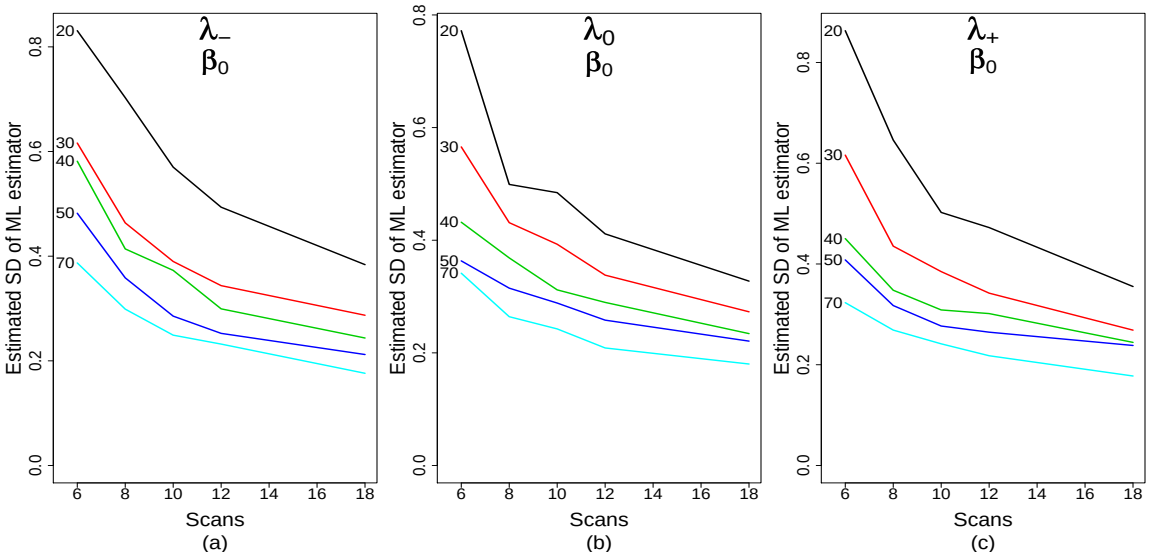

Figure 4.1: $\mathrm{SD}(\widehat{T}_{ML})$ vs. n for varying inter-patient variabilities: each curve represents the estimated standard deviation curve for the specified number of patients

If we consider the $\boldsymbol{\beta}_h$ parameter, we are left with a similar conclusion to that of λ_h. Figures 4.2(a,b) reveal a decreasing estimated standard deviation for the ML estimator for increasing m or n, but it seems that the influence that $\boldsymbol{\beta}_h$ has on $\mathrm{SD}(\widehat{T}_{ML})$ is not strong enough to overcome any possible noise that may be present in our estimates of the standard

deviation of $\widehat{T}_{ML}$ in any well defined way. Every plot with curves representing either number of patients or scans has the same general shape, exhibiting a gentle negative slope, and sits roughly within the same range of standard deviation. The relative inefficacy of the various mean structures in changing the estimated standard deviation of the ML estimator is also reflected in the figures 4.2(a,b) and others where the curves represent a mean structure: the curves of various mean structures have no vertical order. That is, a particular mean structure does not consistently produce a smaller or larger standard deviation than another mean structure. Note that again, this is should not come as a surprise if we recall the results in table A.2 and the relatively small SS value for $\boldsymbol{\beta}_h$.

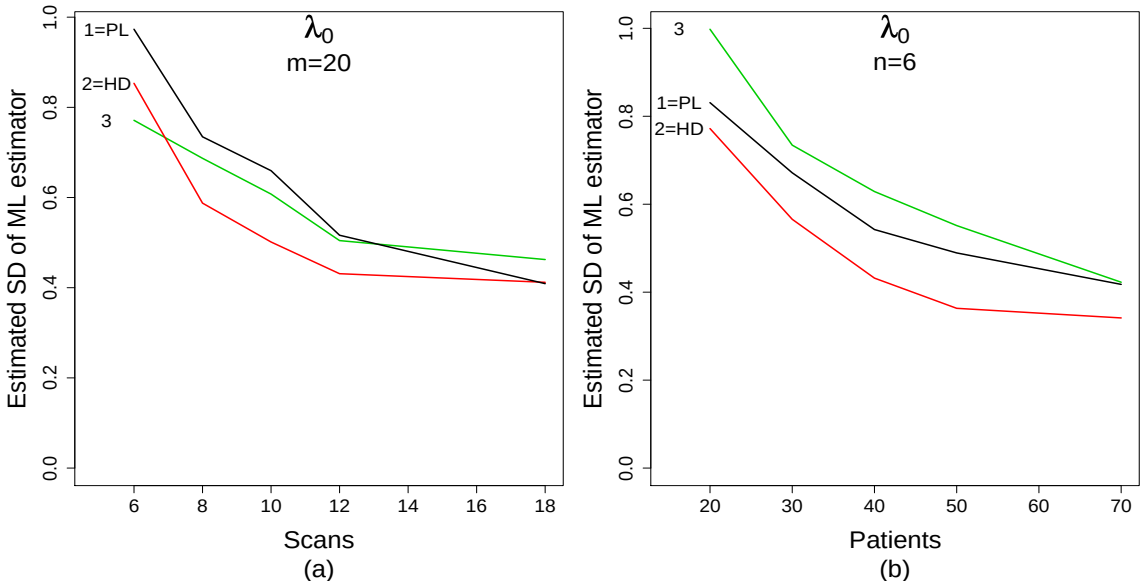

Figure 4.2: SD($\widehat{T}_{ML}$) for varying mean structures: each curve represents the estimated standard deviation curve for a given treatment group with 1 to 3 denoting the mean structures that represent the weakest to strongest treatment effect respectively

We now consider how the number of scans affects the estimated standard deviation of the ML estimator. Looking at figures 4.3(a,b) it is immediately apparent that increasing n results in a decrease in our response. However, in terms of making recommendations based on trying to minimize SD($\widehat{T}_{ML}$), there is considerable noise present. This makes it somewhat difficult to tell which number of scans would be optimal. Looking at figure 4.3(a), substantial gains are evident when the number of scans is increased from $n = 6$ to $n = 8$, but increasing n from 8 to 10 also yields a generous decrease in SD($\widehat{T}_{ML}$). Figures 4.1(a-c) demonstrate the same behavior, which aside from lending credence to our conclusions,

also demonstrates their robustness across inter-patient variability. Assuming that a larger vertical white space between two curves indicates a larger change in SD($\widehat{T}_{ML}$), figure B.18 shows a noticeably larger amount of space on average between the $n = 8$ and $n = 10$ curves when compared to the space between the $n = 10$ and $n = 12$ curves. Thus it seems that $n = 10$ provides the optimal number of scans in terms of minimizing the estimated standard deviation of the ML estimator. The effect of n and the corresponding recommendations can be seen to be robust over the ranges of $\boldsymbol{\beta}_h$ we have discussed by looking at figure 4.3(b) and by comparing the plots in figures B.15-B.18.

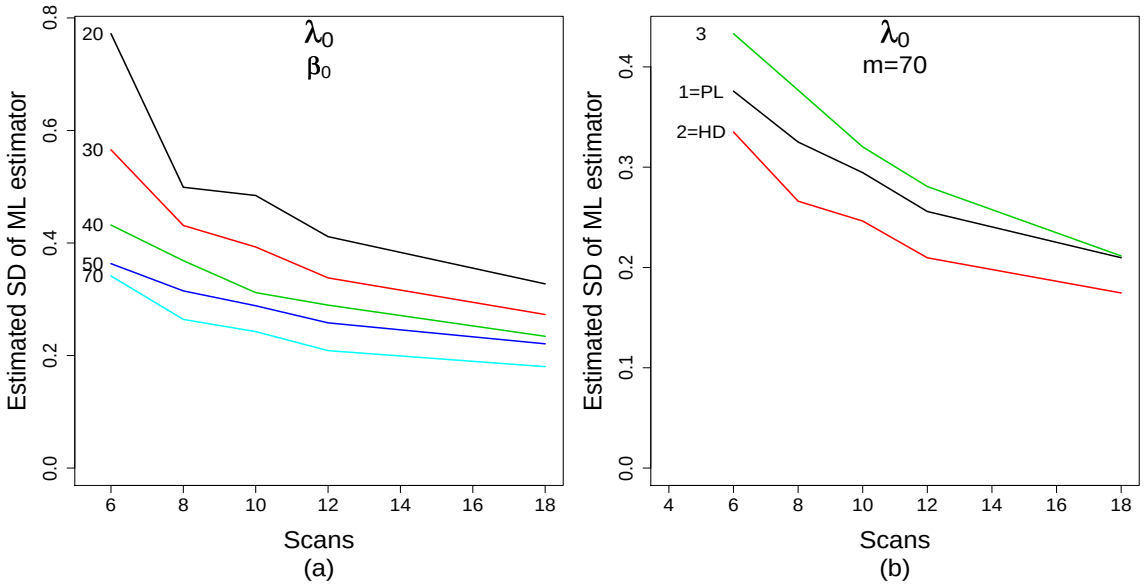

Figure 4.3: SD($\widehat{T}_{ML}$) vs. n for varying levels of patients, (a), or mean structures, (b)

Similar to the effect of number of scans, an increase in the number of patients decreases SD($\widehat{T}_{ML}$). Figures 4.4(a,b) show a decreasing trend as m increases. The slopes of the curves level off when $m \geq 40$, demonstrating that increasing m past 40 does not yield much increase in terms of sensitivity. Again using the idea of space between curves, figures 4.1(a-c) indicate this conclusion to be true and independent of λ_h. This optimal choice of m can be seen to be robust across $\boldsymbol{\beta}_h$ by looking at figures 4.4(b), 4.3(b), or 4.2(b). Like the case of SD($\widehat{T}_{POST}$), it seems there is a fair sized scans:patients interaction effect when considering SD($\widehat{T}_{ML}$). Given a small value of m, SD($\widehat{T}_{ML}$) is more heavily influenced by a change in the value of n than if we considered a larger value of m. This can be seen graphically in figures B.15-B.18 by noticing that as m increases, increasing n produces diminishing returns

in terms of SD($\widehat{T}_{ML}$). Likewise, as n increases, increasing m has less impact on SD($\widehat{T}_{ML}$). These effects are robust across $\boldsymbol{\beta}_h$ and λ_h.

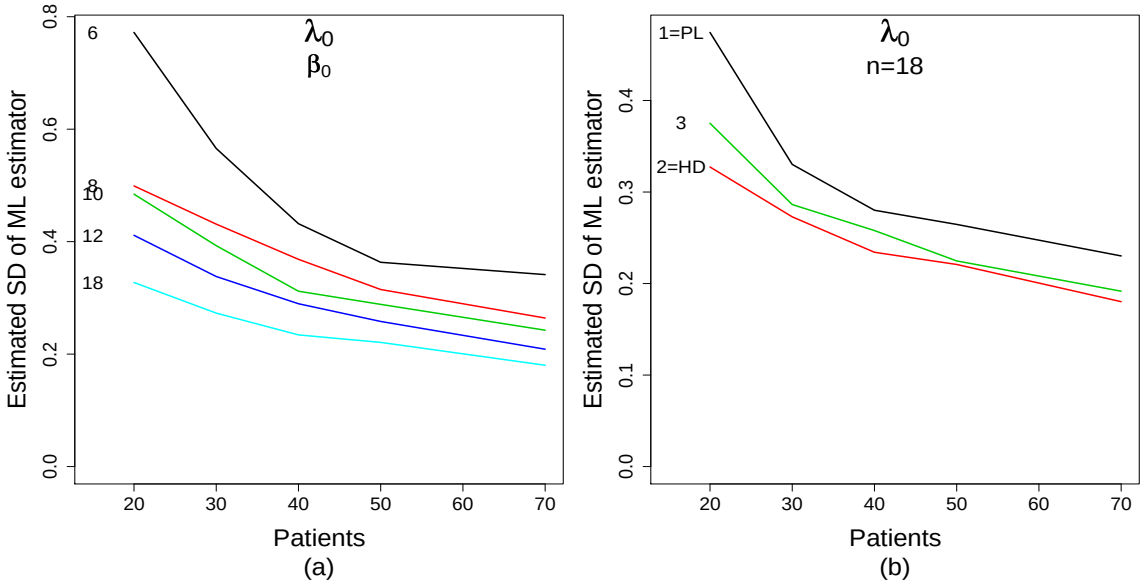

Figure 4.4: SD($\widehat{T}_{ML}$) vs. m for varying levels of scans, (a), or mean structures, (b)

Similar to chapter 3, we analyze the practical situation where there is a fixed budget for a clinical trial, and the trade-off between number of patients and number of scans per patient must be considered. Unfortunately, there is too much noise to make any straightforward and specific conclusions. From figures B.16 and B.18 we can see that similar to the analysis for the POST estimator, the choice would generally depend on the level combination of the model parameters, but for this case no clear general pattern emerges.

In summary, even with the added complexity of working with simulated data from a sophisticated model, some optimal design recommendations are still available: an ideal number of scans would appear to be roughly 10. This choice simultaneously minimizes the estimated SD of the MLE and the cost associated with the number of MRI scans per patient. Moreover, it appears robust across all parameters. And in terms of simultaneously minimizing SD($\widehat{T}_{ML}$) and the overall price per patient, the optimal number of patients seems to be around 40 - a conclusion that is also robust over the parameters we considered within their respective ranges. This chapter also demonstrated that $\widehat{T}_{ML}$ is highly robust to the model parameters.

Chapter 5

Comparing The Estimators

This chapter compares the recommendations based on the POST and ML estimators and evaluates their performance as estimators in relation to their sensitivity and power. It also compares our recommendations for Phase II RRMS/MRI clinical trials with those made by Smith.

5.1 Comparing the sensitivities of and the power based on the estimators

The ML estimator provides great gains in terms of sensitivity over the POST estimator. This is evident when comparing figures 5.1(a,b) below. The curves representing the standard deviation for a fixed λ_h and $\boldsymbol{\beta}_h$ combination over the MLEs sample size parameter ranges are displayed for both the POST and ML estimator. From these figures it is clear that there is a substantial advantage in sensitivity when using $\widehat{T}_{ML}$. The SD range for the curves corresponding to the POST estimator in figure 5.1(a) is between 0.6 and 2.1, whereas the same curves for the ML estimator in figure 5.1(b) lie between 0.3 and 0.8. Though the difference in sensitivity between the two estimators can be greater and smaller than that depicted in figures 5.2(a,b), figures B.19-B.24 show that the difference is substantial and almost always favors the ML estimator.

Looking further at the figures in appendix B.2, the superior sensitivity of the ML estimator appears to be due, in part, to how robust its estimated standard deviation is to the model parameters. Looking at figures B.19-B.24, we see that the SD ranges for the curves

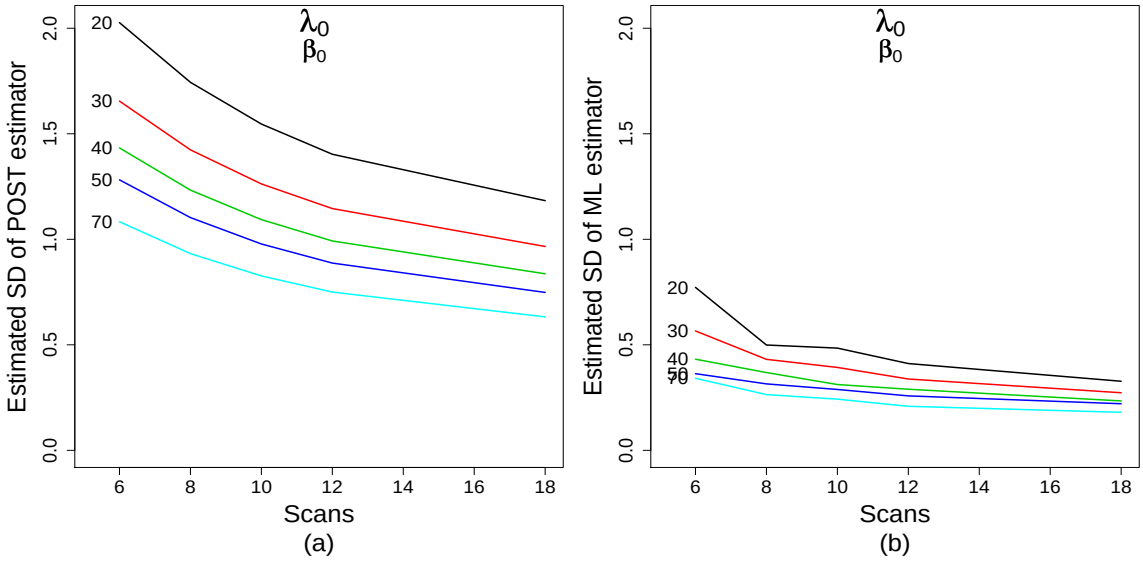

Figure 5.1: Estimated standard deviation vs. n for varying levels of patients

corresponding to $\widehat{T}_{POST}$ change for different levels of λ_h and $\boldsymbol{\beta}_h$, whereas for $\widehat{T}_{ML}$ the SD ranges stay generally constant between 0.2 and 1.0, independent of what underlying mean structure or inter-patient variability is present. This feature of $\widehat{T}_{ML}$ is clearly advantageous, since then precise estimates of λ_h and $\boldsymbol{\beta}_h$ are not required in the planning of a trial.

Note that there is one case that we considered where the sensitivity of the POST estimator actually beats that of the ML estimator, that is, a case where $\mathrm{SD}(\widehat{T}_{POST}) < \mathrm{SD}(\widehat{T}_{ML})$. From figure B.21 we see that the POST estimator performs better than $\widehat{T}_{ML}$ when our model parameter levels are set at λ_- and $\boldsymbol{\beta}_+$. This discrepancy is likely due to errors in the estimation of $\mathrm{SD}(\widehat{T}_{ML})$ caused by too many zeros that result from a highly effective treatment being coupled with a relatively uniform inter-patient variability. (See also figure B.15(g), which shows that $\mathrm{SD}(\widehat{T}_{ML})$ is much greater for this case than for the others considered). Fortunately this situation looks better for higher sample sizes where the SD values are similar to other cases. In any event, this situation is unlikely to arise in a clinical trial since, according to the PRISMS data, the efficacy of the treatment exhibits a strong positive correlation with inter-patient variability.

The relative sensitivities of the estimators are reflected in their power to detect a treatment effect. Using our simulation data, we compute estimates of the power for a variety

of parameter values, assuming a significance level of 0.05. When $\boldsymbol{\beta}_h = \boldsymbol{\beta}_0$ (i.e. there is a moderate treatment effect), the power differs considerably depending on whether we use the POST or ML estimator. Tables A.3-A.5 show that the power estimates for $\widehat{T}_{ML}$ are always substantially better than the power estimates for $\widehat{T}_{POST}$ with a power of at least 80% the majority of the time over the level combinations we considered. And while the performance of $\widehat{T}_{POST}$ improved with a decrease in inter-patient variability, even at its best, the power based on $\widehat{T}_{POST}$ exceeds 80% only once.

Though we were not able to estimate the power for all level combinations of $\boldsymbol{\beta}_h$ and λ_h[1], we were able to gain some insight into the performance of $\widehat{T}_{ML}$ over a fairly broad range of the model parameters. From tables A.3-A.5 we see that when $m \geq 30$, only 10 scans are needed to detect a treatment effect with a power of at least 80%. In fact, if there are 40 or more patients, it is likely that even fewer scans are required for even higher power. However, if only 20 or fewer patients are available, a perhaps prohibitively large number of scans would be required to achieve a reasonable level of power.

Number of patients	Number of scans	Estimated power based on $\widehat{T}_{POST}$	Estimated power based on $\widehat{T}_{ML}$
20	8	0.142	0.642
	10	0.162	0.680
	12	0.179	0.798
30	8	0.173	0.752
	10	0.202	0.850
	12	0.225	0.915
40	8	0.203	0.859
	10	0.239	0.949
	12	0.269	0.970
50	8	0.232	0.938
	10	0.275	0.983
	12	0.310	0.989
70	8	0.284	0.984
	10	0.342	0.995
	12	0.388	0.999

Table 5.1: Power estimates for POST and ML estimators for $\boldsymbol{\beta}_0$ and λ_0

[1]Power estimates were calculated only when appropriate to do so, that is, when the distribution of $\widehat{T}_{ML}$ was approximately normal.

5.2 Comparing recommendations based on our estimators

Despite having substantially less sensitivity than the ML estimator, the POST estimator may still be preferred by some clinicians due to its simplicity. For such clinicians there is an additional benefit to be gained from using $\widehat{T}_{POST}$: its associated optimal sample sizes are relatively small. Summarizing our results from chapters 3 and 4 in table 5.2, at first glance, it appears that the POST estimator has an advantage over $\widehat{T}_{ML}$ since the "optimal" sample size recommendations associated with $\widehat{T}_{POST}$ require half the number of scans and one quarter less patients than those associated with $\widehat{T}_{ML}$. However, it is also important to note that these recommendations are based on different ranges. Therefore, considering a different range for the parameters could conceivably yield different recommendations. This is indeed the case here. Figures 5.2(a,b) show the plotted curves for the standard deviation of the POST estimator under the ranges of the sample size parameters for the ML estimator. Based on these and the related plots in figures B.19-B.24, the recommendations for the POST estimator are very similar to those made for the ML estimator.

	POST Estimator	ML Estimator
Range of n considered	1-18	6-18
Optimal choice for n	5	10
Range of m considered	10-70	20-70
Optimal choice for m	30	40

Table 5.2: Sample size recommendations based on $\widehat{T}_{POST}$ and $\widehat{T}_{ML}$

5.3 Comparing our results with Smith's work

In 1999, Smith created a semi-parametric model for MS lesion count data and used three estimators of the treatment effect, one of which was the POST estimator, and two that incorporated pre-treatment scan data. Smith compared the sensitivity of the three estimators and made sample size recommendations based on the most sensitive of the three. Unfortunately, since the model we considered does not incorporate pre-treatment scans, we are unable to compare our estimators directly. However, we are able to make a rough comparison of Smith's most promising estimator, the ANCOVA estimator, used with his model, with the ML estimator used with Altman and Petkau's model. We can also compare Smith's recommendations based on his ANCOVA estimator with our recommendations

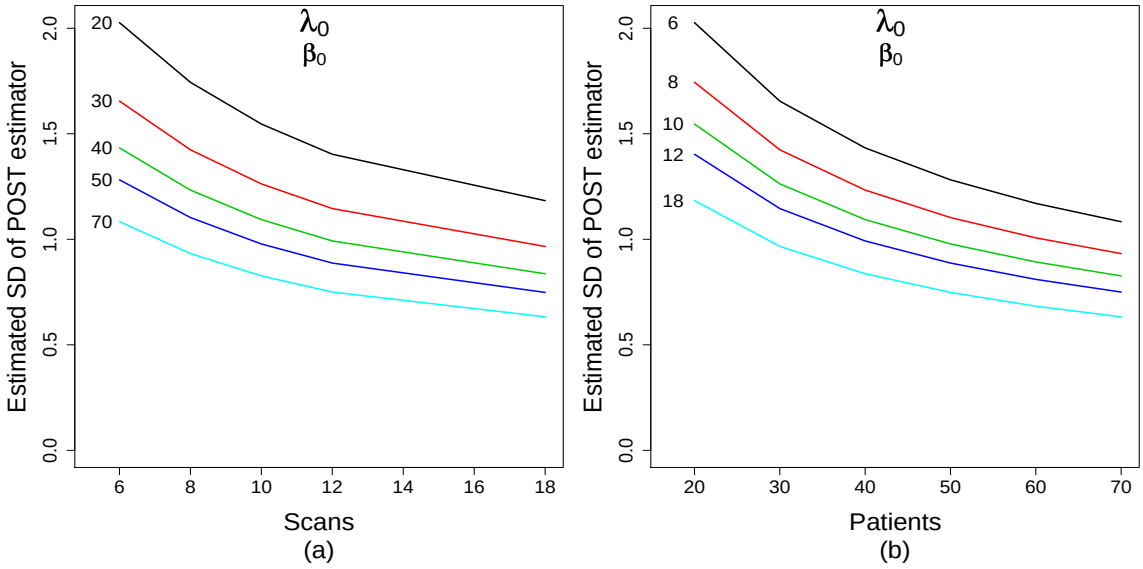

Figure 5.2: SD($\widehat{T}_{POST}$) vs. n for differing numbers of patients (a) and SD($\widehat{T}_{ML}$) vs. m for differing numbers of scans (b)

based on $\widehat{T}_{ML}$. It is important to note that comparing the ANCOVA and ML estimators is not an entirely fair comparison: since we use two different models of the treatment effect, the true values of the two estimators are inherently different. Nonetheless, a study of the the sensitivities of the estimators provides some insight.

Pre-treatment scans	Post-treatment scans	SD($\widehat{T}_{ANCOVA}$)	SD($\widehat{T}_{ML}$)
0	6	-	0.36
	9	-	0.27
1	6	0.24	-
	9	0.22	-
2	6	0.22	-
	9	0.20	-
3	6	0.21	-
	9	0.19	-
4	6	0.21	-
	9	0.18	-

Table 5.3: SD for ANCOVA and ML estimators for the treated group with $m = 60$

For the ANCOVA estimator using Smith's model, the estimated standard deviation

varies depending on the number of pre-treatment scans according to tables 5.3 and 5.4. It seems that the sensitivity of Smith's preferred estimator is decidedly greater when the number of patients is fixed at 60 and the number of post-treatment scans varies between $n = 6$ and $n = 9$ or when the number of post-treatment scans is fixed at 6 and the number of patients varies between $m = 40$ and $m = 70$.

Pre-treatment scans	Number of patients	SD($\widehat{T}_{ANCOVA}$)	SD($\widehat{T}_{ML}$)
0	40	-	0.43
	70	-	0.34
1	40	0.30	-
	70	0.22	-
2	40	0.27	-
	70	0.21	-
3	40	0.26	-
	70	0.20	-
4	40	0.25	-
	70	0.19	-

Table 5.4: SD for ANCOVA and ML estimators for the treated group with $n = 6$

As for recommendations, since Smith only makes optimal design recommendations based on the ANCOVA estimator with a fixed number of patients, the value of comparisons with the design recommendations we made based on the ML estimator will be limited. We are still able to make a crude comparison of the recommendations based on the competing estimators. Using his model and the ANCOVA estimator with the number of patients fixed at 60, Smith's optimal recommendation is to fix the number of post-treatment scans at 7, whereas for the ML estimator and our model, with any number of fixed patients we recommend 10 post-treatment scans.

Chapter 6

Conclusions and Future Work

Notwithstanding challenges brought about by a complex model and estimator, it seems the longitudinal information used by $\widehat{T}_{ML}$ results in an estimator that outperforms the summary statistic-based POST estimator in terms of sensitivity. Our estimator is also far more robust to the wide ranges of the model parameters that we explored – model parameters that reflect population and treatment characteristics encountered when performing RRMS/MRI clinical trials in practice. In addition, the optimal sample size recommendations for a fixed budget made based on the ML estimator fall within standard sample sizes for current Phase II MS/MRI clinical trials.

Since Smith focused largely on the effect of pre-treatment scans and the ML estimator considers only post-treatment scans, the comparison of the ML estimator with the ANCOVA estimator was necessarily indirect. The sensitivity of $\widehat{T}_{ANCOVA}$ currently proves to be a substantial improvement on the POST estimator and, as section 5.3 points out, on the ML estimator as well. These results suggest that the ML estimator based on a model which incorporates pre-treatment scan information would likely be highly sensitive. Such a model would also allow for a more direct comparison of Smith's ANCOVA estimator with the ML estimator.

There are some limitations with our choices of model and the ML estimator. The convergence problems mentioned in section 4.1 complicate estimation of the model parameters, and hence the treatment effect and its standard deviation, unless the sample size parameters are sufficiently large. Throughout this project the estimation of model parameters has been a considerable challenge. This is in part due to the choice of our mean structure, $g_h(t)$. Although our current choice for $g_h(t)$ seems to capture the basic behavior of the lesion count

data, as alluded to in section 4.1, our choice of mean structure was difficult to work with and is perhaps overly sensitive to changes in values of β_{h1} and β_{h2}. With a different choice for $g_h(t)$, it may be easier to estimate the parameters of the model. In particular, further investigation of the reliability of the estimated asymptotic standard deviation is required. Since the method we used to find the MLEs (Quasi-Newton routine) can sometimes produce a negative-definite estimate of the variance-covariance matrix, a different maximization method might produce better estimates of the asymptotic standard deviation. This is desirable since a comparison between the empirical bootstrapped estimates and asymptotic estimates could provide some insight.

Despite the challenges that have surfaced over the course of this project, the ML estimator shows great promise for use in MS/MRI phase II clinical trials and provides a sizable step towards an improved means of investigating multiple sclerosis.

Appendix A

Data

A.1 Experimental design data from ANOVA tables

This section contains the results, as given from the `aov()` function in *R*, of the experimental designs for each of the estimators from chapter 3 and 4. The designs were both full factorial experiments. The tables have been cut and pasted from *R* but have not been modified in any way other than formatting.

	Df	Sum Sq	Mean Sq
scans	1	136.060	136.060
patients	1	130.155	130.155
betas	1	63.699	63.699
lambda	1	275.143	275.143
P	2	0.489	0.244
scans:patients	1	27.726	27.726
scans:betas	1	0.001	0.001
patients:betas	1	12.980	12.980
scans:lambda	1	64.828	64.828
patients:lambda	1	56.068	56.068
betas:lambda	1	48.489	48.489
scans:P	2	0.159	0.080
patients:P	2	0.100	0.050
betas:P	2	0.075	0.038
lambda:P	2	0.341	0.171
scans:patients:betas	1	0.0003	0.0003
scans:patients:lambda	1	13.210	13.210
scans:betas:lambda	1	0.013	0.013
patients:betas:lambda	1	9.881	9.881
scans:patients:P	2	0.032	0.016
scans:betas:P	2	0.010	0.005
patients:betas:P	2	0.015	0.008
scans:lambda:P	2	0.113	0.057
patients:lambda:P	2	0.070	0.035
betas:lambda:P	2	0.045	0.023
scans:patients:betas:lambda	1	0.003	0.003
scans:patients:betas:P	2	0.002	0.001
scans:patients:lambda:P	2	0.023	0.012
scans:betas:lambda:P	2	0.006	0.003
patients:betas:lambda:P	2	0.009	0.005
scans:patients:betas:lambda:P	2	0.001	0.001

Table A.1: ANOVA table of the experimental design for the SD of the POST estimator

	Df	Sum Sq	Mean Sq
scans	1	8.9484	8.9484
patients	1	15.5054	15.5054
betas	1	1.3851	1.3851
lambda	1	5.5663	5.5663
P	2	0.3368	0.1684
scans:patients	1	8.0975	8.0975
scans:betas	1	1.2341	1.2341
patients:betas	1	1.6077	1.6077
scans:lambda	1	3.1012	3.1012
patients:lambda	1	5.7467	5.7467
betas:lambda	1	2.6181	2.6181
scans:P	2	1.0850	0.5425
patients:P	2	0.3146	0.1573
betas:P	2	0.2240	0.1120
lambda:P	2	0.7930	0.3965
scans:patients:betas	1	1.2420	1.2420
scans:patients:lambda	1	3.1400	3.1400
scans:betas:lambda	1	1.3005	1.3005
patients:betas:lambda	1	2.3528	2.3528
scans:patients:P	2	1.0939	0.5470
scans:betas:P	2	0.1470	0.0735
patients:betas:P	2	0.2385	0.1193
scans:lambda:P	2	1.8825	0.9412
patients:lambda:P	2	0.7751	0.3875
betas:lambda:P	2	0.7787	0.3894
scans:patients:betas:lambda	1	1.2364	1.2364
scans:patients:betas:P	2	0.1498	0.0749
scans:patients:lambda:P	2	1.8537	0.9268
scans:betas:lambda:P	2	1.2313	0.6156
patients:betas:lambda:P	2	0.7479	0.3740
scans:patients:betas:lambda:P	2	1.2241	0.6121

Table A.2: ANOVA table of the experimental design for the SD of the ML estimator

A.2 Power comparison tables for POST and ML estimators

This section contains the full results of the power comparison from chapter 5.

Level combination of model parameters	Number of patients	Number of scans	Estimated power based on $\widehat{T}_{POST}$	Estimated power based on $\widehat{T}_{ML}$
$\boldsymbol{\beta}_0$, λ_-	20	6	0.233	0.306
$\boldsymbol{\beta}_0$, λ_-	20	8	0.284	0.413
$\boldsymbol{\beta}_0$, λ_-	20	10	0.326	0.559
$\boldsymbol{\beta}_0$, λ_-	20	12	0.357	0.663
$\boldsymbol{\beta}_0$, λ_-	20	18	0.383	0.765
$\boldsymbol{\beta}_0$, λ_-	30	6	0.300	0.486
$\boldsymbol{\beta}_0$, λ_-	30	8	0.371	0.697
$\boldsymbol{\beta}_0$, λ_-	30	10	0.427	0.852
$\boldsymbol{\beta}_0$, λ_-	30	12	0.469	0.920
$\boldsymbol{\beta}_0$, λ_-	30	18	0.502	0.964
$\boldsymbol{\beta}_0$, λ_-	40	6	0.363	0.493
$\boldsymbol{\beta}_0$, λ_-	40	8	0.450	0.781
$\boldsymbol{\beta}_0$, λ_-	40	10	0.517	0.864
$\boldsymbol{\beta}_0$, λ_-	40	12	0.565	0.961
$\boldsymbol{\beta}_0$, λ_-	40	18	0.602	0.981
$\boldsymbol{\beta}_0$, λ_-	50	6	0.421	0.625
$\boldsymbol{\beta}_0$, λ_-	50	8	0.521	0.875
$\boldsymbol{\beta}_0$, λ_-	50	10	0.596	0.974
$\boldsymbol{\beta}_0$, λ_-	50	12	0.647	0.992
$\boldsymbol{\beta}_0$, λ_-	50	18	0.686	0.996
$\boldsymbol{\beta}_0$, λ_-	70	6	0.527	0.788
$\boldsymbol{\beta}_0$, λ_-	70	8	0.642	0.956
$\boldsymbol{\beta}_0$, λ_-	70	10	0.722	0.993
$\boldsymbol{\beta}_0$, λ_-	70	12	0.773	0.997
$\boldsymbol{\beta}_0$, λ_-	70	18	0.809	1.000

Table A.3: Power estimates for POST and ML estimators for $\boldsymbol{\beta}_0$ and λ_-

Level combination of model parameters	Number of patients	Number of scans	Estimated power based on $\widehat{T}_{POST}$	Estimated power based on $\widehat{T}_{ML}$
$\boldsymbol{\beta}_0$, λ_0	20	6	0.119	0.337
$\boldsymbol{\beta}_0$, λ_0	20	8	0.142	0.642
$\boldsymbol{\beta}_0$, λ_0	20	10	0.163	0.680
$\boldsymbol{\beta}_0$, λ_0	20	12	0.179	0.798
$\boldsymbol{\beta}_0$, λ_0	20	18	0.190	0.871
$\boldsymbol{\beta}_0$, λ_0	30	6	0.142	0.511
$\boldsymbol{\beta}_0$, λ_0	30	8	0.173	0.752
$\boldsymbol{\beta}_0$, λ_0	30	10	0.202	0.850
$\boldsymbol{\beta}_0$, λ_0	30	12	0.225	0.915
$\boldsymbol{\beta}_0$, λ_0	30	18	0.241	0.954
$\boldsymbol{\beta}_0$, λ_0	40	6	0.162	0.707
$\boldsymbol{\beta}_0$, λ_0	40	8	0.203	0.859
$\boldsymbol{\beta}_0$, λ_0	40	10	0.239	0.949
$\boldsymbol{\beta}_0$, λ_0	40	12	0.269	0.970
$\boldsymbol{\beta}_0$, λ_0	40	18	0.288	0.987
$\boldsymbol{\beta}_0$, λ_0	50	6	0.182	0.831
$\boldsymbol{\beta}_0$, λ_0	50	8	0.231	0.938
$\boldsymbol{\beta}_0$, λ_0	50	10	0.275	0.983
$\boldsymbol{\beta}_0$, λ_0	50	12	0.310	0.989
$\boldsymbol{\beta}_0$, λ_0	50	18	0.333	0.993
$\boldsymbol{\beta}_0$, λ_0	70	6	0.220	0.870
$\boldsymbol{\beta}_0$, λ_0	70	8	0.284	0.984
$\boldsymbol{\beta}_0$, λ_0	70	10	0.342	0.995
$\boldsymbol{\beta}_0$, λ_0	70	12	0.388	0.999
$\boldsymbol{\beta}_0$, λ_0	70	18	0.417	1.000

Table A.4: Power estimates for POST and ML estimators for $\boldsymbol{\beta}_0$ and λ_0

Level combination of model parameters	Number of patients	Number of scans	Estimated power based on $\widehat{T}_{POST}$	Estimated power based on $\widehat{T}_{ML}$
$\boldsymbol{\beta}_0$, λ_+	20	6	0.082	0.334
$\boldsymbol{\beta}_0$, λ_+	20	8	0.091	0.497
$\boldsymbol{\beta}_0$, λ_+	20	10	0.100	0.652
$\boldsymbol{\beta}_0$, λ_+	20	12	0.107	0.697
$\boldsymbol{\beta}_0$, λ_+	20	18	0.112	0.819
$\boldsymbol{\beta}_0$, λ_+	30	6	0.091	0.457
$\boldsymbol{\beta}_0$, λ_+	30	8	0.104	0.744
$\boldsymbol{\beta}_0$, λ_+	30	10	0.115	0.845
$\boldsymbol{\beta}_0$, λ_+	30	12	0.125	0.909
$\boldsymbol{\beta}_0$, λ_+	30	18	0.131	0.959
$\boldsymbol{\beta}_0$, λ_+	40	6	0.099	0.676
$\boldsymbol{\beta}_0$, λ_+	40	8	0.115	0.892
$\boldsymbol{\beta}_0$, λ_+	40	10	0.129	0.952
$\boldsymbol{\beta}_0$, λ_+	40	12	0.141	0.959
$\boldsymbol{\beta}_0$, λ_+	40	18	0.149	0.981
$\boldsymbol{\beta}_0$, λ_+	50	6	0.106	0.750
$\boldsymbol{\beta}_0$, λ_+	50	8	0.125	0.935
$\boldsymbol{\beta}_0$, λ_+	50	10	0.143	0.980
$\boldsymbol{\beta}_0$, λ_+	50	12	0.157	0.992
$\boldsymbol{\beta}_0$, λ_+	50	18	0.166	0.992
$\boldsymbol{\beta}_0$, λ_+	70	6	0.120	0.900
$\boldsymbol{\beta}_0$, λ_+	70	8	0.145	0.981
$\boldsymbol{\beta}_0$, λ_+	70	10	0.168	0.995
$\boldsymbol{\beta}_0$, λ_+	70	12	0.187	0.999
$\boldsymbol{\beta}_0$, λ_+	70	18	0.199	1.000

Table A.5: Power estimates for POST and ML estimators for $\boldsymbol{\beta}_0$ and λ_+

Appendix B

Plots

This appendix contains various plots.

B.1 Extra plots for the POST estimator

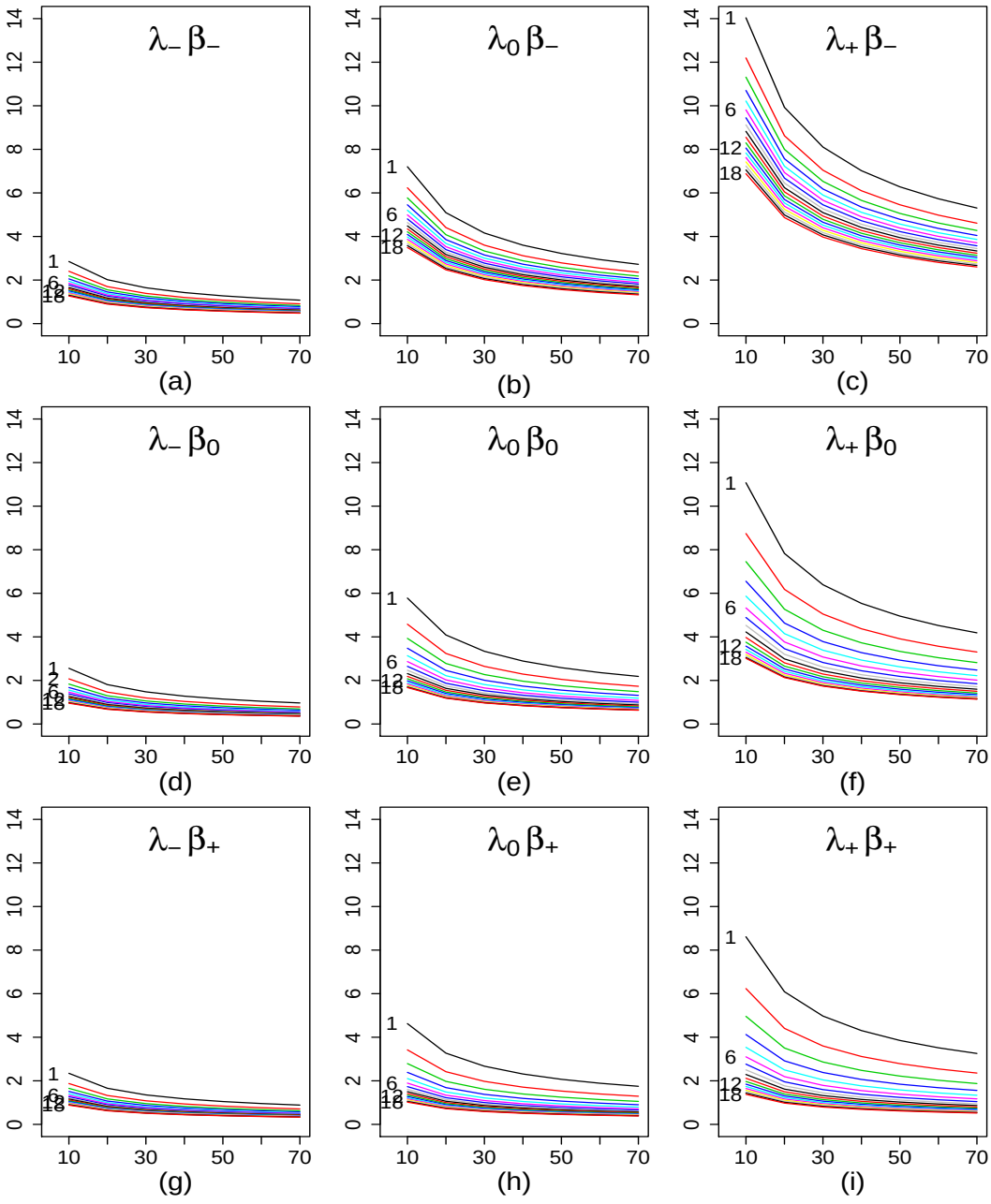

Figure B.1: SD($\widehat{T}_{POST}$) vs. m with varying λ_h and $\boldsymbol{\beta}_h$ with SD axis fixed (curves correspond to the specified number of scans)

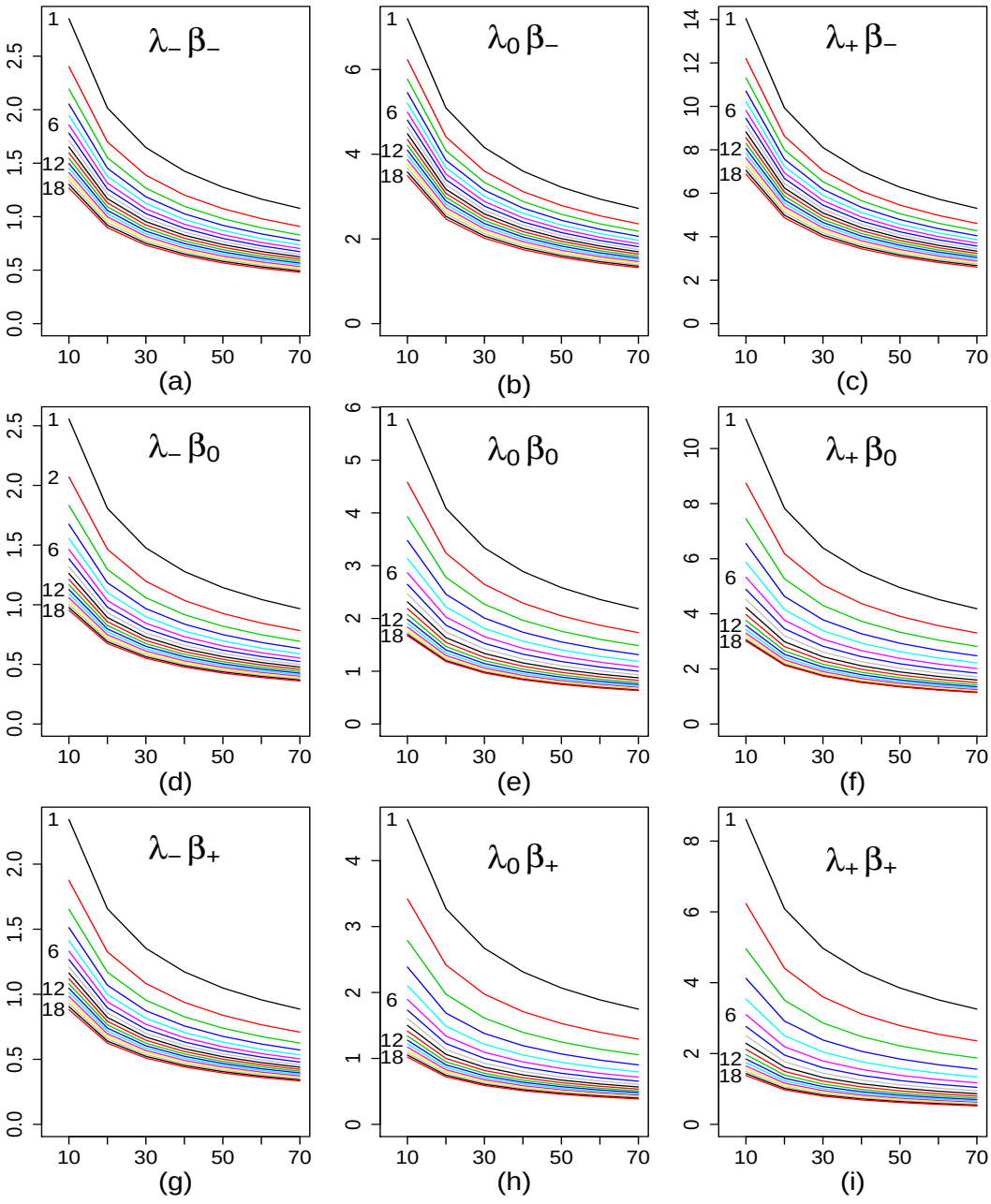

Figure B.2: SD($\widehat{T}_{POST}$) vs. m with varying λ_h and $\boldsymbol{\beta}_h$ with SD axis scaled to best fit plot (curves correspond to the specified number of scans)

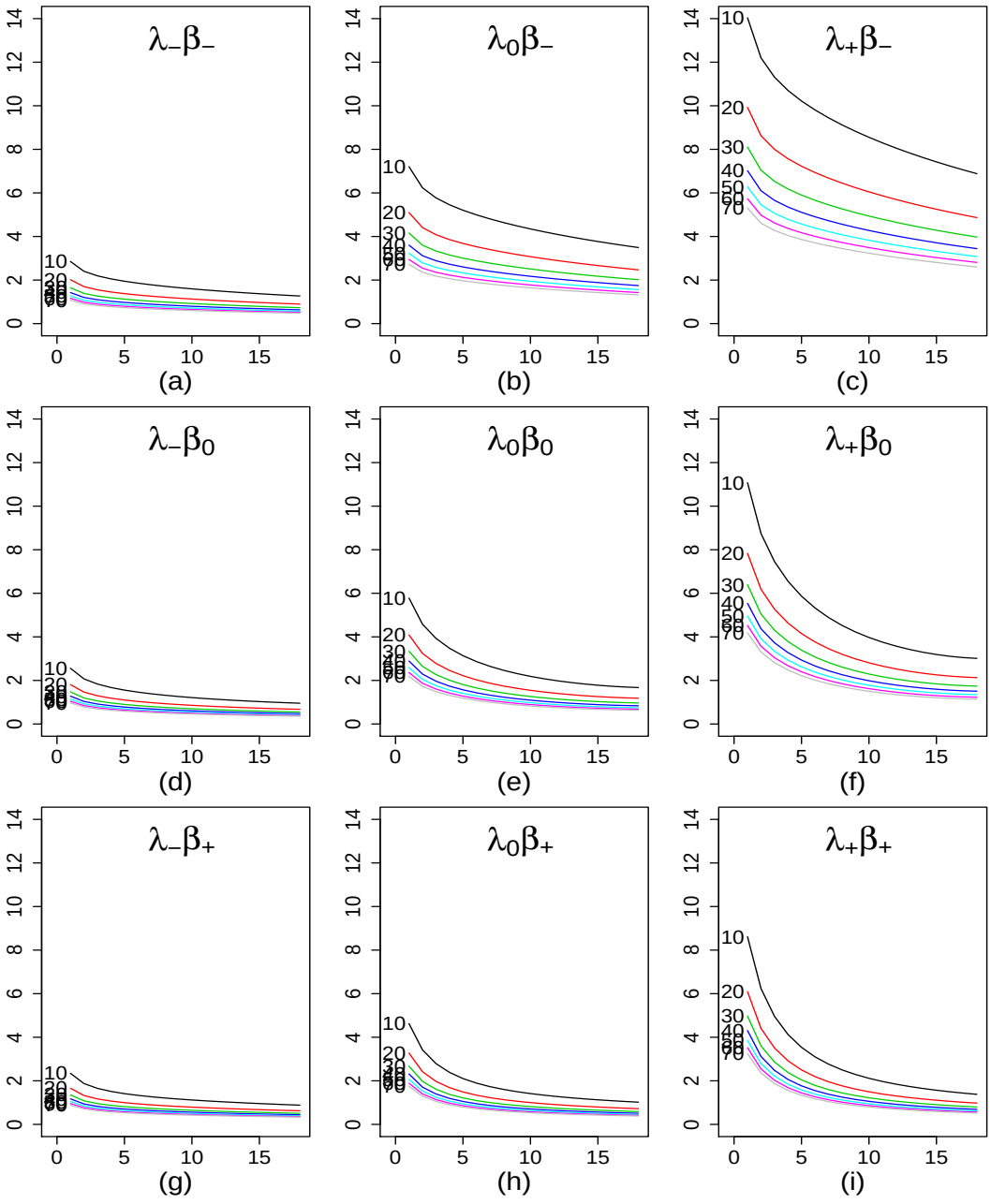

Figure B.3: SD($\widehat{T}_{POST}$) vs. n with varying λ_h and $\boldsymbol{\beta}_h$ with SD axis fixed (curves correspond to the specified number of patients)

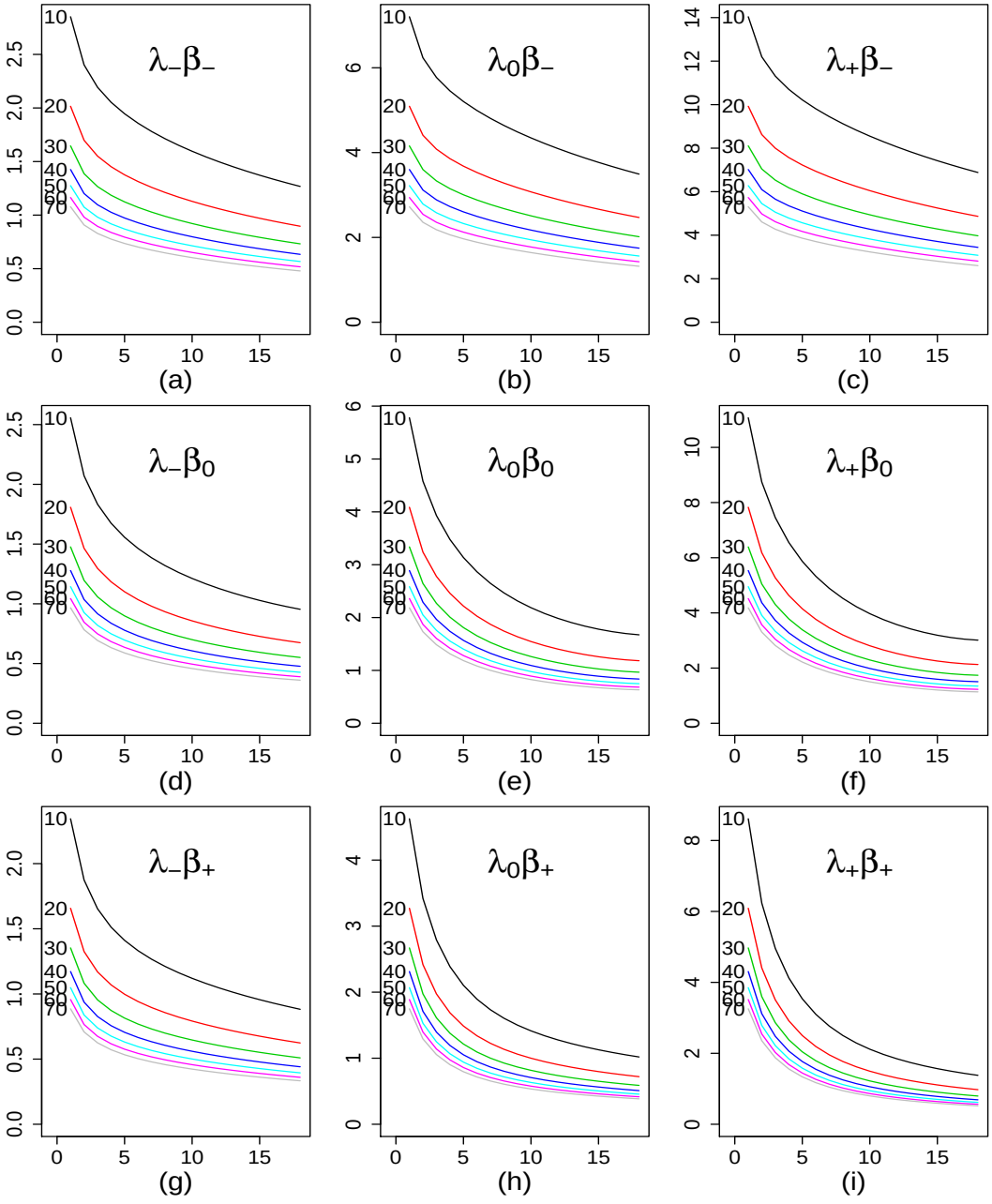

Figure B.4: SD($\widehat{T}_{POST}$) vs. n with varying λ_h and $\boldsymbol{\beta}_h$ with SD axis scaled to best fit plot (curves correspond to the specified number of patients)

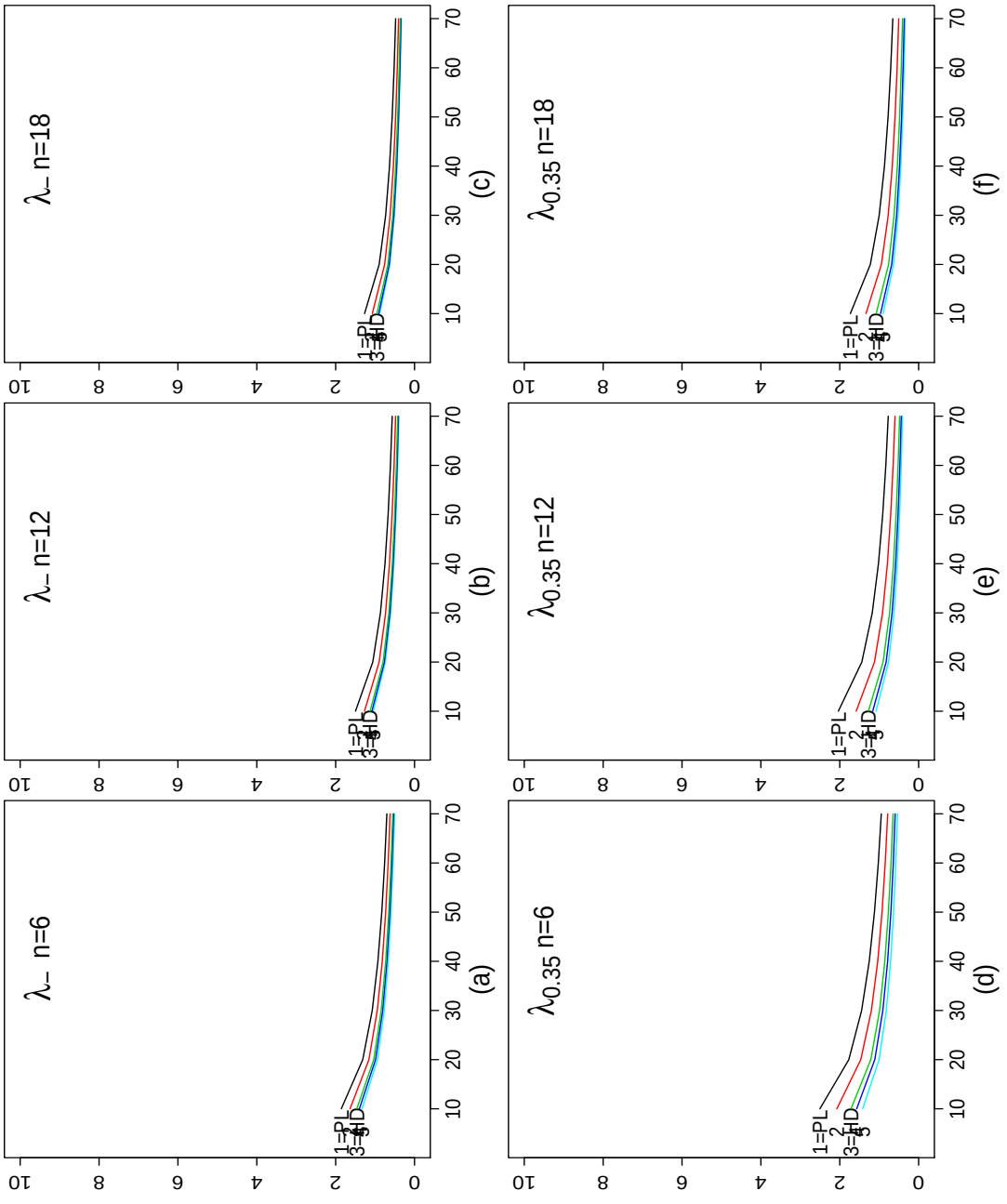

Figure B.5: SD($\widehat{T}_{POST}$) vs. m with varying λ_h and n with SD axis fixed (curves correspond to the specified mean structure)

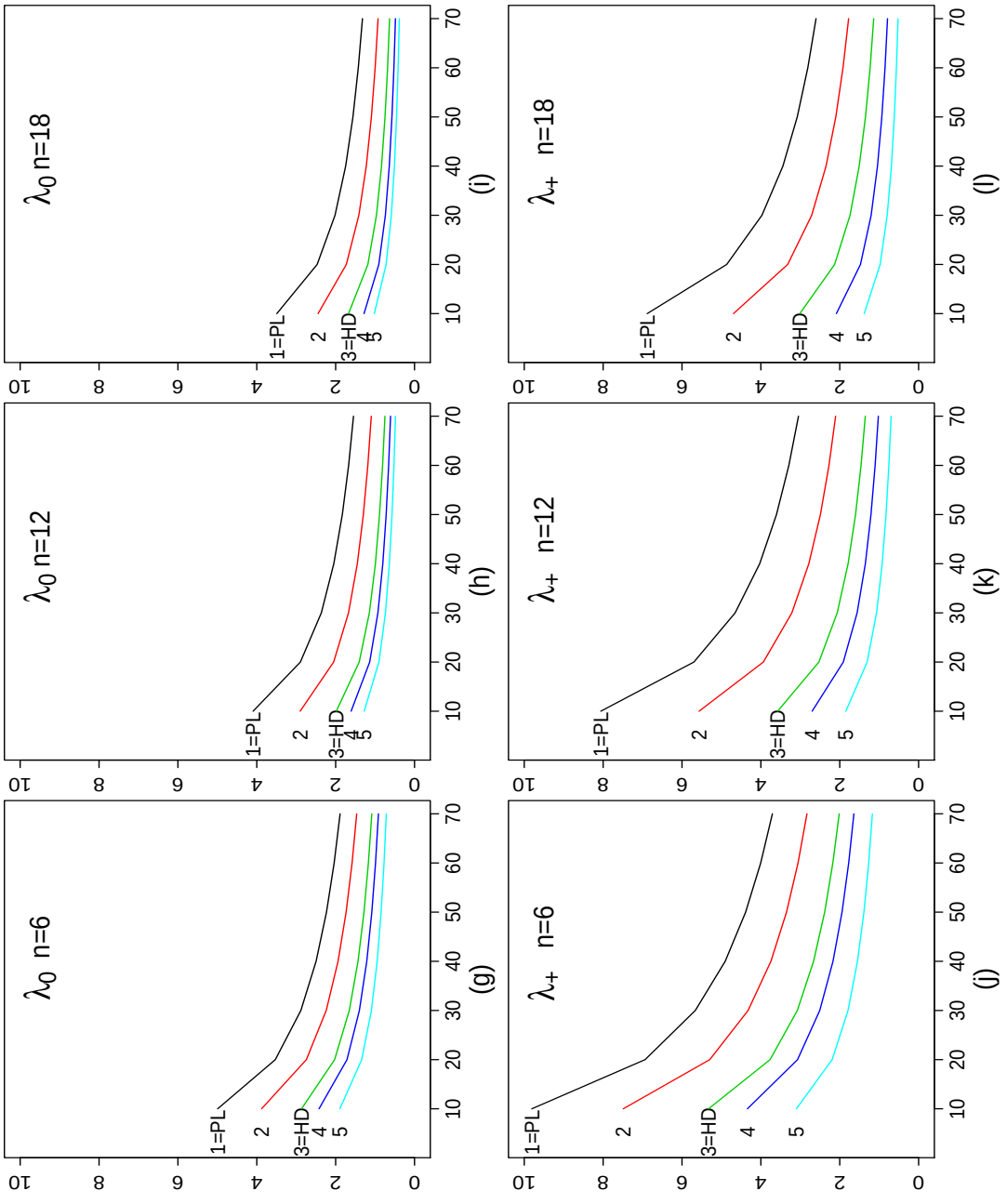

Figure B.6: SD($\widehat{T}_{POST}$) vs. m with varying λ_h and n with SD axis fixed (curves correspond to the specified mean structure)

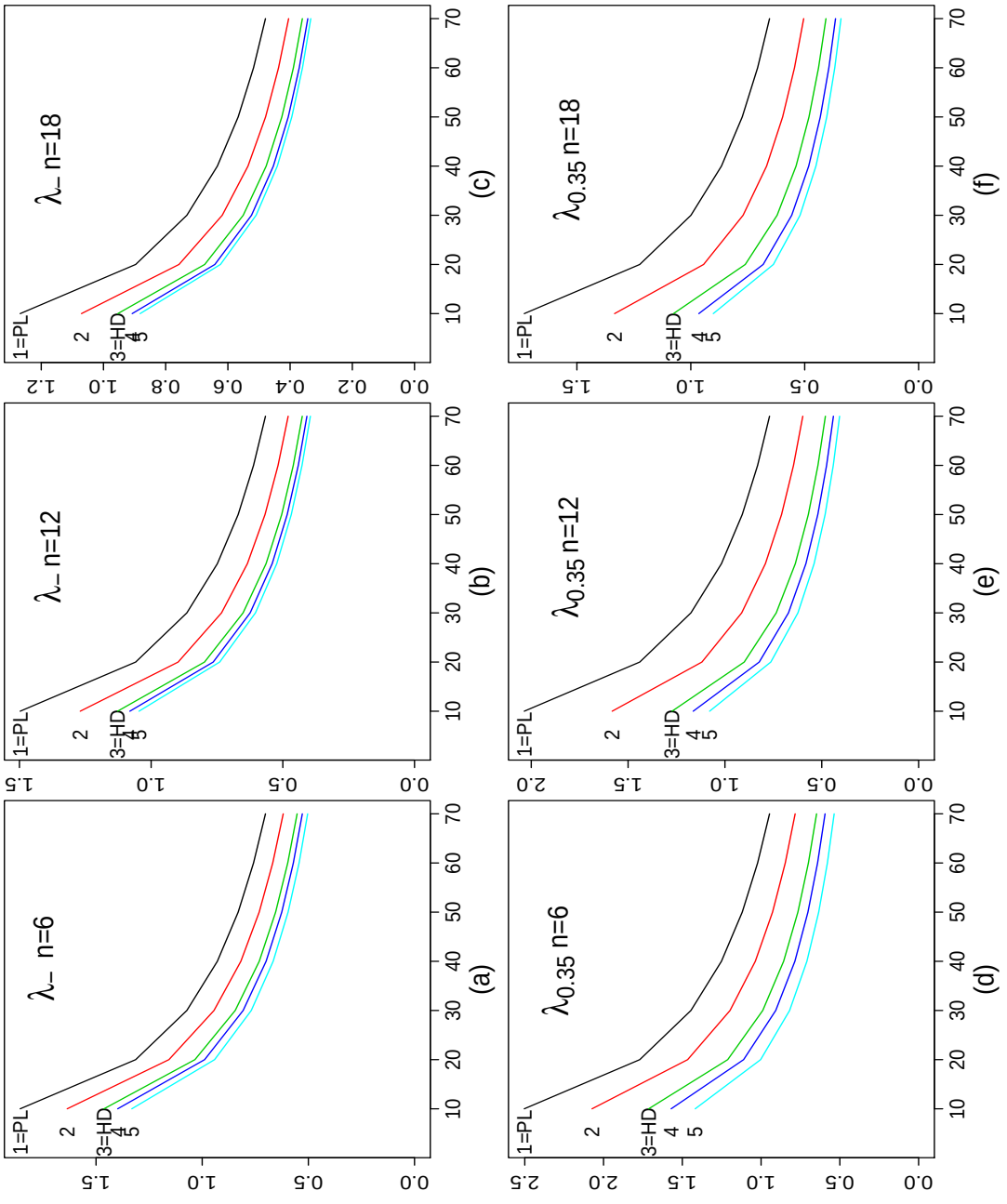

Figure B.7: SD($\widehat{T}_{POST}$) vs. m with varying λ_h and n with SD axis scaled to best fit plot (curves correspond to the specified mean structure)

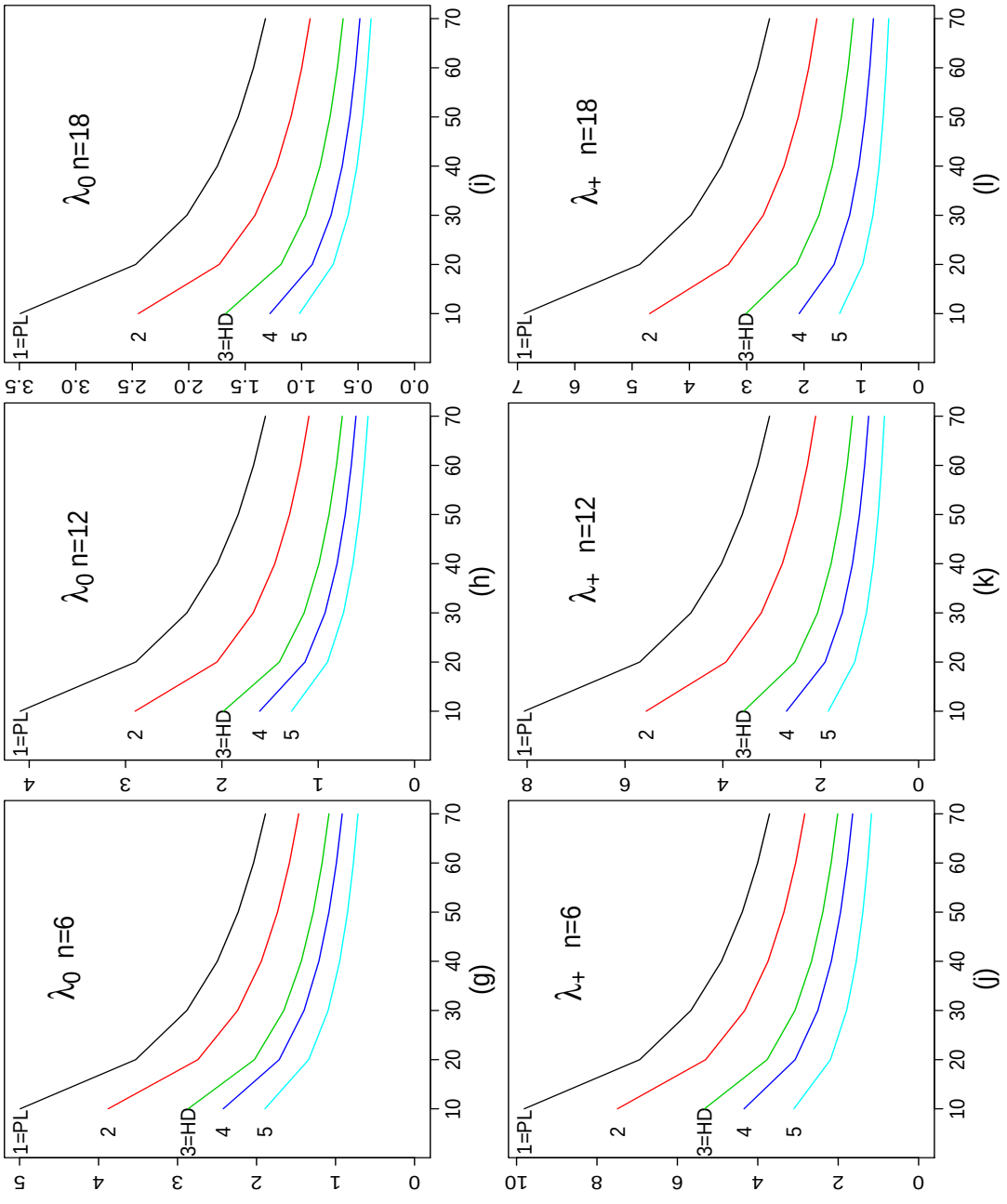

Figure B.8: SD($\widehat{T}_{POST}$) vs. m with varying λ_h and n with SD axis scaled to best fit plot (curves correspond to the specified mean structure)

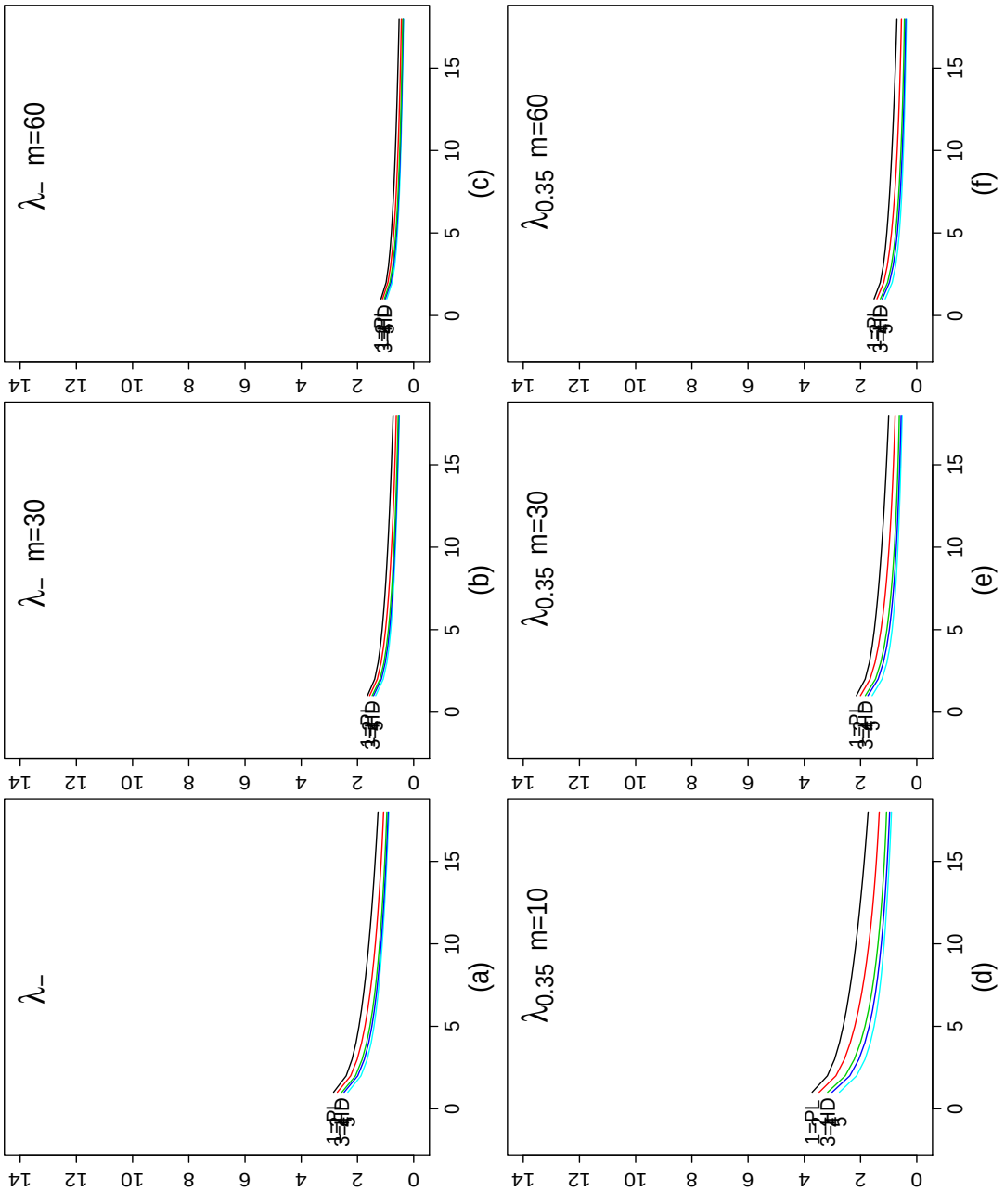

Figure B.9: SD($\widehat{T}_{POST}$) vs. n with varying λ_h and m with SD axis fixed (curves correspond to the specified mean structure)

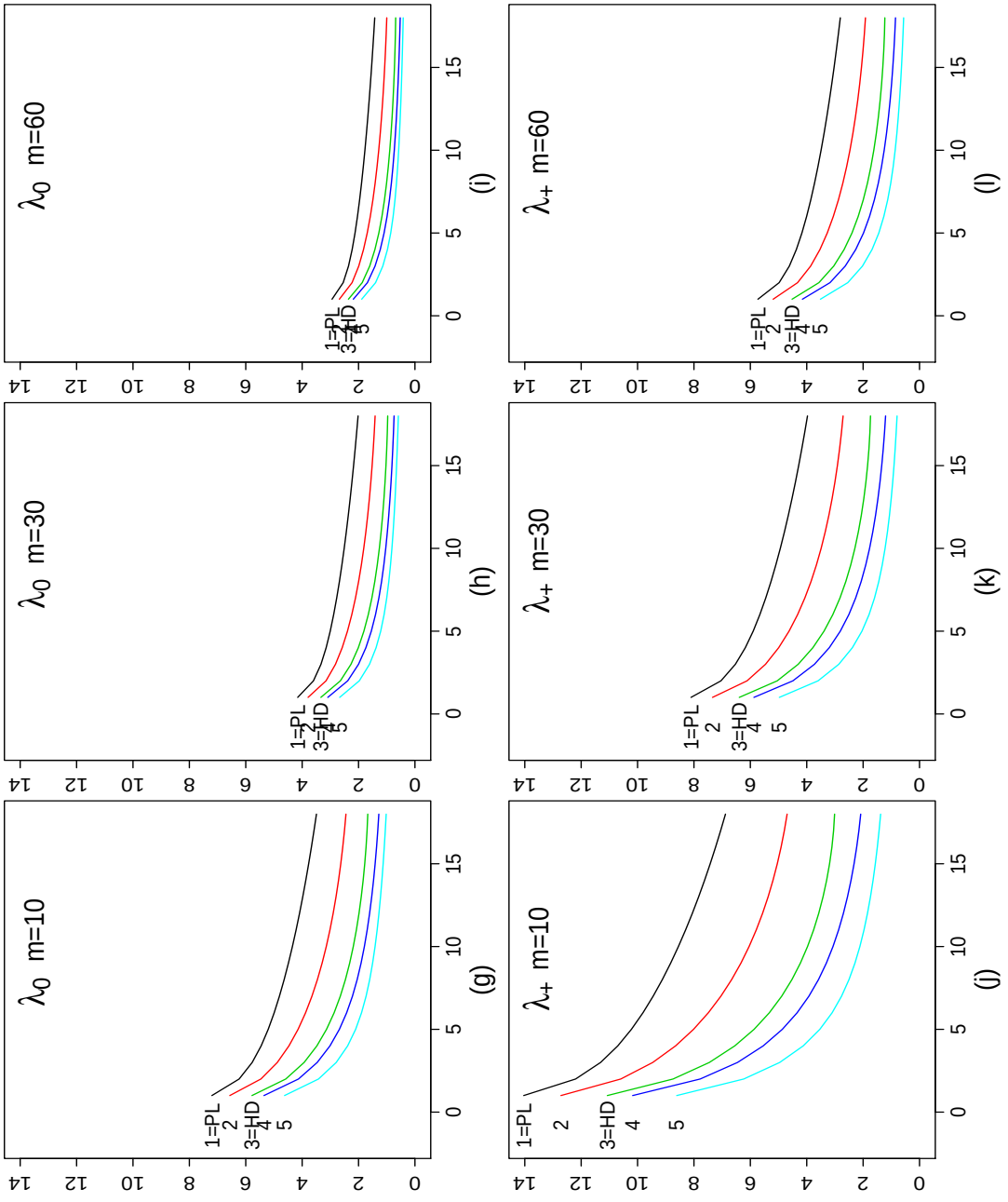

Figure B.10: SD($\widehat{T}_{POST}$) vs. n with varying λ_h and m with SD axis fixed

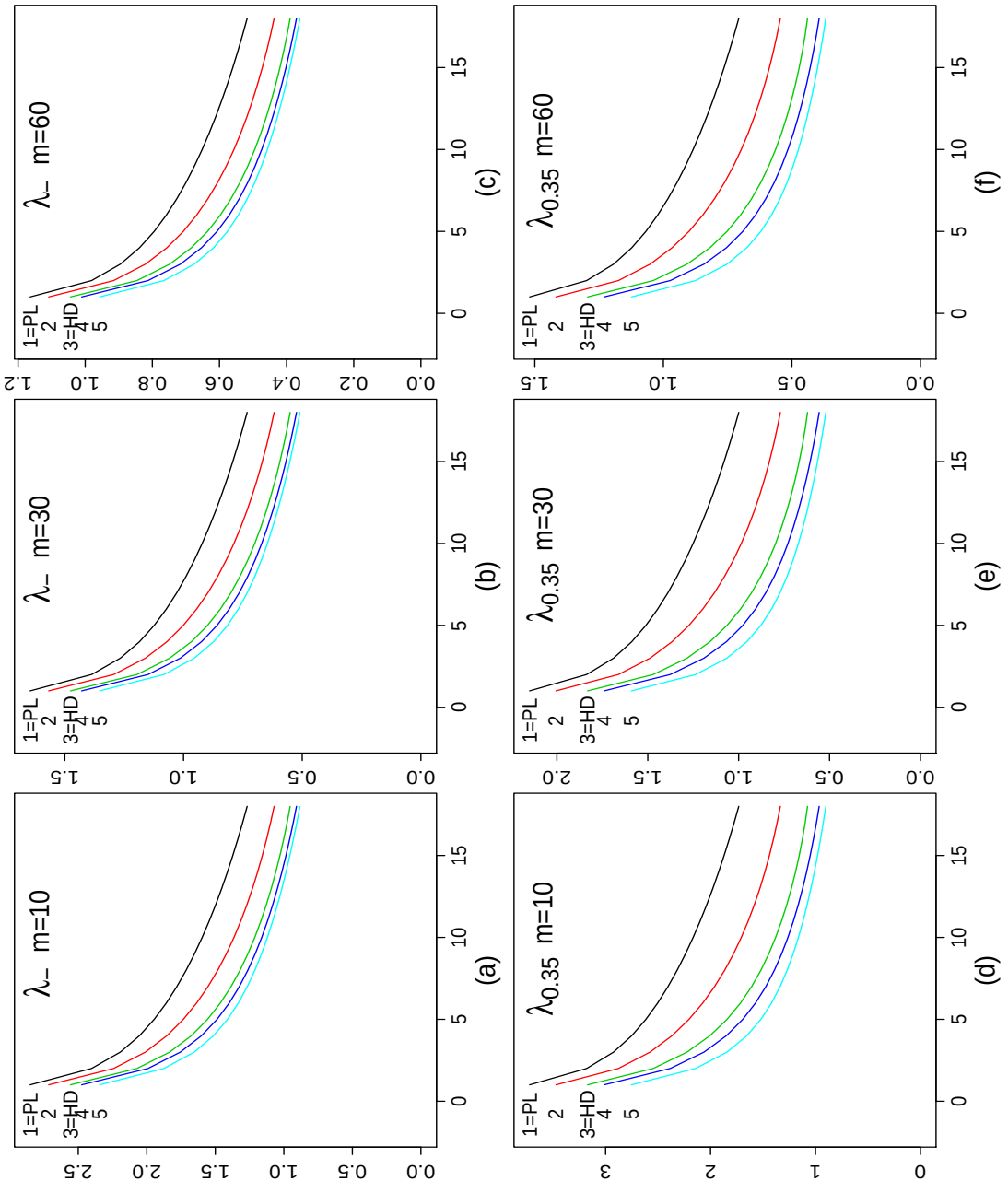

Figure B.11: SD($\widehat{T}_{POST}$) vs. n with varying λ_h and m with SD axis scaled to best fit plot (curves correspond to the specified mean structure)

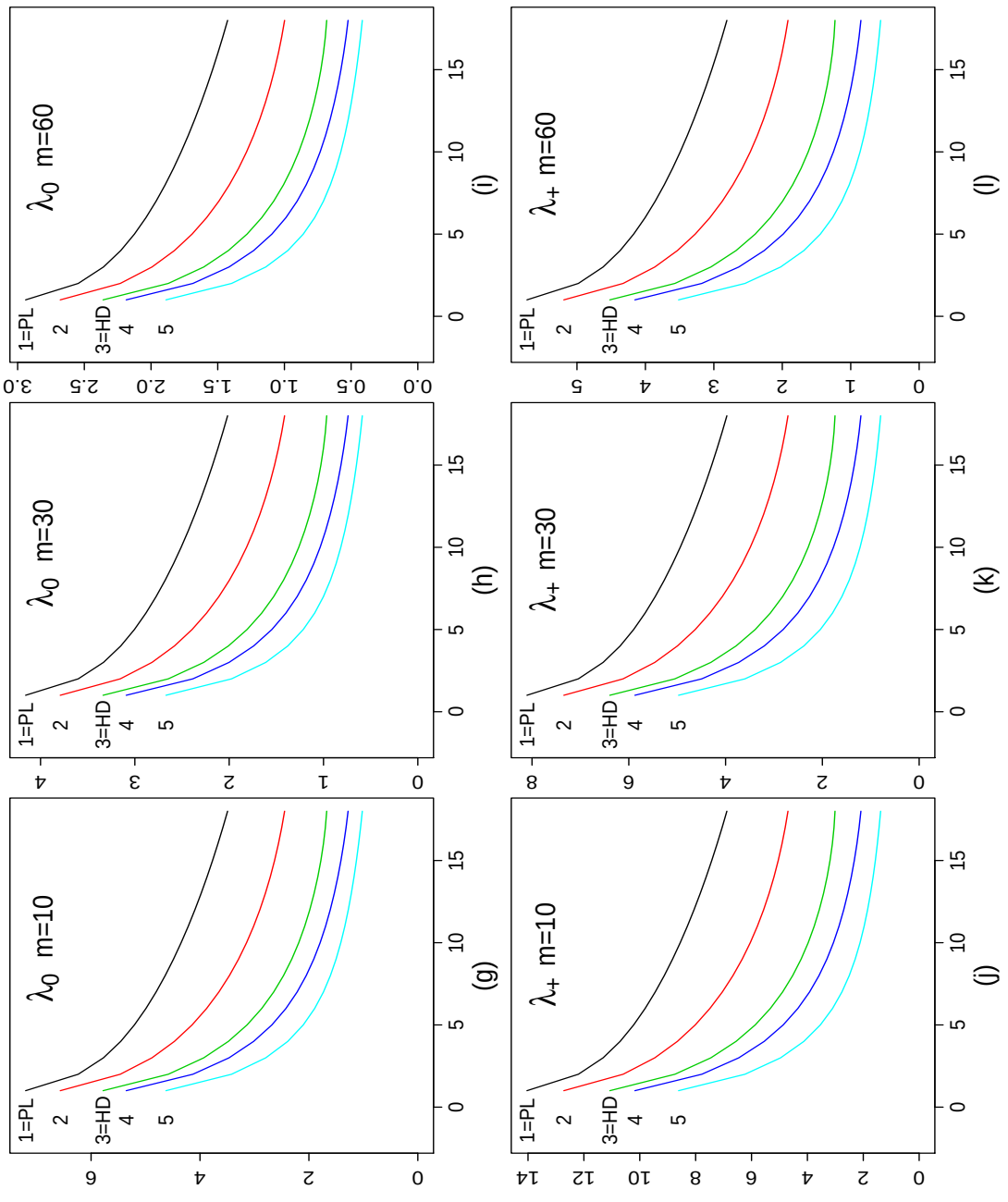

Figure B.12: SD($\widehat{T}_{POST}$) vs. n with varying λ_h and m with SD axis scaled to best fit plot (curves correspond to the specified mean structure)

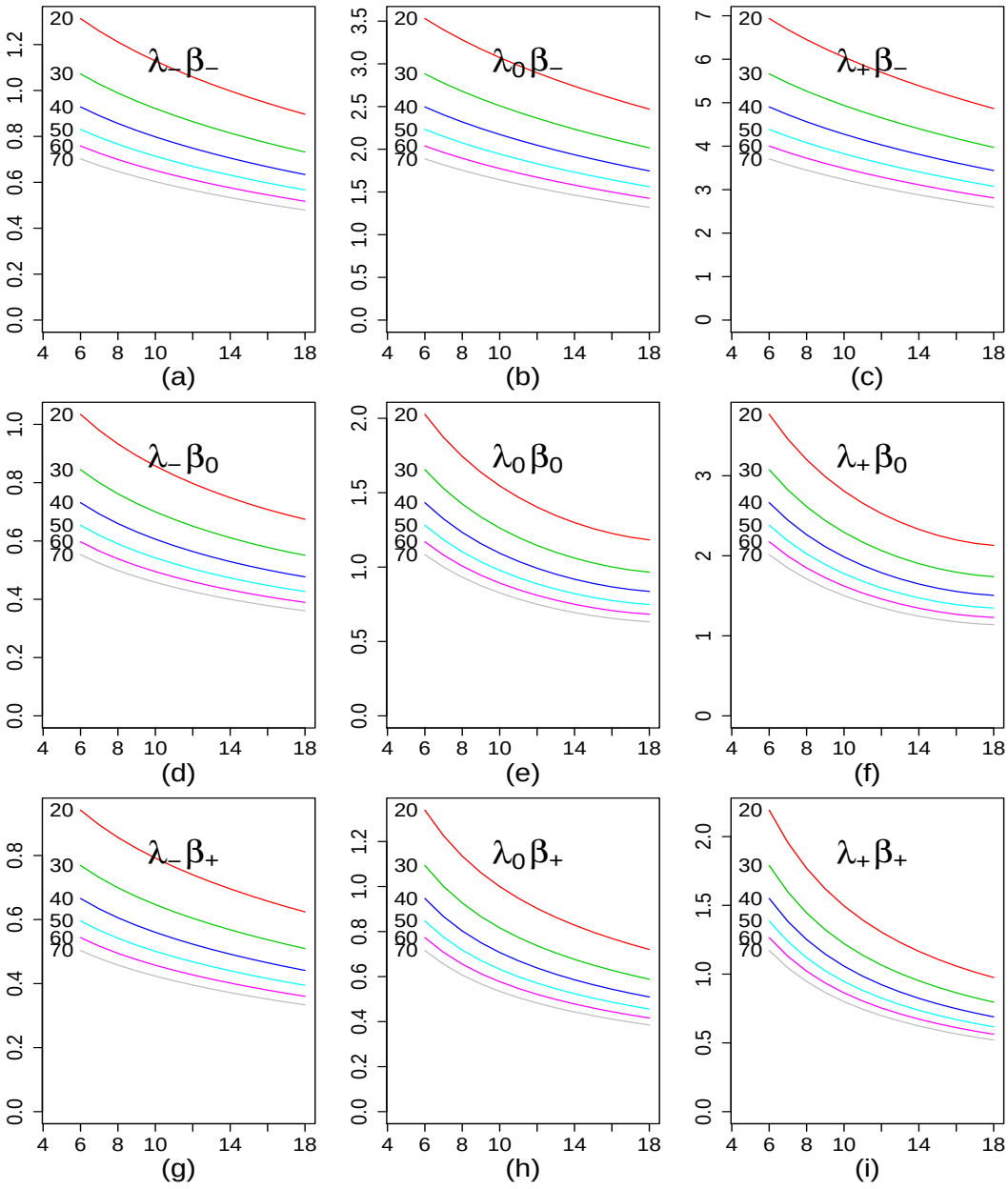

Figure B.13: SD($\widehat{T}_{POST}$) vs. n with varying λ_h and $\boldsymbol{\beta}_h$ with modified sample size parameter range and SD axis scaled to best fit plot (curves correspond to the specified number of patients)

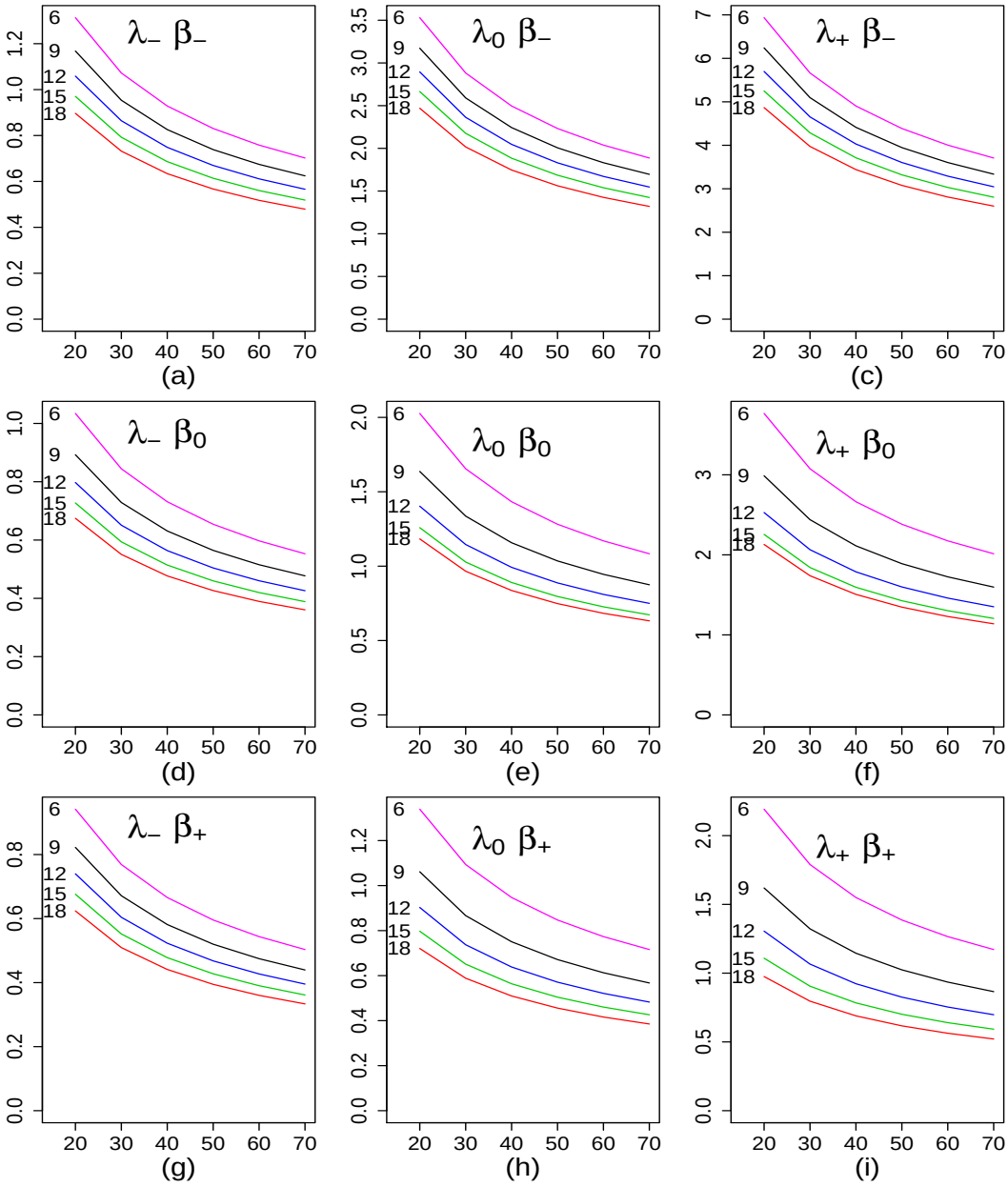

Figure B.14: SD($\widehat{T}_{POST}$) vs. m with varying λ_h and $\boldsymbol{\beta}_h$ with modified sample size parameter range and SD axis scaled to best fit plot (curves correspond to the specified number of scans)

B.2 Extra plots for the ML estimator

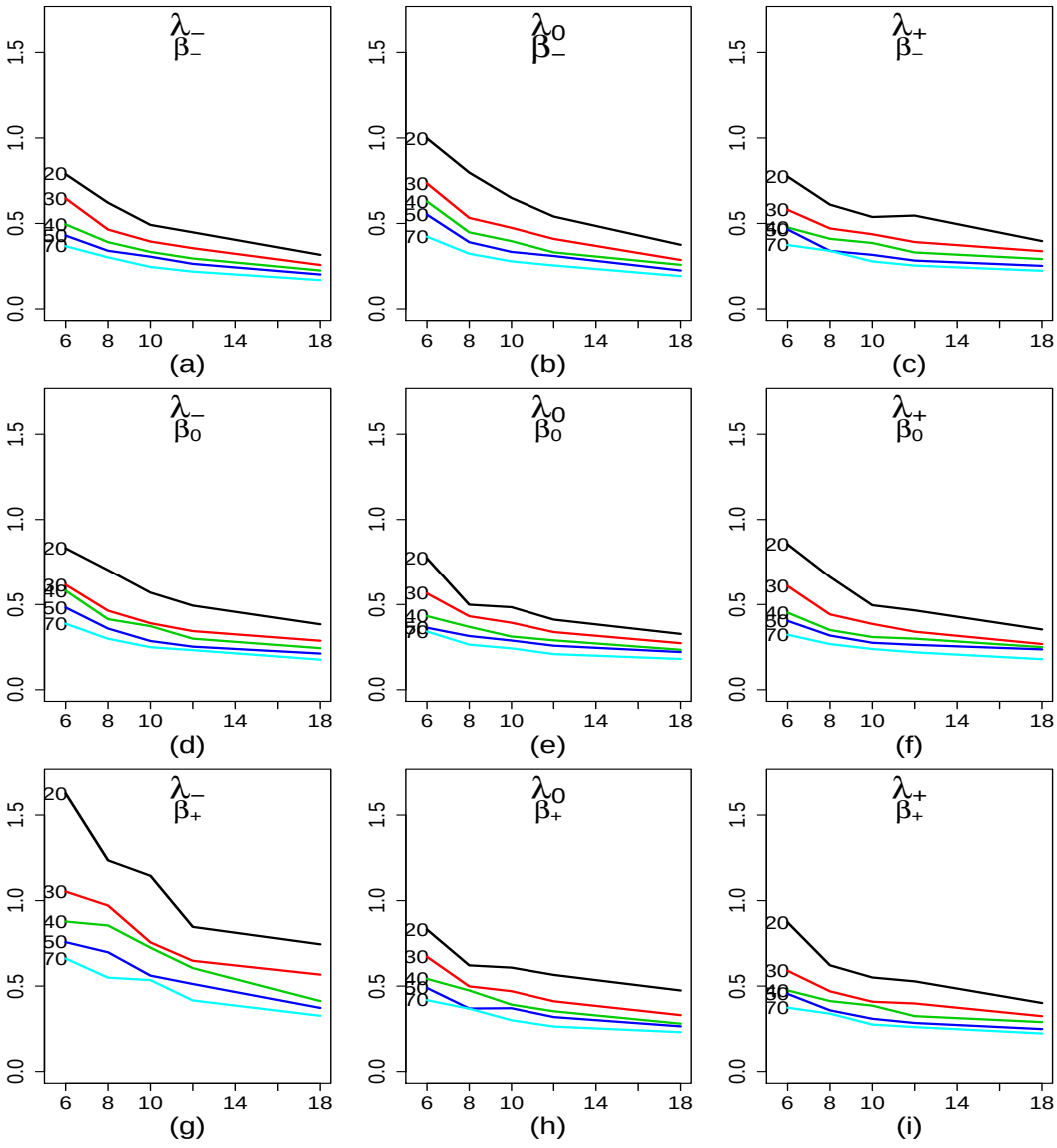

Figure B.15: SD($\widehat{T}_{ML}$) vs. n with varying λ_h and $\boldsymbol{\beta}_h$ with SD axis fixed (curves correspond to the specified number of patients)

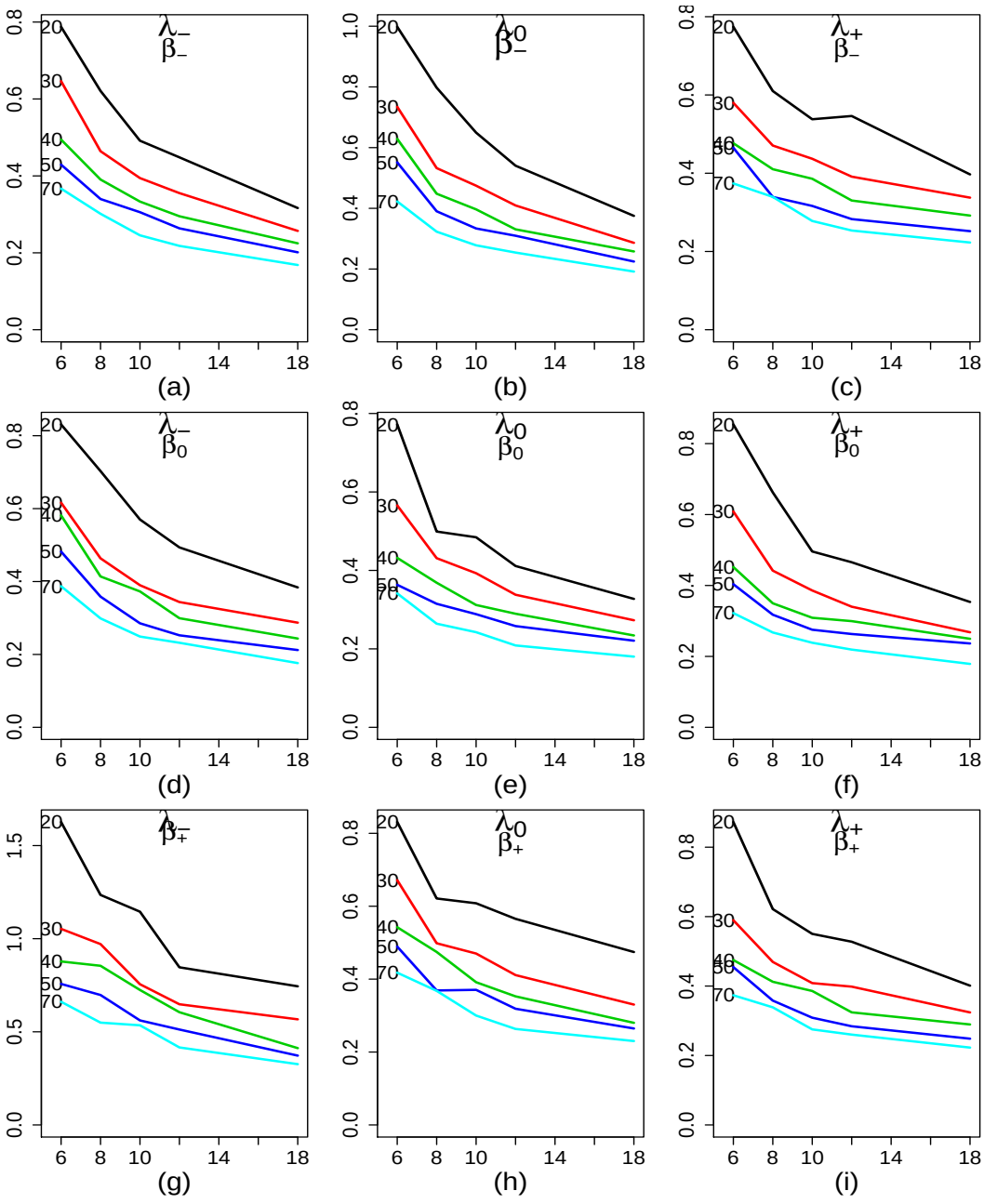

Figure B.16: SD($\widehat{T}_{ML}$) vs. n with varying λ_h and $\boldsymbol{\beta}_h$ with SD axis scaled to best fit plot (curves correspond to the specified number of patients)

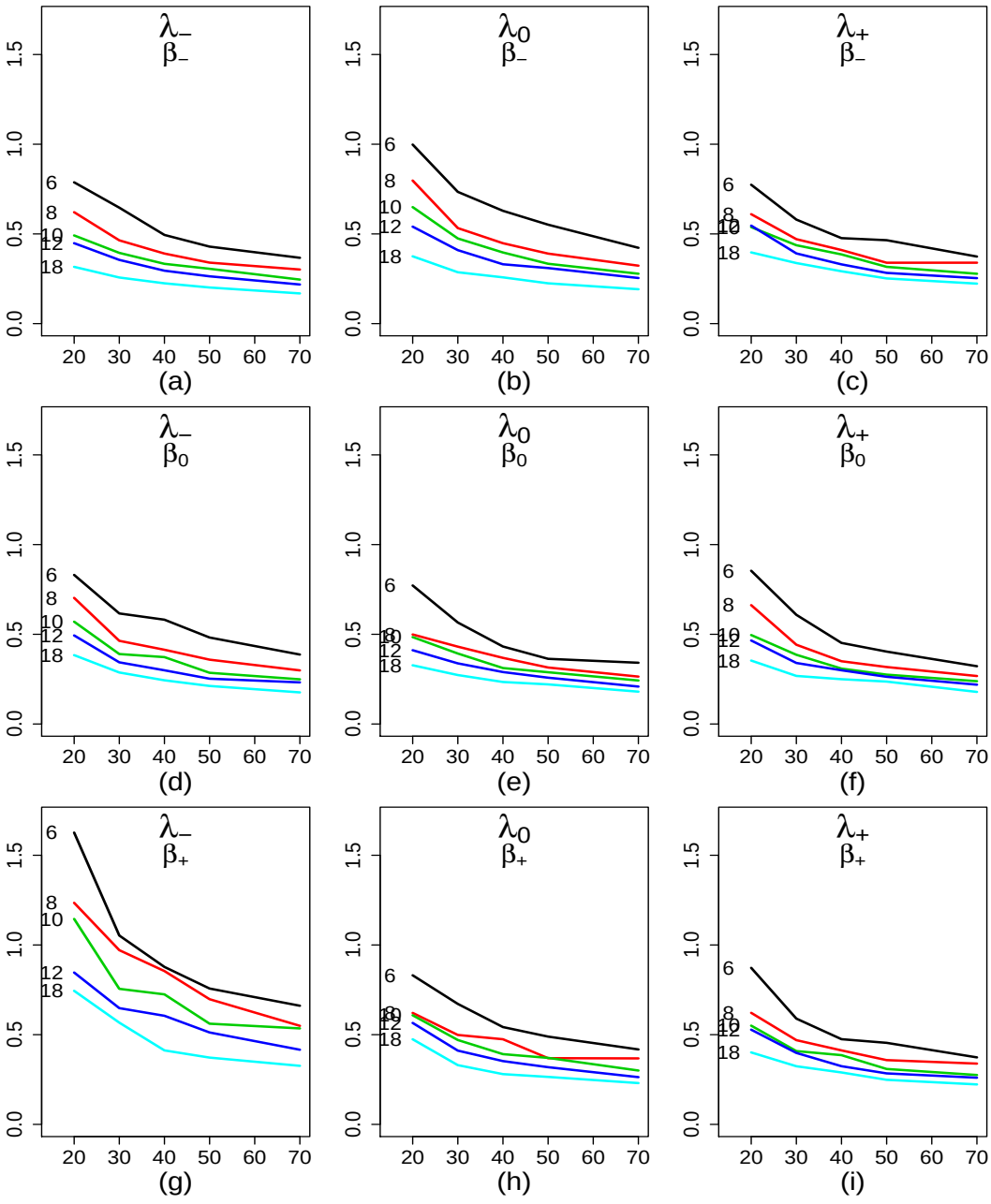

Figure B.17: SD($\widehat{T}_{ML}$) vs. m with varying λ_h and $\boldsymbol{\beta}_h$ with SD axis fixed (curves correspond to the specified number of scans)

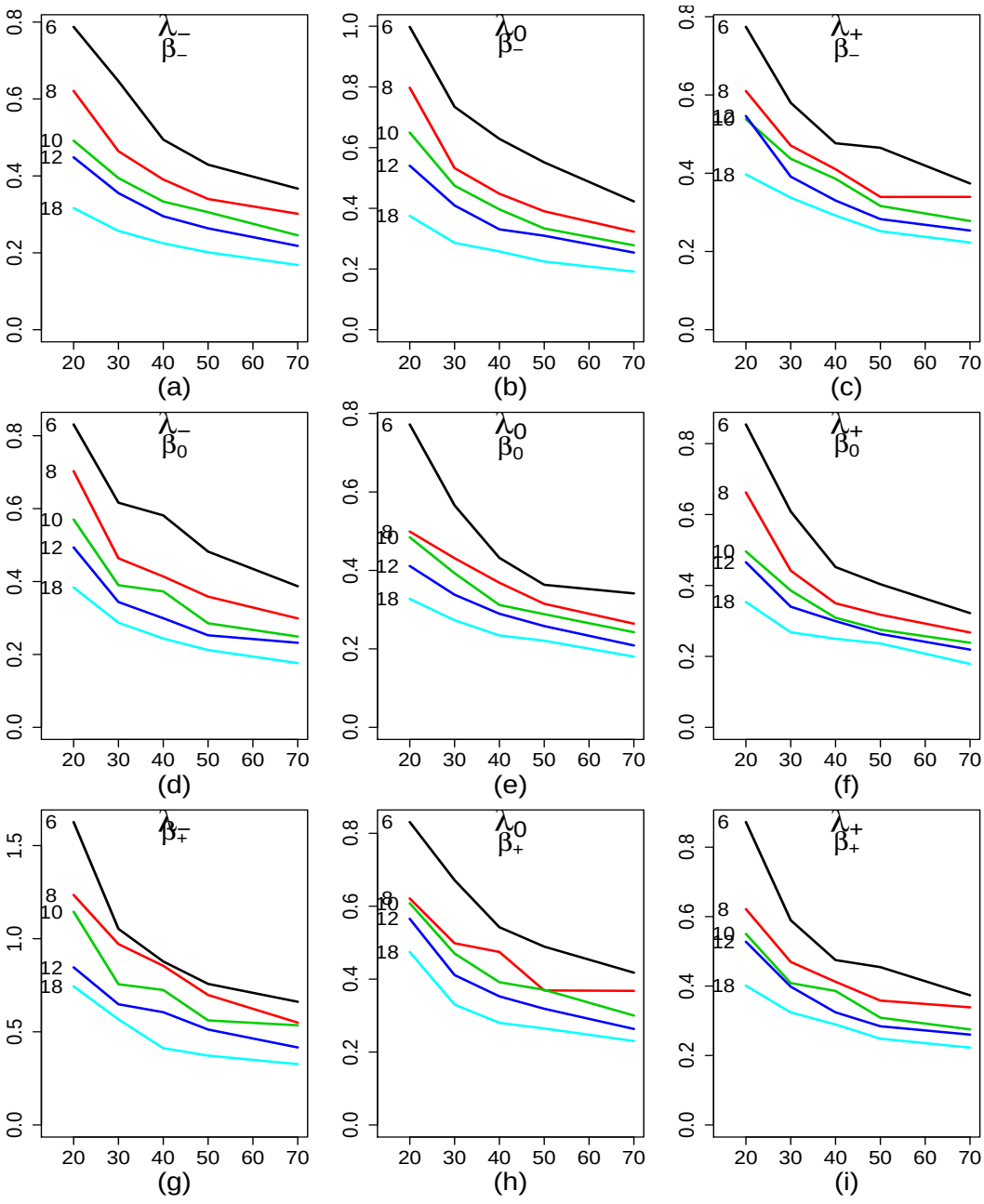

Figure B.18: SD($\widehat{T}_{ML}$) vs. m with varying λ_h and $\boldsymbol{\beta}_h$ with SD axis scaled to best fit plot (curves correspond to the specified number of scans)

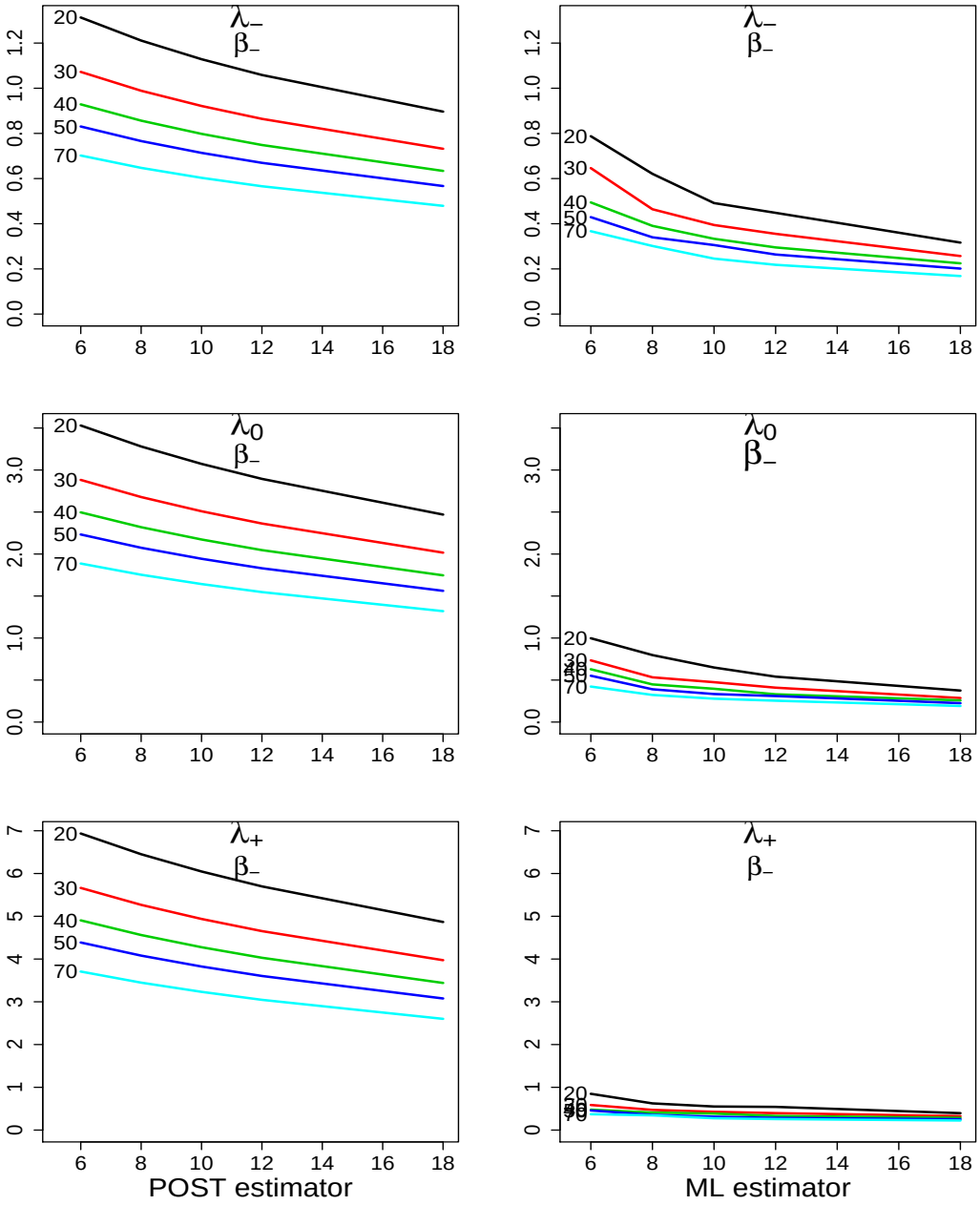

Figure B.19: Estimated standard deviation vs. n for the POST and ML estimators with $\boldsymbol{\beta}_-$ and varying λ_h (curves correspond to the specified number of patients)

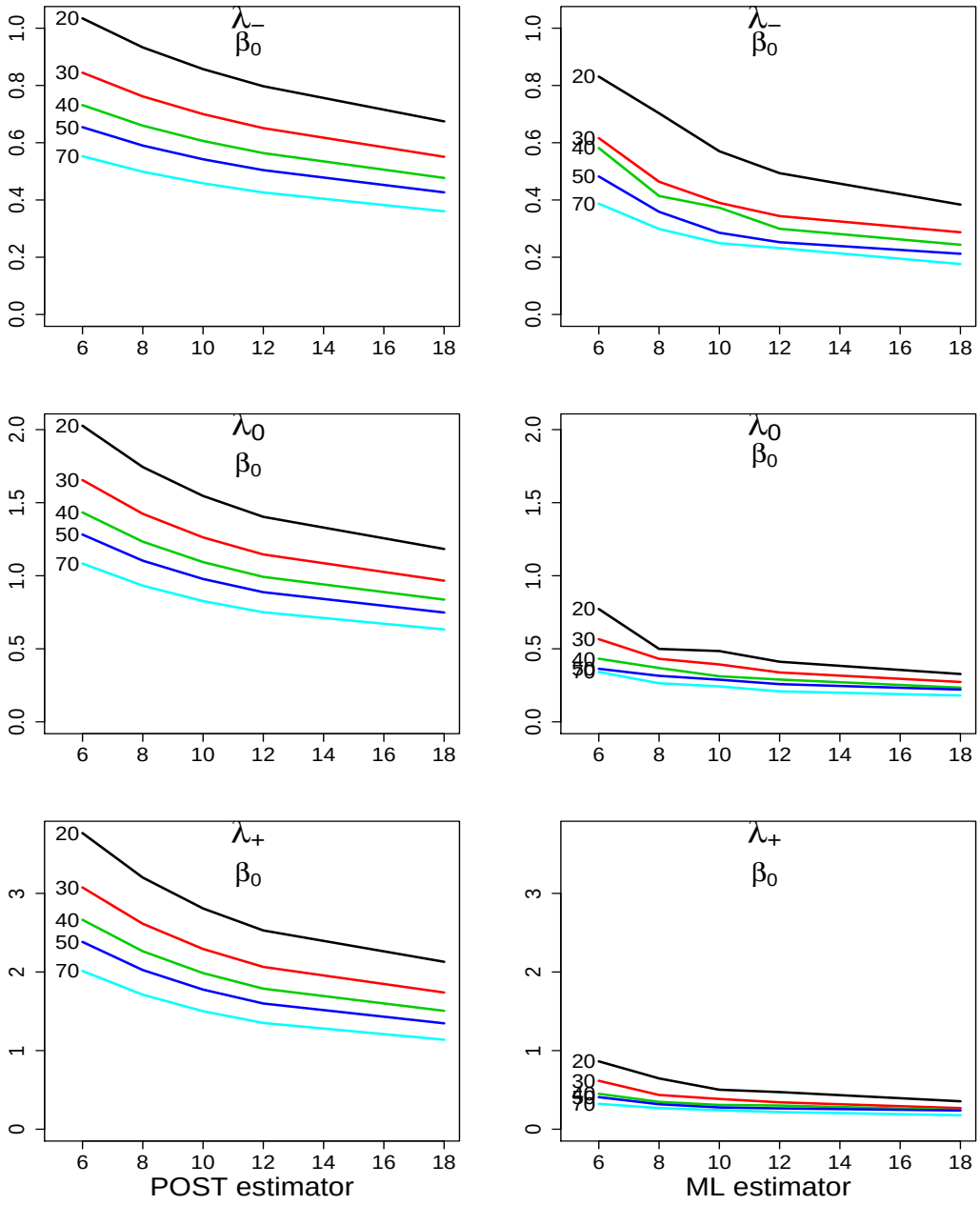

Figure B.20: Estimated standard deviation vs. n for the POST and ML estimators with $\boldsymbol{\beta}_0$ and varying λ_h (curves correspond to the specified number of patients)

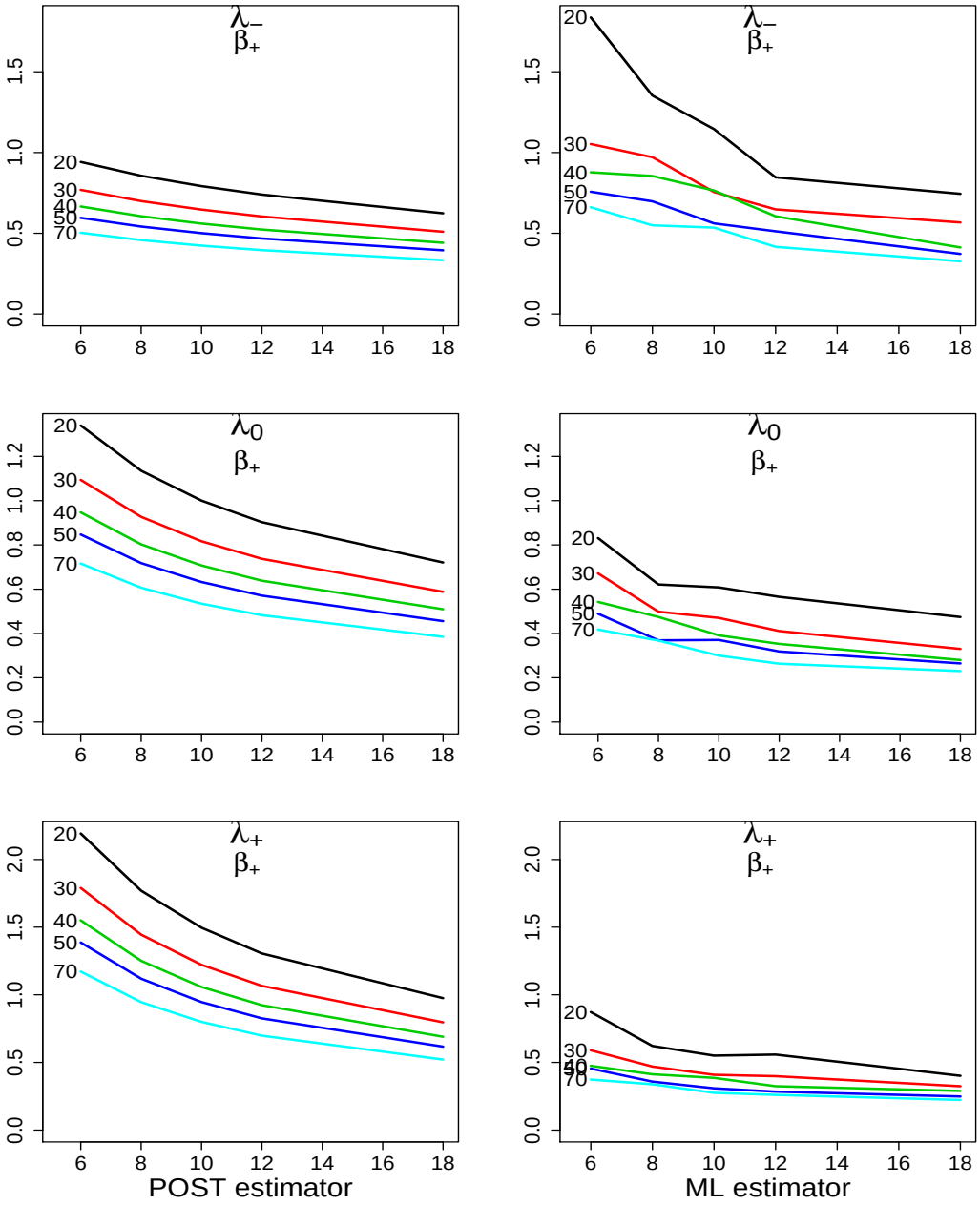

Figure B.21: Estimated standard deviation vs. n for the POST and ML estimators with $\boldsymbol{\beta}_+$ and varying λ_h (curves correspond to the specified number of patients)

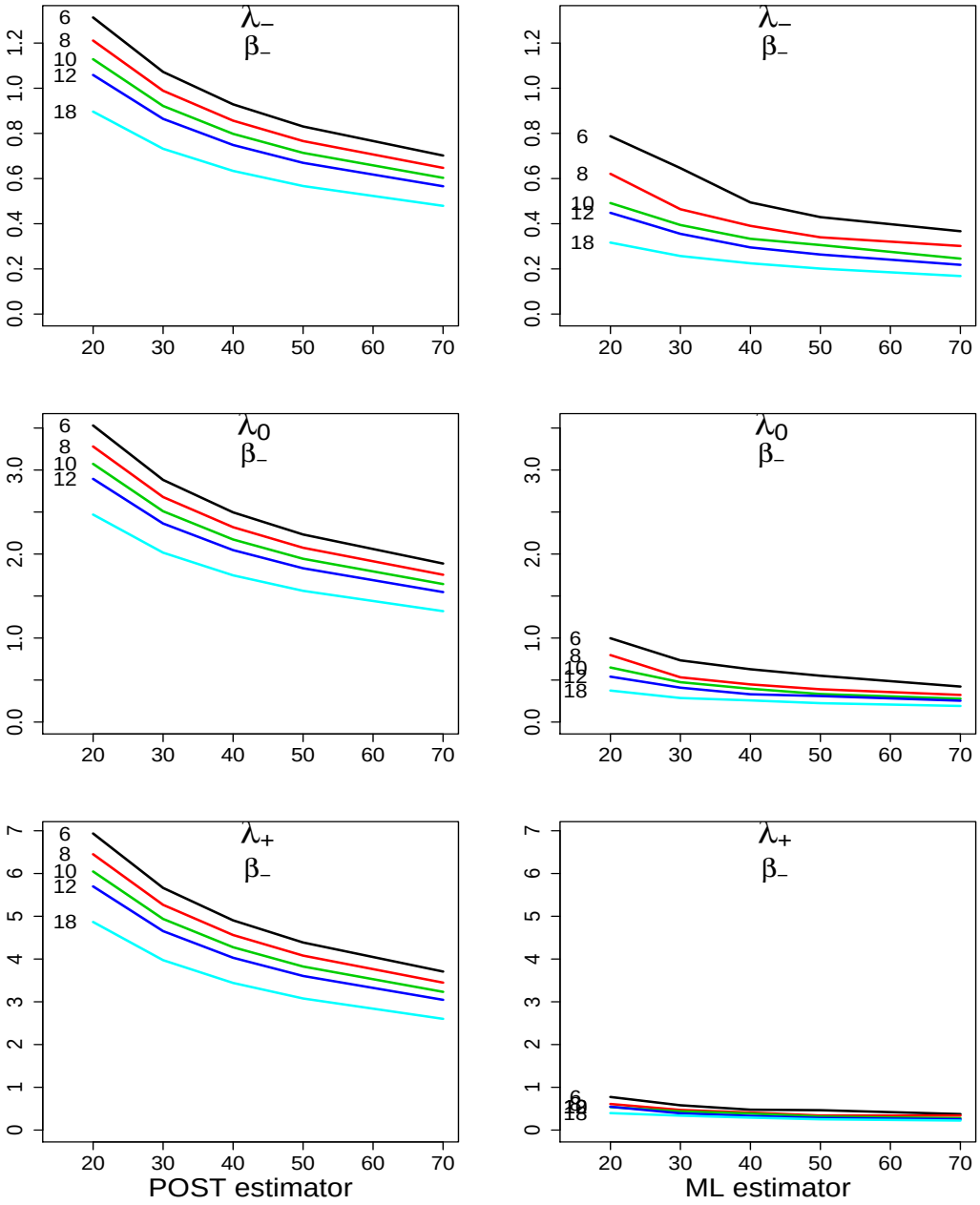

Figure B.22: Estimated standard deviation vs. m for the POST and ML estimators with $\boldsymbol{\beta}_-$ and varying λ_h (curves correspond to the specified number of scans)

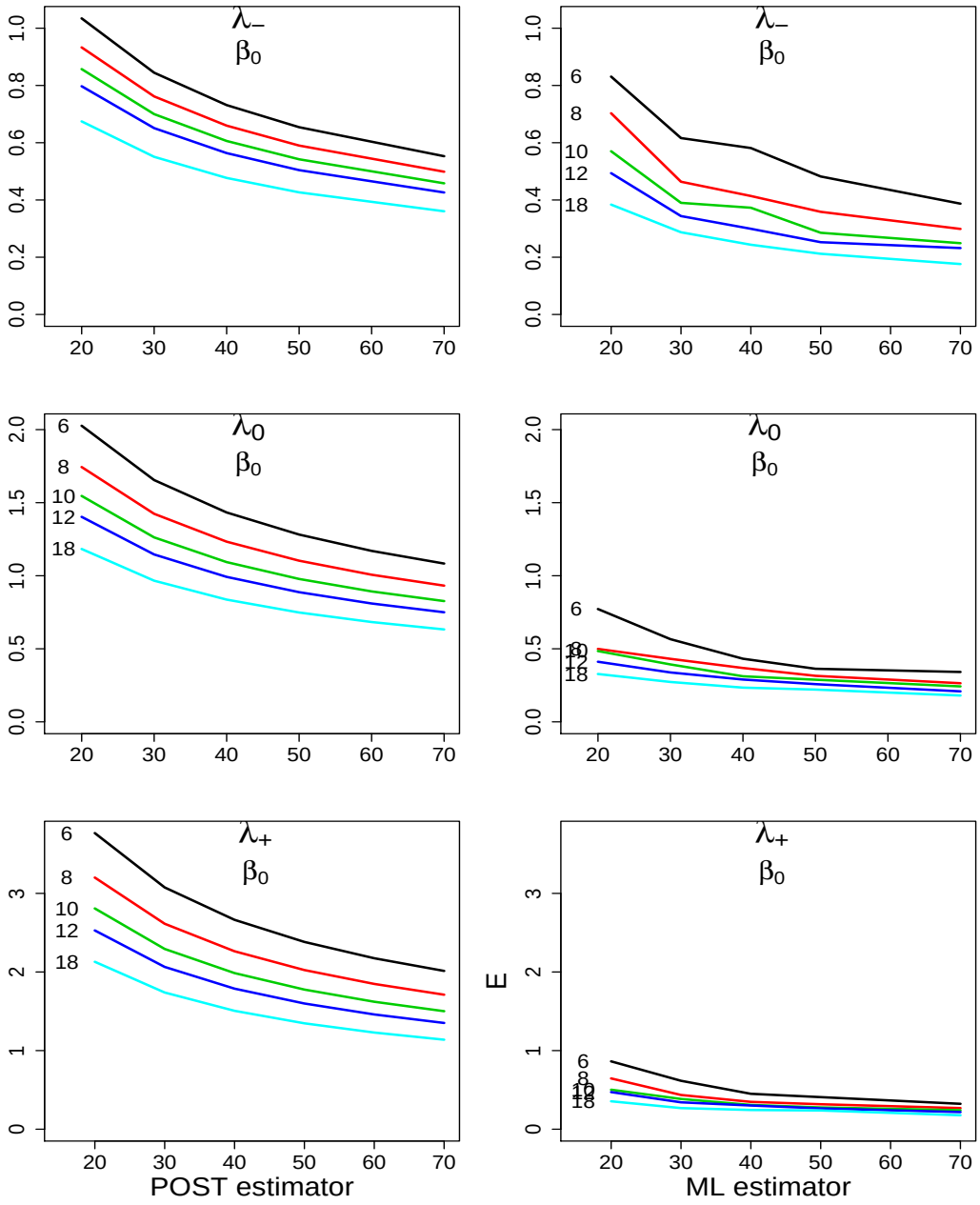

Figure B.23: Estimated standard deviation vs. m for the POST and ML estimators with $\boldsymbol{\beta}_0$ and varying λ_h (curves correspond to the specified number of scans)

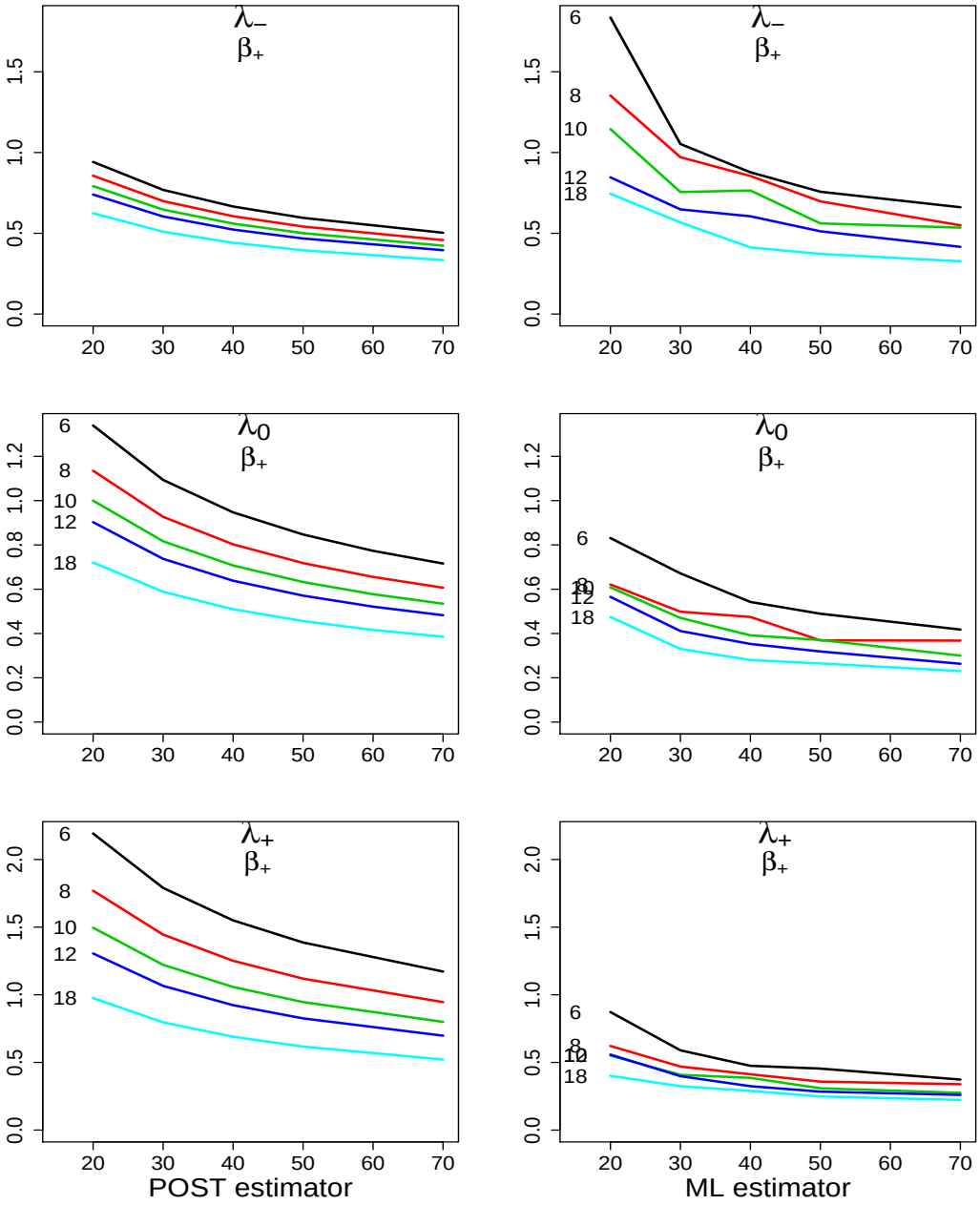

Figure B.24: Estimated standard deviation vs. m for the POST and ML estimators with $\boldsymbol{\beta}_+$ and varying λ_h (curves correspond to the specified number of scans)

Bibliography

Bibliography

[1] Albert, P.S., McFarland, H.F., Smith, M.E., and Frank, J.A (1994). Repeated Measures - Time series for modelling counts from a relapsing-remitting disease: application to modelling disease activity in multiple sclerosis. *Statistics in Medicine*, **13**, 453-466.

[2] Altman, R.M. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *Journal of The American Statistical Association*, **102**, 201210.

[3] Grimaud, S., Barker, G.J., Wang, L., Lai, M., MacManns, D.G., Webb, S.L., Thompson, A.J., McDonald, W.I., Tofts, P.S., and Miller, D.H. (1999). Correlation of magnetic resonance imaging parameters with clinical disability in multiple sclerosis. *Journal of Neurology*, **246**, 961-967.

[4] Nash, J.C. (1979), *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*, New York: John Wiley & Sons.

[5] Paty, D., Li, D., and Zhao, G.J. (1999). *MRI in multiple sclerosis: Implications for diagnosis and treatment* Ares-Serono International S.A., Switzerland.

[6] Smith, A. (1999). Design Strategies for Repeated MRI Scanning in MS Clinical Trials. *Unpublished M.Sc. Thesis* University of British Columbia, Canada, Department of Statistics.

[7] Sormani, M.P. and Filippi, M. (2007). Statistical issues related to the use of MRI data in multiple sclerosis. *Journal of Neuroimaging*, **17**, 56S-59S.

[8] Sormani, M.P., Miller, D.H., Comi, G., Barkhofe, F., Rovarisa, M., Bruzzic, P., and Filippia, M. (2001). Clinical trials of multiple sclerosis monitored with enhanced MRI: new

sample size calculations based on large data sets. *Journal of Neurology, Neurosurgery, & Psychiatry*, **70**, 494-499.

[9] Sormani M.P., Bruzzi, P., Miller, D.H., Gasperini, C., Barkhof, F., and Filippi, M. (1999). Modelling MRI enhancing lesion counts in multiple sclerosis using a negative binomial model: implications for clinical trials. *Journal of the Neurological Sciences*, **163**, 74-80.

Index