An Inverse Problem Statistical Methodology Summary

H.T. Banks, M. Davidian and J.R. Samuels, Jr.
Center for Research in Scientific Computation
and
Center for Quantitative Sciences in Biomedicine
North Carolina State University
Raleigh, NC 27695-8205

August 23, 2007

This summary was presented as part of a lecture series at the workshop *Biomedical Modeling and Cardiovascular-Respiratory Control: Theory and Practice Summer School and Workshop;* A Marie Curie Training Event, Schloss Seggau, Leibnitz, Austria, July 22-28, 2007.

Outline

1. Parameter Estimation: MLE, OLS, GLS
2. Computation of Σ, Standard Errors and Confidence Intervals
3. Model Comparison Techniques

1 Parameter Estimation: MLE, OLS, and GLS

1.1 The Underlying Mathematical and Statistical Models

We consider inverse or parameter estimation problems in the context of a parameterized (with vector parameter $\vec{\theta}$) dynamical system or **mathematical model**

$$\frac{d\vec{x}}{dt}(t) = \vec{g}(t, \vec{x}(t), \vec{\theta}) \tag{1}$$

with **observation process**

$$\vec{y}(t) = \mathcal{C}\vec{x}(t; \vec{\theta}). \tag{2}$$

Following usual convention (which agrees with the data usually available from experiments), we assume a discrete form of the observations in which one has n longitudinal observations corresponding to

$$\vec{y}(t_j) = \mathcal{C}\vec{x}(t_j;\vec{\theta}), \quad j = 1, \ldots, n. \tag{3}$$

In general the corresponding observations or data $\{\vec{y}_j\}$ will not be exactly $\vec{y}(t_j)$ and hence we choose to treat this uncertainty pertaining to the observations with a statistical model for the observation process.

1.2 Description of Statistical Model

We consider a **statistical model** of the form

$$\vec{Y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j, \quad j = 1, \ldots, n, \tag{4}$$

where $\vec{f}(t_j, \vec{\theta}) = \mathcal{C}\vec{x}(t_j;\vec{\theta})$, $j = 1, \ldots, n$, corresponds to the solution of the mathematical model (1) at the j^{th} covariate for a particular vector of parameters $\vec{\theta} \in R^p, \vec{x} \in R^N, \vec{f} \in R^m$, and $\mathcal{C}$ is an $m \times N$ matrix. The term $\vec{\theta}_0$ represents the "truth" or the parameters that generate the observations $\{\vec{Y}_j\}_{j=1}^n$. The term $\vec{\epsilon}_j$ can represent measurement error, "system fluctuations" or other phenomena that cause observations to not fall exactly on the points $\vec{f}(t_j, \vec{\theta})$ from the smooth path $\vec{f}(t, \vec{\theta})$. Since these fluctuations are unknown to the modeler, we will assume $\vec{\epsilon}_j$ is generated from a probability distribution that reflects the assumptions regarding these phenomena. For instance, in a statistical model for pharmacokinetics of drug in human blood samples, a natural distribution for $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ might be the multivariate normal distribution [2].

The purpose of our presentation here is to discuss methodology related to the estimation of the true value of the parameters $\vec{\theta}_0$ from a set Θ of admissible parameters and the variance of the error $\text{var}(\vec{\epsilon}_j)$. We discuss two inverse problem methodologies that can be used to calculate estimates $\hat{\theta}$ for $\vec{\theta}_0$: the ordinary least-squares (OLS) and generalized least-squares (GLS) formulations as well as the popular maximum likelihood estimate (MLE) formulation in the case one assumes the distributions of the error process $\{\vec{\epsilon}_j\}$ are known.

1.3 Known error processes: Normally distributed error

In the introduction of the statistical model we initially made no mention of the probability distribution that generates the error $\vec{\epsilon}_j$. In many situations one readily assumes that the errors $\vec{\epsilon}_j = 1, \ldots, n$, are independent and identically distributed. We discuss a case where one is able to make further assumptions on the error, namely that the distribution is known. In this case maximum likelihood techniques may be used. We discuss first one such case for a scalar observation system, i.e., $m = 1$. If, in addition, there is sufficient evidence to suspect the error is generated by a normal distribution then we may be willing to assume $\epsilon_j \sim \mathcal{N}(0, \sigma_0^2)$, and hence $Y_j \sim \mathcal{N}(f(t_j, \vec{\theta}_0), \sigma_0^2)$. We can then obtain an expression for determining $\vec{\theta}_0$ and σ_0 by seeking the maximum over $(\vec{\theta}, \sigma^2) \in \Theta \times (0, \infty)$ of the likelihood function for $\epsilon_j = Y_j - f(t_j, \vec{\theta})$ which is defined by

$$L(\vec{Y}|\vec{\theta}, \sigma^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}[Y_j - f(t_j, \vec{\theta})]^2\}. \tag{5}$$

The resulting solutions θ_{MLE} and σ^2_{MLE} are the maximum likelihood **estimators** (MLEs) for $\vec{\theta}_0$ and σ_0^2, respectively. We point out that these solutions $\theta_{\text{MLE}} = \theta_{\text{MLE}}(\vec{Y})$ and $\sigma^2_{\text{MLE}} = \sigma^2_{\text{MLE}}(\vec{Y})$ are *random variables* by virtue of the fact that $\vec{Y}$ is a random variable. The corresponding maximum likelihood **estimates** are obtained by maximizing (5) with $\{Y_j\}$ replaced by a given realization $\vec{y} = \{y_j\}$ and will be denote by $\hat{\theta}_{\text{MLE}}$ and $\hat{\sigma}_{\text{MLE}}$ respectively.

Maximizing (5) is equivalent to maximizing the log likelihood

$$\log L(\vec{Y}|\vec{\theta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{j=1}^{n}[Y_j - f(t_j, \vec{\theta})]^2. \tag{6}$$

We determine the maximum of (6) by differentiating with respect to $\vec{\theta}$ (with σ^2 fixed) and with respect to σ^2 (with $\vec{\theta}$ fixed), setting the resulting equations equal to zero and solving for $\vec{\theta}$ and σ^2. With σ^2 fixed we solve $\frac{\partial}{\partial\vec{\theta}}\log L(\vec{Y}|\vec{\theta}, \sigma^2) = 0$ which is equivalent to

$$\sum_{j=1}^{n}[Y_j - f(t_j, \vec{\theta})]\nabla f(t_j, \vec{\theta}) = 0. \tag{7}$$

We see that solving (7) is the same as the least squares optimization

$$\theta_{\text{MLE}}(\vec{Y}) = \arg\min_{\vec{\theta}\in\Theta} J(\vec{Y}, \vec{\theta}) = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n}[Y_j - f(t_j, \vec{\theta})]^2. \tag{8}$$

We next fix $\vec{\theta}$ to be θ_{MLE} and solve $\frac{\partial}{\partial\sigma^2}\log L(\vec{Y}|\theta_{\text{MLE}}, \sigma^2) = 0$, which yields

$$\sigma^2_{\text{MLE}}(\vec{Y}) = \frac{1}{n}J(\vec{Y}, \theta_{\text{MLE}}). \tag{9}$$

Note that we can solve for θ_{MLE} and σ^2_{MLE} separately – a desirable feature, but one that won't arise in more complicated formulations discussed below. The 2^{nd} derivative test (which is omitted here) verifies that the expressions above for θ_{MLE} and σ^2_{MLE} do indeed maximize (6).

If, however, we have a vector of observations for the j^{th} covariate t_j then the statistical model is reformulated as

$$\vec{Y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j \tag{10}$$

where $\vec{f} \in R^m$ and

$$V_0 = \text{var}(\vec{\epsilon}_j) = \text{diag}(\sigma^2_{0,1}, \ldots, \sigma^2_{0,m}) \tag{11}$$

for $j = 1, \dots, n$. In this setting we have allowed for the possibility that the observation coordinates Y_j^i may have different *constant* variances $\sigma_{0,i}^2$, i.e., $\sigma_{0,i}^2$ does not necessarily have to equal $\sigma_{0,k}^2$. If (again) there is sufficient evidence to claim the errors are independent identically distributed and generated by a normal distribution then $\vec{\epsilon_j} \sim \mathcal{N}_m(0, V_0)$. We thus can obtain the maximum likelihood estimators $\theta_{\text{MLE}}(\{\vec{Y}_j\})$ and $V_{\text{MLE}}(\{\vec{Y}_j\})$ for θ_0 and V_0 by determining the maximum of log of the likelihood function for $\vec{\epsilon}_j = \vec{Y}_j - \vec{f}(t_j, \vec{\theta})$ defined by

$$\begin{aligned}\log L(\{Y_j^1, \dots, Y_j^m\}|\vec{\theta}, V) &= -\frac{n}{2}\sum_{i=1}^m \log\sigma_{0,i}^2 - \frac{1}{2}\sum_{i=1}^m \frac{1}{\sigma_{0,i}^2}\sum_{j=1}^n [Y_j^i - f^i(t_j, \vec{\theta})]^2 \\ &= -\frac{n}{2}\sum_{i=1}^m \log\sigma_{0,i}^2 - \sum_{j=1}^n [\vec{Y}_j - \vec{f}(t_j, \vec{\theta})]^T V^{-1}[\vec{Y}_j - \vec{f}(t_j, \vec{\theta})].\end{aligned}$$

Using arguments similar to those given for the scalar case, we determine the maximum likelihood estimators for $\vec{\theta}_0$ and V_0 to be

$$\theta_{\text{MLE}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^n [\vec{Y}_j - \vec{f}(t_j, \vec{\theta})]^T V_{\text{MLE}}^{-1} [\vec{Y}_j - \vec{f}(t_j, \vec{\theta})] \tag{12}$$

$$V_{\text{MLE}} = \text{diag}\left(\frac{1}{n}\sum_{j=1}^n [\vec{Y}_j - \vec{f}(t_j, \theta_{\text{MLE}})][\vec{Y}_j - \vec{f}(t_j, \theta_{\text{MLE}})]^T\right). \tag{13}$$

Unfortunately, this is a coupled system, which requires some care when solving numerically. We will discuss this issue further in Sections 1.4.2 and 1.4.5 below.

1.4 Unspecified Error Distributions and Asymptotic Theory

In section 1.3 we examined the estimates of $\vec{\theta}_0$ and V_0 under the assumption *that the error is normally distributed and is constant longitudinally.* But what if it is suspected that the error is not normally distributed, or the error's distribution is completely unknown to the modeler (as in most applications)? How should we proceed in estimating $\vec{\theta}_0$ and σ_0 (or V_0) in these circumstances? In this section we will review two estimation procedures for such situations: ordinary least squares (OLS) and generalized least squares (GLS).

1.4.1 Ordinary Least Squares (OLS)

The statistical model in the scalar case takes the form

$$Y_j = f(t_j, \vec{\theta}_0) + \epsilon_j \tag{14}$$

where the variance $\text{var}(\epsilon_j) = \sigma_0^2$ is constant in longitudinal data (note that the error's distribution is not specified). If we define

$$\theta_{\text{OLS}}(\vec{Y}) = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^n [Y_j - f(t_j, \vec{\theta})]^2 \tag{15}$$

then θ_{OLS} can be viewed as minimizing the distance between the data and model where all observations are treated as of equal importance. We note that minimizing in (15) corresponds [15] to solving for $\vec{\theta}$ in

$$\sum_{j=1}^{n}[Y_j - f(t_j,\vec{\theta})]\nabla f(t_j,\vec{\theta}) = 0. \tag{16}$$

We point out that θ_{OLS} is a *random variable* ($\epsilon_j = Y_j - f(t_j,\vec{\theta})$ is a random variable); hence if $\{y_j\}_{j=1}^n$ is a realization of the *random process* $\{Y_j\}_{j=1}^n$ then solving

$$\hat{\theta}_{\text{OLS}} = \arg\min_{\vec{\theta}\in\Theta}\sum_{j=1}^{n}[y_j - f(t_j,\vec{\theta})]^2 \tag{17}$$

provides an realization for θ_{OLS}.

Once we have solved for θ_{OLS} in (15), we can replace $\vec{\theta}_0$ in

$$\sigma_0^2 = \frac{1}{n}E[\sum_{j=1}^{n}[Y_j - f(t_j,\vec{\theta}_0)]^2] \tag{18}$$

by $\hat{\theta}_{\text{OLS}}$ to obtain an estimate $\hat{\sigma}^2_{\text{OLS}}$ for σ_0^2.

Even though the error's distribution is not specified we can use asymptotic theory to approximate the mean and variance of the random variable θ_{OLS} [22]. As will be explained in more detail below, as $n \to \infty$, we have that

$$\theta_{\text{OLS}} \sim \mathcal{N}_p(\vec{\theta}_0, \sigma_0^2[\chi^T(\vec{\theta}_0)\chi(\vec{\theta}_0)]^{-1}) = \mathcal{N}_p(\vec{\theta}_0, \Sigma_0) \tag{19}$$

where the sensitivity matrix $\chi(\vec{\theta}) = \{\chi_{jk}\}$ is defined as

$$\chi_{jk}(\vec{\theta}) = \frac{\partial f(t_j,\vec{\theta})}{\partial \vec{\theta}_k}.$$

However, $\vec{\theta}_0$ and σ_0^2 are generally unknown, so one usually will instead use the *realization* $\vec{y} = \{y_j\}_{j=1}^n$ of the random process $\vec{Y}$ to obtain the estimate

$$\hat{\theta}_{\text{OLS}} = \arg\min_{\vec{\theta}\in\Theta}\sum_{j=1}^{n}[y_j - f(t_j,\vec{\theta})]^2 \tag{20}$$

and the *bias adjusted* estimate

$$\hat{\sigma}^2_{\text{OLS}} = \frac{1}{n-p}\sum_{j=1}^{n}[y_j - f(t_j,\hat{\theta})]^2 \tag{21}$$

to use as an approximation in (19).

We note that (21) represents the estimate for σ_0^2 of (18) with the factor $\frac{1}{n}$ replaced by the factor $\frac{1}{n-p}$ (in the linear case the estimate with $\frac{1}{n}$ can be shown to be biased downward and the same behavior can be observed in the general nonlinear case– see Chap. 12 of [22] and p. 63 of [15]). We remark that (18) is true even in the general nonlinear case (it does not rely on any asymptotic theories).

Both $\hat{\theta} = \hat{\theta}_{\text{OLS}}$ and $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{OLS}}$ will then be used to approximate the covariance matrix

$$\Sigma_0 \approx \hat{\Sigma} = \hat{\sigma}^2[\chi^T(\hat{\theta})\chi(\hat{\theta})]^{-1}. \tag{22}$$

We can obtain the standard errors $SE(\hat{\theta}_{\text{OLS},k})$ (discussed in more detail in the next section) for the k^{th} element of $\hat{\theta}_{\text{OLS}}$ by calculating $SE(\hat{\theta}_{\text{OLS},k}) \approx \sqrt{\hat{\Sigma}_{kk}}$. Also note the similarity between the MLE equations (8) and (9), and the scalar OLS equations (20) and (21). That is, under a normality assumption for the error, the MLE and OLS formulations are equivalent.

If, however, we have a vector of observations for the j^{th} covariate t_j and we assume the variance is still constant in longitudinal data, then the statistical model is reformulated as

$$\vec{Y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j \tag{23}$$

where $\vec{f} \in R^m$ and

$$V_0 = \text{var}(\vec{\epsilon}_j) = \text{diag}(\sigma^2_{0,1}, \ldots, \sigma^2_{0,m}) \tag{24}$$

for $j = 1, \ldots, n$. Just as in the MLE case we have allowed for the possibility that the observation coordinates Y^i_j may have different *constant* variances $\sigma^2_{0,i}$, i.e. $\sigma^2_{0,i}$ does not necessarily have to equal $\sigma^2_{0,k}$. We note that this formulation also can be used to treat the case where V_0 is used to simply scale the observations, i.e., $V_0 = \text{diag}(v_1, \ldots, v_m)$ is known. In this case the formulation is simply a *vector OLS* (sometimes also called a weighted least squares (WLS)). The problem will consist of finding the minimizer

$$\theta_{\text{OLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} [\vec{Y}_j - \vec{f}(t_j, \vec{\theta})]^T V_0^{-1} [\vec{Y}_j - \vec{f}(t_j, \vec{\theta})], \tag{25}$$

where the procedure weights elements of the vector $\vec{Y}_j - \vec{f}(t_j, \vec{\theta})$ according to their variability. (Some authors refer to (25) as a generalized least squares (GLS) procedure, but we will make use of this terminology in a different formulation in subsequent discussions). Just as in the scalar OLS case, θ_{OLS} is a *random variable* (again because $\vec{\epsilon}_j = \vec{Y}_j - \vec{f}(t_j, \vec{\theta})$ is); hence if $\{\vec{y}_j\}_{j=1}^n$ is a realization of the *random process* $\{\vec{Y}_j\}_{j=1}^n$ then solving

$$\hat{\theta}_{\text{OLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})]^T V_0^{-1} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})] \tag{26}$$

provides an estimate (realization) $\hat{\theta} = \hat{\theta}_{\text{OLS}}$ for θ_{OLS}. By the definition of variance

$$V_0 = \text{diag}\, E\left(\frac{1}{n}\sum_{j=1}^{n}[\vec{Y}_j - \vec{f}(t_j,\vec{\theta}_0)][\vec{Y}_j - \vec{f}(t_j,\vec{\theta}_0)]^T\right),$$

so an unbiased estimate of V_0 for the realization $\{\vec{y}_j\}_{j=1}^n$ is

$$\hat{V} = \text{diag}\left(\frac{1}{n-p}\sum_{j=1}^{n}[\vec{y}_j - \vec{f}(t_j,\hat{\theta})][\vec{y}_j - \vec{f}(t_j,\hat{\theta})]^T\right). \tag{27}$$

However, the estimate $\hat{\theta}$ requires the (generally unknown) matrix V_0 and V_0 requires the unknown vector $\vec{\theta}_0$ so we will instead use the following expressions to calculate $\hat{\theta}$ and $\hat{V}$:

$$\vec{\theta}_0 \approx \hat{\theta} = \arg\min_{\vec{\theta}\in\Theta}\sum_{j=1}^{n}[\vec{y}_j - \vec{f}(t_j,\vec{\theta})]^T\hat{V}^{-1}[\vec{y}_j - \vec{f}(t_j,\vec{\theta})] \tag{28}$$

$$V_0 \approx \hat{V} = \text{diag}\left(\frac{1}{n-p}\sum_{j=1}^{n}[\vec{y}_j - \vec{f}(t_j,\hat{\theta})][\vec{y}_j - \vec{f}(t_j,\hat{\theta})]^T\right). \tag{29}$$

Note that the expressions for $\hat{\theta}$ and $\hat{V}$ constitute a coupled system of equations, which will require greater effort in implementing a numerical scheme.

Just as in the scalar case we can determine the asymptotic properties of the OLS estimator (25). As $n \longrightarrow \infty$, θ_{OLS} has the following asymptotic properties [2]:

$$\theta_{\text{OLS}} \sim \mathcal{N}(\vec{\theta}_0, \Sigma_0) \tag{30}$$

where

$$\Sigma_0 \quad = \quad \left(\sum_{j=1}^{n} D_j^T(\vec{\theta}_0)V_0^{-1}D_j(\vec{\theta}_0)\right)^{-1} \tag{31}$$

and the $m \times p$ matrix $D_j(\vec{\theta})$ is given by

$$\begin{pmatrix} \frac{\partial f_1(t_j,\vec{\theta})}{\partial\theta_1} & \frac{\partial f_1(t_j,\vec{\theta})}{\partial\theta_2} & \cdots & \frac{\partial f_1(t_j,\vec{\theta})}{\partial\theta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m(t_j,\vec{\theta})}{\partial\theta_1} & \frac{\partial f_m(t_j,\vec{\theta})}{\partial\theta_2} & \cdots & \frac{\partial f_m(t_j,\vec{\theta})}{\partial\theta_p} \end{pmatrix}.$$

Since the true value of the parameters $\vec{\theta}_0$ and V_0 are unknown their estimates $\hat{\theta}$ and $\hat{V}$ will be used to approximate the asymptotic properties of the least squares estimator θ_{OLS}:

$$\theta_{\text{OLS}} \sim \mathcal{N}_p(\vec{\theta}_0, \Sigma_0) \approx \mathcal{N}_p(\hat{\theta}, \hat{\Sigma}) \tag{32}$$

where

$$\Sigma_0 \approx \hat{\Sigma} = \left(\sum_{j=1}^{n} D_j^T(\hat{\theta}) \hat{V}^{-1} D_j(\hat{\theta}) \right)^{-1}. \tag{33}$$

The standard errors can then be calculated for the k^{th} element of $\hat{\theta}_{\text{OLS}}$ $(SE(\hat{\theta}_{\text{OLS},k}))$ by $SE(\hat{\theta}_{\text{OLS},k}) \approx \sqrt{\hat{\Sigma}_{kk}}$. Again, we point out the similarity between the MLE equations (12) and (13), and the OLS equations (28) and (29) for the vector statistical model (23).

1.4.2 Numerical Implementation of the OLS Procedure

In the scalar statistical model (14), the estimates $\hat{\theta}$ and $\hat{\sigma}$ can be solved for separately (this is also true of the vector OLS) in the case $V_0 = \sigma_0^2 I_m$, where I_m is the $m \times m$ identity) and thus the numerical implementation is straightforward - first determine $\hat{\theta}_{\text{OLS}}$ according to (20) and then calculate $\hat{\sigma}^2_{\text{OLS}}$ according to (21). The estimates $\hat{\theta}$ and $\hat{V}$ in the case of the vector statistical model (23), however, require more effort since they are coupled:

$$\hat{\theta} = \arg\min_{\vec{\theta} \in \Theta} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})]^T \hat{V}^{-1} [\vec{y}_j - \vec{f}(t_j, \vec{\theta})] \tag{34}$$

$$\hat{V} = \text{diag}\left(\frac{1}{n-p} \sum_{j=1}^{n} [\vec{y}_j - \vec{f}(t_j, \hat{\theta})][\vec{y}_j - \vec{f}(t_j, \hat{\theta})]^T \right). \tag{35}$$

To solve this coupled system the following iterative process will be followed:

1. Set $\hat{V}^{(0)} = \mathbf{I}$ and solve for the initial estimate $\hat{\theta}^{(0)}$ using (34). Set $k = 0$.
2. Use $\hat{\theta}^{(k)}$ to calculate $\hat{V}^{(k+1)}$ using (35).
3. Re-estimate $\vec{\theta}$ by solving (34) with $\hat{V} = \hat{V}^{(k+1)}$ to obtain $\hat{\theta}^{(k+1)}$.
4. Set $k = k + 1$ and return to 2. Terminate the process and set $\hat{\theta}_{\text{OLS}} = \hat{\theta}^{(k+1)}$ when two successive estimates for $\hat{\theta}$ are sufficiently close to one another.

1.4.3 Generalized Least Squares (GLS)

Although in Section 1.4.1 the error's distribution remained unspecified, we did however require that the error remain constant in variance in longitudinal data. That assumption may not be appropriate for data sets whose error is not constant in a longitudinal sense. A common relative error model that experimenters use in this instance for the scalar observation case [15] is

$$Y_j = f(t_j, \vec{\theta}_0)\,(1 + \epsilon_j) \tag{36}$$

where $E(Y_j) = f(t_j, \vec{\theta}_0)$ and $\text{var}(Y_j) = \sigma_0^2 f^2(t_j, \vec{\theta}_0)$. We will say that the variance generated in this fashion is non-constant variance. The method we will use to estimate $\vec{\theta}_0$ and σ_0^2 can be viewed as a particular form of the Generalized Least Squares (GLS) method.

To define the *random variable* θ_{GLS} the following equation must be solved for the estimator θ_{GLS}:

$$\sum_{j=1}^{n} w_j [Y_j - f(t_j, \theta_{\text{GLS}})] \nabla f(t_j, \theta_{\text{GLS}}) = 0, \tag{37}$$

where Y_j obeys (36) and $w_j = f^{-2}(t_j, \theta_{\text{GLS}})$. The quantity θ_{GLS} is a random variable, hence if $\{y_j\}_{j=1}^n$ is a *realization* of the random process Y_j then solving

$$\sum_{j=1}^{n} f^{-2}(t_j, \hat{\theta})[y_j - f(t_j, \hat{\theta})] \nabla f(t_j, \hat{\theta}) = 0, \tag{38}$$

for $\hat{\theta}$ gives an estimate $\hat{\theta}_{\text{GLS}}$ for θ_{GLS}.

The GLS estimator has the following asymptotic properties [2]:

$$\theta_{\text{GLS}} \sim \mathcal{N}_p(\vec{\theta}_0, \Sigma_0) \tag{39}$$

where

$$\Sigma_0 = \sigma_0^2 \left(F_{\vec{\theta}}^T(\vec{\theta}_0) W(\vec{\theta}_0) F_{\vec{\theta}}(\vec{\theta}_0) \right)^{-1}, \tag{40}$$

$$F_{\vec{\theta}}(\vec{\theta}) = \begin{pmatrix} \frac{\partial f(t_1,\vec{\theta})}{\partial \theta_1} & \frac{\partial f(t_1,\vec{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f(t_1,\vec{\theta})}{\partial \theta_p} \\ \vdots & & & \vdots \\ \frac{\partial f(t_n,\vec{\theta})}{\partial \theta_1} & \frac{\partial f(t_n,\vec{\theta})}{\partial \theta_2} & \cdots & \frac{\partial f(t_n,\vec{\theta})}{\partial \theta_p} \end{pmatrix} = \begin{pmatrix} \nabla f(t_1, \vec{\theta})^T \\ \vdots \\ \nabla f(t_n, \vec{\theta})^T \end{pmatrix}$$

and $W^{-1}(\vec{\theta}) = \text{diag}\left(f^2(t_1, \vec{\theta}), \ldots, f^2(t_n, \vec{\theta}) \right)$. Note that because $\vec{\theta}_0$ and σ_0^2 are unknown, the estimates $\hat{\theta} = \hat{\theta}_{\text{GLS}}$ and $\hat{\sigma}^2 = \hat{\sigma}^2_{\text{GLS}}$ will be used in (40) to calculate

$$\Sigma_0 \approx \hat{\Sigma} = \hat{\sigma}^2 \left(F_{\vec{\theta}}^T(\hat{\theta}) W(\hat{\theta}) F_{\vec{\theta}}(\hat{\theta}) \right)^{-1}$$

. where [15] we take the approximation

$$\sigma_0^2 \approx \hat{\sigma}^2_{\text{GLS}} = \frac{1}{n-p} \sum_{j=1}^{n} \frac{1}{f^2(t_j, \hat{\theta})} [y_j - f(t_j, \hat{\theta})]^2.$$

We can then approximate the standard errors of $\hat{\theta}_{\text{GLS}}$ by taking the square roots of the diagonal elements of $\hat{\Sigma}$. We will also mention that the solutions to (28) and (38) depend upon the numerical method used to find the minimum or root, and since Σ_0 depends upon the estimate for $\vec{\theta}_0$, the standard errors are therefore affected by the numerical method chosen.

1.4.4 GLS motivation

We note the similarity between (16) and (38). The GLS equation (38) can be motivated by examining the weighted least squares (WLS) estimator

$$\theta_{\text{WLS}} = \arg\min_{\vec{\theta}\in\Theta} \sum_{j=1}^{n} w_j [Y_j - f(t_j, \vec{\theta})]^2. \tag{41}$$

In many situations where the observation process is well understand, the weights $\{w_j\}$ may be known. The WLS estimate can be thought of minimizing the distance between the data and model while taking into account unequal quality of the observations [2]. If we differentiate the sum of squares in (41) with respect to $\vec{\theta}$, and *then* choose $w_j = f^{-2}(t_j, \vec{\theta})$, an estimate $\hat{\theta}_{\text{GLS}}$ is obtained by solving

$$\sum_{j=1}^{n} w_j [y_j - f(t_j, \vec{\theta})] \nabla f(t_j, \vec{\theta}) = 0$$

for $\vec{\theta}$. However, we note the GLS relationship (38) does *not* follow from minimizing the weighted least squares with weights chosen as $w_j = f^{-2}(t_j, \vec{\theta})$.

Another motivation for the GLS estimating equation (38) can be found in [12]. In the text the authors claim that if the data is distributed according to the gamma distribution, then the maximum-likelihood estimator for $\vec{\theta}$ is the solution to

$$\sum_{j=1}^{n} f^{-2}(t_j, \vec{\theta}) [Y_j - f(t_j, \vec{\theta})] \nabla f(t_j, \vec{\theta}) = 0,$$

which is equivalent to (38). The connection between the MLE and our GLS method is reassuring, but it also poses another interesting question: What if the variance of the data is assumed to not depend on the model output $f(t_j, \vec{\theta})$, but rather on some function $g(t_j, \vec{\theta})$ (i.e. $\text{var}(Y_j) = \sigma_0^2 g^2(t_j, \vec{\theta}) = \sigma_0^2 / w_j$)? Is there a corresponding maximum likelihood estimator of $\vec{\theta}$ whose form is equivalent to the appropriate GLS estimating equation ($w_j = g^{-2}(t_j, \vec{\theta})$)

$$\sum_{j=1}^{n} g^{-2}(t_j, \vec{\theta}) [Y_j - f(t_j, \vec{\theta})] \nabla f(t_j, \vec{\theta}) = 0 \quad ? \tag{42}$$

In their text, Carroll and Rupert [12] briefly describe how distributions belonging to the exponential family of distributions generate maximum-likelihood estimating equations equivalent to (42).

1.4.5 Numerical Implementation of the GLS Procedure

Recall that an estimate $\hat{\theta}_{\text{GLS}}$ can either be solved directly according to (38) or iteratively using the procedure outlined in Section 1.4.3. The iterative procedure as described in [15] is summarized below:

1. Estimate $\hat{\theta}_{\text{GLS}}$ by $\hat{\theta}^{(0)}$ using the OLS equation (15). Set $k = 0$.
2. Form the weights $\hat{w}_j = f^{-2}(t_j, \hat{\theta}^{(k)})$.
3. Re-estimate $\hat{\theta}$ by solving
$$\sum_{j=1}^{n} \hat{w}_j [y_j - f(t_j, \vec{\theta})] \nabla f(t_j, \vec{\theta}) = 0$$
to obtain $\hat{\theta}^{(k+1)}$.
4. Set $k = k + 1$ and return to 2. Terminate the process when two successive estimates for $\hat{\theta}_{\text{GLS}}$ are "close" to one another.

One finds in practice that the above procedure sometimes does not adequately estimate $\vec{\theta}_0$, so we instead outline a different numerical algorithm with which one often can achieve better results. Recall that the above iterative procedure was formulated by maximizing (over $\vec{\theta} \in \Theta$)
$$\sum_{j=1}^{n} f^{-2}(t_j, \tilde{\theta}) [y_j - f(t_j, \vec{\theta})]^2$$
and then updating the weights $w_j = f^{-2}(t_j, \tilde{\theta})$ after each iteration. Thus, an alternative iterative procedure involves completing the following steps:

1. Estimate $\hat{\theta}_{\text{GLS}}$ by $\hat{\theta}^{(0)}$ using the OLS equation (15). Set $k = 0$.
2. Form the weights $\hat{w}_j = f^{-2}(t_j, \hat{\theta}^{(k)})$.
3. Re-estimate $\hat{\theta}$ by solving
$$\hat{\theta}^{(k+1)} = \arg\min_{\theta \in \Theta} \sum_{j=1}^{n} \hat{w}_j \left(y_j - f\left(t_j, \vec{\theta}\right) \right)^2$$
to obtain the $k + 1$ estimate for $\hat{\theta}_{\text{GLS}}$.
4. Set $k = k + 1$ and return to 2. Terminate the process when two of the successive estimates for $\hat{\theta}_{\text{GLS}}$ are sufficiently close.

One would hope that after a sufficient number of iterations $\hat{w}_j$ would converge to $f^{-2}(t_j, \hat{\theta}_{\text{GLS}})$. Fortunately, under reasonable conditions, if the process enumerated above is continued a sufficient number of times [15], then $\hat{w}_j \to f^{-2}(t_j, \hat{\theta}_{\text{GLS}})$.

2 Computation of Σ, Standard Errors and Confidence Intervals

We return to the case of n scalar longitudinal observations and consider the OLS case of Section 1.4.1 (the extension of these ideas to vectors is completely straight-forward). These n scalar observations are represented by the statistical model

$$Y_j \equiv f(t_j, \vec{\theta}_0) + \epsilon_j, \;\; j = 1, 2, \ldots, n, \tag{43}$$

where $f(t_j, \vec{\theta}_0)$ is the model for the observations in terms of the state variables and $\vec{\theta}_0 \in \mathbb{R}^p$ is a set of theoretical "true" parameter values (assumed to exist in a standard statistical approach). We further assume that the errors ϵ_j, $j = 1, 2, \ldots, n$, are independent identically distributed (*i.i.d.*) random variables with mean $E[\epsilon_j] = 0$ and constant variance $var[\epsilon_j] = \sigma_0^2$, where σ_0^2 is unknown. The observations Y_j are then *i.i.d.* with mean $E[Y_j] = f(t_j, \vec{\theta}_0)$ and variance $var[Y_j] = \sigma_0^2$.

Recall that in the ordinary least squares (OLS) approach, we seek to use a realization $\{y_j\}$ of the observation process $\{Y_j\}$ along with the model to determine a vector $\hat{\theta}^n_{\text{OLS}}$ where

$$\hat{\theta}^n_{\text{OLS}} = \arg\min J_n(\vec{\theta}) = \sum_{j=1}^{n} [y_j - f(t_j, \vec{\theta})]^2. \tag{44}$$

Since Y_j is a random variable, the corresponding estimator $\theta^n = \theta^n_{\text{OLS}}$ (here we wish to emphasize the dependence on the sample size n) is also a random variable with a distribution called the *sampling distribution*. Knowledge of this sampling distribution provides uncertainty information (e.g., standard errors) for the numerical values of $\hat{\theta}^n$ obtained using a specific data set $\{y_j\}$.

Under reasonable assumptions on smoothness and regularity (the smoothness requirements for model solutions are readily verified using continuous dependence results for differential equations in most examples; the regularity requirements include, among others, conditions on *how the observations are taken* as sample size increases, i.e., as $n \to \infty$), the standard nonlinear regression approximation theory ([16, 19, 20], and Chapter 12 of [22]) for asymptotic (as $n \to \infty$) distributions can be invoked. As stated above, this theory yields that the sampling distribution for the estimator $\theta^n(Y)$, where $Y = \{Y_j\}_{j=1}^n$, is approximately a p-multivariate Gaussian with mean $E[\theta^n(Y)] \approx \vec{\theta}_0$ and covariance matrix $cov[\theta^n(Y)] \approx \Sigma_0 = \sigma_0^2[\chi^T(\vec{\theta}_0)\chi(\vec{\theta}_0)]^{-1}$. Here $\chi(\vec{\theta}) = F_{\vec{\theta}}(\vec{\theta})$ is the $n \times p$ sensitivity matrix with elements

$$\chi_{jk}(\vec{\theta}) = \frac{\partial f(t_j, \vec{\theta})}{\partial \theta_k} \qquad \text{and} \qquad F_{\vec{\theta}}(\vec{\theta}) \equiv (f_{1\vec{\theta}}(\vec{\theta}), \ldots, f_{n\vec{\theta}}(\vec{\theta}))^T.$$

That is, for n large, the sampling distribution approximately satisfies

$$\theta^n_{\text{OLS}}(Y) \sim \mathcal{N}_p(\vec{\theta}_0, \sigma_0^2[\chi^T(\vec{\theta}_0)\chi(\vec{\theta}_0)]^{-1}) := \mathcal{N}_p(\vec{\theta}_0, \Sigma_0). \tag{45}$$

There are typically several ways to compute the matrix $F_{\vec{\theta}}$. First, the elements of the matrix $\chi = (\chi_{jk})$ can always be estimated using the forward difference

$$\chi_{jk}(\vec{\theta}) = \frac{\partial f(t_j, \vec{\theta})}{\partial \theta_k} \approx \frac{f(t_j, \vec{\theta} + h_k) - f(t_j, \vec{\theta})}{|h_k|},$$

where h_k is a p-vector with a nonzero entry in only the k^{th} component. But, of course, the choice of h_k can be problematic in practice.

Alternatively, if the $f(t_j, \vec{\theta})$ correspond to longitudinal observations $\vec{y}(t_j) = \mathcal{C}\vec{x}(t_j; \vec{\theta})$ of solutions $\vec{x} \in \mathbb{R}^N$ to a parameterized N-vector differential equation system $\dot{\vec{x}} = \vec{g}(t, \vec{x}(t), \vec{\theta})$ as in (1), then one can use the $N \times p$ matrix **sensitivity equations** (see [4, 9] and the references therein)

$$\frac{d}{dt}\left(\frac{\partial \vec{x}}{\partial \vec{\theta}}\right) = \frac{\partial \vec{g}}{\partial \vec{x}}\frac{\partial \vec{x}}{\partial \vec{\theta}} + \frac{\partial \vec{g}}{\partial \vec{\theta}}$$

to obtain

$$\frac{\partial f(t_j, \vec{\theta})}{\partial \theta_k} = \mathcal{C}\frac{\partial \vec{x}(t_j, \vec{\theta})}{\partial \theta_k}.$$

Finally, in some cases the function $f(t_j, \vec{\theta})$ may be sufficiently simple so as to allow one to derive analytical expressions for the components of $F_{\vec{\theta}}$.

Since $\vec{\theta}_0$, σ_0 are unknown, we will use their estimates to make the approximation

$$\Sigma_0 = \sigma_0^2[\chi^T(\vec{\theta}_0)\chi(\vec{\theta}_0)]^{-1} \approx \hat{\Sigma}(\hat{\theta}^n_{\text{OLS}}) = \hat{\sigma}^2[\chi^T(\hat{\theta}^n_{\text{OLS}})\chi(\hat{\theta}^n_{\text{OLS}})]^{-1}. \tag{46}$$

where the approximation $\hat{\sigma}^2$ to σ_0^2, as discussed earlier, is given by

$$\sigma_0^2 \approx \hat{\sigma}^2 = \frac{1}{n-p}\sum_{j=1}^{n}[y_j - f(t_j, \hat{\theta}^n_{\text{OLS}})]^2. \tag{47}$$

Standard errors to be used in the confidence interval calculations are thus given by $SE_k(\hat{\theta}^n) = \sqrt{\Sigma_{kk}(\hat{\theta}^n)}$, $k = 1, 2, \ldots, p$ (see [13]).

In order to compute the confidence intervals (at the $100(1-\alpha)\%$ level) for the estimated parameters in our example, we define the confidence level parameters associated with the estimated parameters so that

$$P\{\hat{\theta}^n_k - t_{1-\alpha/2}SE_k(\hat{\theta}^n) < \theta^n_k < \hat{\theta}^n_k + t_{1-\alpha/2}SE_k(\hat{\theta}^n)\} = 1 - \alpha, \tag{48}$$

where $\alpha \in [0, 1]$ and $t_{1-\alpha/2} \in \mathbb{R}_+$. Given a small α value (e.g., $\alpha = .05$ for 95% confidence intervals), the critical value $t_{1-\alpha/2}$ is computed from the Student's t distribution t^{n-p} with $n-p$ degrees of freedom. The value of $t_{1-\alpha/2}$ is determined by $P\{T \geq t_{1-\alpha/2}\} = \alpha/2$ where $T \sim t^{n-p}$.

When one is taking longitudinal samples corresponding to solutions of a dynamical system, the $n \times p$ sensitivity matrix depends explicitly on where in time the observations are taken when $f(t_j, \vec{\theta}) = \mathcal{C}x(t_j, \vec{\theta})$ as mentioned above. That is, the sensitivity matrix

$$\chi(\vec{\theta}) = F_{\vec{\theta}}(\vec{\theta}) = \left(\frac{\partial f(t_j, \vec{\theta})}{\partial \theta_k} \right)$$

depends on the number n and the nature (e,g., how taken) of the sampling times $\{t_j\}$. Moreover, it is the matrix $[\chi^T \chi]^{-1}$ in (46) and the parameter $\hat{\sigma}^2$ in (47) that ultimately determine the SE and CI. At first investigation of (47), it appears that an increased number n of samples will drive $\hat{\sigma}^2$ (and hence the SE) to zero as long as this is done in a way to maintain a bound on the residual sum of squares in (47). However, we observe that the *condition number* of the matrix $\chi^T \chi$ is also very important in these considerations and increasing the sampling could potentially adversely affect the inversion of $\chi^T \chi$. In this regard, we note that among the important hypotheses in the asymptotic statistical theory (see p. 571 of [22]) is

$$\frac{1}{n} \chi^T(\vec{\theta}) \chi(\vec{\theta}) \rightarrow \mathcal{X}(\vec{\theta}) \quad \text{as n} \rightarrow \infty$$

for some **nonsingular** matrix $\mathcal{X}(\vec{\theta}_0)$. It is this condition that is rather easily violated in practice when one is dealing with data from differential equation systems, especially near an equilibrium or steady state (see the examples of [4]).

All of the above theory readily generalizes to vector systems with partial, non-scalar observations. Suppose now we have the vector system (1) with partial vector observations given by (3), that is, we have m coordinate observations where $m \leq N$. In this case, we have

$$\frac{d\vec{x}}{dt}(t) = \vec{g}(t, \vec{x}(t), \vec{\theta}) \tag{49}$$

and

$$\vec{y}_j = \vec{f}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j = \mathcal{C}\vec{x}(t_j, \vec{\theta}_0) + \vec{\epsilon}_j, \tag{50}$$

where $\mathcal{C}$ is an $m \times N$ matrix and $\vec{f} \in R^m, \vec{x} \in R^N$. As already explained in Section 1.4.1, if we assume that different observation coordinates f_i may have different variances σ_i^2 associated with different coordinates of the errors ϵ_j, then we have

$$\vec{\epsilon}_j \sim \mathcal{N}_m(\vec{0}, V_0)$$

where $V_0 = diag(\sigma_{0,1}^2, ..., \sigma_{0,m}^2)$ and we may follow similar asymptotic theory to calculate approximate covariances, standard errors and confidence intervals for parameter estimates.

Since the computations for standard errors and confidence intervals (and also the *model comparison tests* outlined in the next section) depend on *an asymptotic limit distribution theory*, one should interpret the findings as sometimes crude indicators of uncertainty inherent in the inverse problem findings. Nonetheless, it is useful to consider the formal mathematical requirements underpinning these techniques. Among the more readily checked hypotheses are those of the statistical model requiring that the errors ϵ_j, $j = 1, 2, \dots, n$, are independent identically distributed ($i.i.d.$) random variables with mean $E[\epsilon_j] = 0$ and constant variance $var[\epsilon_j] = \sigma_0^2$. After carrying out the estimation procedures, one can readily plot the *residuals vs. time* and the *residuals vs. the resulting estimated model (output or observation f) values.* A random pattern for the first is strong support for validity of the independence assumption while a non increasing, random pattern for the latter suggests the assumption of constant variance may be reasonable for the data (measurements) used in the inverse problem calculations. The underlying assumption that the sampling size n must be large (recall the theory is asymptotic in that it holds as $n \to \infty$) is not so readily "verified" and is often ignored (albeit at the user's peril in regard to the quality of the uncertainty findings). Indeed the asymptotic theories are often used in a very heuristic underlying manner to give a loose feeling for the uncertainty involved in the estimates and the level of parametrization used in approximating the underlying mathematical model. It is often the case that the asymptotic results provide remarkably good approximations to the true sampling distributions for finite n. However, in practice there is no way to ascertain whether this holds for a specific example of interest.

3 Model Comparison Techniques

3.1 Motivation

The presentation in this section is motivated by the following questions/needs that arise in modeling studies:

- Frequent QUESTION in modeling studies [7, 8]: can a mathematical model be improved by more detail and/or further refinement?
- EXAMPLE: More detail in a given mechanism (constant rate vs. time or spatially dependent rate–see [1] for questions related to time dependent mortality rates during sub-lethal damage in insect populations exposed to various levels of pesticides).
- EXAMPLE: Does an additional mechanism in the model produce a better fit to data–see [5, 6, 7] for diffusion alone or diffusion plus convection in cat brain transport in grey vs. white matter questions.

Important Remarks

- In model comparison results outlined below, there are really *two models* being compared: the *math model* and the *statistical model*.
- If one embeds the math model in the wrong statistical model (for example, assumes constant variance when it really isn't true), then the math model comparison results will be invalid (e.g., worthless).
- The key to all this is that you must have the math model you want to simplify or improve (e.g., test $\mathcal{V} = 0$ in the example below) embedded in the correct statistical model, so that the comparison really is only with regard to the math model.

EXAMPLE We illustrate the formulation of hypothesis testing by considering a mathematical model for a diffusion-convection process. This model was proposed for use with experiments designed to study substance (labeled sucrose) transport in cat brains. The cat's brain contains grey and white matter[7]. In general, the transport of substance in cat's brains can be described by a PDE describing change in time and space. This model, which is widely discussed in the applied mathematics and engineering literature, has the form

$$\frac{\partial u}{\partial t} + \mathcal{V}\frac{\partial u}{\partial x} = \mathcal{D}\frac{\partial^2 u}{\partial x^2}. \tag{51}$$

Here, the parameter $\vec{\theta} = (\mathcal{D}, \mathcal{V})$, which belongs to some admissible parameter set Θ, denotes the diffusion coefficient $\mathcal{D}$ and the bulk velocity $\mathcal{V}$ of the fluid, respectively. Our problem: test whether the parameter $\mathcal{V}$ plays a significant role in the mathematical model. That is, if the model (51) represents a diffusion-convection process, we seek to determine whether diffusion alone or diffusion plus convection best describes transport phenomena represented in cat brain data sets $\{y_{ij}\}$ for $\{u(t_i, x_j)\}$, the concentration of labeled sucrose at times $\{t_i\}$ and location $\{x_j\}$. We then may take $H_0 : \mathcal{V} = 0$ and the alternative $H_A : \mathcal{V} \neq 0$. Consequently, the restricted parameter set $\Theta_H \subset \Theta$ defined by

$$\Theta_H = \{\vec{\theta} \in \Theta : \mathcal{V} = 0\}$$

will be important. To carry out these determinations, we will need some model comparison tests from statistics.

3.2 ANOVA Type Statistical Tests

In general, assume we have an inverse problem $f(t, \vec{\theta})$ and are given n observations. We define

$$J_n(\vec{\theta}) = J_n(\vec{Y}, \vec{\theta}) = \frac{1}{n}\sum_{j=1}^{n}[Y_j - f(t_j, \vec{\theta})]^2$$

where our statistical model again has the form

$$Y_j = f(t_j, \vec{\theta}_0) + \epsilon_j, \quad j = 1, \ldots, n$$

. Here, $\vec{\theta}_0$ is the "true" value of $\vec{\theta}$ which we assume to exist. We again use Θ to represent the set of all the admissible parameters $\vec{\theta}$.

We take the standard statistical assumptions;

- A1) $\{\epsilon_j\}_{j=1}^{\infty}$ are identical independent distributed with $E(\epsilon_j) = 0$ and $\text{var}(\epsilon_j) = \sigma^2$.

Among other important hypotheses are

- A2) Θ is a compact subset of Eucledian space of R^p and $f(t, \vec{\theta})$ is continuous on $[0, T] \times \Theta$.
- A3) Observations are at $\{t_j\}_{j=1}^{n}$ in $[0, T]$. For some finite measure μ on $[0, T]$,

$$\frac{1}{n}\sum_{j=1}^{n} h(t_j) \longrightarrow \int_0^T h(t)d\mu(t)$$

  as $n \to \infty$, for continuous functions h.

- A4) $J_0(\vec{\theta}) = \int_0^T (f(t, \vec{\theta}_0) - f(t, \vec{\theta}))^2 d\mu(t) = \sigma^2$ has a unique minimizer in Θ at $\vec{\theta}_0$.

Let $\theta^n = \theta^n_{OLS}(\vec{Y})$ be the OLS estimator for J_n with corresponding estimate

$$\hat{\theta}^n = \theta^n_{OLS}(\{y_j\})$$

for a realization $\vec{y} = \{y_j\}$. That is,

$$\theta^n(\vec{Y}) = \arg\min_{\vec{\theta}\in\Theta} J_n(\vec{Y}, \vec{\theta})$$

and

$$\hat{\theta}^n = \arg\min_{\vec{\theta}\in\Theta} J_n(\vec{y}, \vec{\theta}).$$

One can then establish a series of useful results (see [6] for detailed proofs).

- **Result 1:**
  Under A1) to A4), $\theta^n \longrightarrow \vec{\theta}_0$ as $n \to \infty$ with probability 1.
- **Remarks:** In most calculations, we actually use an approximation f^N to f, often a numerical solution to the ODE or PDE for modeling our dynamical system. Here we tacitly assume f^N will converge to f as the approximation improves. There are also questions related to approximations of the set Θ when it is infinite dimensional (e.g., in the case of function space parameters such as time dependent parameters) by finite dimensional discretizations Θ^M. For extensive discussions related to these questions, see [8] as well as [6] where related assumptions A5), A6) on convergences $f^N \to f$ and $\Theta^M \to \Theta$ are given. We will ignore these issues here, keeping in mind that these approximations will also be of importance in the methodology discussed below in most practical uses.

We will need further assumptions to proceed (these will be denoted by A7)–A11) to facilitate reference to [6]). These include:

- A7) Θ is finite dimensional in R^p and $\vec{\theta}_0 \in \Theta$.
- A8) $f : \Theta \to C[0, T]$ is C^2 function.
- A10) $\mathcal{J} = \frac{\partial^2 J_0}{\partial \vec{\theta}^2}(\vec{\theta}_0)$ is positive definite.
- A11) $\Theta_H = \{\vec{\theta} \in \Theta | H\vec{\theta} = c\}$ where H is an $r \times p$ matrix of full rank, and c is a known constant.

In many instances, including the motivating example given above, one is interested in using data to questioning whether the the "true" parameter $\vec{\theta}_0$ can be found in a subset $\Theta_H \subset \Theta$ which we assume for discussions here is defined by the constraints of assumption A11).

Thus, we want to test the *null hypothesis* H_0: $\vec{\theta}_0 \in \Theta_H$.

Define then

$$\theta^n_H(\vec{Y}) = \arg\min_{\vec{\theta}\in\Theta_H} J_n(\vec{Y}, \vec{\theta})$$

and

$$\hat{\theta}^n_H = \arg\min_{\vec{\theta}\in\Theta_H} J_n(\vec{y}, \vec{\theta}).$$

and observe that $J_n(\vec{Y}, \hat{\theta}^n_H) \geq J_n(\vec{Y}, \hat{\theta}^n)$. We define the related non-negative test statistics and their realizations, respectively, by

$$T_n(\vec{Y}) = n(J_n(\vec{Y}, \theta^n_H) - J_n(\vec{Y}, \theta^n))$$

and

$$\hat{T}_n = T_n(\vec{y}) = n(J_n(\vec{y}, \hat{\theta}^n_H) - J_n(\vec{y}, \hat{\theta}^n)).$$

One can establish asymptotic convergence results for the test statistics $T_n(\vec{Y})$, as given in detail in [6]. These results can, in turn, be used to establish a fundamental result about much more useful statistics for model comparison. We define these statistics by

$$U_n(\vec{Y}) = \frac{T_n(\vec{Y})}{J_n(\vec{Y}, \theta_n)}, \tag{52}$$

with corresponding realizations

$$\hat{U}_n = U_n(\vec{y})$$

.

We then have the asymptotic result that is the basis of our ANOVA–type tests

Major Result [6] : Under the assumptions A1)–A11) above and the assumption that H_0 is true,

$$U_n \xrightarrow{\mathcal{D}} \Upsilon(r)$$

as $n \to \infty$ where $\Upsilon \sim \chi^2(r)$, a χ^2 distribution with r degrees of freedom.

An example of this the χ^2 density is depicted in Figure 1 where the density for $\chi^2(4)$ (χ^2 with $r = 4$ degrees of freedom) is graphed.

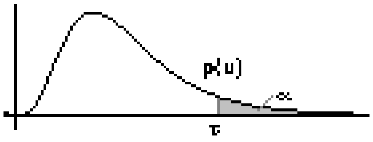

Figure 1: Example of $U \sim \chi^2(4)$ density

In this figure two parameters (τ, α) of interest are shown. For a given value τ, the value α is simply the probability that the random variable U will take on a value greater than α. That is, $Prob\{U > \tau\} = \alpha$ where in hypothesis testing, α is the *significance level* and τ is the *threshold*.

We wish to use this distribution to test the null hypothesis, H_0, for $U_n \sim \chi^2(r)$. If the test statistic, $\hat{U}_n > \tau$, then we *reject* H_0 as false with confidence level $(1 - \alpha)100\%$. Otherwise, we *accept* H_0 as true. For cat brain problem, we use a $\chi^2(1)$ table, which can be found in any elementary statistics text or online.

Table 1: $\chi^2(1)$

α	τ	confidence
.25	1.32	75%
.1	2.71	90%
.05	3.84	95%
.01	6.63	99%
.001	10.83	99.9%

p-value:

The minimum value of α at which H_0 can be rejected, α^*, is called the *p-value.* Thus, the smaller the p-value, the greater the significance of the additional parameters/mechanisms, i.e., the more likely the term should be in the model.

Implementation: Once we compute $\hat{U}_n = \bar{\tau}$, then $p = \alpha^*$ is the value that corresponds to $\bar{\tau}$ on χ^2 graph and so, we reject the null hypothesis at any confidence level, c, such that $c < 1 - \alpha^*$. For example, if for a computed $\bar{\tau}$ we find $p = \alpha^* = .0182$, then reject H_0 at confidence level $(1 - \alpha^*)100\% = 98.18\%$ or lower. For more information, see ANOVA in any good statistics book.

3.3 Alternative statement

To test the null hypothesis H_0, we choose a significance level α and use χ^2 tables to obtain the corresponding threshold $\tau = \tau(\alpha)$ so that $P(\chi^2(s) > \tau) = \alpha$. We next compute $\hat{U}_n = \bar{\tau}$ and compare it to τ. If $\hat{U}_n > \tau$, then we reject H_0 as false; otherwise, we accept the null hypothesis H_0.

3.4 Revisiting the cat-brain problem

There were 3 sets of experimental data examined, under the null-hypothesis $H_0 : \mathcal{V} = 0$.

For the Data Set 1, we found after carrying out the inverse problems over Θ and Θ_H, respectively,

$$J_n(\hat{\theta}^n) = 106.15$$
$$J_n(\hat{\theta}^n_H) = 180.17,$$

which gives us that $\hat{U}_n = 5.579$ (noting that $n = 8 \neq \infty$), for which $p = \alpha^* = .0182$. Thus, we reject H_0 in this case at *any* confidence level less than 98.18%. Thus, we should reject that $\mathcal{V} = 0$, which suggests convection is important in describing this data set.

For Data Set 2, we found

$$J_n(\hat{\theta}^n) = 14.68$$
$$J_n(\hat{\theta}^n_H) = 15.35,$$

thus, in this case, we have $\hat{U}_n = .365$, which implies we accept H_0 with high degrees of confidence (p-value very high). This suggests $\mathcal{V} = 0$, which is completely opposite to the findings for Data Set 1.

For the final set (Data Set 3) we found

$$J_n(\hat{\theta}^n) = 7.8$$
$$J_n(\hat{\theta}^n_H) = 146.71,$$

which yields in this case, $\hat{U}_n = 15.28$. This, as in the case of the first data set, suggests (with $p < .001$) that $\mathcal{V} \neq 0$ is important in modeling the data.

3.5 Conclusions

The difference in conclusions between the first and last sets and that of the second set is interesting.

However, when discussed with the doctors who provided the data, it was discovered that the first and last set were taken from the white matter of the brain, while the other was taken from the grey matter. This later finding was conducive to observed microscopic tests on the various matter (micro channels in white matter that promote convective "flow"). Thus, it can be concluded with a reasonably high degree of confidence, that white matter has convective properties, while grey matter does not.

References

[1] H. T. Banks, J.E. Banks, L.K. Dick and J.D. Stark, Estimation of dynamic rate parameters in insect populations undergoing sublethal exposure to pesticides, CRSC-TR05-22, May, 2005; *Bulletin of Mathematical Biology*, to appear.

[2] H. T. Banks and M. Davidian, SAMSI MA/ST 810 Notes.

[3] H. T. Banks, S. Dediu and H.K. Nguyen, Sensitivity of dynamical systems to parameters in a convex subset of a topological vector space, Center for Research in Scientific Computation Report, CRSC-TR06-25, September, 2006, North Carolina State University; *Math. Biosci. and Engineering,* **4** (2007), 403–430.

[4] H.T. Banks, S.L. Ernstberger and S.L.Grove, Standard errors and confidence intervals in inverse problems: sensitivity and associated pitfalls, CRSC-TR06-10, March, 2006; *J. Inv. Ill-posed Problems* **15** (2006), 1–18.

[5] H. T. Banks and B. G. Fitzpatrick, Inverse problems for distributed systems: statistical tests and ANOVA, LCDS/CCS Rep. 88-16, July, 1988, Brown University; *Proc. International Symposium on Math. Approaches to Envir. and Ecol. Problems*, Springer Lecture Note in Biomath., **81** (1989), 262–273.

[6] H. T. Banks and B. G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, CAMS Tech. Rep. 89-4, September, 1989, University of Southern California; *J. Math. Biol.*, **28** (1990), 501–527.

[7] H. T. Banks and P. Kareiva, Parameter estimation techniques for transport equations with application to population dispersal and tissue bulk flow models (with), LCDS Report #82-13, July 1982, Brown University; *J. Math. Biol.*, **17** (1983), 253–273.

[8] H. T. Banks and K. Kunsich, *Estimation Techniques for Distributed Parameter Systems*, Birkhauser, Boston, 1989.

[9] H. T. Banks and H. K. Nguyen, Sensitivity of dynamical system to Banach space parameters, CRSC Tech Rep. CRSC-TR05-13, N.C. State University, February, 2005; *J. Math. Analysis and Applications*, **323** (2006), 146–161.

[10] P. Bai, H. T. Banks, S. Dediu, A. Y. Govan, M. Last, A. Loyd, H. K. Nguyen, M. S. Olufsen, G. Rempala, and B. D. Slenning, Stochastic and deterministic models for agricultural production networks, CRSC-TR07-06, February, 2007; *Math. Biosci. and Engineering,* **4** (2007), 373–402.

[11] J. J. Batzel, F. Kappel, D. Schneditz and H.T. Tran, *Cardiovascular and Respiratory Systems: Modeling, Analysis and Control*, SIAM Frontiers in Applied Math, SIAM, Philadelphia, 2006.

[12] R.J. Carroll and D. Ruppert. Transformation and Weighting in Regression. Chapman and Hall, New York, 1988.

[13] G. Casella and R. L. Berger, *Statistical Inference,* Duxbury, California, 2002.

[14] J. B. Cruz, ed., *System Sensitivity Analysis*, Dowden, Hutchinson & Ross, Inc., Stroudsberg, PA, 1973.

[15] M. Davidian, Class Notes ST762, NCSU.

[16] M. Davidian and D. Giltinan, *Nonlinear Models for Repeated Measurement Data*, Chapman & Hall, London, 1998.

[17] M. Eslami, *Theory of Sensitivity in Dynamic Systems: An Introduction*, Springer-Verlag, Berlin, 1994.

[18] P.M. Frank, *Introduction to System Sensitivity Theory*, Academic Press, Inc., New York, NY, 1978.

[19] A. R. Gallant, *Nonlinear Statistical Models*, John Wiley & Sons, Inc., New York, 1987.

[20] R. I. Jennrich, Asymptotic properties of non-linear least squares estimators., *Ann. Math. Statist.*, **40** (1969), 633–643.

[21] A. Saltelli, K. Chan and E.M. Scott, eds., *Sensitivity Analysis*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York, NY, 2000.

[22] G. A. F. Seber and C. J. Wild, *Nonlinear Regression,* John Wiley & Sons, Inc., New York, 1989.

[23] K. Thomaseth and C. Cobelli, Generalized sensitivity functions in physiological system identification., *Ann Biomed Eng.*, **27(5)** (1999), 607–616.

[24] D. D. Wackerly, W. Mendenhall III, and R. L. Scheaffer, *Mathematical Statistics with Applications*, Duxbury Thompson Learning, USA, 2002.