Parameter Selection for Ordinary Least Square Estimation of Contact Processes: Revisiting the Dissemination of Scientific Innovation

Caleb Shimberg
Department of Mathematics and Statistics
East Tennessee State University

Abstract

There has recently been great interest in modeling the spread of ideas. It has been found that epidemiological models can be applied to the spread of ideas through a population, modeling them in much the same way as the spread of disease. Applying a subset selection algorithm based on the sensitivity matrix, we will calculate the reliability of optimal reduced parameter vectors for which estimation is to be sought. It should be noted that this algorithm requires prior knowledge of a nominal data set of values for all parameters and constant variances,σ_0^2, in observations. We will also further analyze the sensitivity matrices for the reduced parameter vectors in order to compare the magnitude of change in the number of people in a population who adopt an idea due to relatively equal changes in each of the parameters. In this way we will assess the influence, over time, of several different factors on the dissemination of a scientific idea.

Introduction

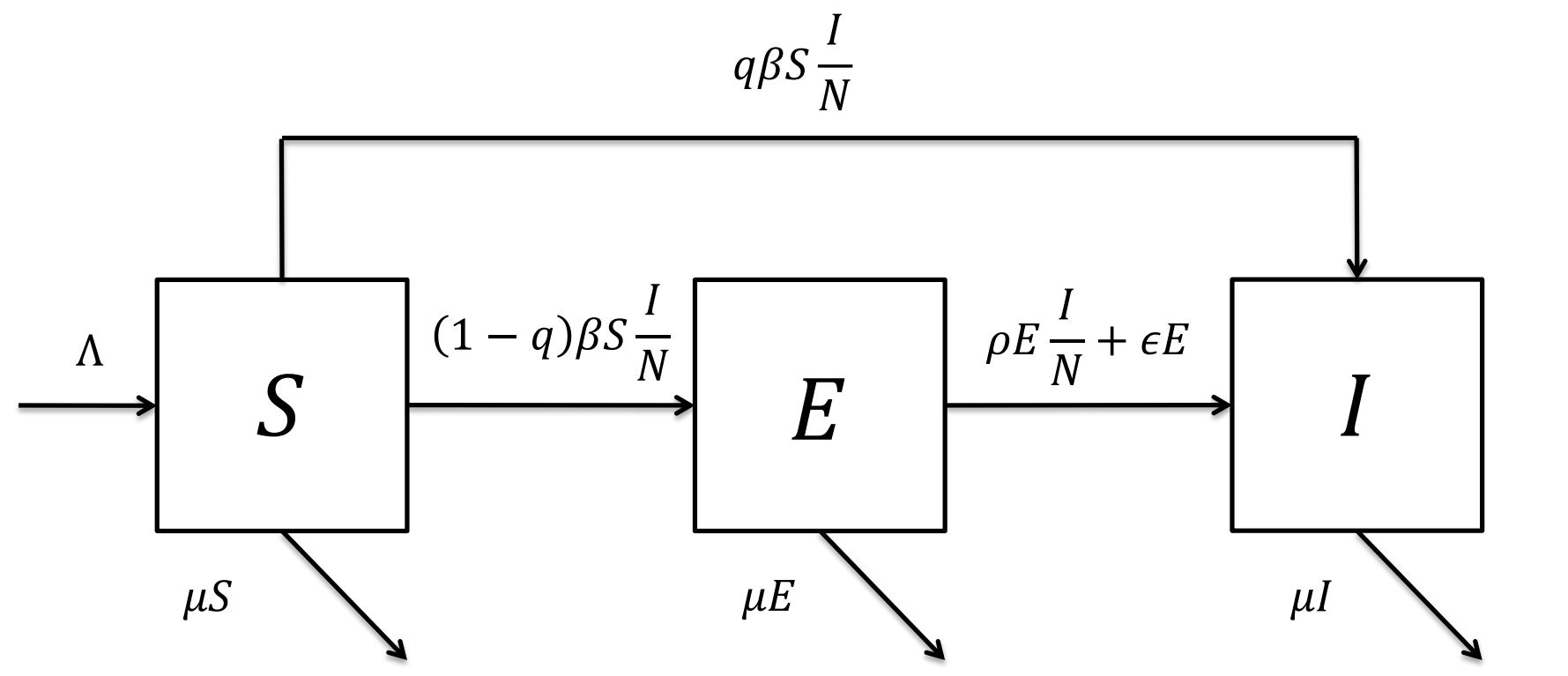
In this poster we seek to explore the influence of various factors on the acceptance of an idea. It is important to first understand what we mean by "ideas". While we use the term literally, we also use it loosely. An idea can be just that: A thought, a belief, or an opinion. However, an idea can also be a technological advancement, such as the use of biodiesel; an unproven theorem in graph theory; or even, as here, the preference for a method of diagramming electromagnetic interactions between subatomic particles. We focus here on the spread of the use of Feynman diagrams because their acceptance is well documented. It is important to note, however, the plethora of possible applications.

We will begin by applying a parameter selection algorithm to a simple epidemiological model that explains the spread of ideas through a population. This parameter selection algorithm will both identify the optimal parameter vector subset for estimation and quantify the uncertainty associated with each possible parameter vector subset.

Once we have reliable estimation for the parameters we choose to estimate, we will analyze their corresponding sensitivity matrices. In this way, we will explore the influence of each of those parameters on the infected class of the population.

Mathematical Model

We propose the following model, with the following state variables and system parameters:



Variable / Parameter	Definition
S	Susceptibles
E	Idea Apprentices
I	Idea Adopters
N	*S+E+I*
Λ	Recruitment Rate
1/μ	Average Lifetime of the Idea
ϵ	Rate of Individual Progression to Adoption
β	Per-capita S-I Contact Rate
ρ	Per-capita E-I Contact Rate
q	S→I Transition Probability Given Contact With Adopters
(1-q)	S→E Transition Probability Given Contact With Adopters

This model implies the following non-linear series of differential equations:

$$\frac{dS}{dt}=\Lambda-\beta S\frac{I}{N}-\mu S=g_1$$
$$\frac{dE}{dt}=(1-q)\beta S\frac{I}{N}-\left(\rho E\frac{I}{N}+\epsilon E\right)-\mu E=g_2$$
$$\frac{dI}{dt}=q\beta S\frac{I}{N}+\left(\rho E\frac{I}{N}+\epsilon E\right)-\mu I=g_3$$

Statistical Model for the Observation Process

We assume that each of the *n* longitudinal observations is a realization of $I(t,\theta_0)$ where θ_0 is the theoretical *true parameter vector* and that each realization is affected by random deviations from the true underlying process. Thus the statistical model is given by

$$Y_i=I(t_i,\theta_0)+\epsilon_i \; for \; i=1,\dots,n\,.$$

We assume that the errors, ϵ_i, have a mean of zero, some common, finite variance, and that they are independent of each other.

Where θ_{OLS}^n minimizes

$$\sum_{i=1}^n[Y_i-I(t_i,\theta)]^2,$$

we know from the central limit theorem that

$$\theta_{OLS}^n\sim N_p(\theta_0,\Sigma_0^n)\,,$$

where $\Sigma_0^n=\sigma_0^2[n\Omega_0]^{-1}\in\mathbb{R}^{p\times p}$ and $\Omega_0=\lim_{n\to\infty}\frac{1}{n}\chi^n(\theta_0)^T\chi^n(\theta_0)$. Asymptotic theory requires both the existence of this limit and the non-singularity of Ω_0.

The $n\times p$ matrix $\chi^n(\theta_0)$ is known as the *sensitivity matrix* of the system and is defined by:

$$\chi_{ij}^n(\theta_0)=\frac{\partial I(t_i,\theta)}{\partial\theta_j}\,, 1\le i\le n\,, 1\le j\le p$$

If we define $\theta=(S_0,E_0,I_0,\Lambda,\mu,\epsilon,\beta,\rho,q)$ and $X=(x_1,x_2,x_3)=(S,E,I)$, then numerical values of $\chi^n(\theta)$ can be readily calculated for a given θ by solving

$$\frac{dx}{dt}=g(t,x(t,\theta);\theta)$$
$$\frac{d}{dt}\frac{\partial x}{\partial\theta}=\frac{\partial g}{\partial x}\frac{\partial x}{\partial\theta}+\frac{\partial g}{\partial\theta}$$

from $t=t_0$ to $t=t_n$.

Parameter Selection Algorithm

It is standard practice to reduce the number of parameters in a statistical model. For example, say we have a model with $\theta=(\theta_1,\theta_2,\theta_3,\theta_4,\theta_5)$. If literature suggests reasonable nominal values, *a* and *b*, for θ_1 and θ_4, then we can set $\theta_1=a$ and $\theta_4=b$ leaving only the parameters $(\theta_2,\theta_3,\theta_5)$ to be estimated. We would call the three remaining parameters the "active parameters." In more general terms, we would say that we are interested in all vectors with 3 active parameters from a 5-vector, for which there are $\binom{5}{3}=10$ possible choices.

The parameter selection algorithm used here can be thought of as doing three things:

1. It generates all possible reduced parameter vectors.
2. It discards all vectors for which χ is rank-deficient.
3. It calculates "selection scores" $\alpha(\theta)$ as follows.

To calculate the selection scores we will first find the covariance matrix Σ for a given parameter vector θ:

$$\Sigma(\theta)=\sigma_0^2[\chi(\theta)^T\chi(\theta)]^{-1}$$

Next we will define a vector v, analogous to a vector of coefficients of variation:

$$v_i(\theta)=\frac{\sqrt{(\Sigma(\theta))_{ii}}}{\theta_i}$$

To determine the selection score, we simply take the norm of that vector:

$$\alpha(\theta)=\|v(\theta)\|$$

Analysis of the χ Sensitivity Matrix

In order to better understand the influence of each parameter on the number of "Idea Adopters", *I*, we will analyze the χ sensitivity matrix. For each active parameter there is a corresponding column of the sensitivity matrix that represents the instantaneous rate of change in the infected population *I* with respect to that parameter, as a function of time. We say that, at a given time, one parameter is more "influential" than another if, at that time, a change in that parameter corresponds to a larger change in the infected population than an equal change in another parameter.

Results

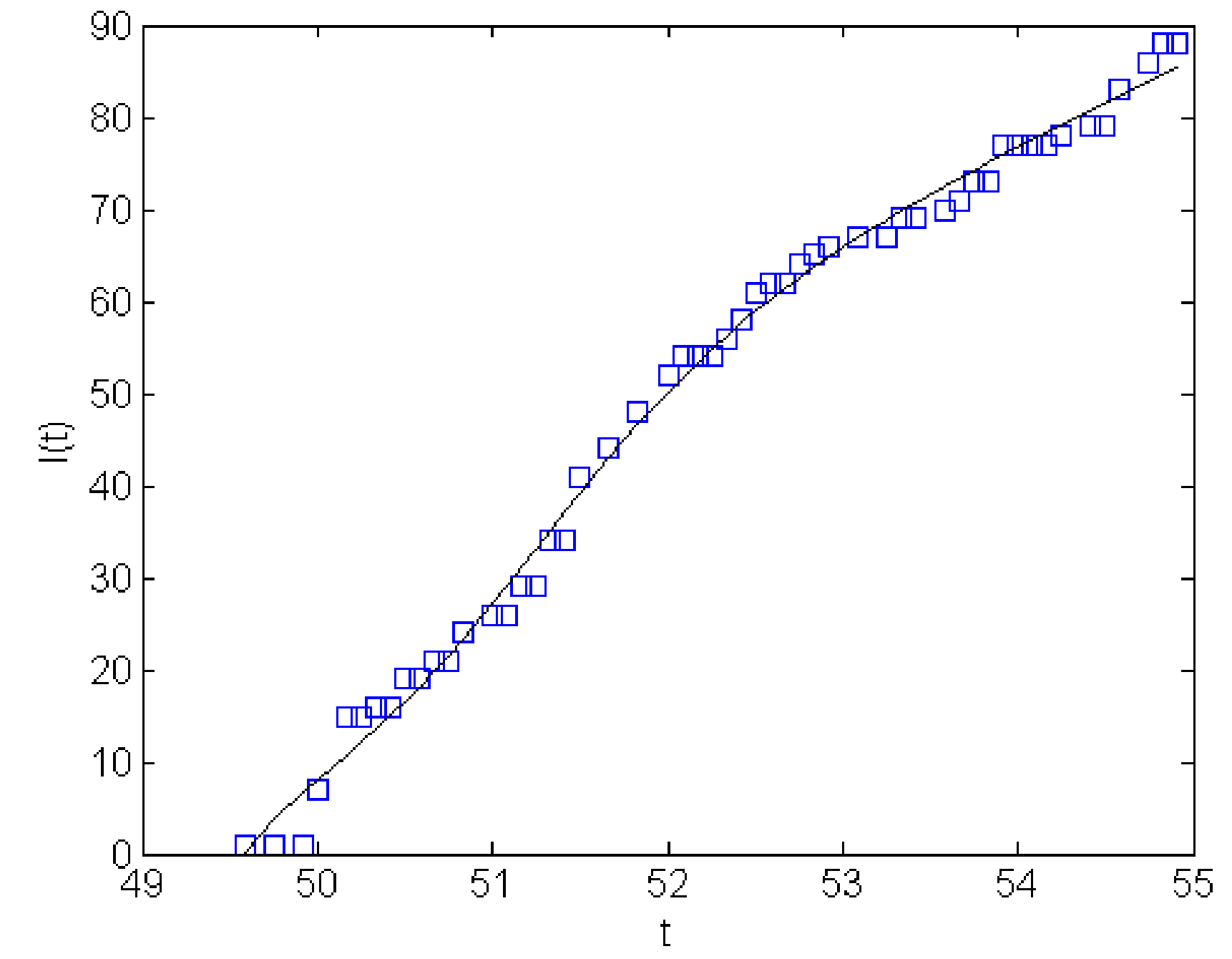
Below is a table of the smallest selection scores for each possible number of active parameters.

# of active parameters	1	2	3	4	5	6	7	8	9
smallest possible α(θ)	0.0431	0.1774	0.3306	1.0105	8.4530	18.7093	47.4071	114.2555	327.3306

Clearly the uncertainty inherent in estimating more than four parameters precludes the use of an active parameter vector so large. For the sake of brevity, we will proceed to further analyze only active parameter vectors of length 4.

First we will plot $I(t,\hat{\theta}_{OLS})$, the "curve of best fit" for 4 active parameters.

Figure 1



We now plot the χ sensitivity matrices for the first and second "best" choices for active parameter vectors of length 4, along with some statistics associated with the parameters.

Figure 2

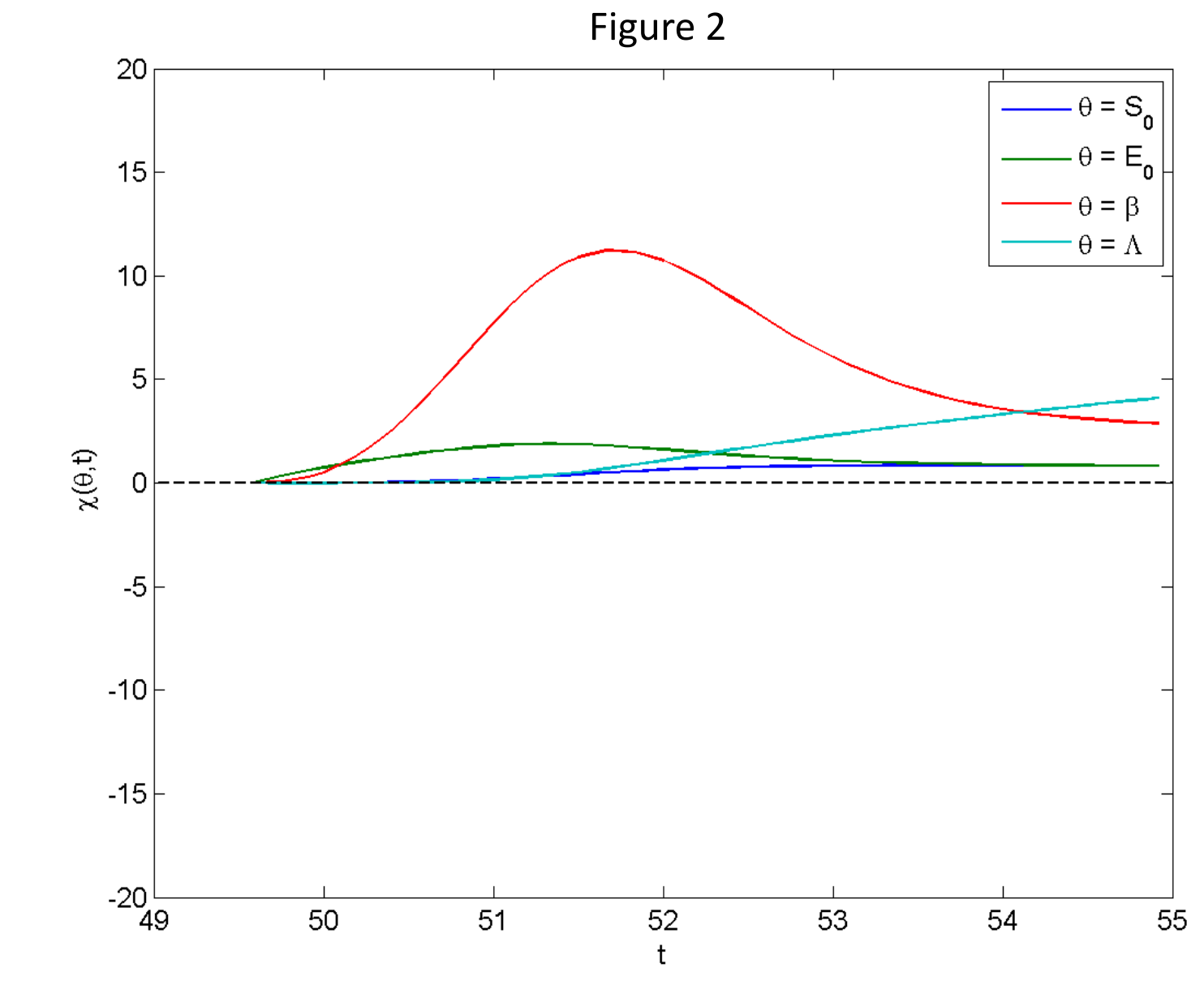
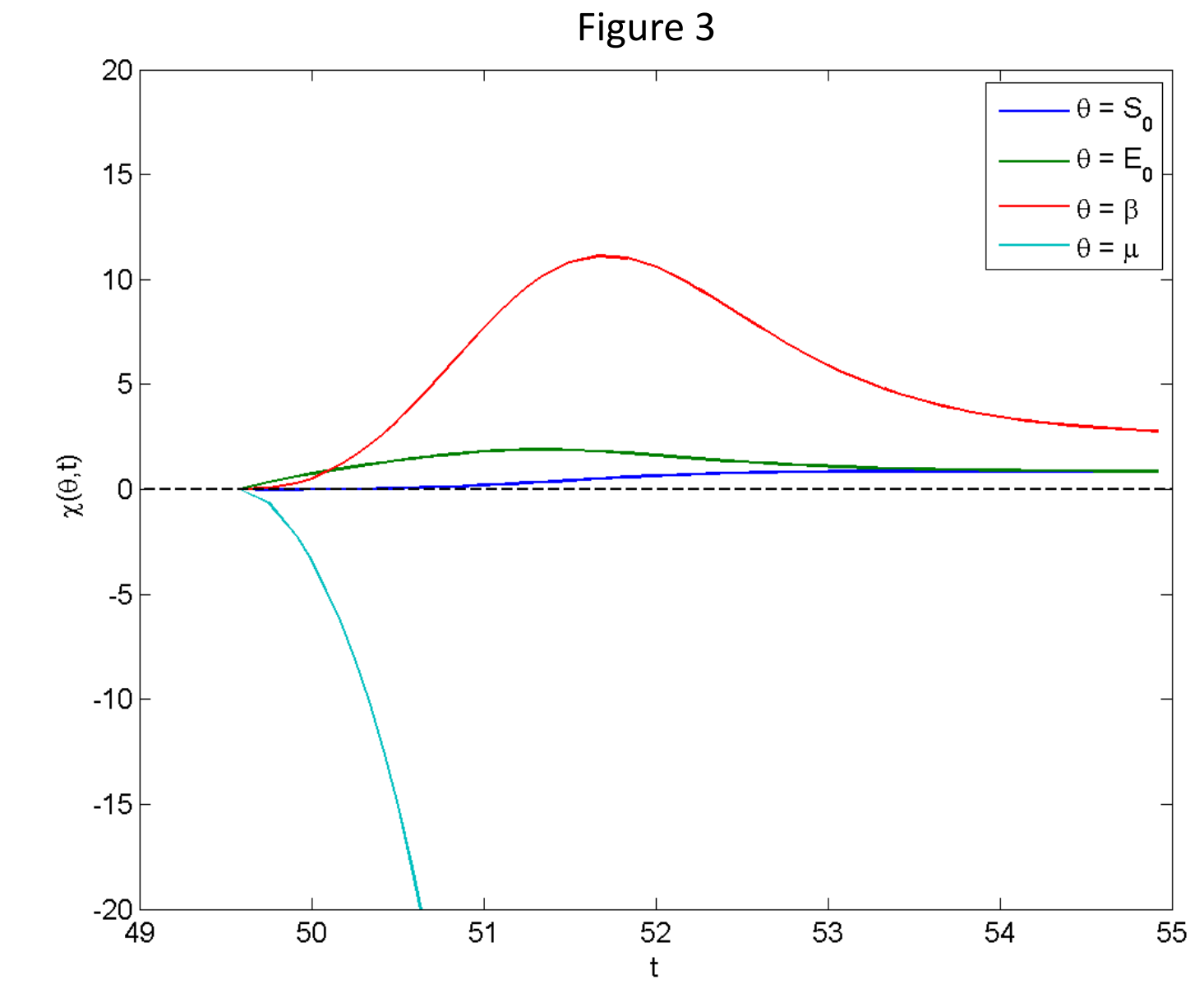


Figure 3



	S_0	E_0	I_0	β	ρ	ϵ	q	Λ	μ
Nominal Value			0.1		2	2.36	0.1		0.039
Estimate	30.4	10.6		2.34				12.5	
Standard Error	4.1	0.933		0.275				0.935	
Coefficient of Variation	0.135	0.0877		0.118				0.0746	

	S_0	E_0	I_0	β	ρ	ϵ	q	Λ	μ
Nominal Value			0.1		2	2.36	0.1	12.1	
Estimate	30.4	10.6		2.33					0.0322
Standard Error	3.4	0.932		0.238					0.0106
Coefficient of Variation	0.112	0.0882		0.102					0.33

In Figure 3 we notice that, at time $t=51$, a single unit increase in μ corresponds to a greater than 20 unit decrease in the size of the infected population, while a single unit increase in β (the next most influential parameter) corresponds to a less than 10 unit increase in the size of the infected population. This is largely because a single unit increase in μ is much more meaningful than a single unit increase in β. In order to more meaningfully compare the influence of the parameters, we define a standardized sensitivity matrix $\zeta(\theta_{OLS})$ such that changes in parameter values as well as corresponding changes in the size of the infected population are measured in percent increases and decreases. The standardized sensitivity matrix is calculated as follows:

$$\zeta_{ij}^n(\theta_{OLS})=\frac{\theta_j}{I(t_i,\theta)}\frac{\partial I(t_i,\theta)}{\partial\theta_j}\quad 1\le i\le n, 1\le j\le p.$$

We now plot the ζ standardized sensitivity matrices for the first and second best choices for active parameter vectors of length 4.

Figure 4

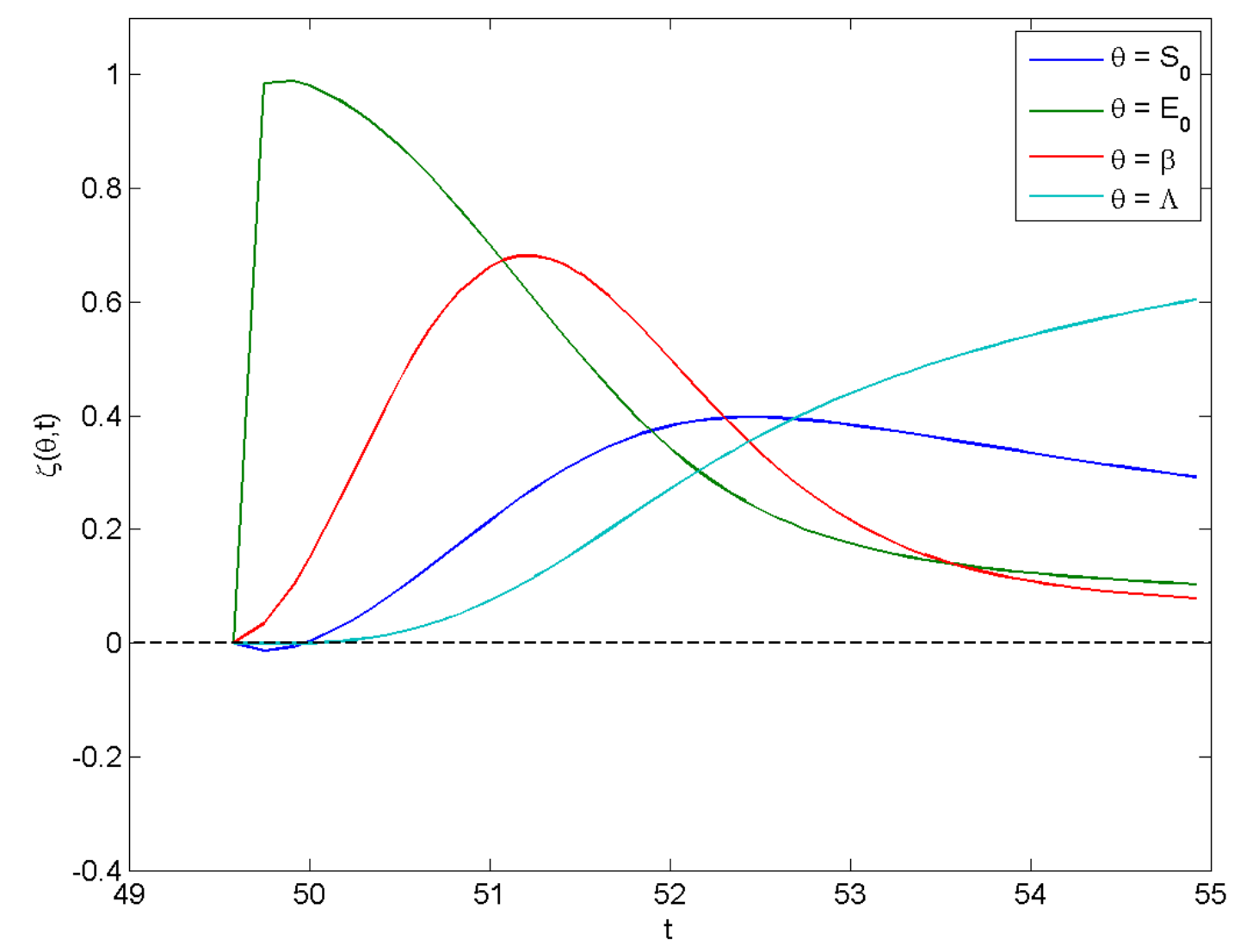
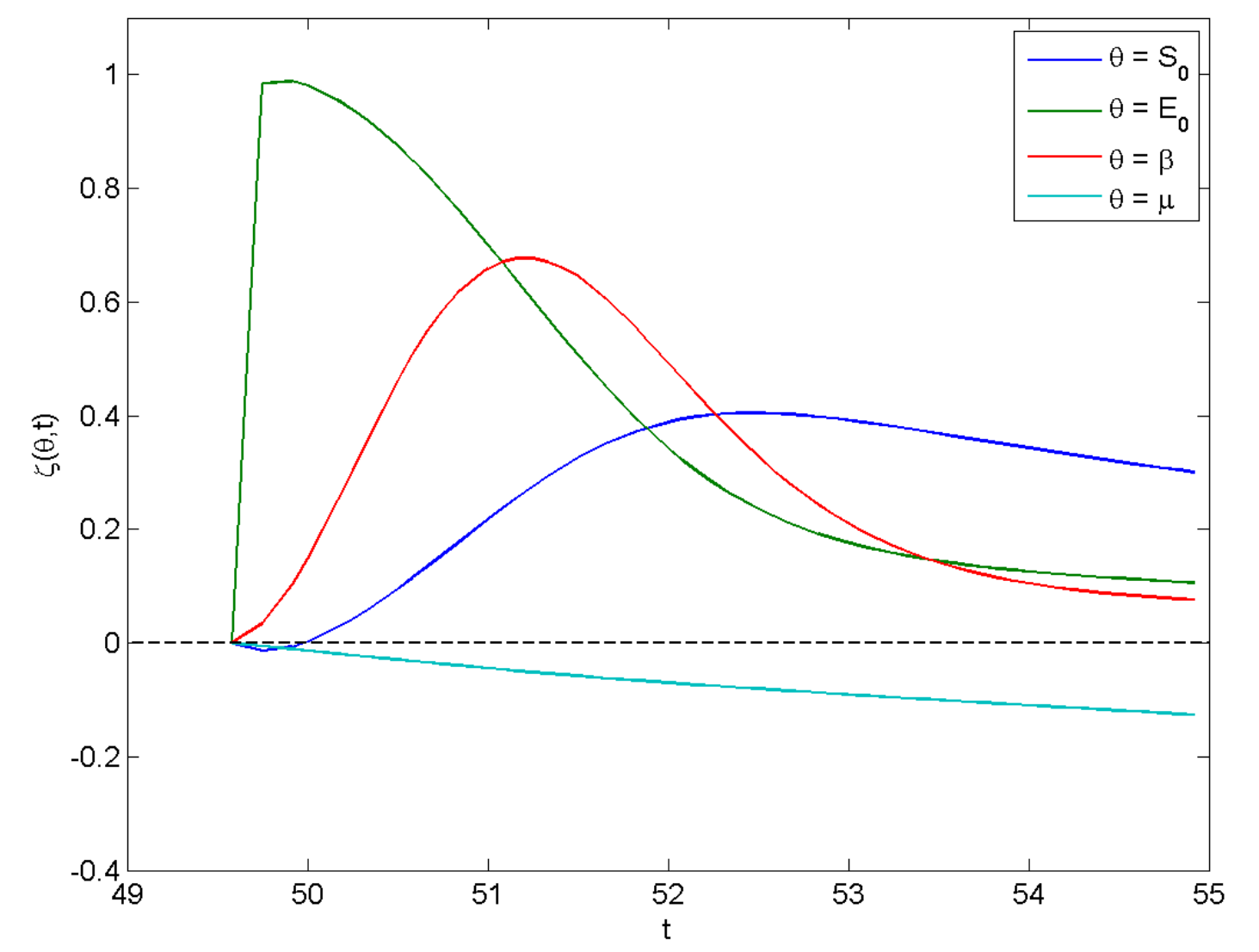


Figure 5



Summary

We have applied a parameter selection algorithm to an epidemiological model designed to model the spread of ideas through a population. We have quantified the uncertainty inherent in each active parameter vector and, based on that, selected both the optimal reduction sizes and the parameters to be estimated for a reduced parameter vector of given length. We have further analyzed the sensitivity matrices for each of these reduced parameter vectors, and through that analysis quantified not only the rate of change in the infected population due to a given parameter at a given time, but also standardized the sensitivity matrices in order to more meaningfully compare each parameter's influence on the infected population over time.

References

[1] A. Cintron-Arias, H.T. Banks, A. Capaldi and A.L. Lloyd *A sensitivity matrix based methodology for inverse problem formulation*, J. Inv. Ill-Posed Problems **17** (2009), 545-564

[2] A. Capaldi, S. Behrend, B. Berman, J. Smith, J. Wright, and A.L. Lloyd *Parameter Estimation and Uncertainty Quantification for an Epidemic Model*, (2009)

[3] L. M.A. Bettencourt, A. Cintron-Arias, D.I. Kaiser, and C. Castillo-Chavez *The Power of a Good Idea: Quantitative Modeling of the Spread of Ideas from Epidemiological Models* Physica A: Statistical Mechanics and its Applications, 364 (May 2006), 513-536

[4]D. Kaiser, K. Ito, and K. Hall *Spreading the Tools of Theory: Feynman Diagrams in the USA, Japan, and the Soviet Union* Social Studies of Science 36 (2004), 533-564