Reliability and Uncertainty in Diffusion MRI Modelling

Christopher Ned Charles

B.S. (Purdue University), M.S. (University of Sydney)

A thesis submitted in fulfilment

of the requirements for the degree of

Doctor of Philosophy

Faculty of Health Sciences

University of Sydney

2016

Title and Publication Info

This thesis was written by a non-native Australian attempting to follow the Australian rules of spelling and grammar. Post-publication, the software to replicate the experiments in this thesis will be available at http://diffusioned.github.io. The software is also available upon request by contacting the author via email at ccha4217@uni.sydney.edu.au.

Table of Contents

Acronyms/Symbols/Notation

ADC – Apparent Diffusion Coefficient

DWI – Diffusion Weighted Imaging

IVIM – Intra-voxel Incoherent Motion

K-L – Kullback-Leibler (Distance)

LLS – Linear Least Squares

LOOCV – Leave-one-out Cross-Validation

MRI – Magnetic Resonance Imaging

MSD – Mean Squared Displacement

MSE – Mean Squared Error

NLLS – Nonlinear Least Squares

PDF – Probability Density Function

ROI – Region (or Regions) of Interest

RSS – Residual Sum of Squares

SER – Standard Error of Regression

SNR – Signal-to-Noise Ratio

SD – Standard Deviation

Definition of Terms

These definitions come from either the Oxford Dictionary of English (ODE) [1] or the Oxford Dictionary of Statistics (ODS) [2] as specifically noted below.

Reliability – The quality of being trustworthy or of performing consistently well. The degree to which the result of a measurement, calculation, or specification can be depended on to be accurate. (ODE)

Uncertainty – The state of being uncertain - not able to be relied on; not known or definite. (ODE)

Estimator – A statistic used to estimate a parameter. The realized value of an estimator for a particular sample of data is called the estimate (or point estimate). (ODS)

Expected Value – The expected value of a random variable X, is denoted by E(X) and may be interpreted as the long-term average value of X. (ODS)

Bias – If the expected value of the statistic is equal to the parameter then it is described as being an unbiased estimator and the realized value is referred to as an unbiased estimate. If T is an estimator of the parameter θ and the expected value of T is $\theta+b$, where $b \neq 0$, then b is called the bias and the estimator is a biased estimator. (ODS)

Variance – A measure of the variability in the values of a random variable. It is defined as the expected value of the squared difference between the random variable and its expected value. (ODS)

Consistent – An estimator is said to be a consistent estimator if, as the number of samples increases indefinitely, the estimator converges in probability to the true parameter value (ODS).

Abstract

Current Diffusion MRI studies often utilise more complex models beyond the single exponential decay model used in clinical standards. As this thesis shows, however, two of these models, biexponential and kurtosis, experience mathematical, ill-conditioning issues that can arise when used with regression algorithms, causing extreme bias and/or variance in the parameter estimates. Using simulated noisy data measurements from known truth, the magnitude of the bias and variance was shown to vary based on signal parameters as well as SNR, and increasing the SNR did not reduce this uncertainty for all data. Parameter estimate reliability could not be assessed from a single regression fit in all cases unless bootstrap resampling was performed, in which case measurements with high parameter estimate uncertainty were successfully identified. Prior to data analysis, current studies may use information criteria or cross-validation model selection methods to establish the best model to assess a specific tissue condition. While the best selection method to use is currently unclear in the literature, when testing simulated data in this thesis, no model selection method performed more reliably than the others and these methods were merely biased toward either simpler or more complex models. When a specific model was used to generate simulated noisy data, no model selection method selected this true model for all signals, and the ability of these methods to select the true model also varied depending on the true signal parameters. The results from these simulated data analyses were applied to ex vivo data from excised prostate tissue, and both information criteria measures and bootstrap sample distributions were able to identify image voxels whose parameter estimates had likely reliability issues. Removing these voxels from analysis improved sample variance of the parameter estimates.

Acknowledgements

I would like to thank my Mom and Dad for providing me with support and encouragement during my PhD and for giving me a warm home back in Alaska to write. I would also like to thank the rest of my family for also supporting me with fun things to do, great ideas, and the occasional place to stay. I can't thank enough all of my friends in Sydney, especially on the Northern Beaches, where I'll always have a home, mentally and/or physically. Thanks for helping a struggling student with the occasional beer or dinner, and thanks for believing in me and my work. Special thanks goes to my friends Marcus Wraight and Catrina McCallum for letting me stay at their wonderful home while in various periods of transition. I'd also like to thank my friends Dr. Dan Swan and Sarah Swan for their support and culinary delights – I started writing this thesis while dog sitting in their home in the Byron Hinterland and also worked on it during a few sessions at their 100 Mile Table café. Thanks to my long-time friend Scott "Scooter" Robinson for letting me stay at his house for several months, where I wrote and edited most of this thesis in his Colorado "writing retreat". Thanks to the rest of my fellow PhD students in the Medical Radiation Sciences group – hope we all successfully get our degrees and have continued success in the future. Finally, I'd like to thank my supervisors Dr. Roger Bourne and Prof. Mark McEntee for reviewing portions of this work.

Chapter 1
Introduction and Literature Review

Since its invention over 40 years ago [3], the capabilities of MRI as a tool to non-invasively obtain anatomical details and identify anomalies and lesions without any ionizing radiation delivered to the patient have improved dramatically. MRI works by measuring the magnetic properties of atomic nuclei present in human tissue in the presence of a large magnetic field, and is often used to measure either the rate for nuclear spins to return to equilibrium with their surroundings after a transmitted electromagnetic pulse (T1-weighting) or the rate that a group of spins lose their phase coherence (T2-weighting) [4]. The parameters for an MRI scan can be adjusted to weight a given image to detect differences in T1 or T2 in order to highlight or suppress different tissues for medical diagnosis. Adding extra field gradients to a T2-weighted spin echo pulse sequence, for example, gives an MRI scanner the ability to measure the mean displacement of an ensemble of water molecules moving in tissues – a process known as Diffusion MRI, or Diffusion Weighted Imaging (DWI) [5]. In DWI measurements, analysing the distance that molecules move in a given volume of interest (voxel) during a given time period can provide insight into the underlying tissue properties.

1.1 Molecular Diffusion and Brownian Motion

When a molecule is suspended in a liquid or gas, it moves in a random fashion due to thermal energy and collisions with neighbouring molecules, a stochastic process known as Brownian motion. Although the exact path of each molecule in a given volume is different, when combined over an entire ensemble, the random motions can be quantified via a displacement probability distribution that is measurable, such as a Gaussian distribution in the case of self-diffusion in the absence of a concentration gradient.

1.1.1 Gaussian Self-Diffusion

Diffusion refers to the process of how molecules or particles disperse through a substance with a typical example being dispersion of coloured dye added to a glass of water. This can be characterized using a simple relation known as Fick's First Law of Diffusion, which is given by [6]

$$J = -D\nabla\varphi\,, \tag{1}$$

where J is the diffusion flux, $\nabla\varphi$ the concentration gradient of the molecular ensemble in space, and D the diffusivity or diffusion coefficient. The diffusion coefficient refers to how the solute and solvent molecules move from a higher to lower concentration in the case of two substances. In the case of self-diffusion, movement of particles of a single substance, a self-diffusion coefficient describes the molecular motion. The value of this diffusion coefficient can be calculated by the Einstein-Sutherland relation [7]

$$D = \frac{kT}{f}, \tag{2}$$

where k is the Boltzmann constant, f the friction coefficient, and T the temperature. The friction coefficient is dependent on the molecular substance itself, meaning a self-diffusion coefficient of a given substance will change only on its relation to temperature. For water at a temperature of 20 °C, the value is $2.02\text{x}10^{-9}$ m^2/s, whereas at 37 °C (as in the human body) it has a diffusivity of $3.03\text{x}10^{-9}$ m^2/s [8]. As Brownian motion is a stochastic process, the exact path that each molecule travels is unknown. However, the *average* displacement distance of a molecular ensemble will be longer if the molecules have a higher diffusivity. If a molecular ensemble is made up of freely-diffusing water, the random displacements in a single direction will assume a Gaussian Probability Density Function (PDF)

$$P(r,t) = \frac{1}{\sqrt{4\pi Dt}} \exp\left(-\frac{r^2}{4Dt}\right), \tag{3}$$

where r is the displacement in a given direction, D the diffusivity, and t time. This Gaussian PDF equation describes a displacement centred about r = 0 with variance equal to Dt. If a molecular ensemble with a fixed diffusivity of $3.0\text{x}10^{-9}$ m^2/s is evaluated for two diffusion times, 40 and 80 ms, the variance of the 80 ms PDF is equal to $2.4\text{x}10^{-10}$ m and the variance of the 40 ms PDF equal to $1.2\text{x}10^{-10}$ m. These can be better represented as displacement standard deviations, equal to 15.5 μm and 11.0 μm for the 80 and 40 ms distributions, respectively, illustrating that the molecules will displace further on average when given more time. The average displacement will also increase if the diffusivity is increased, and in both cases, the Gaussian distribution becomes wider. The displacement of the ensemble in n-dimensional space can also be defined by the relation

$$\langle r^2 \rangle = 2nDt\,. \tag{4}$$

where $\langle r^2 \rangle$ is known as the Mean Squared Displacement (MSD) with the angle brackets specifying an ensemble average. In the case of free diffusion, MSD has a linear relation to time.

1.1.2 Hindered Diffusion

If obstacles are present that hinder the paths of molecules as they diffuse, the MSD will decrease, modifying the relation in Equation 4. If Equation 4 is inverted and solved for a new, *effective* value of D, the result is

$$D_{eff}(t) = \frac{\langle \big(r(t) - r(0)\big)^2 \rangle}{6t}, \tag{5}$$

where r(t) – r(0) is the displacement of one molecule from time zero to time t. If multiple molecules collide with obstacles as they diffuse through space, their total displacements over time will be shorter, and the effective diffusion coefficient will be less than the free diffusivity. In the case of hindered diffusion, the displacements will still assume a Gaussian distribution as per Equation 3,

but the distribution will be narrower than a free diffusion distribution to reflect the decrease in MSD.

1.1.3 Anomalous Diffusion

When a molecular ensemble diffuses through porous media, percolation clusters, or fractal-like structures, the relationship of MSD to time becomes nonlinear and assumes a power law [9]

$$\langle r^2 \rangle \propto Dt^{\alpha}. \tag{6}$$

If the value of α is less than one, the process is called *subdiffusion* and if α is greater than one, the process is *superdiffusion*, and when α is exactly one, the diffusion assumes the linear relationship in Equation 4. While hindered diffusion is due to obstructions that randomly obstruct diffusing molecules, anomalous diffusion involves obstructions with periodic patterns or clusters at similar scales, causing the molecular displacements to be highly correlated with each other [10]. Anomalous diffusion has been observed in excised human tissue using DWI measurements [11].

1.1.4 Restricted Diffusion

When molecules are confined within a space such that some of the molecules are reflected during the diffusion time measurement, the diffusion is called restricted. For molecules confined within a closed pore, the MSD will increase linearly at short diffusion times, but as a larger proportion of molecules bounce off the walls of the pore, the mean displacement no longer scales linearly with time, and as the diffusion time increases, the mean displacement eventually reaches a maximum value. In this scenario, the relation between MSD and time varies depending on the length of time, and if the measurement time is sufficiently long, the MSD no longer changes with time and instead its value is dependent on the pore geometry. For some basic restricted geometries, such as a sphere or two opposing planes, closed-form expressions have been determined for the diffusion coefficient with respect to the physical geometry [12]. In most complex systems, such as biological tissue, the geometries are complex and the molecular diffusion properties can exhibit various combinations of the previously mentioned diffusion types. These molecular displacement distributions are difficult to define with a closed-form expression and are inadequately described by a Gaussian PDF or any anomalous relationship.

For some cellular tissues, there may be structural order present, but cellular walls are also permeable, adding an additional dimension to the relationship between molecular displacement and geometry. At short measurement times, DWI measurements are sensitive to molecular displacements on the order of micrometres, which are close to the size of a single cell in the human body, so molecular collisions with internal cellular structure must also be accounted for. Additionally, in some cellular structures such as the cylindrical shaped white matter fibres in the brain, diffusion restrictions in one spatial dimension can also be different than the other two such that the diffusion is anisotropic [13]. A detailed discussion of the various restricted diffusion relationships won't be presented further in this thesis, however, a comprehensive review of the MRI literature on restricted diffusion can be found in [14], discussing various methodology,

experiments, and models with well-defined relations between geometries, measurement times, and diffusion coefficients, with additional updated review found in [7].

1.2 Modelling DWI Data

DWI measurements can be acquired in an MRI scanner using a variety of pulse sequences, and the most common model in present use is the pulsed gradient spin-echo (PGSE) sequence developed by Stejskal and Tanner [15]. The PGSE relates the attenuation in the measured signal to the diffusion coefficient through the equation

$$\frac{S}{S_0} = \exp\left(-\gamma^2 G^2 \delta^2 \left(\Delta - \frac{\delta}{3}\right) D\right), \tag{7}$$

where S is the signal with the diffusion gradient applied, S_0 the signal with no diffusion gradients applied, G the diffusion gradient amplitude, δ the gradient pulse width, Δ the time between pulses, γ a constant (gyromagnetic ratio), and D the variable of interest – the diffusion coefficient. If two signals are measured, one with diffusion gradients, and one without, the scanner parameters can all be inserted into the equation and the value for D calculated. For most DWI measurements, the value of Δ and δ are fixed, with the value of $(\Delta - \delta/3)$ known as the diffusion *time*, and the gradient amplitude is adjusted to different values. In one of his papers, Le Bihan suggested to simplify the equation by combining the scanner acquisition parameters into one single factor, called a "b factor" or b-value [16], which simplifies the equation to

$$\frac{S_b}{S_0} = \exp(-b \cdot D)\,. \tag{8}$$

Real DWI data measured by an MRI scanner has added noise associated with the acquisition process, so to reduce the variance of the value of D, multiple signal measurements are taken at different sequence parameters and a curve of best fit calculated for the data points. The single exponential decay model of Equation 8 is often termed a monoexponential decay model and is widely used in DWI analysis. Other commonly found models are the multiexponential decay model, kurtosis model, and stretched exponential model, which expand off of the monoexponential decay model, along with even more complex models that can assess diffusion in multiple dimensions.

1.2.1 Monoexponential Decay Model

If Equation 8 is used to assess a pure liquid undergoing free diffusion, for example, the calculated value of D will be the diffusion coefficient, equal to the variance of a Gaussian PDF. For more complex environments where there are combinations of hindered and restricted diffusion, or where there are combinations of substances diffusing freely, the measured signal will no longer be a single exponential decay [17]. If the model in Equation 8 is still applied to measurements with non-Gaussian diffusion, then D is instead termed an Apparent Diffusion Coefficient (ADC)

$$S_b = S_0 \exp(-b \cdot ADC)\,. \tag{9}$$

Because the *ADC* and *b*-value parameters are contained within the exponential function, their relationship to the measured signal S_b is nonlinear. However, if a logarithmic transformation is applied to both sides of the equation, for any two *b*-values, the *ADC* value can be determined by the equation

$$ADC = \frac{\log\big(S(b_2)\big) - \log\big(S(b_1)\big)}{b_1 - b_2}. \tag{10}$$

This relationship can be solved exactly with only two measurements (the endpoints of a line) and for $b_1 = 0$, the value $S(b_1) = S_0$, as in Equation 9. This allows for a simple calculation of the *ADC* value, but sampling only two data points can make this value highly susceptible to noise. To reduce the bias and variance of the estimated *ADC*, multiple *b*-value measurements can be acquired, logarithmically transformed, and fit with a linear regression method.

The monoexponential model has been in use since the development of DWI imaging and the early focus of most DWI anatomical measurements was structures in the water-rich tissue of the human brain. An early review showed that the diffusion coefficient was able to identify structures in the brain, such as white matter, grey matter, and cerebrospinal fluid [18]. This review also reports on using DWI acquisitions for the detection of ischemia in the brain due to blood vessel blockage as in a stroke, as well as the detection of cancerous lesions. DWI researchers have the ability to detect minute changes to the white matter structure [19], and the ability to create whole mappings of the brain through tractography [20, 21]. DWI measurements of decreasing *ADC* are now also a standard recommendation for the imaging of acute ischemic stroke [22, 23].

DWI imaging has since moved beyond the brain to various tissues found throughout the entire body. It has been used in the detection of cancerous tumours inside the visceral organs such as liver, kidney, and pancreas [24], along with lesions found in lymphatic tissue and bone marrow [25]. Application of the simple monoexponential decay model on DWI data currently performs well at the detection of cancers at the whole body level [26, 27], and on cancers in the pelvic region [28] along with other visceral organ cancers [29]. Whole body DWI, where the entire body is imaged in one scan, is now being implemented in clinical use, due to its capabilities for identifying and monitoring metastatic disease [30]. DWI has also been included as part of a standard, multiparametric prostate imaging protocol for the detection of prostate cancer in the international standard PI-RADS (version 2) [31, 32]. For many DWI measurements, especially in the brain, the *ADC* has been superseded by the application of more complex models to distinguish structural brain features more accurately [20]. However, the *ADC* in many measurements may not actually be measuring the molecular MSD and the meaning of *ADC* as applied to DWI data can be unclear [33]. Hence, current DWI research often focuses on more complex methods and models to detect and identify intricate tissue structure and diffusion restrictions in images.

1.2.2 Multiexponential Decay Model

A natural extrapolation of the single monoexponential decay model is a model that includes more than one decay component. If a given measurement voxel of a diffusion MRI measurement is made

up of two or more separate structures with different diffusion coefficients, the result will be a signal that is a sum of exponential decays,

$$S(b) = S(0)\sum_{i=1}^{N} f_i exp(-bD_i)\,, \tag{11}$$

where f_i is the signal fraction of each component of the overall signal with the total of all N signal fractions equal to 1. This first main application of this model was made by Le Bihan et al [34] where a two decay component equation was used to estimate the diffusion of water through tissues and the *perfusion* of blood moving through vessels. This biexponential model use is referred to as IntraVoxel Incoherent Motion (IVIM) and this study established the theoretical basis for the biexponential model. The IVIM/biexponential model can be used with regression fitting using either Equation 11 with two decay components and four fitting parameters or by substituting the signal fraction of the second component with the remainder from 1 of the first signal fraction component,

$$S_b = S_0(\, SF_1 \exp(-bD_1) + (1 - SF_1)\exp(-bD_2)\,)\,. \tag{12}$$

In this equation, the value of D_1 is customarily known as the "fast" component as it has a larger decay rate and decays faster as the b-value increases, while the D_2 component is known as the "slow" component. In the IVIM model, the fast component is assumed to be perfusion, as the moving blood causes a faster decay, while diffusion is assumed to be the slow component. This model has four parameters and requires a nonlinear regression method to estimate these parameters with noisy data. This model is also currently used for a variety of measurements outside of IVIM imaging, and Chapter 2 of this thesis contains a detailed analysis of the current uses of the biexponential model in the literature and examines its reliability when using it to fit noisy data.

1.2.3 Kurtosis

If a stochastic variable has a Gaussian PDF, the first two cumulants are the well-known *mean* or *expected value*, labelled μ, and the *variance*, σ^2. This PDF is also known as the normal distribution and is often identified mathematically as $N(\mu, \sigma^2)$. The third cumulant of a PDF is the *skewness* and the fourth cumulant is known as the *kurtosis*, which refers to the degree that a distribution is either peaked or rounded [35]. When measuring molecular displacements in complex tissue, the kurtosis is a way to determine how the displacement distribution differs from a Gaussian one produced by free diffusion. Comparing the kurtosis value this way measures the *excess kurtosis*, so a Gaussian distribution has excess kurtosis of zero. If a distribution is sharper or more peaked, there is a positive excess kurtosis, and if it is more rounded, the excess kurtosis is negative. Estimating the excess kurtosis in DWI measurements can be made with the equation [36]

$$S_b = S_0 \exp\left(-bD_{app} + \frac{1}{6}b^2 D_{app}^2 K_{app}\right). \tag{13}$$

In this equation, an extra fit parameter K_{app} is also estimated, making this a three parameter model, and the D and K parameters have a subscript of *app* to reflect that they are apparent coefficients. If the measured diffusion of a given set of noisy data is free diffusion, then the value of K is zero, and Equation 13 is identical to Equation 9. This model only focuses on the change in kurtosis from a Gaussian distribution, which has a mean and skewness of zero. In the DWI literature, excess kurtosis is commonly referred to as just "kurtosis" and the model in Equation 13 called the "kurtosis model", which is the convention that will be followed in the rest of this thesis.

The kurtosis does not have a known, direct biophysical interpretation in human tissue like the IVIM model [37], but focuses instead on measuring the shape of the molecular displacement distribution. While it may not be directly modelling any physical properties of tissue, it is still a *statistical* model, and will be referred to as a model in the rest of this thesis for consistency. The kurtosis model has appeared in the last ten years in the DWI literature and has been used in the analysis of ischemic stroke and neurodegenerative diseases [38]. Kurtosis has also been used to identify gliomas [39], tissue changes in the human kidney [40], and more recently in studies of abnormalities and lesions in the prostate [41, 42]. These studies report that the kurtosis model was better at assessment of specific tissue anomalies than the conventional, monoexponential model, identifying its usefulness in DWI analysis. Chapter 3 examines the kurtosis model in detail, including its reliability when fitting the model to data.

1.2.4 Stretched Exponential Model

The stretched exponential model is used with fitting algorithms to assess anomalous diffusion in a given voxel. From Section 1.1.3, anomalous diffusion has a power law scaling parameter shown in Equation 6, and this diffusion scaling has been modified to an equation of [43]

$$S_b = S_0 \exp[-(b \cdot DDC)^{\propto}] . \tag{14}$$

The DDC value for this equation is called the Distributed Diffusion Coefficient. Hall and Barrick [44] modified this equation to better reflect the change in diffusion dynamics,

$$S_b = S_0 \exp(-Ab^{\gamma}) . \tag{15}$$

In this case, the value of A is the same as DDC^{α}, and the focus of this equation is on the stretching parameter γ, which can be directly related to the fractal dimension of the tissue structure. This is another three parameter fitting model, and here the stretching parameter can assume values between 0 and 1. When the stretching parameter value approaches 1, the diffusion becomes more like Gaussian diffusion, and eventually produces the same relation as Equation 9. The stretched exponential model is also relatively recent, but has been used in DWI analysis of liver fibrosis [45] and ex vivo prostate tissue samples [37]. These two studies also reported that the stretched exponential model provided more information about the respective tissue than the monoexponential model. This model, however, will not be examined further in this thesis.

1.2.5 Three-dimensional Diffusion Modelling

The previous models are designed to measure the signal over multiple b-values with the scanner gradients setup for the same single direction in three-dimensional Cartesian space. Producing a three-dimensional diffusion assessment of tissue requires multiple scans in different orientations and combining the resulting diffusion coefficient values from these orientations. The most popular method of doing this is the construction of a diffusion tensor, which assesses the changes in multiple diffusion measurement vectors [46]. In three dimensions, a tensor requires a 3x3 matrix of nine elements which covers all of the different combinations. For Diffusion Tensor Imaging (DTI), this matrix would be

$$\boldsymbol{D} = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix}. \tag{16}$$

The cross product values are symmetrical in this tensor, i.e. $D_{yx} = D_{xy}$, so only six unique diffusion directions need to be measured. The six diffusion values are usually selected to be as evenly spaced over the directional sphere as possible. In a typical DTI measurement, one PGSE acquisition with no applied directional gradient is taken (b=0) for a voxel, and six measurements taken with the gradient applied and a b-value between 600-1000 s/mm^2. The b=0 value is used with each gradient signal b measurement to calculate the *ADC* values based on Equation 10, and with those values and the (x, y, z) gradient information for each signal, the six directional diffusion values needed for Equation 16 can be calculated. The diffusion tensor can also be decomposed into three eigenvectors $\boldsymbol{\varepsilon}$, which form three orthogonal axes in (x, y, z) space. These eigenvectors have three eigenvalues $\boldsymbol{\lambda}$ associated with them, with the primary eigenvalue (λ_1) having the value with the largest diffusion coefficient, and the other two eigenvalues having the secondary diffusivities. These eigenvectors can be used to visualize the directional difference in diffusivity for the voxel, by creating an ellipsoid with ε_1 as the primary axis as shown in Figure 1.

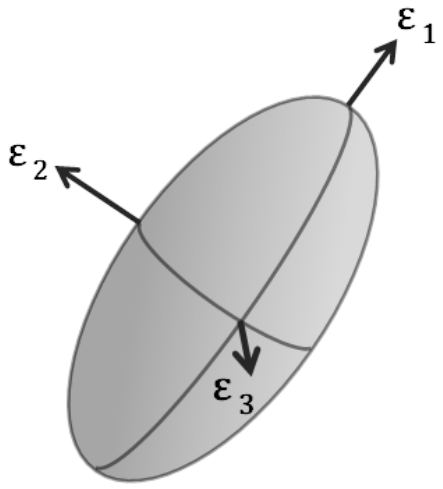

Figure 1 – Ellipsoid representing the orientation of the three diffusion tensor eigenvectors

The diffusion tensor eigenvalues also produce two other useful measurements. The Mean Diffusivity (*MD*) is the average of the three diffusivity eigenvalues, which because it's the average of multiple values, is more robust to noise than a single *ADC* measurement. The other measure is the fractional anisotropy (*FA*), which is a value between 0 and 1 that describes the degree of diffusion anisotropy in the voxel, and an example equation to calculate this measure is

$$FA = \frac{\sqrt{(\lambda_1 - \lambda_2)^2 + (\lambda_2 - \lambda_3)^2 + (\lambda_3 - \lambda_1)^2}}{\sqrt{2(\lambda_1^2 + \lambda_2^2 + \lambda_3^2)}}. \quad (17)$$

Fractional anisotropy is a widely used measure to quantify anisotropy in diffusion measurement over a wide variety of applications. Diffusion tensor measurements in the MR literature are widespread, especially in relation to the brain. For many measurements in the brain, though, DTI has been reported to be inadequate for measuring certain structures, with more complex models demonstrated to have better performance. While, analysis of three-dimensional models won't be specifically discussed in this thesis, the previously described one-dimensional models can be applied to each axis of a set of multi-axis data to assess specific differences in each axis.

1.2.6 Other Models

Additional DWI models have been developed that are combinations of the aforementioned models, for example, a biexponential model, with either one or both decay components replaced by a stretched exponential decay component, was assessed on prostate tissue data [47]. A model called VERDICT, consisting of three separate components that assess the intracellular, extracellular, and perfusion components of prostate tissue, has been used to assess both healthy and cancerous tissues [48]. For brain tissue measurements, dozens of combinations of a set of fundamental three-dimensional models were compared and assessed on their fit quality [49]. Improvements in MRI scanner technology lead to better SNR and higher image resolution, giving increased ability to identify more structures in tissue. In response to these improvements, more flexible models with added parameters that assess additional information in the tissue are introduced. As the presented studies for the biexponential, kurtosis, and stretched exponential models reported, these models can better fit data than the simpler monoexponential model, demonstrating their clinical usefulness. The drawback of a more complex model, though, is that while an increased number of model parameters may make it more sensitive to changes in tissue structure, it also becomes more susceptible to noise and the model parameters can be harder to interpret.

1.3 Estimation of Model Parameters

Actual DWI data will always be corrupted with some amount of noise contributions from various physical phenomena in the acquisition process. Assessing the underlying properties of the tissue structure in noise-corrupted data requires mathematical and statistical methods that attempt to quantify both signal and noise to produce the best model parameter estimates. A probability density function describes the probability that a stochastic variable will assume a value given the parameters of the distribution. In this section, the reverse problem is analysed through the introduction of *likelihood* – what the most likely parameter values are for a model, given the

measured data. Determination of the most likely parameter values for a model and data is usually done using least squares regression or maximum likelihood estimation.

1.3.1 Least Squares Regression

Finding the most likely parameter values involves studying the relationship between the independent (sometimes called predictor) variable(s) and the dependent (response) variable [50]. This relationship usually is constructed through a mathematical function with an additional stochastic variable (ε) representing the addition of error or noise

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\beta}) + \varepsilon \,. \tag{18}$$

The independent variable $\boldsymbol{x}$ and dependent variable $\boldsymbol{y}$ are bolded as they can be a vector of multiple values and likewise the vector of parameters $\boldsymbol{\beta}$ is bolded as there can also be more than one parameter in the equation. One of the simplest relationships between x and y is a linear relationship dependent on only two parameters: β_0, the y-intercept of the line, and β_1, its slope. When fitting that model to a set of n observed data, the linear regression model is written as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \dots, n \,. \tag{19}$$

Least squares (LS) regression is a method that determines the best fit values of the β parameters based on the sum of squared deviation between the observed data y and the line of best fit based on those parameters and the values of x [51]. These deviations are also called *residuals* and the mathematical expression for minimizing a general least squares problem is

$$\min \sum_{i=1}^{n} (y_i - f(x_i\,, \boldsymbol{\beta}))^2 \,. \tag{20}$$

To evaluate the linear model in Equation 19, $f(x_i, \boldsymbol{\beta})$ can be substituted with $\beta_0 + \beta_1 x_i$. An illustration of Linear Least Squares (LLS) regression using this model on simulated noisy data is shown in Figure 2. The plot there shows that with the addition of noise to the true signal, the LLS regression algorithm is only able to get the best estimate of the true signal given the available information and model, but can't recover the true underlying signal. Thus, in parametric regression, the goal is to find a *maximum likelihood* estimate which should give the set of parameters with the *minimum* deviance between a specified model and the data. Fortunately, least squares regression does give the maximum likelihood *if* the errors are independent, normally distributed, and have equal variance [52]. LLS algorithms make things very easy as the mathematical formulation is based on straightforward matrix algebra and gives an *exact* solution for the parameters

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \,. \tag{21}$$

The hat on the $\boldsymbol{\beta}$ parameter vector indicates that these are the parameter *estimates*. As demonstrated in Section 1.2, using parametric models beyond a simple linear model is often desired. Fortunately, more complicated models can be used with LLS regression as long as all of the

parameters have a linear relationship in the equation. Thus, if the specified model was a third-order polynomial model like

$$y = \beta_3 x^3 + \beta_2 x^2 + \beta_1 x + \beta_0 + \varepsilon\,, \tag{22}$$

an LLS algorithm could be used as the estimated parameters ($\boldsymbol{\beta}$) all have a linear relationship to y in the model equation.

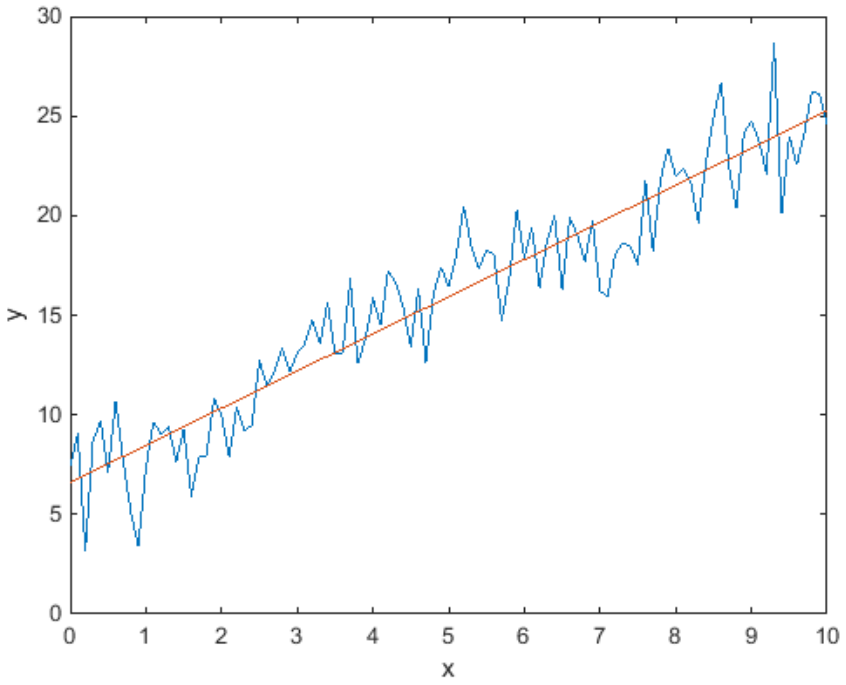

Figure 2 – Linear Least Squares Fit (red line) to simulated data (blue)

The simulated data is generated by the linear model y = 2x + 6 + ε, where ε is Gaussian noise with zero mean and standard deviation of 2. The line of best fit has a slope of 1.87 and an intercept of 6.6, indicating that there is bias between the best estimates and the true values.

1.3.2 Nonlinear Least Squares Regression

As shown previously in Section 1.2, unless the model was mathematically transformed, all models discussed therein were nonlinear models, as one or more of the parameters had a nonlinear relationship to the measured signal. This requires the use of Nonlinear Least Squares (NLLS) algorithms to estimate the model parameters for a given set of data. NLLS algorithms typically calculate a linear approximation of the fit for a set of parameters, and then iteratively refine that solution to find the minimum deviance between model and data. Different algorithms perform different operations to find this minimum deviance, but all NLLS algorithms must be provided with a set of starting values for the model parameters. For a given model, however, there can be multiple local minima, but the minimum deviance is found at what is called the global minimum. For certain models, multiple parameter start points should be used to attempt to find a global

minimum, as the algorithm may get hung up at a local minimum for a given set of starting values. There is also no guarantee that a particular algorithm can find a minimum deviance less than a specified stopping criteria, so sometimes an algorithm may just stop and return whatever values it stopped at.

Residual Sum of Squares (RSS)

An essential measure in least squares regression, both linear and nonlinear, is the Residual Sum of Squares (RSS), which is the sum of all individual residuals squared, and is a standard output from NLLS regression algorithms. This quantity is the equivalent of the expression in Equation 20, with each residual $r = y_i - f(x_i, \boldsymbol{\beta})$, and is the value that is minimized as a NLLS algorithm attempts to find a minimum. A lower RSS means less deviation between the model fit and data, so the RSS can be used as a measure to compare the closeness of a fit. For a given DWI data voxel, comparing the RSS value obtained when fitting different models on that voxel can show which model gives the closest fit to the data.

Jacobian

The key matrix to be aware of when using NLLS algorithms is the Jacobian matrix (with respect to the model function). If the goal is minimizing the deviance between a model equation with m parameters to be evaluated and a set of n noisy data observations, then the Jacobian matrix $\boldsymbol{J}$ is an m x n matrix made up of the partial derivatives of each parameter versus each independent variable data point. So, if x is the independent variable, and $f(x)$ the model function, as in Equation 20, the individual elements of $\boldsymbol{J}$ are

$$J_{ij} = \frac{\partial f(x_i)}{\partial \beta_j} \text{ where } i = 1, \dots, m \text{ and } j = 1, \dots, n, \tag{23}$$

meaning J_{ij} is the partial derivative of $f(x)$ with respect to the regression parameter β_j, evaluated at x_i. Evaluation of the Jacobian is important in NLLS regression algorithms for two reasons. The first is that many NLLS algorithms accept user calculated Jacobian functions. Many NLLS algorithms that are gradient-based attempt to approximate the local Jacobian elements based on the nearby numerical conditions. If the Jacobian derived from the model function is provided to the algorithm, the algorithm can find a minimum more rapidly. The downside of NLLS algorithms is the significantly slower speed in finding the minimum, often taking 30, 40, or 50+ iterations. Providing a specific, function-based Jacobian to the algorithm can significantly reduce the number of iterations needed and therefore the time needed to find a minimum.

The second advantage of using the function Jacobian is for the calculation of statistical diagnostics for a specific regression fit. After a minimum is found by the NLLS algorithm, the Jacobian matrix evaluated at that minimum can be returned from the algorithm and this Jacobian can be used to obtain additional diagnostic information about the regression fit itself. This information includes the correlation and covariance matrices for the model parameters, along with the estimated parameter variance, at a given fit [53]. Although the Jacobian is usually obtainable from a NLLS

algorithm, the other measures derived from it are often not standard in regression software. While the RSS and its derivations, i.e. R^2, provide a measure to determine the closeness of fit at the signal level, they do not provide any information about the parameter estimates themselves.

1.3.3 Maximum Likelihood Estimation

As mentioned in Section 1.3.1, if the errors are independent, normally distributed, and have equal variance, then least squares regression is equivalent to the maximum likelihood. As the next section will show, most DWI measurements have a Rician distribution. This distribution is not normally distributed over all measurements, and also has unequal variance over the measurements, a condition known as *heteroskedasticity*. To account for changes in regression fitting due to this Rician distribution, Gudbjartsson and Patz created a bias reduction scheme to adjust for the differences in the distribution [54]. This can be used to obtain a corrected signal A, using the measured data M and the best estimate of the noise variance σ^2, via the equation

$$A = \sqrt{|M^2 - \sigma^2|}\,. \tag{24}$$

To get the best estimate of the maximum likelihood (ML), however, a custom equation that takes into account the full Rician distribution needs to be used. One such equation was created by Sjibers and den Dekker for ML analysis as a log likelihood function [55]. The function maximizes the log likelihood for the signal values A, again given the magnitude data M and estimated variance σ^2 with the full log likelihood equation for N data points being

$$\log\, L = \sum_{i=1}^{N} \log\left(\frac{M_i}{\sigma^2}\right) - \sum_{i=1}^{N} \frac{M_i^2 + A^2}{2\sigma^2} + \sum_{i=1}^{N} \log I_0\left(\frac{A \cdot M_i}{\sigma^2}\right), \tag{25}$$

where I_0 is the zeroth order, modified Bessel function of the first kind. The variables M and σ^2 can be plugged into the equation and a function minimization algorithm can be used to minimize the negative of the log likelihood function (equivalent to maximizing log likelihood) to get the ML values of A. Sijbers and den Dekker showed that using this ML function with noisy data generally gave a reduction in fitting error over a least squares algorithm. Walker-Samuel et al also showed that by adding the monoexponential decay model (Equation 9) to the ML equation, the bias in the estimation of the ADC was reduced compared to a least squares algorithm fit [56]. While the literature suggests that ML algorithms may improve error while attempting to fit noisy data, the computational complexity using Equation 25 increases considerably compared to a standard NLLS algorithm, requiring additional computation time to analyse all data. While either of these methods can be used to assess DWI data, the simplicity and speed of NLLS regression means it is often chosen over custom ML analysis.

1.4 Assessment and Analysis of Noise and Error

When acquiring magnitude DWI data, the added noise from the MRI scanner is not normally distributed, but instead has a Rician distribution, and this noise corruption affects the values of the model parameter estimates. To assess noise and error when fitting a model to data, the metric

often used is Mean Squared Error (MSE), which accounts for both the variance in the parameter estimates as well as the bias between the estimates and the true parameter values. Both bias and variance need to be minimised when assessing the performance of a model over repeated sample measurements.

1.4.1 Rician Distribution of Magnitude MRI Measurements

A typical PGSE measurement acquires complex valued data with real and imaginary image components, and the noise associated with these components has a Gaussian distribution [57]. Often, these real and imaginary images are combined into a magnitude and phase image and the phase portion usually discarded to remove phase artefacts in the image, leaving the magnitude measurement of the signal which is calculated with the equation [58]

$$S(b) = \sqrt{(S_r + n_r)^2 + n_i^2}\ . \tag{26}$$

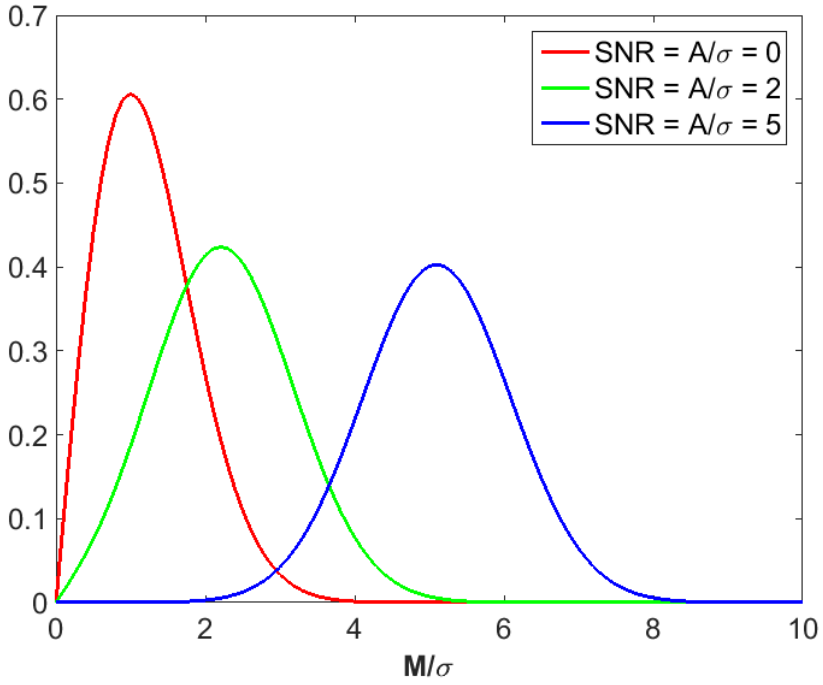

Figure 3 – Rician PDF's for three different SNR values

The noise standard deviation (σ) is set to 1 in this plot, so the true signal (*A*) is equal to the SNR value. The *x*-axis is the resulting magnitude measurement (*M*) divided by σ for the PDF of each true signal. For an SNR of 5, the Rician PDF is a Gaussian with mean of 5 and standard deviation of 1 (the added noise). At an SNR of 2, the left tail of the distribution is rectified and there is a slight positive bias to the mean value of the distribution. When there is zero signal, the resulting PDF is a Rayleigh distribution with a mean value of 1.

The second term under the square root would normally be $(S_i + n_i)^2$, but Equation 26 reflects the discarding of the imaginary measurement S_i. The total magnitude measurement now consists of a nonlinear combination of the real part of the signal plus two individual, Gaussian-distributed, real and imaginary noise components. The combination of the signal and noise components now has a distribution known as the Rice distribution, where its PDF is determined by [59]

$$p(M|A) = \frac{M}{\sigma^2}\exp\left(\frac{M^2 + A^2}{2\sigma^2}\right) I_0\left(\frac{MA}{\sigma^2}\right). \tag{27}$$

This equation gives the conditional PDF of the magnitude measurement M given the unknown true signal A, with σ^2 the associated noise variance and I_0 the modified zeroth-order Bessel function. As Equation 26 shows, if the two Gaussian noise components are squared, even if the particular noise components are negative, the resulting values will always be positive, leading to a "noise floor" phenomenon. This is reflected in the Rice distribution, as it is effectively a Rayleigh distribution when there is no signal (A) present [54]. Alternatively, when the SNR (A/σ^2) is ≥ 5, the distribution is effectively Gaussian with variance σ^2. A plot illustrating the changes in the Rician PDF at different SNR values is shown in Figure 3.

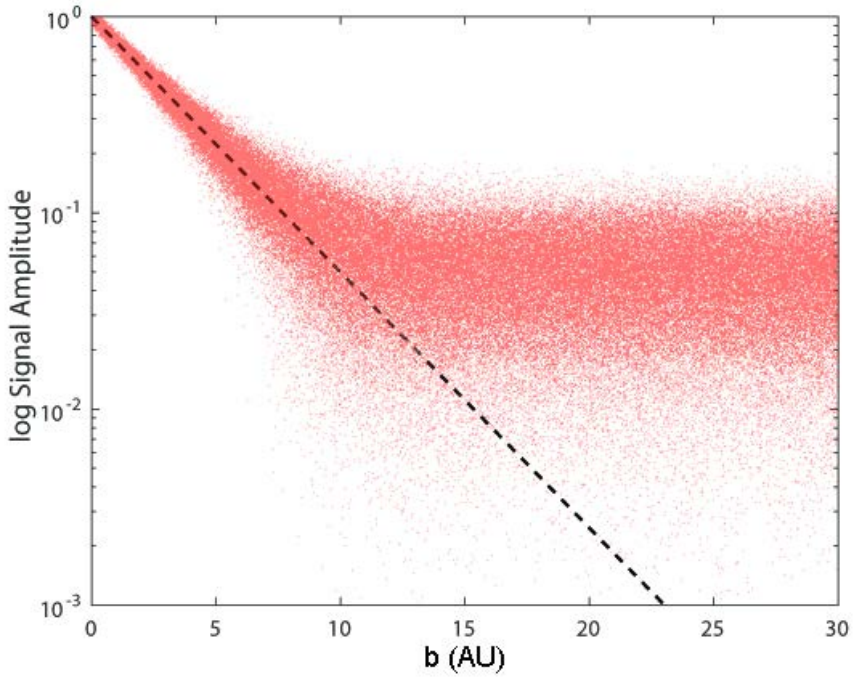

Figure 4 – A semilog plot of a noise-free exponential decay signal (black dashed line) and a "cloud" (scattered red dots) representing possible measurements of the resulting magnitude signal with added noise

Noisy signals were created by adding noise per Equation 26 with a standard deviation of 0.04 to the noise-free signal of exp(-0.3*b*). The value of *b* in this case is an Arbitrary Unit (AU). At high values of *b*, the measurements no longer track the true signal, but instead level off at the noise value of 0.04.

The Rician distribution of magnitude MRI images causes issues with DWI signal measurements due to the exponential decay relation in Equation 8. As shown in Figure 4, as the value of b increases, the true, noise-free signal of exp(-0.3b), indicated by the dashed black line, decreases. Simulated Gaussian noise components (η_r, η_i) are added to the noise-free signal per Equation 26, and as the noise-free signal decreases into the noise floor, the noisy signal measurements deviate significantly and level off at a near constant value. Thus, as the b-value increases, the SNR of the noisy signal not only decreases as the noise-free signal approaches the level of the noise, but the noisy signal also becomes biased compared to the true underlying signal value. If more additional measurements are taken at these higher b-values, the averaged signal will still not converge to the true, noise-free signal value, but instead one that is inherently biased, illustrating a significant issue with DWI measurements at higher b-values [60].

1.4.2 Mean Squared Error

To assess the fit quality in the case of a regression estimate based on a least squares model, an often-used measure is the mean squared error [61]. The MSE is useful as it incorporates both bias and variance, and for a regression estimate $\hat{\theta}$ of the true value θ, the MSE can be calculated by

$$MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = Var(\hat{\theta}) + \left[Bias(\hat{\theta}, \theta)\right]^2 . \tag{28}$$

If a particular regression estimate of noisy data is completely unbiased and converges asymptotically toward the true value with more measurements, then the bias would be zero and the MSE would be due solely to the noise variance. As was just shown with the Rician signal bias, the noise distribution may cause a regression estimate to be biased, and therefore the MSE will have both variance and bias components. With DWI measurements of tissue where there are various restrictions to diffusion, a true unbiased estimator is unobtainable because the regression models introduced previously don't completely describe the unknown, true signal. The MSE will then have variance and bias due to the Rician bias *plus* additional bias due to deviation between the selected model and the true underlying signal, demonstrating the importance of having more complex models with additional parameters – to reduce the bias between model and truth.

1.4.3 Complexity and Overfitting

While having too simple of a model for the data leads to bias in the parameter estimates since the model doesn't describe the signal well, a simpler model is more resistant to variance in the data. In the opposite case, when a model is very flexible and has lots of parameters, the model has less bias, but is more affected by variance in the data. Figure 5 shows a basic example of this phenomenon where noise-free data from a single exponential decay of exp(-1.0b) was fitted with Gaussian noise with a standard deviation of 0.1 at eleven data points. Two regression fits were then performed on the noisy data, a linear fit with slope and intercept parameters and a cubic polynomial fit with four parameters (as in Equation 22). Neither model completely accounts for the underlying signal, but the linear fit is more biased with respect to the true signal, especially at the lowest and highest b-values. The cubic polynomial fits the true data curve well for the first five data points, however, when the noisy signal deviates significantly from the true curve at higher b-values, the regression fit

is more affected by these deviations, increasing its MSE due to the variance. This phenomenon is known as overfitting [62]. When a model is overfit, it may provide the closest fit to one set of data, but when testing on different sets, it may give poor estimates that are largely affected by noise. Selection of a model that minimises MSE over all possible data values involves a compromise between finding a model complex enough to reduce the bias between model and data, but one not too complex to then increase the variance, a dilemma known as the bias-variance trade-off [63]. Balancing this bias-variance trade-off selects a model with the best performance, not just for a single acquisition, but over repeated samples of the same signal. Basing model selection only on how well a model fits a set of data is inadequate, and error analysis of DWI data requires more comprehensive methods that account for overfitting.

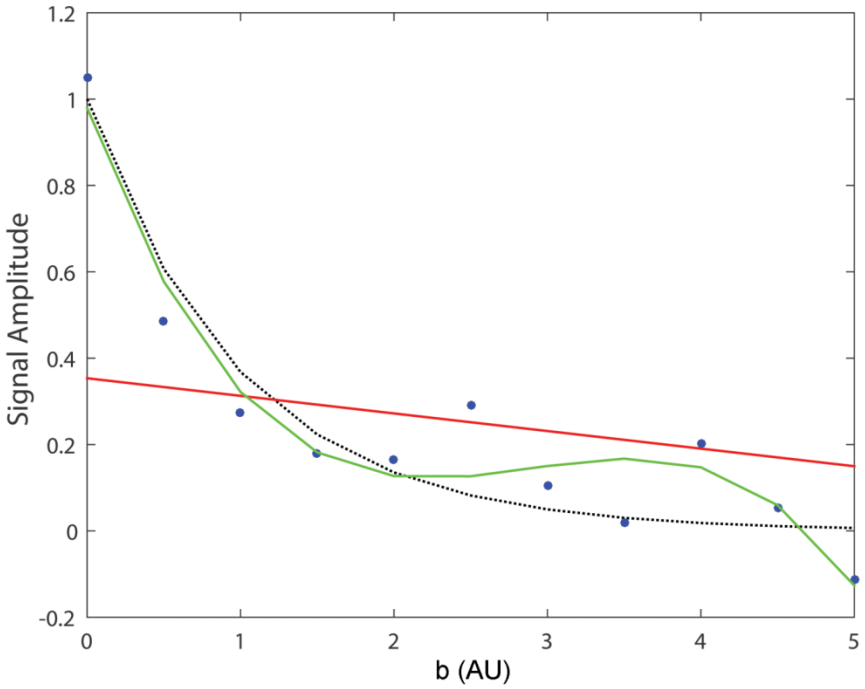

Figure 5 – Simple line/curve of best fit using a linear model (red) and a cubic model (green)

Linear model equation is $S = \beta_1 b + \beta_0$, and cubic model equation is $S = \beta_3 b^3 + \beta_2 b^2 + \beta_1 b + \beta_0$. The true signal (black dotted line) is $S = \exp(-b)$ with Gaussian noise of standard deviation equal to 0.1 added to create the noisy data (blue dots). The cubic model fits the data points better, but is more affected by noise, especially at higher values of b.

1.5 Model Selection

Balancing the bias-variance trade-off requires a model to have just enough complexity to assess features in the data, but not too much to be affected by noise when those features are minimal, a phenomenon known as model *parsimony* [62]. Applying the model with the best minimisation of the error between model and data is trivial if the process that generated the data is straightforward and known, such as measuring the diffusivity of water using a monoexponential model. When

assessing complex tissue with various restrictions and hindrances to the diffusing molecules, however, the true signal can have hundreds of dimensions and is effectively unknown. The model that best minimises the error between data and truth over repeated measurements should still be used in this case. When acquiring DWI data, however, this best model is often unknown and can be completely different for certain types of tissue. Hence, researchers want to determine the model that gives the best performance for a given tissue study. This requires model selection methodology that can compare multiple models to a given set of data.

1.5.1 Statistical Tests

A common practice seen in the DWI literature is the use of statistical hypothesis testing such as the *F*-test [64, 65], a likelihood ratio test [66], or an approximation of it like a χ^2 goodness-of-fit test [67]. These statistical tests are popular because they produce an identifier of statistical *significance*, usually a *p*-value, which is the probability that the observed value would be obtained strictly by chance if the null hypothesis was true. If this *p*-value meets a specified significance level, then one model is decreed significantly better than the other. The *F*-test is used to distinguish between nested models, and whether the more complex model is better than a simpler one, so it can be used effectively to compare the monoexponential model to the biexponential, kurtosis, or stretched exponential models, where the monoexponential model is nested inside. It cannot, however, be used to compare any of the three complex models to each other. If these models were able to be compared using a statistical test, comparing four models would require six separate tests, and the chance for random errors propagating through multiple tests is much higher than a single test between two models.

1.5.2 Information Criteria

The Akaike Information Criteria (AIC) is a model selection method that balances the model's complexity with a factor that penalises models based on the number of parameters [68]. The AIC assesses an important element of information theory, the Kullback-Leibler (K-L) divergence or distance [69], which is a measure of the difference between two probability distributions. Put another way, this is the information lost when one distribution approximates another, and can assess any two probability distributions, including the unknown true distribution. This is an important factor when selecting the best approximating model with DWI tissue data, as the truth is often unknown. The AIC compares two or more models to the unknown true distribution via estimation of the likelihood,

$$\mathrm{AIC} = -2\log\left(L\left(\widehat{\boldsymbol{\theta}}\middle|\boldsymbol{y}\right)\right) + 2k\,, \tag{29}$$

where $L\left(\widehat{\boldsymbol{\theta}}\middle|\boldsymbol{y}\right)$ is the maximum likelihood of the parameter estimates $\widehat{\boldsymbol{\theta}}$ given the data $\boldsymbol{y}$, and k is a variable representing the number of parameters being estimated in a particular model. Multiple models can be compared on the same set of noisy data this way, and the model that has the lowest value or score is the one with the lowest K-L distance to the true distribution. The likelihood function of a more complex model will be lower than a simpler model, but the $2k$ factor in the AIC equation acts as a penalty against complexity as it grows as the number of parameters increase. In

the case of fitting a model using least squares regression, the likelihood function in Equation 29 can be replaced by the RSS measure returned by the least squares fit, giving

$$\mathrm{AIC} = n \log\left(\frac{\mathrm{RSS}}{n}\right) + 2k\,, \tag{30}$$

where n is the number of data measurements and k, in this LS equation, includes the number of model parameters plus 1, which accounts for the variance in the estimation [70]. The change in sign in the first term of Equation 30 reflects that least squares algorithms attempt to *minimise* the RSS value, while likelihood methods attempt to *maximise* the likelihood function. This equation allows for multiple models to be easily compared using the results from a LS regression, with the lowest AIC value still selecting the model with the lowest divergence from the truth. There is also a corrected version of the AIC called the $\mathrm{AIC_c}$, which is more appropriate if the sample size is low, i.e. $n/k < 40$ [62],

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}\,. \tag{31}$$

The AIC is an appropriate tool for DWI analysis using least squares regression as it uses the RSS value from the fitting algorithm, or it can also be calculated using any custom maximum likelihood function. In the DWI literature, the AIC has recently been used as a way to rank multiple models on the same data set [37, 47, 71], as has the corrected $\mathrm{AIC_c}$ version [72]. However, the AIC and $\mathrm{AIC_c}$ information criteria are not without their drawbacks. The AIC can rank multiple models indicating the one with the lowest score is the best fitting model, but it cannot report that the models all have a large divergence from the truth. AIC provides only a relative value between two or more models *for a specific measurement*, and is not an absolute score that assesses fit quality of a model across all data. A related criterion called the Bayesian Information Criterion (BIC) has also been used to rank models in DWI data [73]. While this measure is valid for use in comparing DWI models, literature examples on comparison of AIC and BIC suggest that neither method has significant performance improvements over the other [74] and both are good approximations in most circumstances [75].

1.5.3 Cross-Validation

The other model selection method that will be examined in this thesis is cross-validation. Cross-validation involves removing one or more data points from a data set, and then using a model and the remaining data points to predict the value at the omitted point. Multiple data points can be left out, but since most diffusion MRI measurements don't have many data points for each voxel, typically only one data point is left out at a time, which is called Leave-one-out Cross-Validation (LOOCV). Cross-validation techniques are often used to validate a model by dividing up the data set into section, using one section of the data as a *training* set to get the best predictions from the model, and compare how these predictions perform on the other section, which is known as the *test* set [76]. Cross-validation can also be used for model selection [77], like the AIC, and it has been shown that the AIC and cross-validation are asymptotically equivalent for maximum likelihood estimation for large sample sizes, and the AIC does have some cross-validation properties when maximum likelihood estimation is used with the models [78].

Cross-validation has been used in the DWI literature to estimate the best model for a given set of data [79-81]. It provides another method of model selection using the information only from the given models and the data to be examined, meaning no additional information is needed. Because each point in a given voxel data measurement must be left out, the fitting time for the regression method for an entire measurement is multiplied by the number of data points in the set, i.e. if measurements are made at eleven different *b*-values, each one will have to be left out in turn, so the total time to perform regression on the data set is twelve times more than a single regression (including the original regression fit). This is a significant drawback when measuring several thousand voxels from a DWI data set. However, because it performs multiple tests on a single voxel data set, there may be an increase in selection performance compared to the approximations calculated using information criteria.

1.6 Statistical Inference

Statistical inference refers to the process of analysing observed data and attempting to infer the underlying properties of the phenomena that created them [82]. The topic of statistical inference is at the heart of DWI data analysis and ties together all of the topics introduced thus far in this chapter. DWI is a diagnostic imaging tool and the aim of most DWI analyses is to find phenomena in the measured signal that infer that a particular condition or abnormality is present in a volume of interest in human tissue. This is the process of causation and the forward problem – the presence of this condition *causes* this phenomenon in the measured signal. DWI analysis, like most medical imaging modalities, investigates the reverse problem – a signal phenomenon is measured, but the desired knowledge is whether a specific condition in a volume of interest caused the measurement. The reverse problem in DWI data analysis is difficult both because the signal is confounded by noise from the data acquisition process, but also when measuring human tissue, there can be several conditions present in a given volume at once. When making inferences about a given measurement, signal phenomena are then *correlated* with a particular tissue condition, often by selecting a region of interest (ROI) via pathological classification after tissue excision, and registering that same ROI in the DWI data. While the correlation of signal phenomena and tissue properties does not guarantee causation, statistical analysis gives researchers information to make inferences that should be the most likely given the data and model.

1.6.1 Uncertainty in Model Parameter Estimates

The methodology of most parametric model fitting in the DWI literature references in this thesis usually involve selecting a set of data to be measured, examining and cleaning that data by eliminating outliers or measurements that don't reflect reality, and preparing the data so each voxel in the data set is fit using a regression algorithm (often NLLS) with one or more of the previously mentioned models. Regression fitting is performed on all data and the results are often analysed by correlating the returned parameter estimates with regions of the data, selected either by visual identification or confirmation via histopathology results (or both). For each model tested, the parameter estimates for a selected ROI are often grouped as a distribution and the mean and standard deviation of this distribution reported. If there are multiple regions selected for a given acquisition, say a region of cancerous tissue and a region of normal tissue, the parameter estimates

from these two regions are often compared by a statistical test. If a test for a given parameter rejects the null hypothesis that there is no difference between the voxels from the two regions at a given level of significance, the parameter is often reported as being *significantly* higher or lower in one region vs. another. The interpretation of this inference is that the parameter difference in the two populations is likely due to changes inherent in the tissue structure, and is often correlated with pathology or a different imaging modality to bolster this evidence. Based on this knowledge, these differences in model parameters should allow researchers to make better diagnostic predictions of a given tissue anomaly or condition *in future studies*.

In addition to studies on fitting models to tissue data, there are many studies in the literature on noise affecting the MRI acquisition process, uncertainty in DWI signal measurements, and methods to minimise noise, with a few examples presented in Section 1.4. However, a gap in the literature exists between these two portions of the data analysis process, namely, a detailed study of the reliability of current DWI models when used with NLLS regression algorithms. Modern computational languages make NLLS regression algorithms easy to use and are essentially "plug and play", where both data and model are entered into the algorithm and parameter values returned in seconds. Because of this, it's easy for a researcher to overlook the model fitting step as a source of error, and under this assumption, any significant errors or outliers in the parameter estimates may be attributed back to noise from the physical MRI acquisition process. While the monoexponential model is well established in clinical use, it provides a rough estimate of tissue microstructure, so introducing more complex models to clinical DWI analysis would expand medical researchers' capabilities to better identify structural details [83]. For two of these models, biexponential and kurtosis, the monoexponential model is nested within their mathematical equations, and the assumption is that these more complex models may supersede it, since they can assess any monoexponential signal, as well.

The assumption of similar model performance of the biexponential and kurtosis models is largely based on current statistical studies of data combined from multiple patients. This gives a large range of parameter estimates to accommodate most possibilities, however, these values are mostly obtained from single acquisition studies. While statistical methods produce an estimate of what parameter values are likely to be obtained in future studies, there is no guarantee that these values will be obtained unless tested. What is missing in the literature is a detailed assessment of how parameter estimates vary across repeated sample measurements *of the same patient*. This may be due to the prohibitively high cost of acquiring MRI data, especially in vivo, so studies that investigate repeated sample measurements from the same voxel or voxels are rarely performed. With no indications of regression analysis problems in the DWI literature, performing such an experiment may seem pointless to begin with. Testing how models perform when fitting repeated measurements can instead be achieved by generating synthetic data via computer simulation. While an assessment of simulated data cannot be directly related to the results of actual empirical studies, using these data can isolate whether uncertainty in the regression fitting process is due solely to noise or to algorithmic issues, as well. Chapter 2 and Chapter 3 of this thesis present an assessment of the reliability of the biexponential and kurtosis models, respectively, using such data.

1.6.2 Reliability of Model Selection Methods

Section 1.5 presented a sample of DWI literature studies where multiple models were applied to a set of tissue data, and a specific model was reported to have been selected by a model selection method as the best for those tissue data. The inference here is that that model would then give better predictive ability in identification of a tissue type or condition, assisting researchers who will choose models in future studies. These studies often use information criteria or cross-validation to compare multiple models, with the models ranked based on how often they are selected as the best model for all voxels of a specific tissue or ROI. While these studies are important contributions to the DWI literature, the choice of model selection method in these studies varies. AIC, BIC, and AIC_c all have well-established theoretical foundations in the literature via the references presented in Section 1.5.2, but no method has been clearly established as being better than the others with a lack of detailed studies performed on when or why a particular criterion should be used for DWI measurements.

With these information criteria, in-depth analysis of their selection performance is lacking, most likely, due to the statistical literature presenting these information criteria with little regard to uncertainty in the model selection process. When comparing models on a single voxel from a set of DWI data in one acquisition, inferences are made for one specific noisy measurement of a true signal, but this disregards the possibility that the same model selection method may choose a completely different model for that voxel in another acquisition. While model selection uncertainty has been noted in a few places in the model selection literature, e.g. Section 1.7 of [62] as well as [84, 85], there is little evidence in the DWI literature on how reliably model selection criteria select models over repeated measurements. The biexponential and kurtosis models each have the monoexponential model nested within them, so if a signal is effectively monoexponential, all three of these models should fit it similarly. The expectation is that model selection methods would select the monoexponential model as the best model in this case, but if the kurtosis or biexponential model happen to be selected, these models should also deliver similar results for the parameter estimates. As noted in the previous section, however, equal performance of these models when assessing monoexponential data is still based on assumption. What is missing in the literature, then, is a detailed assessment of the common model selection methods when analysing repeated measurements from various signals, and whether there is any cost when a model selection method chooses a more complex model on a signal better described as monoexponential or vice versa. A detailed study of how information criteria and cross-validation reliably selected models on synthetic data is presented in Chapter 4.

1.6.3 Aims of this Thesis

Detailed analysis on the parameter estimates from the biexponential and kurtosis models and their sensitivity to noise is lacking in the literature. On complex tissue where the truth is unknown, the justification for using more complex models on a set of data is often from previous studies that used model selection methodology. These methods are applied under the assumption that they will always determine the best model without providing any explanation of what “best” means, or, if there is any uncertainty associated with the model selection process.

Based on this analysis, the aims of this thesis were:

- Quantify the uncertainty in the biexponential model parameter estimates when using NLLS regression algorithms across possible parameter values, and assess the effects of measurement noise on the parameter estimates.
- Quantify the uncertainty in the kurtosis model parameter estimates with NLLS algorithms.
- Introduce additional diagnostic measures to NLLS regression analysis that could allow researchers to identify whether there are large uncertainties in their parameter estimates.
- Present a detailed analysis of the effects of measurement noise and varying acquisition parameters on the AIC, AIC_c, and LOOCV model selection methods.
- Investigate how the uncertainty in the parameter estimates varied as a model was applied to data where the model did and didn't describe the underlying signal.
- Combine the findings from these simulated parameter estimate and model selection analyses and revisit a multimodel, DWI analysis of excised prostate tissue.

Chapter 2
Performance of the Biexponential Model Using Simulated DWI Data

2.1 Introduction and Background

In 2008, Le Bihan wrote an updated review [86] on the usage of the biexponential model in IVIM imaging, highlighting many successful studies and technological improvements made in the twenty years since his initial IVIM publication. One specific in vivo liver study [87] was highlighted there, where the authors reported that the mean perfusion component was significantly lowered in measurements from patients with cirrhotic livers compared to healthy liver tissue, while the mean diffusion component was slightly higher. This decrease in liver perfusion concurred with the results of other liver perfusion studies, and this biexponential model result stood in contrast to earlier monoexponential model DWI studies that reported a decrease in *ADC* value and attributed it to overall decreased diffusion. This is just one specific study that demonstrates the additional measurement capabilities of the IVIM model, since it was able to assess both perfusion and diffusion components, and indicates why so much research has gone into establishing it as a reliable model for analysing DWI data.

In addition to the decay parameters representing diffusion or perfusion rates, the amplitude parameters are often correlated with changes in the vasculature of tissue, demonstrating that the signal fraction of the amplitude components correlates with the volume fraction of a given voxel [88]. In less well-perfused tissues, or in ex vivo measurements, biexponential analyses have attempted to resolve two diffusion components instead. These two diffusion components are often attributed to water in intra and extracellular compartments, although this finding is now regarded as overly simplistic [89, 90], not least because non-monoexponential (non-Gaussian) diffusion has been reported from the cytoplasm of a single cell [91]. As well as free and restricted diffusion compartments, multiexponential decay has been correlated with other factors including exchange between restricted diffusion compartments [92] and T2 relaxation effects [93]. The biexponential model has also been used in the Kärger model [94] in the absence of a compartment exchange component, as well as an approximation in a mesoscopic effective medium theory model [95]. Biexponential models have also been demonstrated to fit T2 in vivo brain data better than a monoexponential model [96-99]. The theoretical and biophysical basis for the biexponential/IVIM model has been well established in the literature, and the large body of empirical research indicates that the biexponential model is useful and should be a good addition to clinical research.

Estimating biexponential model parameters with modern NLLS regression fitting methods requires computational algorithms, but there are many examples in the computational literature that are highly *critical* of the biexponential model. In the book *Numerical Methods That Work*, written in 1970 by Forman Acton [100], the chapter titled "Interlude: What *Not* to Compute" has a section on exponential fitting with the following quote, "For it is well known that an exponential equation of this type in which all four parameters are to be fitted is *extremely* ill conditioned. That is, there are many combinations of <*the fitting parameters*> that will fit most exact data quite well indeed (will you believe four significant figures?) and when experimental noise is thrown into the pot, the entire

operation becomes hopeless." The problem of fitting sums of exponentials was demonstrated before Acton's work, with Lanczos demonstrating in 1956 that a data set from a sum of three exponential decay components can be closely fit by completely different sets of parameter estimates [101]. There have since been many computational literature examples on methods to reduce the fitting bias and variance of multiexponential decay models, for example, a separable least squares algorithm that regressed separately on the linear and nonlinear components of the model equation [102]. This algorithm has been updated and improved for modern programming languages, with separate software packages for *R* [103] and *MATLAB* [104]. These two packages both provide enhanced error reporting and algorithm analysis capabilities along with the fitted parameter results, but the examples provided within this documentation also demonstrate how easily multiexponential decay models can return significantly different parameter estimates from the same signal.

These warnings from the computational literature have been noticed in a few DWI studies. A review paper [105] noted these warnings and specifically acknowledged Acton's work, but contrasted these with examples in the DWI literature where the biexponential model returned results that were relatively consistent across different research groups and experiments. A study using the IVIM model to examine breast cancer by Sigmund et al [88] specifically noted that the biexponential model can be *ill-conditioned,* and rather than use the entire model in a regression algorithm, performed a segmented analysis that was "more numerically stable". Another recent paper noted ill-conditioning present when fitting with the biexponential model, classified these measurements as outliers, and recorded the frequency of these outliers while adjusting the number and spacing of *b*-value measurements [106]. A third recent paper also reported an increase in outliers and extreme parameter estimate values from the biexponential model and acknowledged that these results may be due to ill-conditioning [107]. These few examples have acknowledged the algorithmic issues of the biexponential model and have identified specific problematic aspects when analysing DWI data, indicating that there can be problems with the IVIM model.

The monoexponential decay model is used in clinical measurements and its use in regression analysis of DWI data and issues with noise and parameter selection have been widely acknowledged and well-characterized [56, 108]. The potential effects of noise on NLLS approaches to biexponential analysis, however, have only been estimated for a number of specific cases [67, 109, 110]. Thus, there is a need to provide DWI researchers with a more complete analysis of the effects of noise on biexponential model parameter estimates by performing NLLS regression fits on simulated DWI data where the truth is known, giving a complete assessment of bias and variance in the parameter estimates. Such an error assessment should also explain ill-conditioning, the scenarios where it occurs in the biexponential model, the severity of its effects, and possibly include an expanded array of statistical tools to help researchers better identify possible problems in their parameter estimates.

2.1.1 Parameter Estimation Errors

A comprehensive review of multiexponential analysis in physical phenomena can be found in Istratov and Vyvenko [111], which explains the mathematical theory of the model's difficulties and

presents several computer algorithms and their performance on various data sets. In the section on NLLS regression fitting, the authors reported several studies that had difficulty resolving two decay components with a decay ratio of 2 at an SNR of 1000. Reporting a minimum decay ratio suggests that the uncertainty of the parameter estimates decreases as the decay ratio increases and vice versa. This is consistent with the aforementioned DWI study by Sigmund et al [88], which stated specifically, "…finite data sampling/precision, small perfusion fraction and/or *similar compartmental diffusivities* make an unconstrained fit ill-conditioned" (Emphasis added). These reports also suggest that the reliability of the parameter estimates from a biexponential model regression fit can vary based on the characteristics *of the measured phenomena themselves*.

In order to produce simulated DWI data that replicate data that are typically used for fitting, a literature survey was performed to examine the range of parameter values that have been reported with actual tissue data. Table 1 shows the results of this survey, listing the reported mean biexponential parameter values for SF_1, D_1, D_2, the resulting ratio of D_1/D_2, the SNR of the data (if reported), and the type of tissue measured along with any specific categorical conditions.

Table 1 – Reported parameters from DWI studies using a biexponential model in regression fitting

Shaded studies are ex vivo and all studies are human unless otherwise specified. A plot of this data is presented in Figure 6.

Tissue	SF_1	D_1	D_2	D_1 / D_2	$SNR_{b=0}$	Reference
Muscle (rat) - Edematous	0.89	2	0.27	7.4	30-50	[112]
- Control	0.84	1.3	0.16	8.1		
Liver (mouse)	0.76	1.23	0.27	4.6	> 90	[113]
Liver – Normal Tissue	0.19	16.1	1.1	14.6		[114]
- Cancer	0.14	51.9	1	51.9		
Breast	0.9	N/A	1.4		69	[115]
Liver - Benign	0.031	27.6	1.56	17.7		[116]
- Malignant Lesions	0.064	21.7	1.29	16.8		
Prostate (fixed) – Normal Tissue	0.74	1.56	0.23	6.8	160	[117]
- Cancer	0.6	0.87	0.1	8.7		
Kidney - Enhancing mass	0.28	14.1	1.47	9.6	> 15	[118]
- Non-enhancing mass	0.06	11.1	2.4	4.6		
Breast - Lesion, normal fit	28.4	8.67	1.01	8.6		[119]
- Segmented fit	13.3	15.3	1.323	11.6		
Prostate - Benign	0.21	8.03	1.21	6.6		[120]
- Cancer	0.14	8.36	0.84	10.0		
Liver	0.13	136	1.11	122.5	52	[121]
Muscle (rabbit, heart, fixed)	0.82	0.72	0.06	12.0		[122]
Brain - Grey Matter	0.74	1	0.3	3.3		[123]
- White Matter	0.69	1	0.1	10.0		
Liver	0.24	160	1.38	115.9		[124]

Kidney Cortex	0.17	14.2	1.6	8.9		[125]
Liver (rat) - Fibrosis Category 0	0.22	37.99	0.98	38.8		[126]
- Category 1	0.18	30.38	0.94	32.3		
- Category 2	0.17	29.15	0.79	36.9		
- Category 3	0.15	27.22	0.8	34.0		
- Category 4	0.14	27.07	0.81	33.4		
Liver - Fibrosis Category 0	0.25	76.2	0.91	83.7		[127]
- Category 1	0.25	75.7	0.9	84.1		
- Category 2	0.24	67.3	0.87	77.4		
- Category 3	0.25	60.7	0.84	72.3		
- Category 4	0.22	55.6	0.88	63.2		
Liver - Normal	0.17	70.6	1.02	69.2		[128]
- Carcinoma	0.17	28.2	1.07	26.4		
Pancreas - Pancreaitis	0.16	N/A	1.07	N/A		[129]
- Pancreatic Cancer	8.2	N/A	1.09	N/A		
Pancreas	0.16	64.5	1.58	40.8	>12	[130]
Pancreas - Head	0.39	14.05	0.92	15.3		[131]
- Body	0.4	15.2	0.91	16.7		
- Tail	0.33	15.2	0.87	17.5		
Brain - Grey Matter	0.49	1.5	0.5	3.0	100	[132]
- White Matter	0.7	1.2	0.1	12.0		
- Thalamus	0.72	1.3	0.3	4.3		
- Putamen	0.65	1.1	0.3	3.7		
Brain (rat, thalamus)	0.7	1.12	0.33	3.4	> 85	[133]
Prostate - Central gland	0.73	2.68	0.44	6.1		[134]
- Peripheral Zone	0.73	2.52	0.23	11.0		
Brain - White Matter, MS, Lesion	0.88	14.3	0.76	18.8		[135]
Liver - Normal	0.32	39.6	1.17	33.8		[136]
- Cirrhotic Tissue	0.25	27.9	1.04	26.8		
Kidney - Lesions	0.16	N/A	1.4	N/A		[137]
Prostate - Central Gland	0.18	10.9	1.3	8.4		[138]
- Peripheral Zone	0.23	21.2	1.3	16.3		
- Rectal Wall	0.24	31.3	1.1	28.5		
- Tumor	0.15	25.2	0.82	30.7		
Prostate - Peripheral Zone	0.7	2.9	0.7	4.1		[139]
- Transition Zone	0.6	2.9	0.7	4.1		
- Cancer	0.5	1.7	0.3	5.7		
Breast - Lesion	0.1	15.1	1.15	13.1		[88]
Breast - Lesion	0.16	98.2	0.7	140.3		[140]
Breast - Benign Cyst	0.72	2.12	0.19	11.2		[141]
- Malignant Lesion	0.67	2.1	0.18	11.7		

Kidney - Review paper	0.35	23.9	1.6	14.9		[142]
Larynx - Normal	0.33	N/A	1.6	N/A		[143]
- Cancer	0.11	N/A	1.13	N/A		
Kidney	0.44	12.7	1.4	9.1		[65]
Liver - Normal/Fibrosis 1	0.31	59.7	1.11	53.8		[144]
- Fibrosis 2-3	0.25	41.8	1.1	38.0		
- Fibrosis 4	0.25	32.27	0.98	32.9		
Lung - Squamous Cell Carcinoma	0.26	15	1.02	14.7		[145]
- Adenocarcinoma	0.23	15.25	1.12	13.6		
- Malignancy	0.23	14.14	1.09	13.0		
- Benign Lesion	0.26	12.8	1.34	9.6		
Kidney - Cortex	0.31	14.2	1.8	7.9		[146]
- Medulla	0.34	11.3	1.5	7.5		
- Kidney	0.32	18.2	1.7	10.7		
- Cyst	0.22	8.9	1.9	4.7		

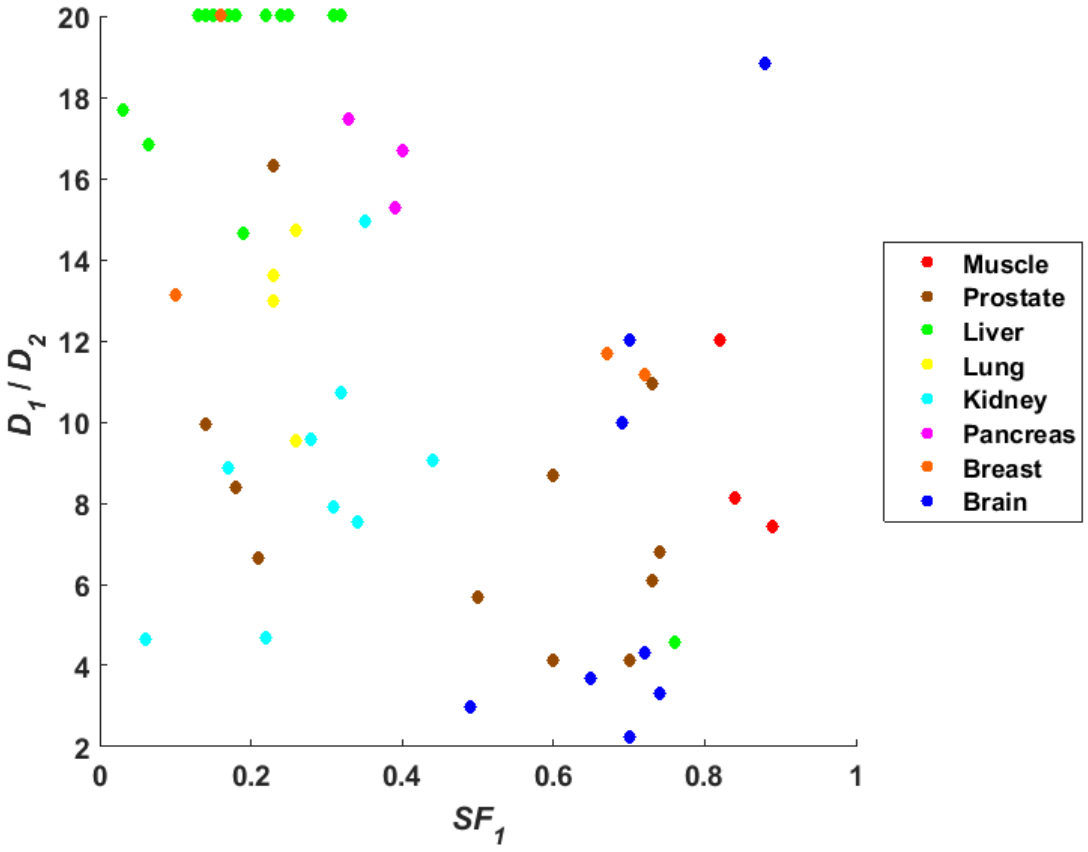

Figure 6 – Reported parameter values for biexponential studies in the DWI literature

SF_1 and D_1/D_2 for the studies in Table 1. D_1/D_2 ratios higher than 20 were set to 20.

The parameter estimates in Table 1 can be more easily visualised as a scatter plot, as shown in Figure 6, using the signal fraction component, SF_1 as the x-axis and the decay component ratio, D_1/D_2 as the y-axis. All measurements were categorised into different tissue types, and each data point has a colour corresponding with its tissue type shown in the legend. Additionally, all decay ratios higher than 20 were set to 20 to keep the graph scaled to a smaller area for ease of viewing. Figure 6 shows that the reported values of SF_1 vary widely from 0 to 1 and the D_1/D_2 ratio from 2 to 20. The results in the highly perfused tissues, such as liver, generally show a much higher decay ratio, but there are a few brain studies that reported a mean D_1/D_2 ratio between 2 and 4. The only SNR value reported for one of these low decay ratio studies was 100, which is considerably lower than the minimum resolution SNR reported in Istratov and Vyvenko of 1000, so it is possible that some of these studies could contain ill-conditioned fits. These studies reported parameter estimates by combining the results from all voxels in a ROI and reporting a mean and standard deviation. The ROI parameter distributions are usually compared categorically (e.g. normal tissue vs. cancerous) for statistical significance using a statistical test that may assume these distributions are normal (e.g. *t*-test or ANOVA). Reporting parameter estimates this way simplifies the process of statistical inference and reduces sampling error by including more measurements, but masks whether there are a significant number of outliers or whether these estimates are actually normally distributed.

Some recent studies in Table 1 have actually gone further and included error estimates of the model parameters themselves. Andreou et al [114] reported very large errors in some parameter estimates, including an upper bound of a 95% confidence interval (CI) for D_1 (perfusion) of 2,120% over the nominal estimate and an upper bound of 240% for an SF_1 CI. Bailey et al [115] added a 68% CI to measurements based on a χ^2 goodness-of-fit test, and Dyvorne et al [121] reported large errors in the perfusion decay component based on inter-scan reproducibility error tests. Cho et al [119] reported parameter standard errors with in vivo patient estimates, and also included simulated data where the bias from the true parameter values could be assessed. When discussing these large parameter errors, the authors have associated them with noise in the data, tissue heterogeneity, or other random effects, so isolating specific sources of these large estimated values should be another goal of a simulation study.

2.1.2 Conditioning, Collinearity, and Correlation

Linear least squares regression problems have an exact solution that can usually be obtained for a linear regression based on matrix operations per Equation 21. These matrices can be used to calculate the *condition number* of a numerical problem, which measures the magnitude of changes in the solution related to the magnitude of changes in the data [147]. If small changes in the data result in similar changes in the solution, then the system is well-conditioned, but if small changes in the data result in much larger changes, then the system is *ill-conditioned*. Parameter estimates from an ill-conditioned system can have large errors due to computational rounding problems, numerical precision issues, and/or instability in the mathematical results. In a linear regression, the presence of ill-conditioning is often used as a diagnostic for *collinearity* (sometimes referred to as multicollinearity), and is often seen where there are two or more independent variable components that have nearly the same slope in a linear model [148]. As these components become more nearly

collinear, the algorithm can no longer identify them as separate as $(\boldsymbol{X}^T\boldsymbol{X})$ approaches singularity, and when this matrix is inverted, its elements can have extremely large or small values [149].

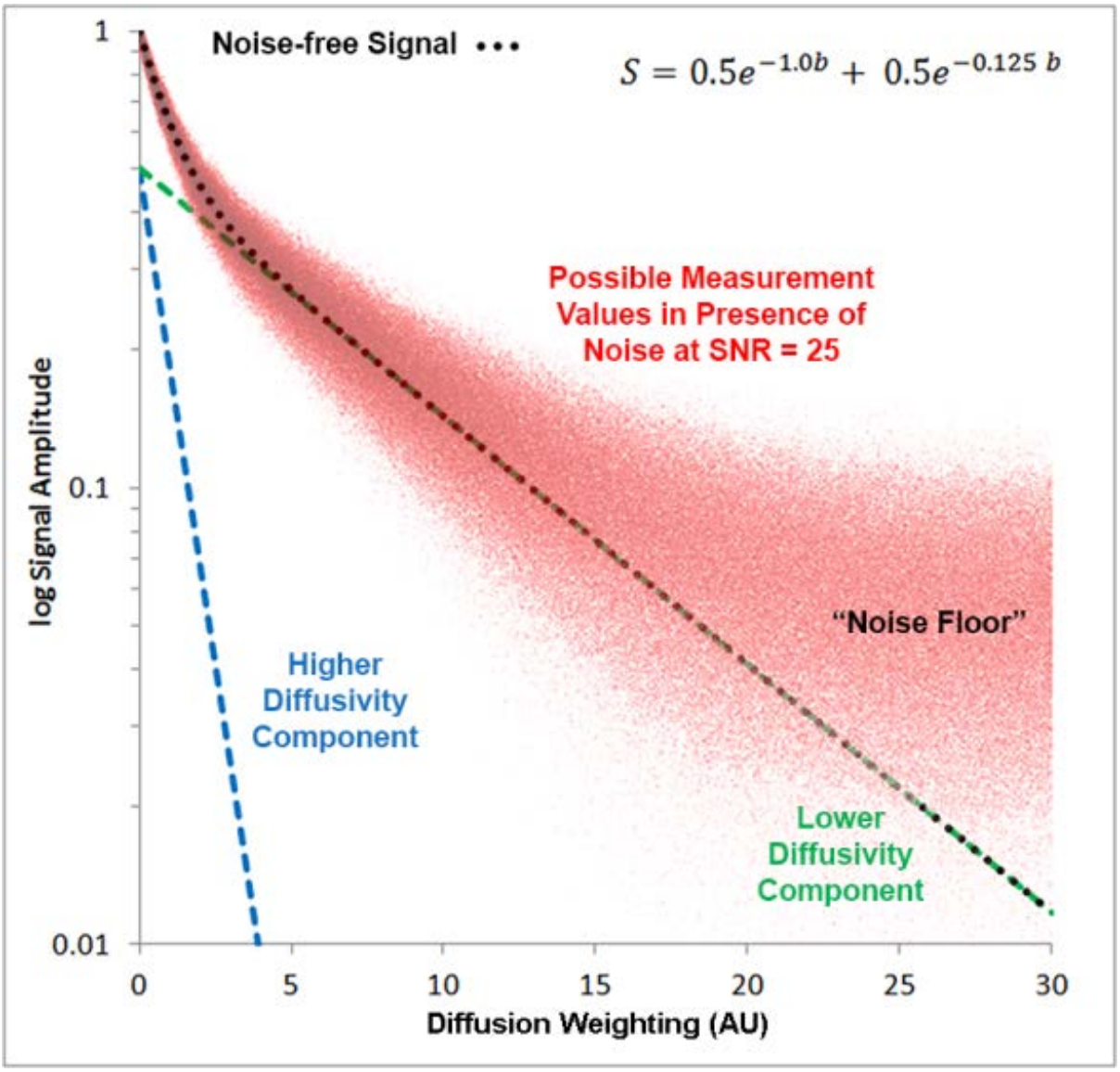

Figure 7 – Measurement of biexponential signal decay in noisy magnitude MR images

In this example, the inherent "noise-free" signal (black dots) is composed of two signals of equal signal fractions that sum to 1 with decay constants D_1 and D_2 differing by a factor of 8. The red "cloud" represents the possible magnitude image voxel values obtained by a measurement at SNR = 25.

With most DWI measurements, there is usually just one independent variable, the *b*-value, so there aren't any collinearity issues with the data in this case. However, there can be highly collinear components *in the regression model itself* [150], and as was discussed in Section 1.3.2, most NLLS algorithms operate by solving a local linear approximation of the function based on the local gradient values of the function found in the Jacobian matrix . In this approximation equation, the

Jacobian appears just like the matrix $\boldsymbol{X}$ does for linear equations, as the transpose of the Jacobian multiplied by itself ($\boldsymbol{J^T J}$) [151]. If there is significant collinearity between the regression model parameters, the Jacobian can also have near-singular values and become ill-conditioned, translating to increased parameter estimation errors. Due to its nonlinear nature with respect to the b-value and the decay parameters, collinearity in the biexponential model in Equation 32 is not evident. If a simulated biexponential signal with two diffusivity components is plotted logarithmically in the y-axis, however, the linear aspect of the decay components is now visible, as shown in Figure 7.

If the value of D_2 (the slow component) increases while D_1 is fixed at one, the individual components will become more and more close to parallel until perfectly collinear. Hence, if D_1 and D_2 have a lower ratio, there will be more collinearity between the decay components, and there will most likely be larger errors in the parameter estimates. When collinearity is present in a regression model, the usual effect is increased variance in one or more of the parameter estimates, often much greater than the variance in the data [152], although it is possible for parameter variance to also decrease [153]. Parameter estimates from NLLS regression often have some amount of bias present [154], but ill-conditioning in the regression model can significantly increase this bias [155]. An additional effect of collinearity is that occasionally the bias in the parameter estimates can increase while the variance of the parameter estimates *decreases* [156].

While it is possible for the biexponential model to have significant model collinearity, it is rarely mentioned or detected in nonlinear regression modelling of DWI data. It has been mentioned in a DTI study in reference to avoiding collinearity by ensuring that the three-dimensional gradient directions are not made in parallel [157]. For DWI nonlinear regression, however, one study stated that the biexponential model was robust and reported no problems with collinearity [135]. An additional known problem with collinearity, however, is that it is usually not detected by the usual model selection methods based on RSS values like AIC, goodness-of-fit tests, F-tests, etc [158]. It also may not show up, or be severely underestimated, when examining the distribution of the fit residuals [159]. For example, if the true decay rates in a signal being fitted by a biexponential model were identical, the contribution to the model error or RSS value would be the same as a single exponential decay model with one of the biexponential decay terms essentially undetected. Thus, while the measured data itself may appear to be well-fitted by such a model, the parameter estimate calculations are often unstable, and the parameter values can be unreliable and have little relation to what the model represents [160].

Parameter Identifiability

Perfectly collinear decay parameters in data being fit by a model are no longer *identifiable* [52]. An explanation of this statistical concept can be shown using the biexponential model equation with individual amplitude parameters

$$S_b = A_1 \exp(-bD_1) + A_2 \exp(-bD_2)\,. \tag{32}$$

Assume a signal has total signal amplitude at b = 0 of 1 and A_1 and A_2 are both non-zero. If the decay components of the signal, D_1 and D_2, are effectively equal, then $D_1 \approx D_2 \approx D$ and Equation 32 reduces to

$$S_b = (A_1 + A_2)\exp(-bD)\,. \tag{33}$$

Equation 33 is then effectively a monoexponential decay equation with two amplitude components and the amplitude portion of the equation is effectively solving $A_1 + A_2 = 1$, which is an *ill-posed* problem as it has an infinite number of solutions for A_1 and A_2. Likewise, if one of the amplitude components is close to zero, Equation 32 reduces to a single exponential decay equation, but the regression algorithm still attempts to estimate two parameters from a non-existent decay component. These identifiability problems are also referred to as the model having *redundant* parameters [161], with the parameters attempting to resolve signal components that aren't there. Identifiability can be a significant problem in estimation of the parameter estimates for a model, since a non-identifiable model means that there can be many parameter values that equally fit the data [162]. When this happens, there is no longer a single maximum likelihood estimate since there are multiple parameter combinations that are equally likely, so the model is *inconsistent* and repeated measurements will not converge to the true value and/or may not converge at all.

These parameter identification scenarios reflect the warning in Sigmund et al [88] that ill-conditioning can be present in signals with similar diffusivities and/or low perfusion fraction, but these signal fractions or diffusivities are not known a priori and are actually the parameters of interest. In these worst-case scenarios, where the signal decay components are identical or one amplitude component is zero, while some regression algorithms may stop and report an error, most will still report a solution with parameter estimates. When measurement noise is added to these problematic true signals, noise may affect the signal such that the algorithm doesn't fail completely, but there is severe ill-conditioning present that causes unstable parameter estimates. Hence, detection of ill-conditioning in the biexponential model requires expansion beyond the standard fare of diagnostic tools for NLLS regression. Seber and Wild [150] note these "problems of approximate non-identifiability, correlated estimates, and poor precision of estimation" in nonlinear models, and they refer to the combination of these problems as the result of "ill-conditioning", a convention that will be followed in the rest of this thesis.

2.1.3 Regression Diagnostics

To determine if ill-conditioning is present in a given regression fit, the best measure to start with is the condition of the Jacobian matrix. The Jacobian is a gradient matrix that is used to find the minimum squared residual value of a given fit, and is returned by most standard NLLS regression algorithms. This Jacobian matrix will be of the dimensions, *Number of parameters x Number of* b-*values*, and to calculate the condition of this matrix, the singular value decomposition (SVD) of the Jacobian is taken, which returns the singular values for each parameter, and the matrix condition is then calculated by the ratio of the largest singular value divided by the smallest. A large condition number indicates that there may be more instability with the matrix, with a good summary of this issue given in [103]. While the references in the previous section demonstrate the usefulness of the condition number in diagnosing ill-conditioning in *linear* LS fitting, it's effectiveness in NLLS fitting is unclear.

Standard Error of Regression

Another measure used is the standard error of regression (SER). This formula is calculated by dividing the RSS value of a given fit by the degrees of freedom (number of diffusion weightings – number of parameters), and taking the square root,

$$SER = \sqrt{\frac{RSS}{n-p}}, \tag{34}$$

where n is the number of diffusion weighted measurements or b-values, and p the number of parameters being estimated in the tested model. The square of the SER is effectively a normalised value of the RSS, and the SER is then the estimate of the error in the regression problem, i.e., ε in Equation 18. The SER is used for the value of σ in the following measures.

Covariance and Correlation Matrices

Two other diagnostic measures that can possibly indicate problems with a regression fit are the covariance matrix of the parameter fits and the related correlation matrix, since a high degree of correlation between two parameters implies collinearity (though the converse is not true) [158]. The covariance matrix for the parameters can easily be calculated using the transpose of the Jacobian multiplied by itself ($\boldsymbol{J^T J}$) and the equation

$$\boldsymbol{C_v} = \sigma^2 \left(\boldsymbol{J^T J}\right)^{-1}, \tag{35}$$

where σ^2 is the estimated variance associated with the regression fit. The covariance matrix is a $n \times n$ square, symmetric matrix, where n is the number of parameters in the regression model, with the biexponential model having a 4x4 matrix. The diagonal elements in the matrix are the estimated variances for each parameter while the off-diagonal elements contain the covariance between two specific parameters respectively, allowing for comparison of variance at the parameter level, which can be compared to the overall variance in the regression fit. The covariance elements can be used to assess which combination of parameters cause the regression fit to vary together, and this can be also visualised by calculating the Pearson correlation coefficients among the parameters. This can be done by calculating the correlation matrix using the covariance matrix, giving a normalized correlation value between two parameters from -1 to 1, where the correlation equals the covariance between the two parameters divided by their estimated individual standard deviations. The equation to calculate the correlation matrix is

$$\boldsymbol{C_r} = \left(diag(\boldsymbol{C_v})\right)^{-0.5} \cdot \boldsymbol{C_v} \cdot \left(diag(\boldsymbol{C_v})\right)^{-0.5}, \tag{36}$$

where $diag(\boldsymbol{C_v})$ means the diagonal elements of the covariance matrix. The correlation coefficients indicate where any two parameters are highly correlated, implying a high degree of collinearity between those parameters.

Variance Inflation Factor (VIF)

An additional measure calculated from the correlation matrix that is the standard method to assess collinearity in linear regression in the parameter estimates is the *variance inflation factor* (VIF) [163]. The VIF for a given parameter is

$$VIF_k = \left(1 - R_k^2\right)^{-1}, \tag{37}$$

where R_k^2 is the coefficient of determination of the regression of the kth parameter on the other parameters. A simpler way to calculate the values of VIF is to use the diagonal values of the inverse of the correlation matrix,

$$VIF_p = diag_p\left(\boldsymbol{C_r}^{-1}\right), \tag{38}$$

where p is the parameter of interest. The VIF estimates the multiplication factor for the parameter variance due to collinearity against the variance with no collinearity present.

Parameter Standard Deviation and Confidence Intervals

To assess the variance for the estimated parameters in a regression fit, the standard deviation of each parameter can be calculated which can then be used to give a confidence interval for each estimated parameter value. Reported in a few DWI studies earlier in this chapter, confidence intervals provide more information than point estimates for the parameters in one fit, giving researchers an estimated range that the parameter estimate would likely assume in future samples. A confidence interval is also useful to assess the true value of a parameter in a given model, but it does not always contain the true value within it. For example, 95% confidence intervals indicate that upon repeated noisy acquisitions of the same true signal, 95% of the confidence intervals will contain the true value. The standard deviation for each parameter estimate can be calculated by taking the square root of the individual parameter variance (diagonal) values in the covariance matrix.

$$\sigma_p = \sqrt{diag_p(\boldsymbol{C_v})}, \tag{39}$$

where p is the parameter of interest. To calculate a two-sided 95% confidence interval based on the t-distribution, the value of α is set to 0.05, the *degrees of freedom* equal to the number of measurements n minus the number of parameters p, and the multiplication factor $t_{(\alpha,n-p)}$ can be sampled from the PDF via a lookup table or calculated by a variety of computer algorithms. The interval is then calculated by

$$\begin{gathered} Pr(-C < \beta < C) = 0.95 \\ where\ C = \sigma_p \cdot t_{(\alpha,n-p)}, \end{gathered} \tag{40}$$

β the parameter value estimate, and σ_p the standard deviation for the parameter. Confidence intervals based on the t-distribution return estimates based on normally distributed errors, but as is often the case in nonlinear regression, errors in the parameter values are often not normally distributed [150]. t-distribution intervals also don't take into account that the parameter estimates

in a NLLS regression fit may be bounded, so the returned values don't reflect reality, i.e. a negative value for the decay rate or amplitude parameters in DWI data. While the confidence intervals may not be reliable in these cases, confidence intervals still provide an estimate of the variance in the parameter estimates across repeated samples.

Perturbation Analysis and Bootstrapping

A comprehensive review of collinearity and conditioning diagnostics, which discusses many of the problems with collinearity and ill-conditioning presented here, along with analysis of the regression diagnostic methods, can be found in [164]. This book also produced an effective diagnostic measure for ill-conditioning from a related paper by the same author [165] termed perturbation analysis. Perturbation analysis is a simple simulation test that adds a small amount of noise to the input signal and seeing whether the resultant noise in the parameter estimates is of a similar magnitude, essentially identical to the definition of ill-conditioning made earlier in this chapter. If the magnitude of the noise in the parameter estimates is much greater, then it can be said that a particular regression fit suffers from conditioning problems. This process is very much akin to the common statistical technique called bootstrapping, particularly the *parametric bootstrap*, which involves taking the data from a given fit and adding a small amount of noise to that data by resampling the residuals from that fit [166].

The parametric bootstrap takes the set of all residuals returned from a specific regression fit and randomly samples, with replacement, new sets of residuals, say 1000 of them. The parameter estimates are plugged into the model and the resulting signal is then added to each of the 1000 new sets of residuals, creating 1000 new data sets. A new regression fit is then performed on the new data sets, and the resulting parameter estimates from all sets can be combined into parameter distributions for examination of the distribution shape and variance. Because the standard deviation of the residuals if of a similar magnitude to the noise inherent in the signal, the parametric bootstrap is nearly identical to perturbation analysis, as it adds a small amount of noise to the fitted signal and produces parameter distributions that can be examined for errors. An added benefit of creating estimated parameter distributions is that confidence intervals can then be created on these parameter distributions via percentile calculations, giving the bootstrapped confidence intervals for each parameter. Because each bootstrap fit also can be bounded, these confidence intervals should better reflect the variance in each parameter estimate.

Graphical Analysis

The nonlinear regression literature (e.g. [103]) has many examples of nonlinear functions where an algorithm can get "trapped" in a local minimum and thus never find this global minimum. This is why multiple start points are often used with NLLS regression, and is also why exhaustive, global optimization methods like simulated annealing or genetic algorithms have been developed [103]. While an NLLS algorithm may return a solution, it doesn't know whether this minimum is global or the algorithm happened to get trapped in a local minimum, and hence multiple start points across the parameter space are used to make sure the global minimum is found. Many graphical methods have been developed to assist researchers in learning more about the results from regression algorithms [167]. Some of these methods can be used to create a picture about the regression

algorithm's results in the *neighbourhood* of a solution, providing an assessment of all nearby solutions where any values less than the current minima can be found. With this additional information, researchers can ascertain whether their solution is indeed the global minimum, and if there is a real need to switch to a more exhaustive algorithm.

In particular, Chapter 3 in Seber and Wild [150] introduces the graphical concept of *sum-of-squares contours* which shows how the RSS is distributed in the neighbourhood of a solution by creating a discrete array of points near the solution and iteratively comparing the RSS value at each point to the RSS value at the algorithm's solution. For example, if a NLLS regression solution is found for the monoexponential model with S_0 = 1 and ADC = 0.5, then a two-dimensional discrete grid of 121 points could be created with eleven values from 0.95 to 1.05 via 0.01 steps in the S_0 axis and eleven values from 0.45 to 0.55 via 0.01 steps in the ADC axis. If the RSS values from all points are plotted versus the parameter values as two-dimensional contours, they can illustrate the change in RSS like a topographic map. The minimum RSS value over this parameter space is akin to the bottom of a valley, topographically, and represents the minimum that the NLLS regression algorithm attempts to find. This information can also give a visual assessment of the confidence interval around a given parameter, or in the case of two-dimensional plots for two parameters, confidence *regions*, as termed by Bates and Watts. This method will be used in this chapter for a few NLLS regression solutions, to determine its potential to investigate minima in the biexponential model.

2.1.4 Rician Bias

In addition to the effects of ill-conditioning, the parameter estimates in the biexponential model can also be affected by Rician bias. In the presence of noise, the magnitude measurement used in biexponential NLLS regression may deviate significantly from the inherent underlying signal. There are clearly two problems with using a biexponential regression to the measured signal in an attempt to extract the decay constants and relative signal fractions of the underlying components. First, the rapid decay of the faster D_1 component means that it makes no significant contribution to the measurement during most of the observed signal decay. If the selection of b-values does not include multiple measurements that cover the faster D_1 decay, then its estimate may be imprecise and/or biased. The second problem is that the presence of Rician bias in the measurement leads to the appearance that the slow component (D_2) decays much more slowly than its actual signal, since the noise floor "lifts" the signal at high b-values. Even with a large number of measurements in the high b-value range the estimate of the component characteristics of D_2 will be biased by the noise and, asymptotically, the parameter will converge toward an inherently biased value.

Figure 8 illustrates this issue with results from biexponential fitting of a sample noisy measurement obtained from the system illustrated in Figure 7. In this example the fitted biexponential has the equation

$$S_b = 0.62\exp(-0.74b) + 0.38\exp(-0.09b)\,. \tag{41}$$

The signal fraction of the fast diffusivity component is 24% overestimated, the fast diffusivity component value is 26% underestimated, and the slow diffusivity component is 28% underestimated, showing the lifting effect that the Rician bias can have at higher b-values. In

addition to investigating the effects of ill-conditioning on biexponential parameter estimates, the effects of Rician bias on DWI data must be incorporated into this simulation study.

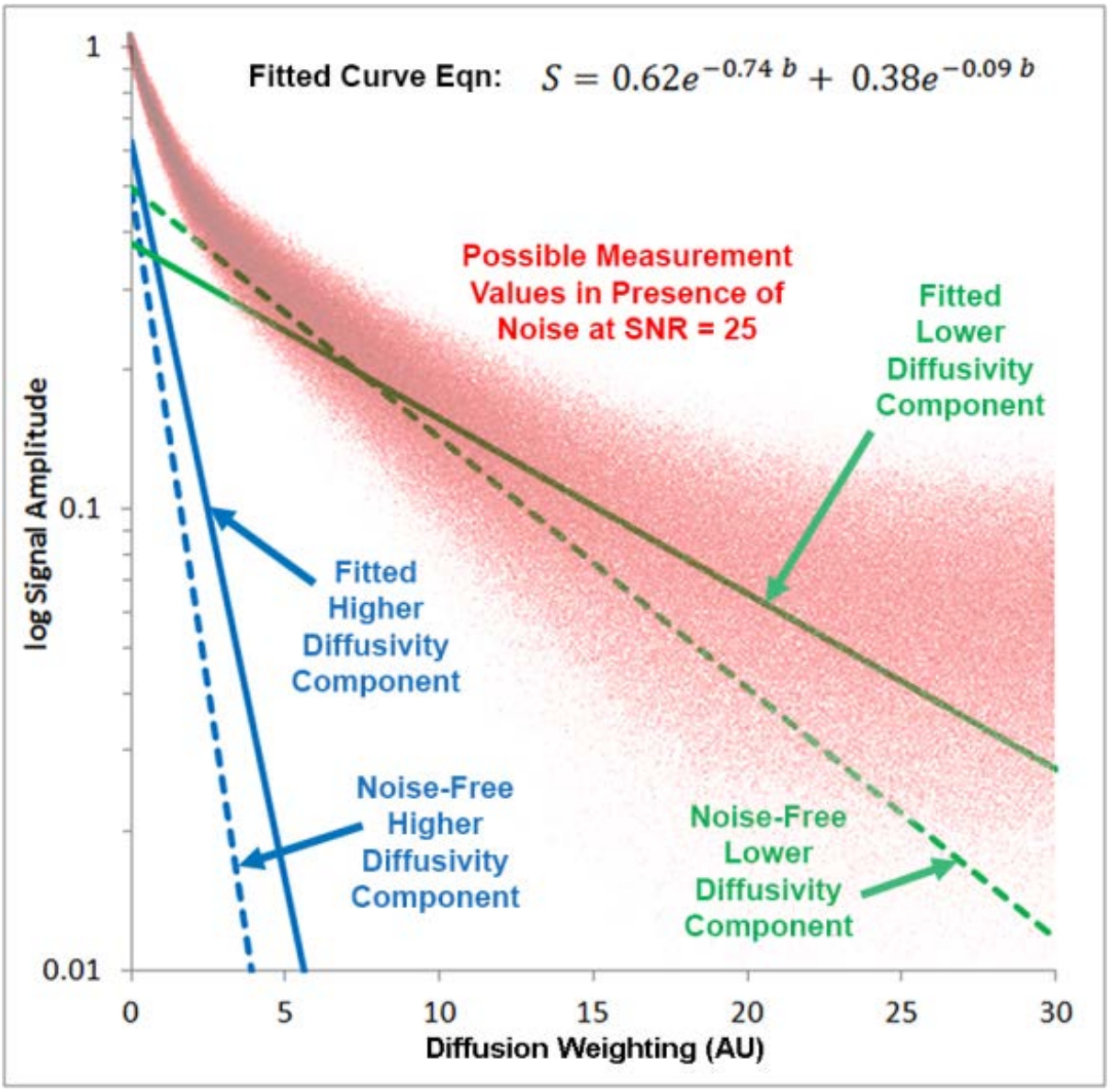

Figure 8 – Components of a biexponential regression fit to the true signal in Figure 7 plus noise

The higher and lower diffusivity components from a regression fit are plotted (solid lines) along with their true values (dashed lines). In this example the noise floor has strongly biased the estimate of the lower diffusivity component (green), but there are also significant errors in the estimate of the higher diffusivity component and relative signal fractions.

2.1.5 Chapter Aims

The concept of ill-conditioning presented in this chapter, along with the presented references that indicate that the biexponential model can be affected by it, demonstrate a need for a detailed study to determine what its actual effects on parameter estimates are when assessing DWI data in NLLS regression. By studying simulated DWI data, the magnitude of the effects of noise on the bias and variance in the parameter estimates can be determined. To simulate DWI data, artificial measurement noise can be repeatedly sampled and added to a given noise-free signal, creating hundreds of noisy sample measurements for each noise-free signal. If the parameter estimates from all noisy signal fits are combined, a detailed distribution can be created from which the bias and variance of the estimates of each parameter can be determined for each noise-free signal. Multiple noise-free signals also need to be created to determine the effects on the parameter estimates when the true signal differs. Since the monoexponential model is well established in clinical use, the monoexponential model will also be compared on the same noisy signals to compare the variance in its estimates.

The aims of this chapter were to:

- Determine the bias and variance in the biexponential model parameter estimates by combining the fits of repeated noisy samples from each noise-free signal.
- Examine how this bias and variance changes as the true parameter values of the noise-free signals vary over the parameter space in Figure 6.
- Examine the effects on the parameter estimates as the measurement SNR changes by adjusting the magnitude of the simulated noise added to the noise-free signals.
- Compare the effects of noise on monoexponential model parameter estimates to assess any differences to biexponential model estimates.
- Assess the effects of Rician bias on biexponential model estimates.
- Assess whether the diagnostic measures based on the Jacobian matrix from the NLLS regression fits (condition number, covariance matrix, correlation matrix, VIF, parameter standard error) can indicate when ill-conditioning is present in a given fit.
- Determine what the effects of ill-conditioning have on confidence intervals derived from the NLLS Jacobian matrix.
- Investigate the results from the parametric bootstrap perturbation analysis and determine its effectiveness in detecting ill-conditioning and large variance in the parameter estimates.
- Assess if there are any possible solutions that can help remedy ill-conditioning when using the biexponential model.

2.2 Methods

Simulated data was created using the computer language MATLAB (Mathworks, Natick, MA, USA) with Equation 12 as the basis for signal creation. Since this analysis is strictly mathematical, the values for the diffusivity components were normalised, with the value of D_1 for all experiments set to 1. The total signal amplitude, S_0, for all experiments was also normalised to 1. To obtain the values for the other two parameters, 50,000 random values were uniformly sampled from a range

of 2 – 20 for the ratio of D_1/D_2, with an additional 50,000 uniform random samples from a range of 0 – 1 for SF_1, matching the bounds of the parameter space in Figure 6. The parameter combinations were then used to create 50,000 noise-free signals, at eleven simulated b-values or diffusion weightings (0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6), with the term "diffusion weighting" emphasised to distinguish that the units were arbitrary and not related to actual scanner acquisitions. These weightings were chosen to represent a typical or feasible acquisition strategy, with enough spacing and sampling to cover the rapid drop off at the maximum decay value (100% D_1) while keeping a few weightings for the minimum decay value (100% D_2), with a graph of the weightings and the decay curves on a logarithmic scale shown in Figure 9.

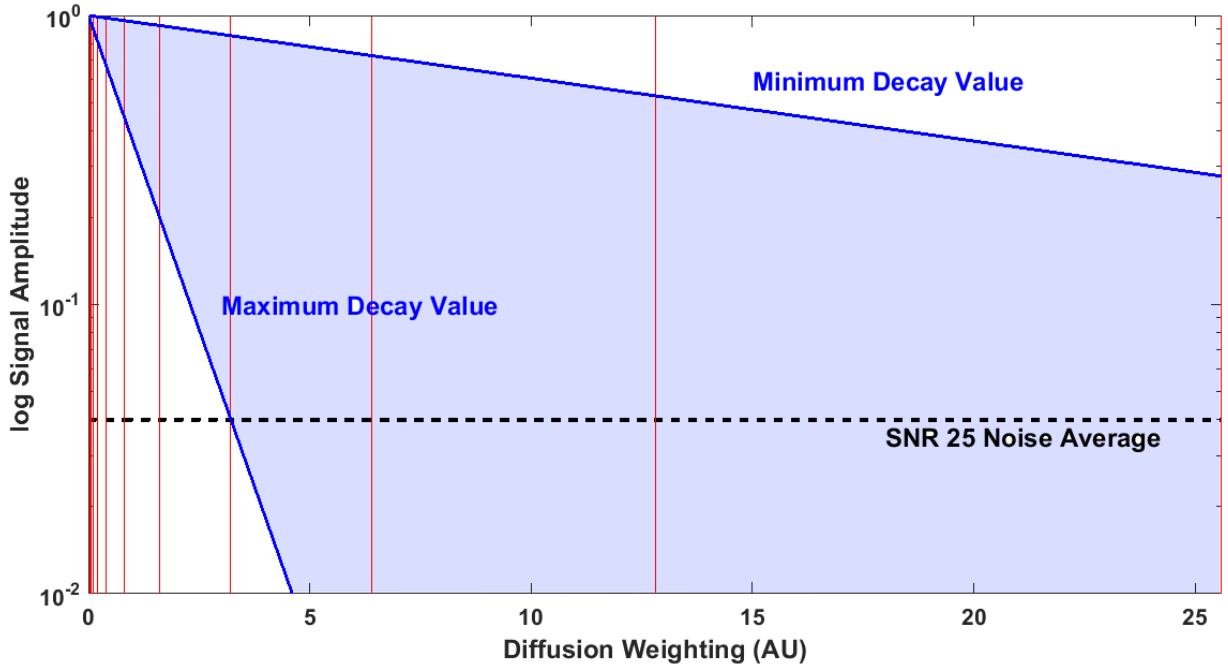

Figure 9 – Range of test decay signals and diffusion weightings

Blue lines indicate the maximum (100% D_1 component at a decay rate of 1) and minimum (100% D_2 component at a decay rate of 0.05) decay signals in the artificial data set, with the blue shaded area indicating the range of simulated decays. Red lines indicate the 11 diffusion weightings (arbitrary units) for the measurement protocol. The SD of the noise level at an SNR of 25 is indicated by the black dashed line, with the SNR 100 level equal to the x-axis at 0.01 and the SNR 200 level (0.005) not visible.

Noise associated with a magnitude measurement was added at three SNR values of 25, 100, and 200, approximately covering the range from the studies in Table 1. For each noise-free signal, 200 noisy measurements were created, giving a total test set of ten million measurements at each SNR. Each simulated noisy measurement was created using Equation 26, with the two noise components individually sampled randomly from a normal distribution with a standard deviation equal to the inverse of the SNR value, since the noise-free signal amplitude is always 1. To reduce the effect of

Rician bias on each measurement, the bias reduction formula in Equation 24 was applied to all noisy signals before fitting, using the known standard deviation from the SNR used to create the data, plus a small random perturbation of 10% since its exact value would not be exactly obtained in a physical DWI measurement.

2.2.1 Biexponential Model Regression Fitting

Fitting the biexponential model to the simulated noisy data was performed using a NLLS algorithm (*lsqcurvefit* in MATLAB) with a trust-region-reflective optimization option. The individual amplitude model in Equation 32 was used as the regression model, the individual amplitudes were bound in the algorithm to a range between 0 and the maximum amplitude in each signal measurement, and the individual decay values were bound to a range between 0 and 4. This allowed for a positive overhead (from the true values) on the parameter estimates, but kept them from going negative. Because NLLS regression can sometimes have local minima that can trap the regression algorithms, for each signal measurement, five separate regression fits were performed using random start values within these parameter bounds. The regression fit with the minimum RSS value was kept and the others discarded. From this minimum fit of the five, all parameter estimates and residuals (including the RSS value) were saved for analysis. The parameter estimates were then compared to the known true parameter values from their respective noise-free signals, with SF_1 estimates calculated by the equation $SF_1 = A_1 / (A_1 + A_2)$.

2.2.2 Monoexponential Model Regression Fitting

For comparison, the monoexponential model in Equation 9 was also fit to all noisy signals, with S_0 bound in the NLLS algorithm between 0 and 2 times maximum amplitude in each measurement, and ADC between 0 and 4. Since the monoexponential model is assessing biexponential signals, only the variance in the parameter estimates was assessed, since there is no true parameter value to be compared.

2.2.3 Rician Bias and Low SNR Rejection Strategy

The different decay rates for the signals in Figure 9 were also chosen to assess how Rician bias affects the model parameter estimates. For example, signals at the Minimum Decay Value shown there stay well above the SNR 25 Noise Average level for all diffusion weightings. However, signals at the Maximum Decay Value reach the noise level at the seventh diffusion weighting with three more measurements sampled where the signal is well below it. Thus, the signal lifting effects of the noise floor, as seen in Figure 4 will definitely be present. Common practice in the literature is to limit acquisitions to lower b-values (around 800 s/mm^2) so that all measurements remain well above the noise floor (e.g. [168]). Since many of the noise-free biexponential signals tested here had at least one diffusion weighted measurement at or below the noise floor, a selective fitting strategy was also applied to each magnitude measurement to remove all higher weighted measurements with low SNR. The strategy followed was similar to the protocol in [109] on each of ten million noisy measurements in the test set, i.e. for each measurement, the first diffusion weighted data point with an SNR < 2 was removed along with all measurements from higher

diffusion weightings. The effects of employing this measurement strategy were compared to the full eleven diffusion weighting strategy for both the biexponential and monoexponential models.

2.2.4 Regression Diagnostics

From each regression fit, the final Jacobian matrix returned by the algorithm was also saved and used to calculate the diagnostic measures from Section 2.1.3, including the Jacobian condition number, standard error of regression, covariance matrix, correlation matrix, parameter standard errors, parameter VIF, and parameter confidence intervals based on *t*-distribution estimates. MATLAB code to construct these additional diagnostics was largely derived from the freely available, variable-projection, NLLS algorithm code in [104].

2.2.5 Bootstrap Analysis

Confidence intervals were also created using the parametric bootstrap from 1000 bootstrap samples for each regression fit. However, this was only done for a very limited subset of the noisy signals since the regression fits for the 10 million noisy signals took 48 hours of computing time to perform, even with distributed code using 12 parallel processors. Additionally, sampling the residuals for the parametric bootstrap resulted in the addition of normally distributed noise to the fitted data results, but since the artificial data was created using Rician distributed noise, the residual values were also added to the signal using Equation 26 to create a magnitude measurement. Otherwise, it was possible for there to have been negative signal data at high diffusion weightings, which didn't represent a realistic acquisition.

2.2.6 Graphical Analysis

To assist further in determining the effects of ill-conditioning and collinearity on the problem of NLLS regression fitting with the biexponential model, sum-of-squares contours were examined for two noise-free signals sampled from the test set in Section 2.2. One signal had a D_1/D_2 ratio equal to 20, and the other a ratio equal to 2, with $A_1 = A_2 = SF_1 = 0.5$ and $D_1 = 1$ for both signals. 200 discrete values were chosen for SF_1 equally distributed between 0 and 1, and 200 discrete values were chosen for D_1 equally distributed between 0 and 4. Using the true value of D_2 for each grid point signal (0.05 and 0.5, respectively), along with an amplitude value $S_0 = 1$, Equation 12 was used to calculate the signal for the 40,000 discrete SF_1, D_1 combinations with the same eleven diffusion weightings. For all grid points, the eleven residual values were calculated as the difference between that grid point signal and the test signal, and then squared and summed to determine the RSS value. These RSS values at each of the discrete points were then plotted as a two-dimensional contour map. Additionally, to determine the effects when NLLS regression fitting a biexponential model with perfect collinearity between decay components, a monoexponential signal was created using the biexponential model equation, with $D_1 = D_2 = 1$, $SF_1 = 0.5$, and $S_0 = 1$. The RSS values were also plotted there as a two-dimensional contour map using the same discrete values.

2.3 Results and Discussion

2.3.1 Bias and Variance in Biexponential Model Parameter Estimates

All noise-free signals were sorted into an array of 100 bins based on true SF_1 by 100 bins based on true D_1/D_2 ratio. For all noisy data created from each noise-free signal in each bin, the error between the fitted parameter estimates and the known true parameter values were grouped as a distribution. These errors for SF_1, D_1, and D_2 are each shown as a two-dimensional pseudocolour plot for each SNR in Figure 10, with the mean absolute error shown for the SF_1 estimates, and the mean percent error shown for the D_1 and D_2 estimates. Overlaid on each pseudocolour plot is a contour plot showing either the standard deviation for the SF_1 estimates or the coefficient of variation (CV) for the D_1 and D_2 estimates. The variance values in the contour plots were smoothed with a 3x3 averaging filter for display. Due to the random uniform distribution of the true parameter values, several bins ended up with no noise-free signals in them, which are shown as the white pixels scattered throughout the pseudocolour plots.

Figure 10 shows that the bias and variance of the parameter estimates do vary depending on the true parameter values that created the signal. All plots show an area where the bias and variance of the parameter estimates is low, these areas are generally centred at intermediate values of SF_1 (near equal signal contribution from each component) and high ratios of D_1/D_2, and they enlarge as the SNR increases. Conversely, as the signal fraction of either component goes to zero, or the D_1/D_2 ratio goes towards the minimum tested ratio of 2, with the signal approaching monoexponential decay, the parameter estimate bias and variance increase rapidly and nonlinearly. While the area of parameter space with large bias and variance diminishes as SNR increases, many signals are affected even at an SNR of 200, a ratio that would be obtainable only in a long ex vivo tissue study, demonstrating that high SNR does *not* guarantee reliable parameter estimates using the biexponential model.

The SF_1 parameter estimates are positively biased for low true SF_1 values and negatively for high values. The D_1 parameter estimates are positively biased in areas of high uncertainty, except where the fast decay component is basically non-existent ($SF_1 \approx 0$). The estimates of D_2 are negatively biased for most of the true parameter space, except where its signal fraction is close to zero, where they abruptly shift to a large positive bias. At an SNR of 25, there is measurable variance in the D_1 and D_2 estimates across the entire parameter space, with the lowest contour value for the coefficient of variation equal to 0.3, and this increases sharply as the true signal becomes more monoexponential. The standard deviation of the SF_1 estimates are fairly low in the "reliable area", but at a true SF_1 value of 0.1 or less, the standard deviation of the SF_1 estimates is 0.35 or more, nearly ten times the standard deviation of the added signal noise (0.04). To investigate the nature of these errors in more detail, several noise-free signal samples across the parameter space were selected and histograms of the parameter estimates were created for all regression fits of each noise-free signal. The estimates of the 200 noisy signals from three separate noise-free signals, with three different SF_1 values of approximately 0.5, 0.25, and 0.1 and a D_1/D_2 ratio of approximately 15 for all, were selected to illustrate the effects of decreasing signal fraction, with the resulting distributions shown in Figure 11.

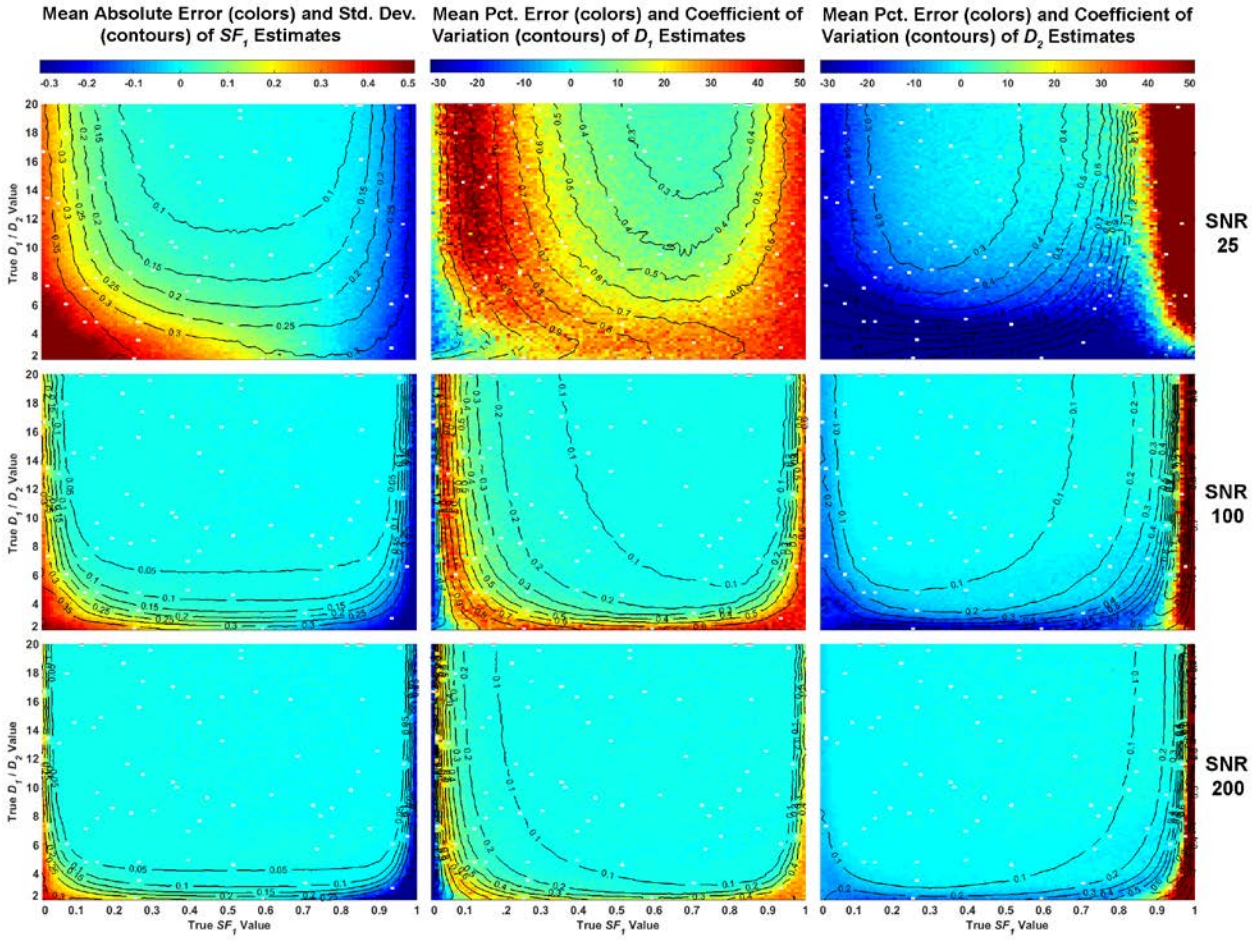

Figure 10 – Uncertainty in biexponential model parameter estimates for three SNR levels

Bias (pseudocolour plots) and variance (contours) of the parameter estimate distributions in each bin. White pixels on the plots indicate bins with no signals in them. Bias and variation in the parameter estimates increases rapidly when the true signal is closer to monoexponential, and this phenomenon still persists even when the simulated SNR is increased to 200.

At a true signal fraction of 0.5, the histograms of the amplitude estimates (A_1 & A_2) are centred on their true values of 0.5 and have a well-formed, approximately normal distribution. As the true signal fraction SF_1 decreases to 0.25 and then 0.1, the amplitude distributions are centred closely to their true values, however they both exhibit severe skewness (A_1 right-skewed, A_2 left-skewed) and there is also a small amount of bimodality in the SF_1 = 0.1 estimates. The D_1 estimates at SF_1 = 0.5 are generally centred on the true value of 1, however there are several estimates that are greater than 2. At SF_1 = 0.25, the majority of the D_1 estimates are centred around 0.8, but the distribution is definitely not Gaussian and there are many values found all the way up to 4, which was the estimate upper bound set on the regression fit. At SF_1 = 0.1, the D_1 estimates are widely dispersed over the range between 0 and 4, with the majority of the values found close to 0 and a considerable number of values found at the upper bound of 4. The D_2 estimates are not as widely dispersed as the D_1

estimates, and are generally centred on the true value of 0.067, other than several estimates found near zero for SF_1 = 0.1.

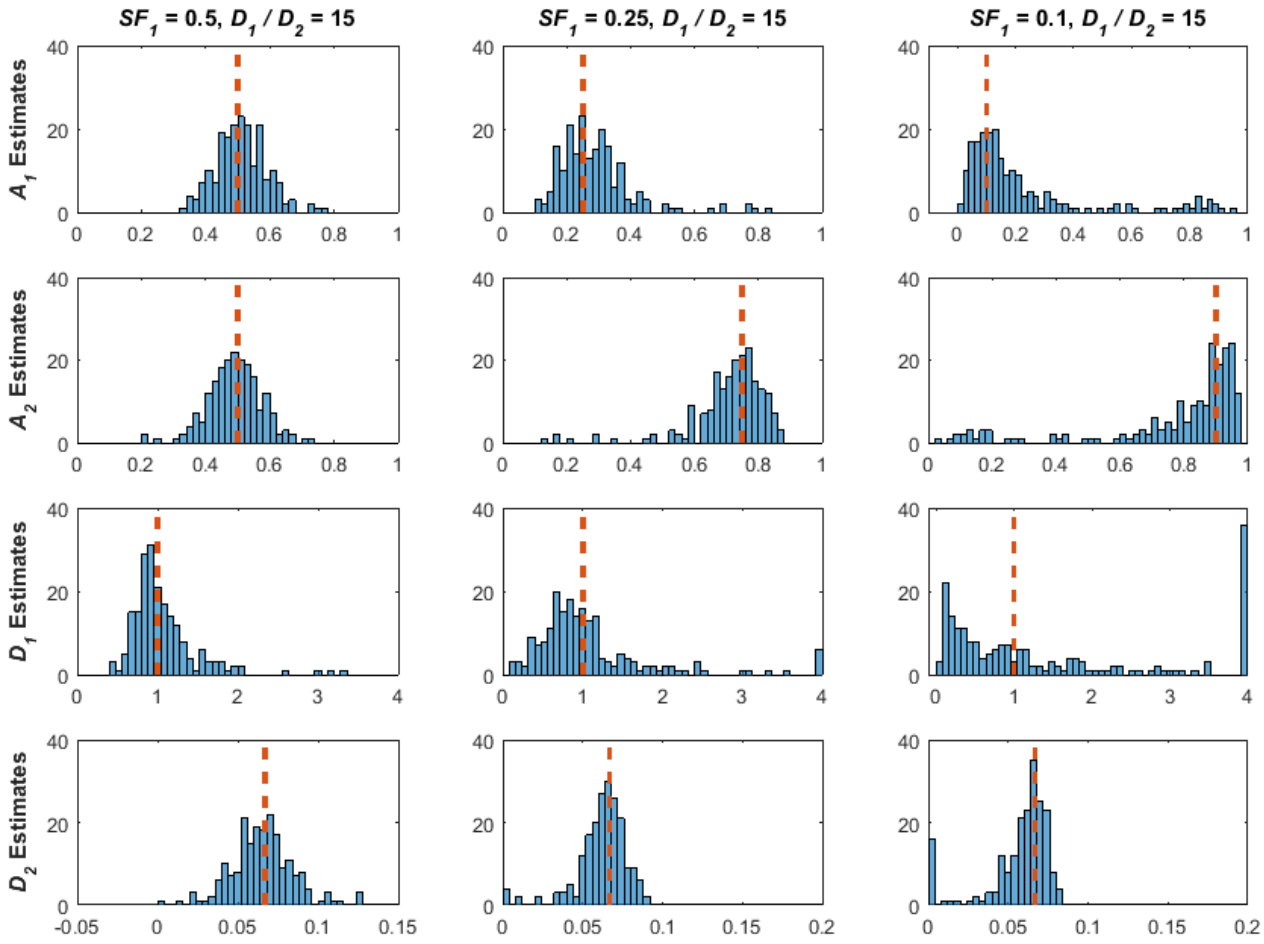

Figure 11 – Parameter estimate histograms for three different true SF_1 values

The individual estimates (rows) from 200 regression fits to noisy signals at an SNR of 25 are displayed for three different noise-free signals (columns). True D_1/D_2 ratio for all three signals is 15, true D_1 = 1, and true SF_1 is (L to R) 0.5, 0.25, and 0.1. The *x*-axis for each histogram is the individual parameter value, the *y*-axis is the bin count, and the dashed red line indicates the true parameter value. As SF_1 decreases, such that the true signal is closer to monoexponential, the bias and variance in the parameter estimates increase.

To illustrate the effects of a decreasing D_1/D_2 ratio, three noise-free signals were investigated, each with an SF_1 value of approximately 0.5 and three different D_1/D_2 ratio values of 8, 4, and 2 (true D_2 = 0.125, 0.25, and 0.5), with the resulting parameter estimate distributions from the noisy signal fits shown in Figure 12. The effects on the amplitude estimates are different when the true D_1/D_2 ratio is decreased. The true amplitude values should remain at 0.5 for all three ratios, however, as the ratio decreases, the amplitude dispersion increases, but the centre of the A_1 component shifts to the right and at a ratio of 2, the centre of the distribution is around 0.95 – giving a large positive bias

between the A_1 fit estimates and the true value. The large bias is also seen for the A_2 component, only the centre of the distribution is biased negatively to around 0.05.

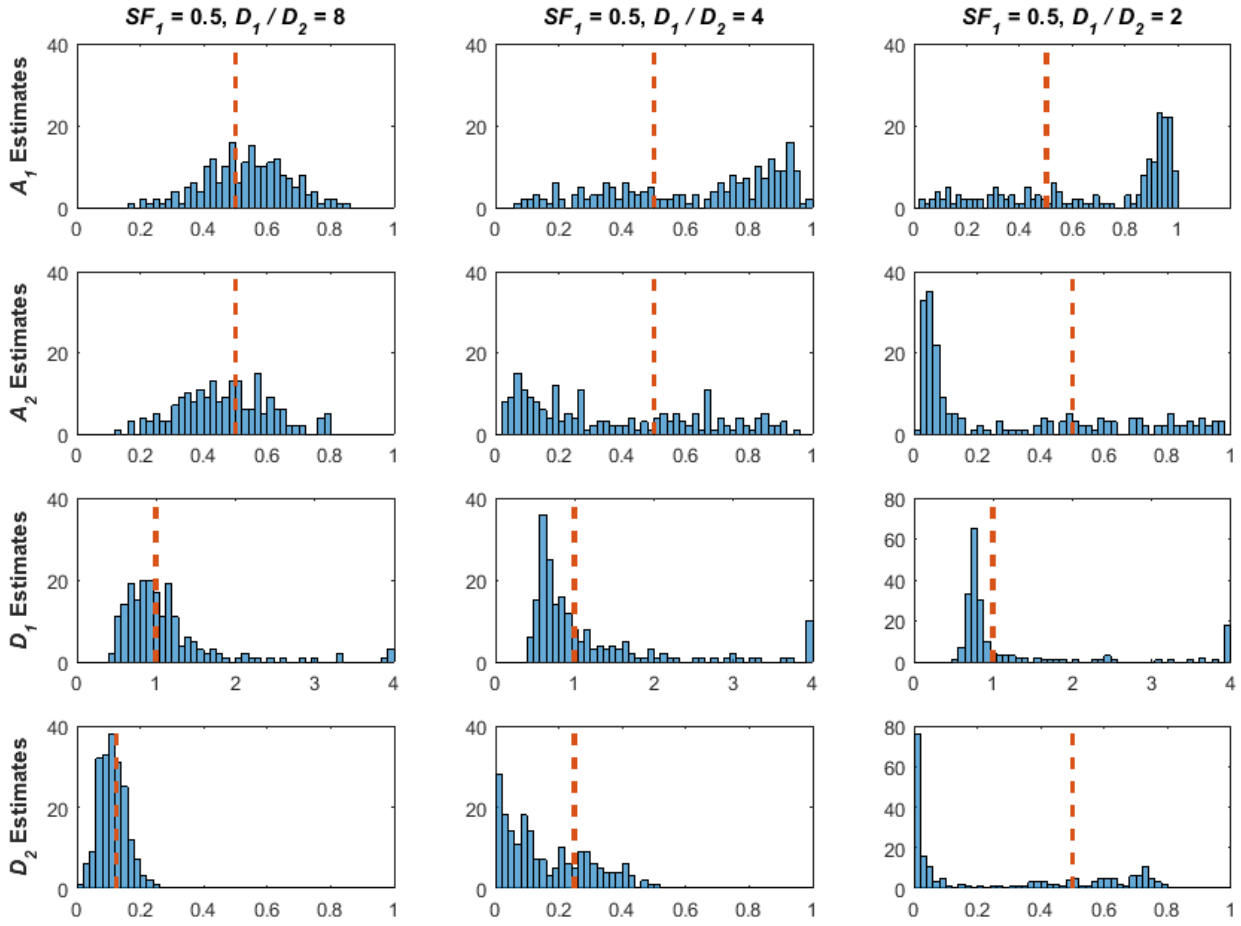

Figure 12 – Parameter estimate histograms for three different true D_1/D_2 ratios

The individual estimates (rows) from 200 regression fits to noisy signals at an SNR of 25 are displayed for three different noise-free signals (columns). True SF_1 value for all three signals is 0.5, true D_1 = 1, and true D_1/D_2 ratio is (L to R) 8, 4, and 2. The *x*-axis for each histogram is the individual parameter value, the *y*-axis is the bin count, and the dashed red line indicates the true parameter value. Again, as the true signal is closer to monoexponential (decreased D_1/D_2 ratio), the bias and variance in the parameter estimates increase.

The D_1 estimates are centred mostly on the true value of 1 for a D_1/D_2 ratio of 8, but the distribution is right-skewed and several estimates have values greater than 2. At a ratio of 4, the centre of the estimates is around 0.7 and the distribution is more peaked, and this peaked trend increases even more at a ratio of 2. The D_2 estimates are centred on the true value of 0.125, for a D_1/D_2 ratio of 8, but at a ratio of 4, the values are widely dispersed with several estimates located near zero. At a ratio of 2, most of the D_2 estimates are found at values less than 0.05 with many found at zero, which is a very large bias, since the true D_2 value at this ratio is 0.5. The effects of decreasing the

D_1/D_2 ratio on the parameter estimates seems to be more severe, since the minimum ratio of 2 not only shows increased dispersion for all parameters, but it also produces noticeable bias between all four parameter estimates and their true values. The distribution of the parameter estimates shown in both Figure 11 and Figure 12 illustrate that even with a limited sample of the possible true parameter values in the biexponential test set, there can be large bias between the parameter estimates and their true values, the variance of the estimates can also be very large, and these distributions are not normal or well-formed and can even be bimodal. In the biased cases, the estimators are no longer consistent, and adding more signal acquisitions to reduce sampling error still does not make the estimates converge to the true value.

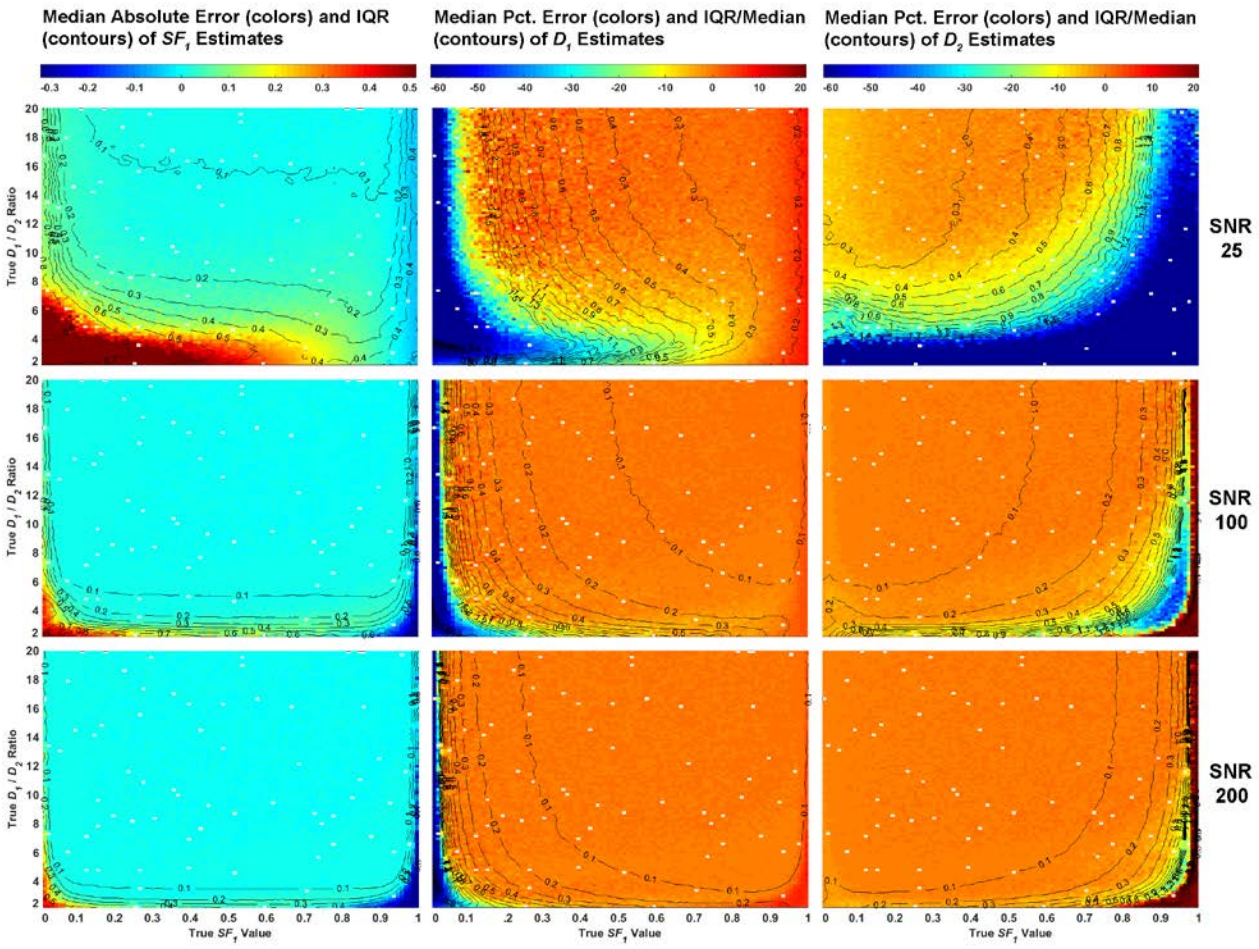

Figure 13 – Robust uncertainty measures for biexponential model parameter estimates at three SNR

Plots showing median error (colour) and IQR (contours) measures of the parameter estimates in each bin as described at the top of each column. Even when using these more robust measures, the same phenomenon of increased bias and variance when the signal is effectively monoexponential is still present.

Since the values for mean and standard deviation can be affected considerably by outlier values, the images of the errors in Figure 10 can be reproduced using statistical measures more robust to

extreme values. Instead, the central tendency of the errors can be measured using the absolute median (SF_1) or percentage median (D_1 & D_2) values, and the variance or deviation can be measured using the interquartile range (SF_1) or the interquartile range divided by the median (D_1 & D_2). The parameter estimate errors analysed with these robust statistics are shown in Figure 13. With normally distributed data, the mean equals the median, and the IQR should be equal to 1.34 times the standard deviation [148]. The robust measure plots show that the areas of high bias in the mean SF_1 parameter estimates at an SNR of 25 are greatly reduced and not as noticeable in the median errors, suggesting the presence of a significant amount of outlier estimates that caused a large bias in the mean. There is also a reduction of the areas of significant bias and deviation in both the D_1 and D_2 estimates, especially when the signal fraction of that specific component is close to 1. The differences between the set of parameter plots in Figure 10 and the robust parameter plots in Figure 13 suggest that there are large areas in the parameter space where the parameter estimate distributions are not normal and/or have a significant number of outlier values.

2.3.2 Variance in Monoexponential Model Parameter Estimates

The monoexponential parameter estimate errors for the same noisy biexponential signals at an SNR of 25 were also grouped by bins based on the same true parameter values of each noise-free signal. The standard deviation of the estimates for the amplitude, S_0, and decay component, *ADC*, from Equation 9 are displayed as pseudocolour plots, as shown in the top row in Figure 14. The letters A – E on the parameter colour plots match the labelled signals in the bottom plot. Figure 14 also shows that the standard deviations of both the amplitude and decay coefficient estimates varied based on the true biexponential parameter values, with a range of 0.014–0.027 for the S_0 estimates and 0.004–0.089 for the *ADC* estimates. These estimates also varied across the true parameter space, but the magnitude of the change in standard deviation for either parameter wasn't nearly as large as for the biexponential parameter estimates. For example, the standard deviation of the amplitude parameter S_0 in the top left corner of Figure 14 is 5–10 times less than the standard deviation of the SF_1 parameter shown via contour in the top left corner of Figure 10. The monoexponential *ADC* parameter was compared to the two individual decay components in the biexponential model by calculating the CV, standardising the variance in *ADC* by dividing the standard deviation (top right, Figure 14) by the mean *ADC* value for each measurement, as shown in Figure 15. The highest CV from Figure 15 is around 0.16, which is also considerably less than the highest CV values of 1.2 in the D_1 and D_2 estimates shown in the top row plots in Figure 10.

This was important, since it illustrated that the normalized variance penalty of incorrectly applying a model was 7.5 times higher when applying the biexponential model to a monoexponential signal versus the opposite application. The plot in Figure 15 also shows an inverse relation to the CV plots for the D_1 and D_2 estimates – where the CV for the biexponential parameter estimates was lowest when the D_1/D_2 ratio was highest with close to equal signal fraction for each component, the CV for the *ADC* is highest at these values. This was largely due to the goodness-of-fit that the monoexponential model had when fitting the simulated biexponential measurements. This is shown using the averaged SER in Figure 16, indicating that for effectively monoexponential signals, the model fit well, as expected, but as the D_1/D_2 ratio increased and the signal fraction between the

two components became equal, the monoexponential model had a worse fit, and so the bias between model and data increased.

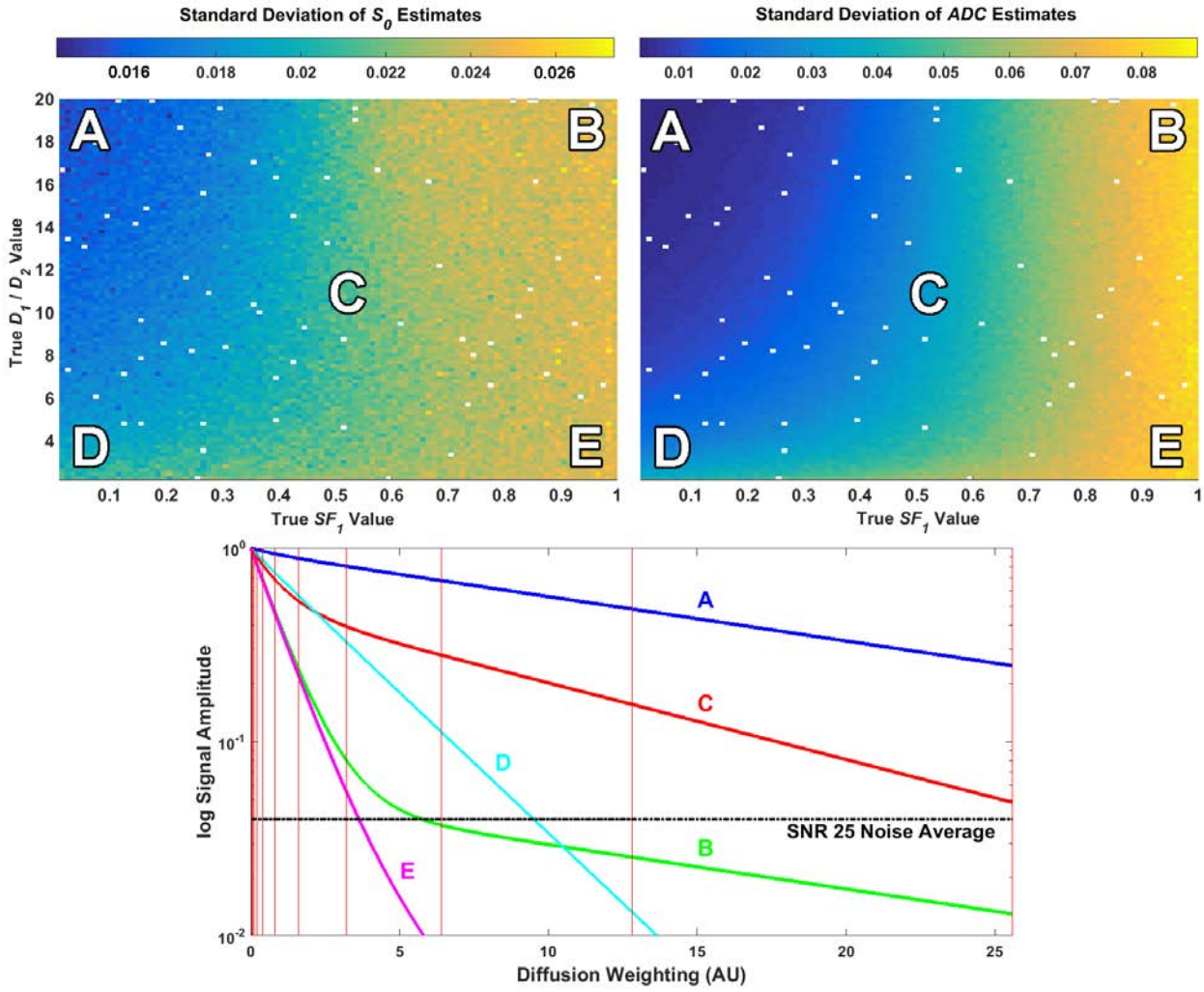

Figure 14 – Colour plots of errors in monoexponential estimates with five example biexponential signals

Colour plots display the standard deviation for the monoexponential amplitude (S_0) and decay (*ADC*) estimates based on the true biexponential parameter values. To illustrate the effects of noise across the parameter space, five example signals are identified (A-E) on the plots with their corresponding signals on the log-linear line plot (bottom). SNR for all noisy signals was 25. The SD of these two monoexponential parameter estimates increases for signals that decay into the noise floor more rapidly.

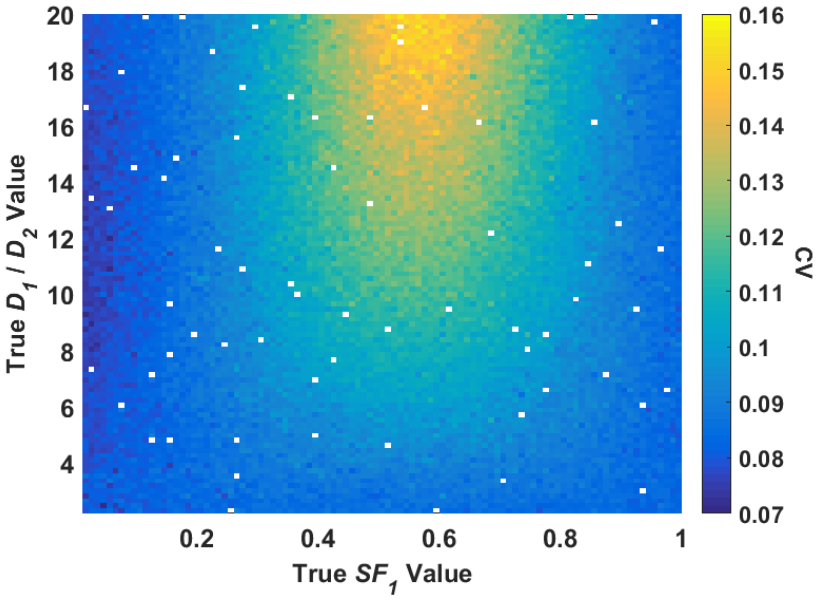

Figure 15 - Coefficient of Variation for monoexponential *ADC* parameter estimates from true biexponential signals (SNR = 25)

When examining the normalised variation of the *ADC* estimates, the CV is highest at signals with equal signal fraction and highest D_1/D_2 ratio and lowest when the signal is effectively monoexponential.

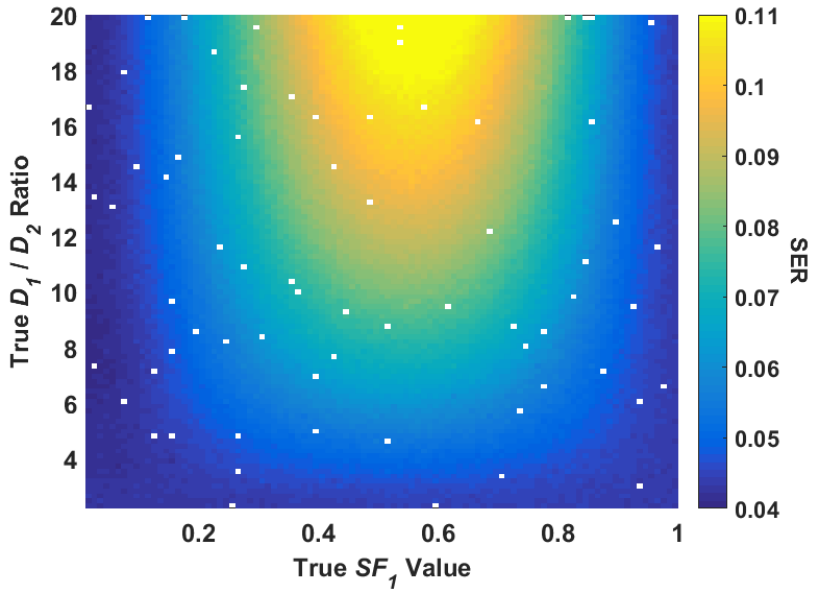

Figure 16 –Standard Error of Regression (SER) for monoexponential model fits to true biexponential signals averaged over all fits in each bin

This pattern in the SER correlates well with the CV pattern of the *ADC* estimates in Figure 15, suggesting a direct relationship between the monoexponential model's parameter estimates and how well this model fits the data.

Residuals in Biexponential Fitting

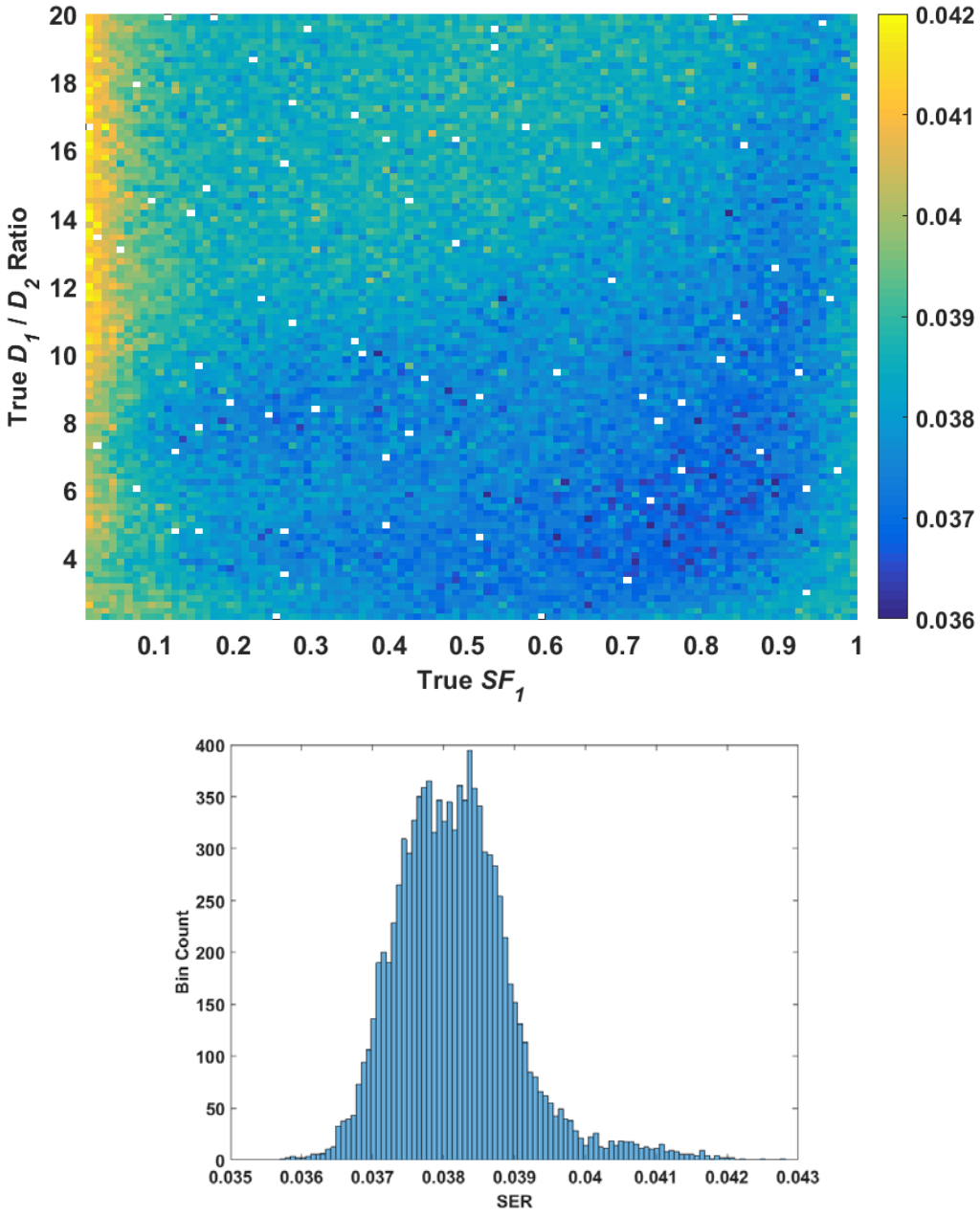

Figure 17 – Colour plot (top image) and histogram (bottom image) of the standard error of regression for all SNR 25 fits of the biexponential model to simulated biexponential data

The top colour plot displays the SER distribution in each bin whereas the bottom shows the bin SER distribution by value. Noisy signal SNR is 25. These SER values are all close to the simulated noise value of 0.04, indicating that the biexponential model fits all data fairly well.

The SER was also calculated for each of the biexponential fits and a colour plot of the mean SER for all regression fits in each bin is shown in Figure 17 along with a histogram of the values of all

regression fits for the test set. Compared to the SER of the monoexponential model fits in Figure 16 that varied between 0.04 and 0.11, depending on the parameter values of the biexponential signal, the SER for the biexponential fits was found between 0.036 and 0.042, a much tighter range. This range of values is unsurprising, since the fitting model was the same as the true signal, and the added noise was equal to 0.04, but there were slight deviations across the entire parameter space. More importantly, however, these values illustrate one of the major issues with collinearity and correlation discussed in Section 2.1.2, namely, most standard methods of measuring regression fit based on the residuals do *not* detect large uncertainty in parameter estimates.

As discussed in Section 1.3.1, an assumption for least squares regression to return maximum likelihood parameter estimates is that the errors are normally distributed, which were examined using the residual values from a given regression fit. A Kolmogorov-Smirnov one sample test of normality was performed on the binned regression fits across the parameter space, and at a p-value significance level of 0.05, all residuals were found to be distributed normally. An assessment for a single regression fit was also performed visually using a Q-Q plot, where the residuals from the regression fits in an area of high uncertainty (true SF_1 = 0.01, D_1/D_2 = 2.06) and one from low uncertainty (true SF_1 = 0.5, D D_1/D_2 1/D2 = 19.85) were grouped together and each plotted in separate Q-Q plots as shown in Figure 18.

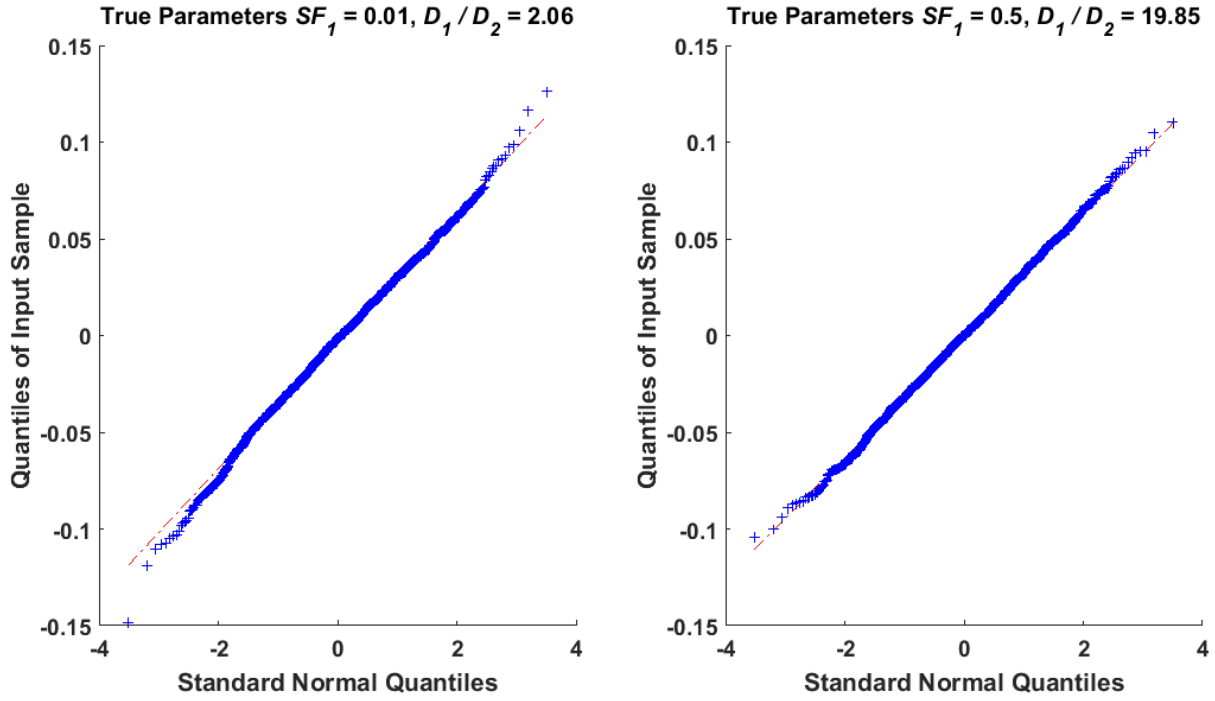

Figure 18 - Q-Q Normality Plot of Residuals

Left plot shows the distribution of residuals for fits to 200 noisy signals on a noise-free signal that has large errors in the parameter estimates. Right plot shows the distribution for fits on a noise-free signal that has small errors in the parameter estimates. These plots show that the residuals are normally distributed regardless of high bias and variance in the biexponential parameter estimates.

These two plots show that while the parameter estimates were non-normally distributed, the residuals from the fits themselves were normally distributed, and thus there was no indication that the maximum likelihood assumptions of normality for NLLS regression fits don't hold. This also showed that large uncertainty in the parameter estimates did not manifest themselves in the distribution of the residuals, either.

2.3.3 Low SNR Rejection Strategy

While the CV showed the normalised variance in the *ADC* estimates, and how its value was closely related to the error in regression fitting to biexponential signals, it didn't resolve why the standard deviation of the two monoexponential parameter estimates in Figure 14 varied across the possible true signal parameter space. This instead was attributed to the amplitude for each diffusion weighted measurement and how it compared to the noise. The line plot at the bottom of Figure 14, shows five selected noise-free signals on a line plot similar to the signal range in Figure 9. Signals "A" and "E" show signals selected near the minimum and maximum possible decay rates respectively. Additionally, the three signals B – D show selected intermediate signals in the line plot between the two extremes. These signal labels match the labels in the top left S_0 plot and top right *ADC* plot, indicating where these signals are approximately located on those plots. The parameter values of these five signals are given in Table 2.

Table 2 - Parameter values for the selected signals in Figure 14

Signal	SF_1	D_1/D_2
A	0.05	19
B	0.95	19
C	0.5	11
D	0.05	3
E	0.95	3

The minimum standard deviation for both parameter estimates is close to point A, the top left corner of the plot, which coincides where the signal fraction of the D_1 component is zero, so the signal is made up purely of the slow component, D_2, which has a value of 0.05 (1/20). As the line plot at the bottom of Figure 14 shows, signal A stays well above the noise floor (SNR = 25) for the entire diffusion weighting range, and the SNR for all measurements is greater than 7. The maximum standard deviation for both parameter estimates is near signal E, which is the bottom right corner of the plot where the signal becomes purely the fast component, D_1, with a value of 1. The line plot shows that signal E decreases more rapidly, such that the last four diffusion weighting measurements have an SNR of 1 or below, effectively measuring noise. Due to these variations in measurement amplitudes compared to the noise floor, strictly reporting the SNR at b = 0, as is common practice in the DWI literature, was inadequate here, since its value was 25 for all noisy measurements. Calculating a signal-averaged SNR instead, averaging the SNR over all eleven b-

value/diffusion weighting measurements, gave a better assessment of each noisy measurement, a plot of which is shown for the biexponential measurements in Figure 19.

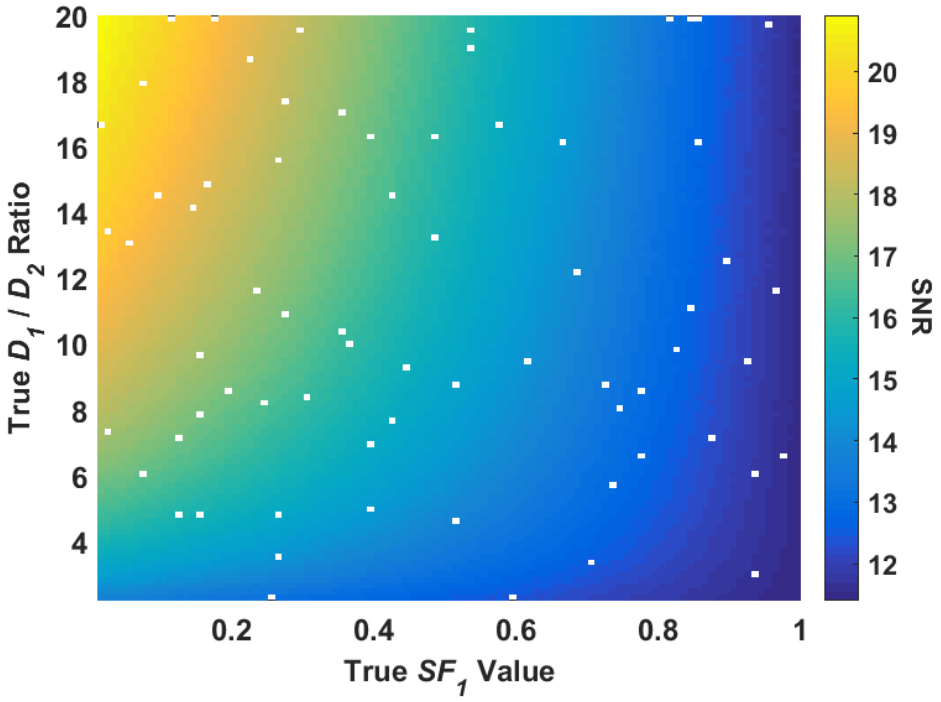

Figure 19 – Mean signal-averaged SNR for noise-free biexponential signals ($SNR_{b=0}$ = 25) in each bin

This signal-averaged SNR has an inverse relationship to the monoexponential parameter estimates shown in Figure 14, indicating that as the SNR increases, the SD in the parameter estimates decrease.

This signal-averaged SNR plot illustrates the increase in standard deviation of the monoexponential parameter estimates was proportional to the decrease in SNR of the true signals. Thus, if Signal E in Figure 14 was being used for fitting, its value at the eighth diffusion weighted measurement would be less than two times the standard deviation at an SNR of 25 (0.04), meaning measurements 8 through 11 would be disregarded and only measurements 1-7 used for fitting. After removing all measurements below an SNR of 2 for the biexponential signals, the updated signal-averaged SNR is shown in Figure 20. This shows an increase in the signal-averaged SNR for the entire test set compared to Figure 19. There are also bands on this plot, which are caused by discrete jumps in the number of diffusion weightings used in the signal measurements. These jumps are illustrated inFigure 21, which displays the median number of weightings used in the signals in each bin.

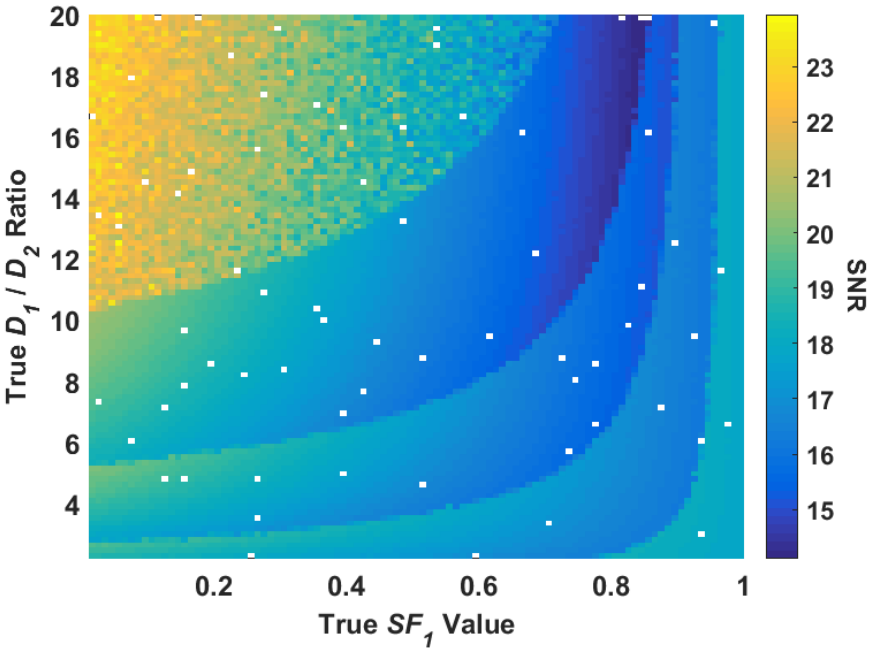

Figure 20 – Mean signal-averaged SNR for noise-free signals in each bin after rejecting all measurements with SNR < 2

This shows that the low SNR rejection strategy improved the signal-averaged SNR when compared to the results in Figure 21.

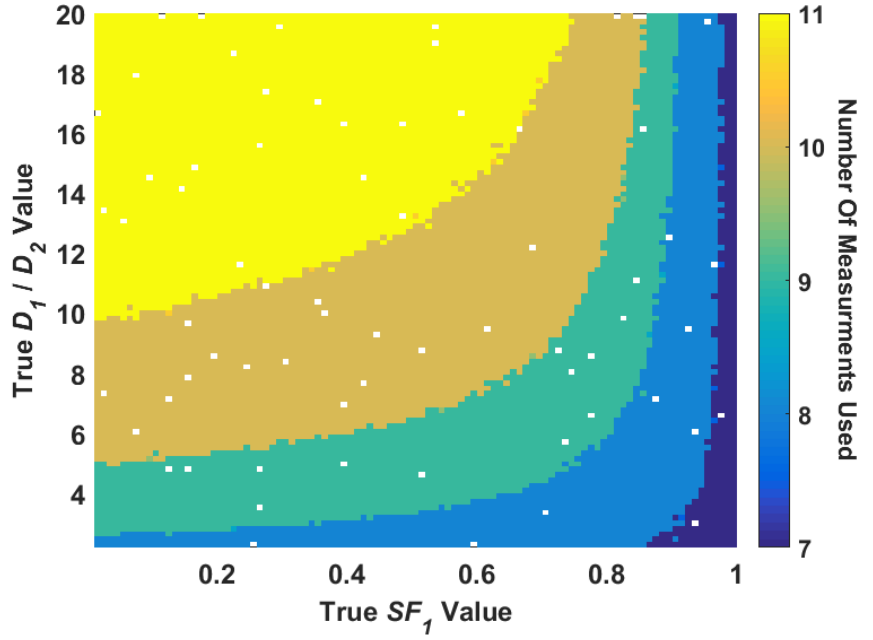

Figure 21 – Median number of signals used for the biexponential test set when rejecting measurements with SNR < 2

After applying this low SNR rejection strategy to the entire biexponential test set, and fitting a monoexponential model to each signal, there was very little change to the standard deviation of the S_0 and *ADC* parameter estimates, and the variance in these estimates actually *increased*. The difference in all estimates was less than 0.002 and 0.0005 for the S_0 and *ADC* estimates in each bin in Figure 14, respectively, and less than 0.004 for the CV of the *ADC* estimates in Figure 15. There was, however, a significant decrease in fitting error (SER) for portions of the test set using the low SNR rejection method, as seen in Figure 22, since rejecting the noisiest measurements improved the closeness of the fit to the data and decreased the RSS value from which the SER is based. The SER decrease is most noticeable at the lowest signal-averaged SNR measurements in Figure 19, where the improvement of the low SNR rejection over fitting all data points is about 0.01, compared to the added noise of 0.04 at an SNR of 25. Thus, when fitting a monoexponential model on this biexponential test set and with this specific distribution of diffusion weightings, removing the noisiest diffusion weighted measurements improved the regression fit to the data but did not significantly improve the model's parameter estimates.

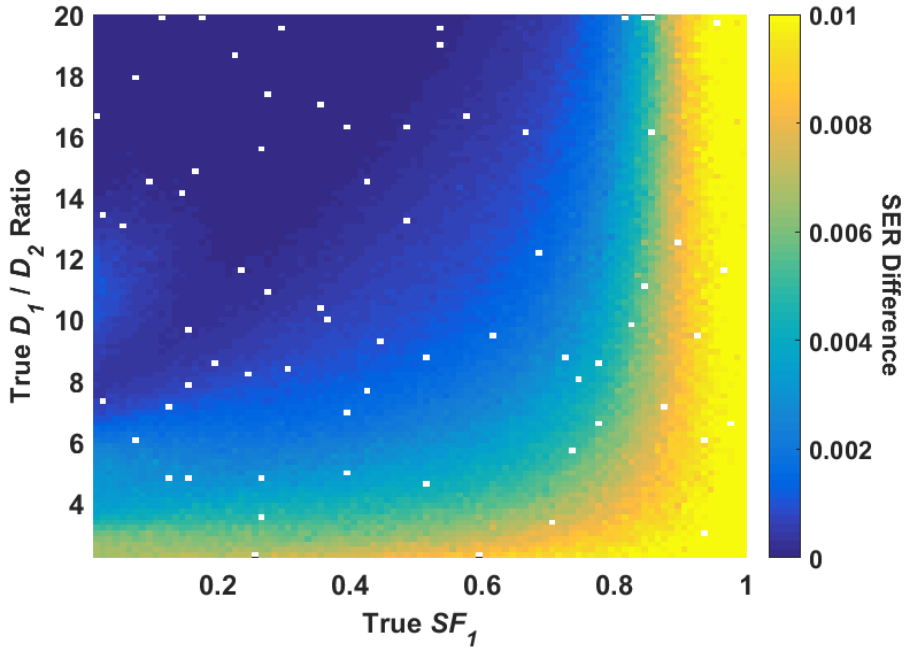

Figure 22 – Improvement in SER between fitting with all data points and rejection of data less than an SNR of 2

Positive value indicates a decrease in SER using the SNR < 2 data rejection strategy. This strategy improved monoexponential fitting only when the true biexponential signal consisted mostly of the fast decay component.

Using the same low SNR rejection strategy with the biexponential model showed distinctly different results in the parameter estimates with the reduction of diffusion weighting measurements causing large *increases* in both the bias and variance of the parameter estimates. The top row of Figure 23 shows the errors in the parameter estimates at an SNR of 25 from the top row from Figure 10

where all eleven measurements are used for signal fitting. The bottom row of Figure 23 shows the errors in the parameter estimates using the updated low SNR rejection strategy. There is a slight increase in the bias and dispersion of the SF_1 estimates where the true SF_1 values are close to one, but there is a major increase in the bias of the D_1 estimates at these values as well as an expansion of the high bias area in the D_2 estimates. The CV values for the D_1 and D_2 estimates have also increased slightly across most of the parameter space, as well.

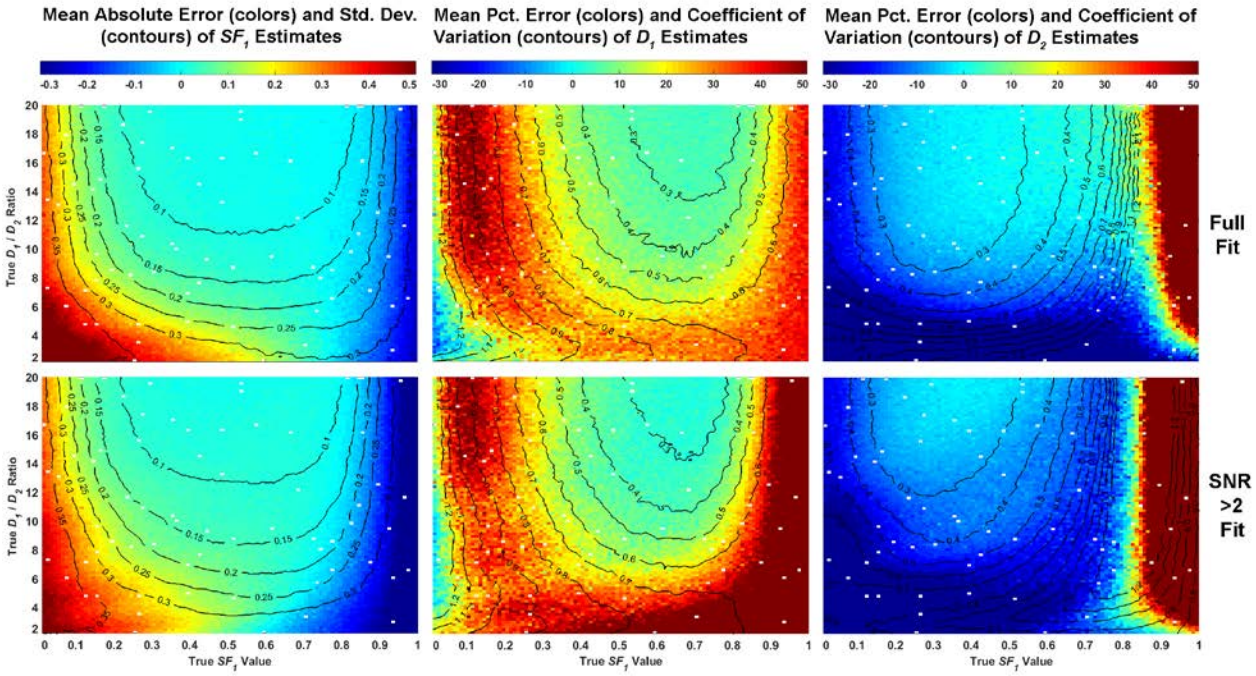

Figure 23 – Parameter estimate errors when fitting all eleven measurements (top row) and fitting only measurements with SNR ≥ 2 (bottom row). $SNR_{b=0}$ = 25

This shows that the low SNR rejection strategy actually increased the bias and variance in the parameter estimates, regardless of the improvement in signal-averaged SNR.

The error increase for all three parameter estimates was correlated with the areas of a reduction in the number of measurements used in the fitting. Figure 21 shows the median number of measurements that are used in the biexponential test set when only using measurements where the SNR ≥ 2. The top left corner shows the area that uses the full 11 diffusion weighted measurements, and there is little increase in error in this area on the parameter estimate error plots in the SNR ≥2 fits. The areas of large bias in the parameter estimates are seen where the total measurements are reduced to 7 or 8. The increase in biexponential parameter estimate errors indicates that attempting to reduce the noise by removing low SNR measurements was counterproductive for a large portion of the biexponential test set. Broad conclusions about the effects on biexponential model parameter estimates by reducing the number of diffusion weightings cannot be drawn here, though, since removing the three or four highest measurements left very little weighting on the slow decay component in this simulated setup. However, due to this significant increase in the

parameter estimates, the remainder of biexponential model fitting results in this chapter used the full eleven diffusion weighted measurements.

2.3.4 Regression Diagnostics

Condition Number of the Algorithm Matrices

The condition numbers of the Jacobian matrices from all noisy signal regression fits at an SNR of 25 were analysed for any correlations with the increased errors seen in the parameter estimates. All regression fits were again grouped by the same 100x100 bins as the results in Figure 10, and a count of the percentage of regression condition numbers that were above 100 were calculated with the results shown in Figure 24. This image shows a pattern similar to the areas of uncertainty in Figure 10, where more condition numbers above 100 were found where the true signal fraction of either component is low, or the D_1/D_2 ratio was also low. The most significant finding on this map is the large area of the graph on the left hand side ($SF_1 < 0.2$) where nearly 100% of all regression fits had a condition number greater than 100. This is significant because this large area of ill-conditioned regression fits happened to coincide where the signal-averaged SNR in Figure 19 was highest, illustrating why high SNR does not necessarily lead to better parameter estimates when fitting with the biexponential model.

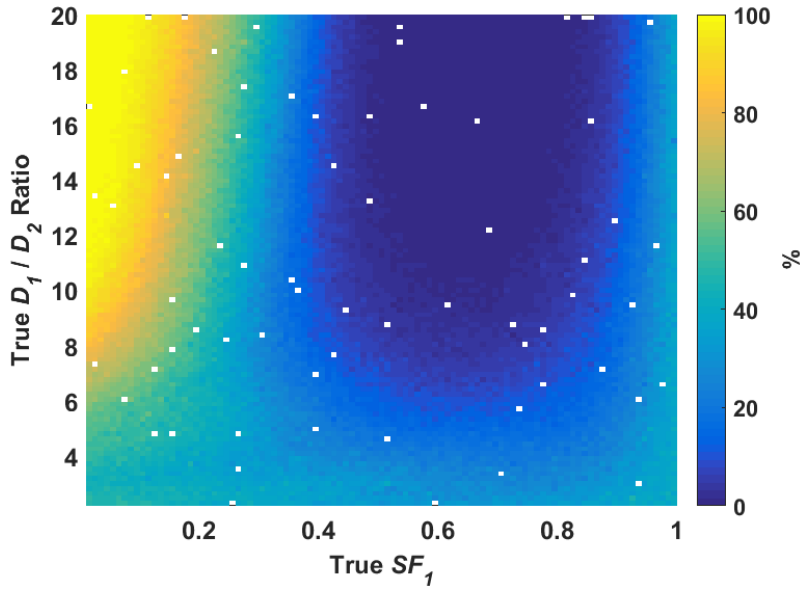

Figure 24 – Percentage of signal fits in each bin with a Jacobian condition number greater than 100

When examining the values of all regression fit condition numbers, several fits with condition numbers of 10^6 or higher were found as shown in Figure 25. The area with the greater number of extreme condition numbers is again found in areas of the parameter space where the true signal fraction and/or the D_1/D_2 ratio is low. While these maps suggest that extreme condition number may be a good measure for indicating regression fits with high parameter estimate error, the noisy measurements from a few noise-free signals were analysed for direct correlation between high condition number and parameter estimates that deviated significantly from the true value, but the

correlation was poor. While, it's not a measure that can directly indicate problems with the parameter estimates, *on average*, a high condition number was correlated with large bias and/or variance in the parameter estimates. This condition number analysis confirmed the hypothesis posited in Section 2.1.2 that the biexponential model could have ill-conditioning issues when the true signal was approximately monoexponential.

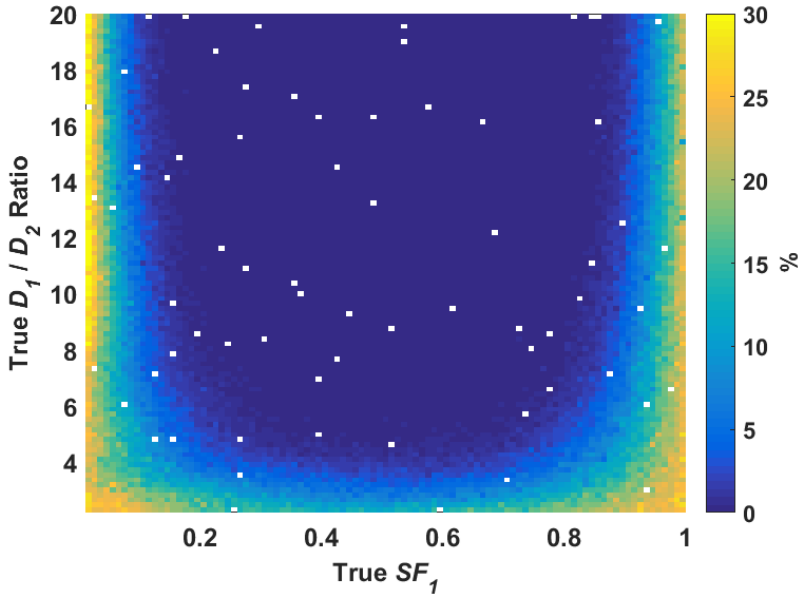

Figure 25 - Percentage of signal fits in each bin with a Jacobian condition number greater than 10^6

Note that the maximum value of the colour scale is 30% in this graph. These two figures show that the condition number is highest for monoexponential signals.

Parameter Standard Deviations and Other Jacobian-based Methods

The parameter standard deviations (SD) were determined by the square root of the diagonal elements of the covariance matrix calculated for each regression fit and used to estimate the error of each parameter estimate. The mean values of all parameter SD values were grouped by bin and are displayed in Figure 26. This figure shows four pseudocolour plots for each parameter where the mean SD of the parameter estimates all go up as the signal becomes more like a monoexponential. However, the upper range of these plots is 10, way beyond the value that would be expected for estimated parameter values between 0 and 1. Again, the reason for these large mean values was the presence of extreme calculated values that skewed the mean. Displaying the median values, instead, as in Figure 27 illustrates a more reasonable picture as the values there are scaled between 0 and 0.2. In this plot, the SD of the two amplitude parameters are very high in the area where SF_1 is very low, as well as a similar area around a decay ratio of 6 that juts out from the left side. The median SD values are lower over the rest of the parameter space, and decrease further where the signal fraction of the slow component is near zero. The two decay components also have the same areas where the errors differ, have very high parameter errors where SF_1 is very low, and slightly increased error where SF_1 is very high.

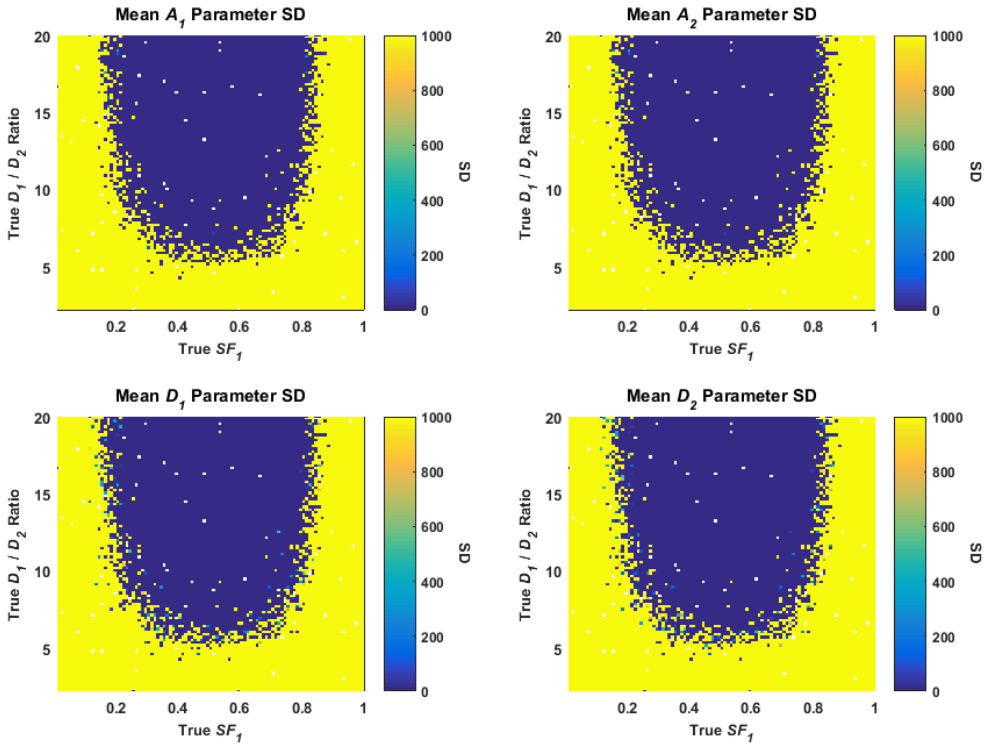

Figure 26 - Mean parameter standard deviation calculated from each signal regression fit

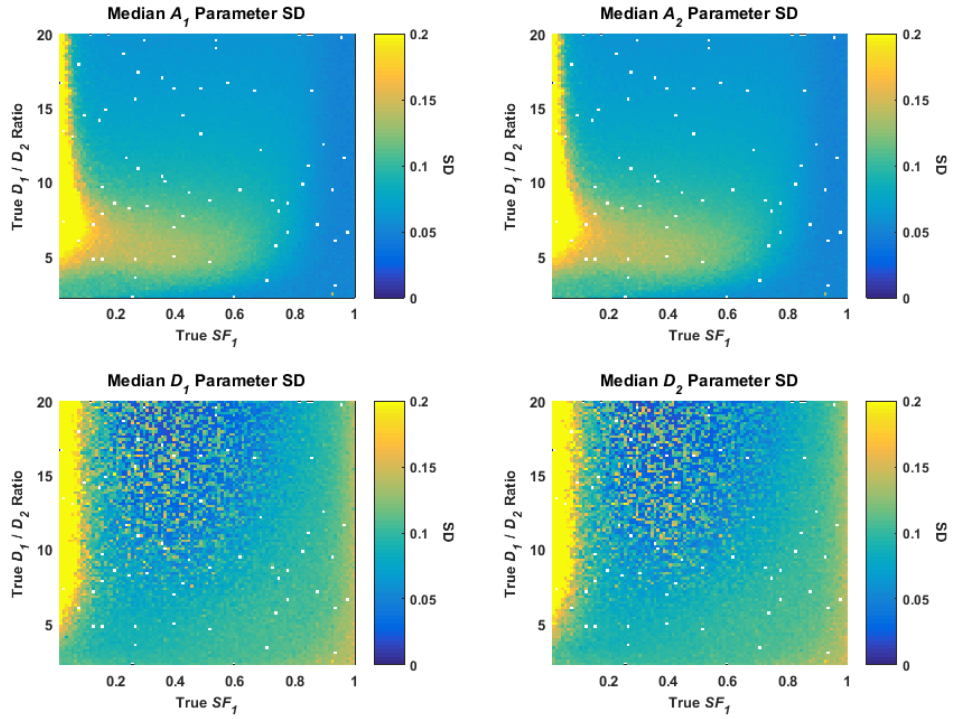

Figure 27 - Median parameter standard deviation calculated from each signal regression fit

Compared to Figure 10, parameter SD is a poor indicator of high uncertainty in the parameter estimates.

The area in the top centre of the parameter space appears to have completely random median standard deviation values between 0 and 1. Figure 27 shows that even after eliminating the effects of outlier estimates, the diagnostic value of parameter SD estimates was poor. While these values of SD indicated some areas of large error in the parameter estimates, they performed poorly even in areas correlated with low parameter estimate bias and variance, and were random in some instances. Not surprisingly, the parameter SD performed poorly as it was derived from the NLLS algorithm Jacobian matrix, and on average, there were extreme values of parameter SD found where there were extreme condition numbers of that matrix. The other diagnostic measures derived from the Jacobian matrix, namely, the correlation matrix, covariance matrix, and VIF, also performed poorly as indicators of high bias and variance in the parameter estimates. Thus, when using an NLLS algorithm, it appears that the ill-conditioning in the biexponential model affected these diagnostic measures as well as the parameter estimates.

Confidence Intervals

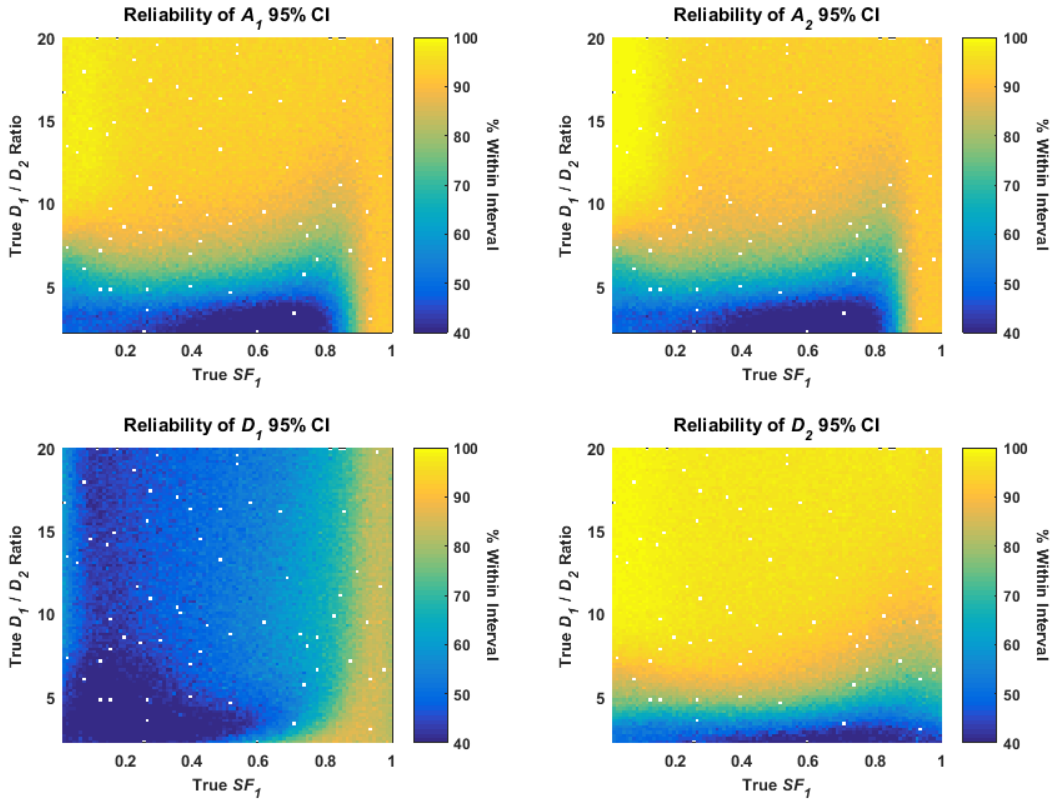

Figure 28 – Percentage of 95% confidence intervals that contain the true value for all four parameter estimates

These confidence intervals, based on the Jacobian, are also poor indicators of high uncertainty in the parameter estimates.

The confidence intervals were calculated based on the parameter SD values and the *t*-distribution values in Equation 40. As Figure 11 and Figure 12 both showed, the parameter estimates for

several of true parameter combinations were definitely *not* normally distributed, violating the normality assumptions for *t*-distribution based confidence intervals. To see the effects of these non-normal distributions, along with the effects of ill-conditioning on the parameter SD values, the 95% confidence intervals were determined from each regression fit and examined for how often the true value was contained within each interval. The selection rates of how often the true parameter values correctly fell within their respective parameter estimate confidence intervals were grouped by bins and plotted across the entire parameter space, as seen in Figure 28. These plots show that for most of the parameter space, the 95% confidence intervals for both amplitude parameter estimates, as well as the estimates for D_2, encompassed the true value at least 95% of the time. There is a small area in the lower left corner of the parameter space for the amplitude estimates where the reliability of the confidence intervals dropped below 60%. The estimates for D_1, however, have large areas where the reliability was below 60-70%, and the reliability only meets the 95% level at the very highest SF_1 values. The reliability for these intervals may be overstated however, since the confidence intervals could have a very wide distance between the minimum and maximum values that easily encompassed the true value. Estimated 95% confidence intervals calculated using Equation 39, with 7 degrees of freedom for a *t*-distribution, gave a factor of 2.36 times the estimated parameter errors in Figure 27. If the amplitude parameter estimates were bounded between 0 and 1 in a regression fit, but the parameter standard deviation for that parameter is 1, then the minimum and maximum values for the confidence interval would be well outside of those bounds, including negative minimum values, which were physically impossible. A two-sided, 95% confidence interval for a Gaussian distribution is shown in Figure 29 to illustrate the area that is typically encompassed.

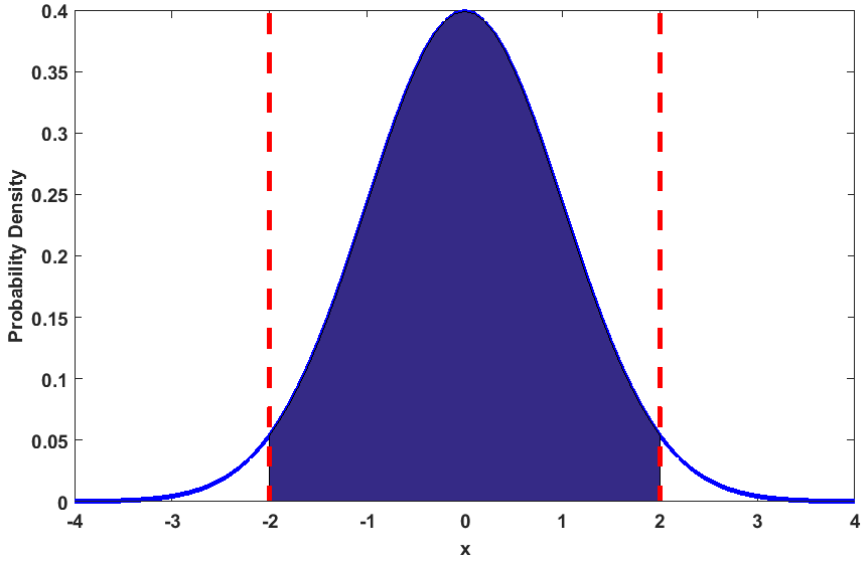

Figure 29 – Probability density function of a Gaussian distribution (blue line) with shaded area indicating the area encompassing 95% of the distribution

Red lines indicate the values that would be reported for the confidence interval (-2, 2). This Gaussian distribution has a mean of 0 and a standard deviation (σ) of 1, so the values of x correspond to multiples of σ and 95% of the distribution is within 2σ of the mean.

The case studies of the parameter estimate histograms shown in Figure 11 and Figure 12 were updated to add in the median confidence intervals taken from all noisy regression fits that make up the histogram. While the confidence intervals had differing values for each fit, examining the median value of the minimum and maximum confidence intervals from all noisy signal fits for a given true signal showed whether the confidence intervals approximately encompassed 95% of the parameter estimate histograms. The updated histograms with the added confidence intervals are shown in Figure 30 and Figure 31. For the decreasing signal fraction histograms in Figure 30, for SF_1 = 0.5, the confidence intervals approximated the 95% area of the well-formed parameter estimates, other than the D_2 estimates which were much wider than the distribution. As SF_1 decreased, the confidence intervals did even worse at encompassing the parameter estimates, or in the case of the D_2 estimates, they encompassed a much larger area than the distribution itself.

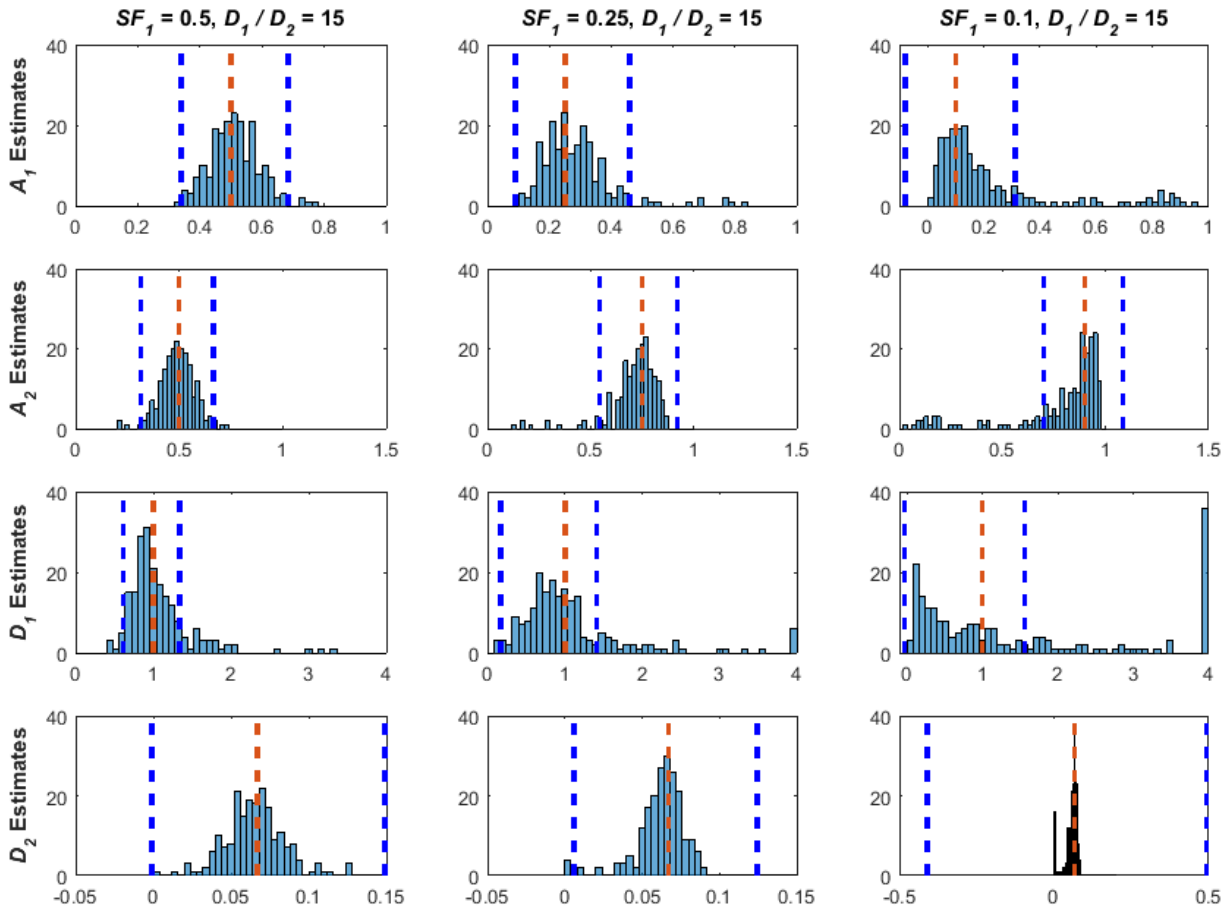

Figure 30 – Histograms of noisy signal parameter estimates (rows) with median estimated 95% confidence intervals for three different noise-free signals (columns) of different true SF_1 values

Median confidence intervals (dashed blue lines) calculated from the parameter errors of the noisy signal regression fits in each histogram. Dashed red lines indicate true parameter values.

Similar trends are seen for the confidence intervals in Figure 31, where the decreasing D_1/D_2 ratio increasingly corrupted the ability of the confidence intervals to encompass the parameter estimate

distributions, and at the lowest ratio of 2 for some parameter estimates, the median confidence intervals didn't even encompass the true parameter values. While these confidence intervals were reliable when the parameter estimate errors were low, when ill-conditioning considerably affected the biexponential regression fits, the confidence intervals became unreliable. The effects of this unreliability were the interval both over- and underestimating the possible range of values that the estimates could take when testing repeated sample measurements. These complications of ill-conditioning in the biexponential model confidence intervals may help to explain the extreme interval values seen in the literature examples presented in Section 2.1.1.

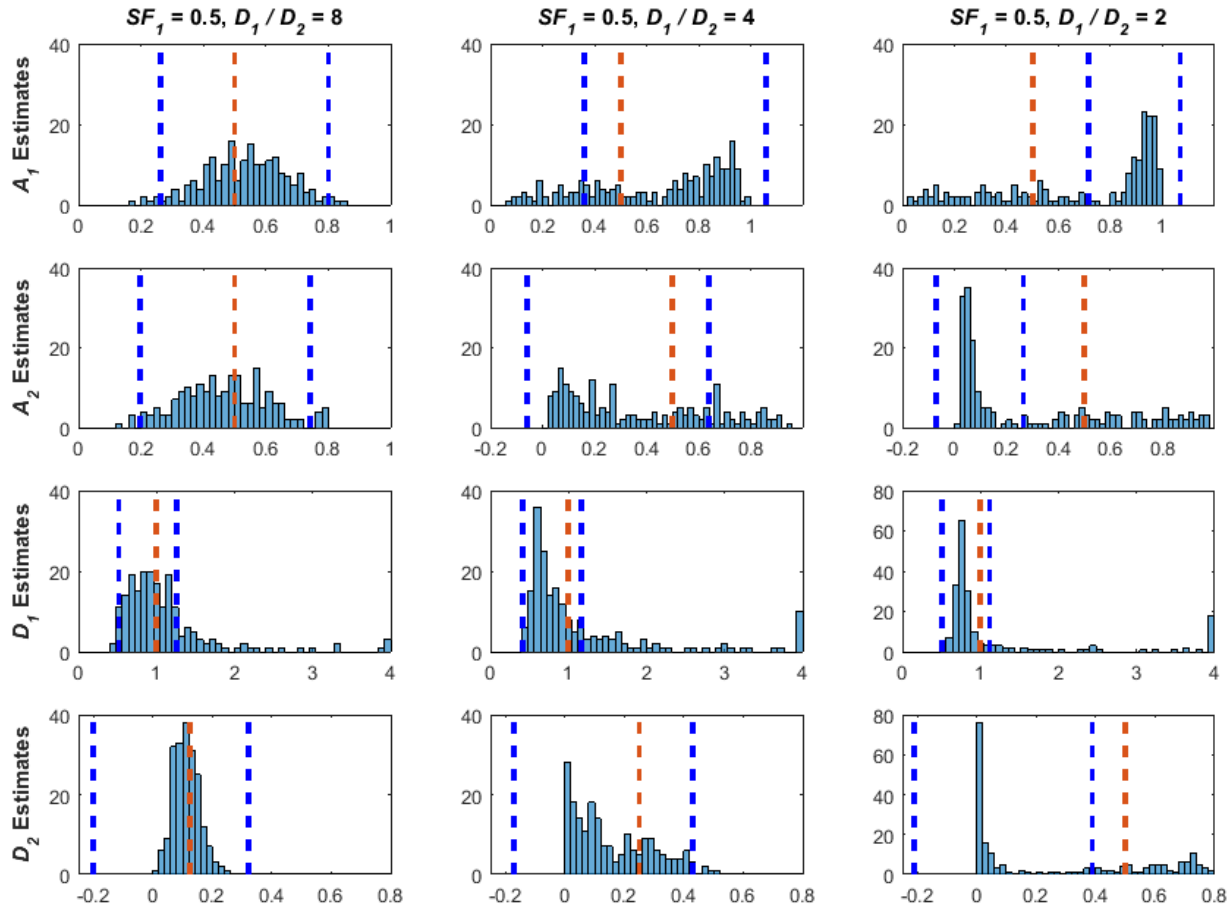

Figure 31 - Histograms of noisy signal parameter estimates (rows) with median estimated 95% confidence intervals for three different noise-free signals (columns) of different true D_1/D_2 ratios

Median confidence intervals (dashed blue lines) calculated from the parameter errors of the noisy signal regression fits in each histogram. Dashed red lines indicate true parameter values. These two figures show that the Jacobian-based confidence intervals poorly predict the range of values for fits with large uncertainty in the estimates.

2.3.5 Bootstrap Analysis

Bootstrap analysis consisted of fitting 25 noisy signals for a limited subset of noise-free signals, and for each noisy signal fit, additional regression fits were performed on 1000 parametric bootstrap samples derived from that signal (i.e. fits of a fit). All bootstrap parameter estimates for each signal were combined by parameter to analyse the distribution of the bootstrap sample estimates and the original signal fit from which they were derived. From these results, three noise-free signal cases were selected for display: one in the area of low parameter estimate errors with true SF_1 = 0.525 and true D_1/D_2 = 15.05, another signal with low signal fraction with SF_1 = 0.075 and D_1/D_2 = 15.05, and finally a low decay ratio signal with SF_1 = 0.5 and D_1/D_2 = 2.45. For the noisy signal regression fits from each of these three noise-free signals, one fit was chosen where the parameter estimates closely match the true values, and a second where the parameter estimates deviated significantly from the true values. The 1000 bootstrap parameter estimates for each regression fit are displayed as histograms for each noisy signal fit in Figure 32, Figure 33, and Figure 34, along with two values indicating the original signal fit parameter estimate and the true value for that parameter from the noise-free signal.

The bootstrap samples for the regression fits in Figure 32 show that for all parameters from both signals, the distributions are close to normal, with no significant outliers noticeable. The distribution of these estimates change significantly for both of the noisy regression fits from the low SF_1 true signal in Figure 33. For Signal 1, the distributions are mostly centred on the parameter estimates, but the amplitude component distributions have noticeably long tails, and D_1 is widely dispersed. Signal 2's parameters are distributed differently and the amplitude distributions were widely dispersed with no distribution structure. Finally, for the true signal with a low decay ratio in Figure 34, Signal 1 has A_1, A_2, and D_2 distributions partially clustered far from the original parameter estimates and the true value, but still has considerable variance in the distribution. Signal 2, however, has the bootstrap samples distributed around the original fit's parameter estimates, but these estimates all deviate significantly from their true values. These bootstrap distributions show that when the true signal was from an area of low uncertainty from the parameter space in Figure 10, regardless of how close the estimates were to the true values, the distribution was well-formed and has a relatively normal distribution shape. If the true signal was from an area where there was significant ill-conditioning in the parameter estimates, regardless of how close a parameter estimate was to its true value, the bootstrap distribution will most likely be widely dispersed and irregularly shaped. The shapes of these distributions were also similar to the distribution of the parameter estimates in Figure 11 and Figure 12, showing how bootstrap resampling can assess the reliability of a *single* regression fit.

Table 3 – True parameter values and regression fit estimates for Figure 32.

	A_1	A_2	D_1	D_2
True values	0.53	0.48	1.00	0.066
Noisy Signal 1 Estimates	0.51	0.49	1.00	0.075
Noisy Signal 2 Estimates	0.67	0.31	0.65	0.036

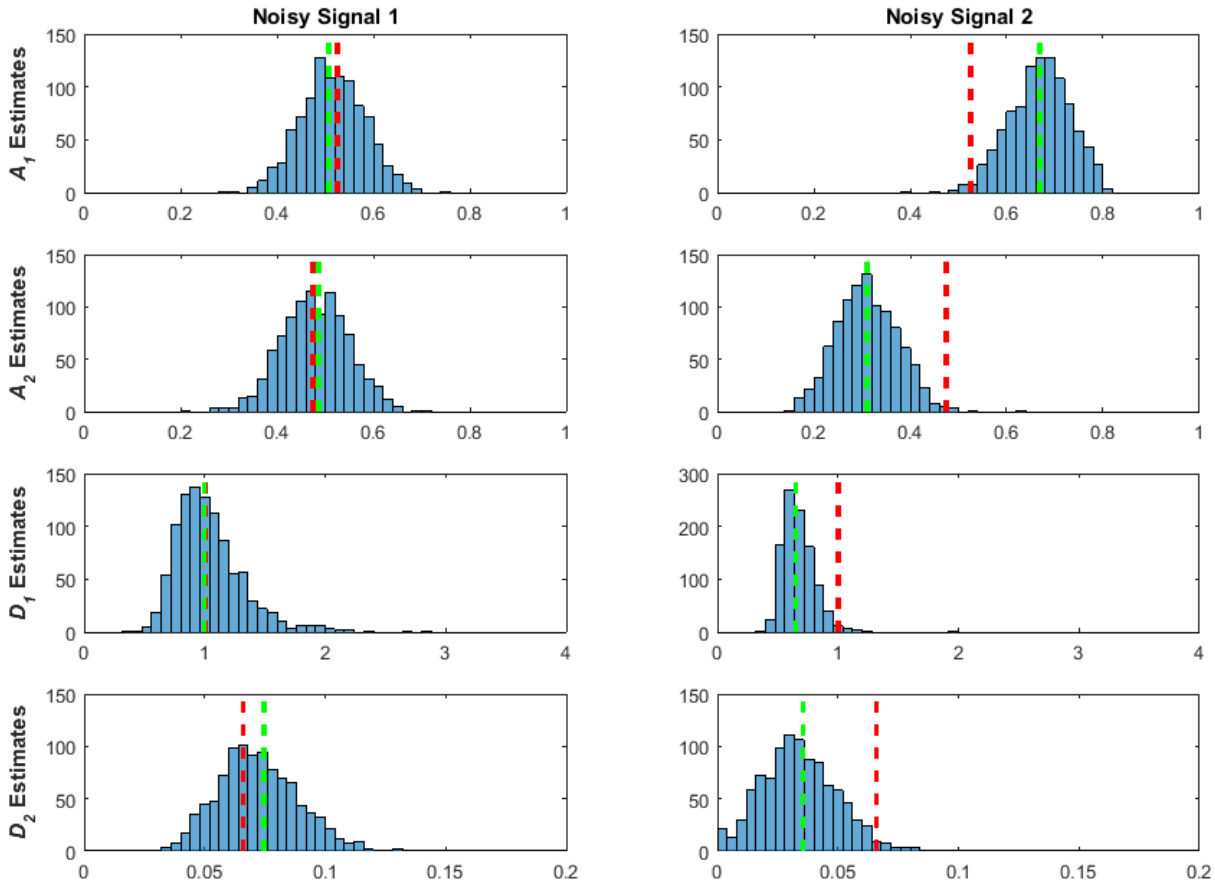

Figure 32 – Histograms of bootstrap samples from parameter estimates for two noisy regression fits from the same noise-free signal in an area of "low uncertainty"

True parameter values of the noise-free signal (dashed red lines on the histograms) along with the parameter estimates from the two noisy regression fits (dashed green lines) are displayed in the table at the top of the image. Even for the fit with the highest deviation in the parameter estimates, the bootstrap distributions are well-formed and encompass the true value.

Table 4 – True parameter values and regression fit estimates for Figure 33.

	A_1	A_2	D_1	D_2
True values	0.075	0.93	1.00	0.066
Noisy Signal 1 Estimates	0.090	0.90	1.01	0.057
Noisy Signal 2 Estimates	0.55	0.44	0.19	0.024

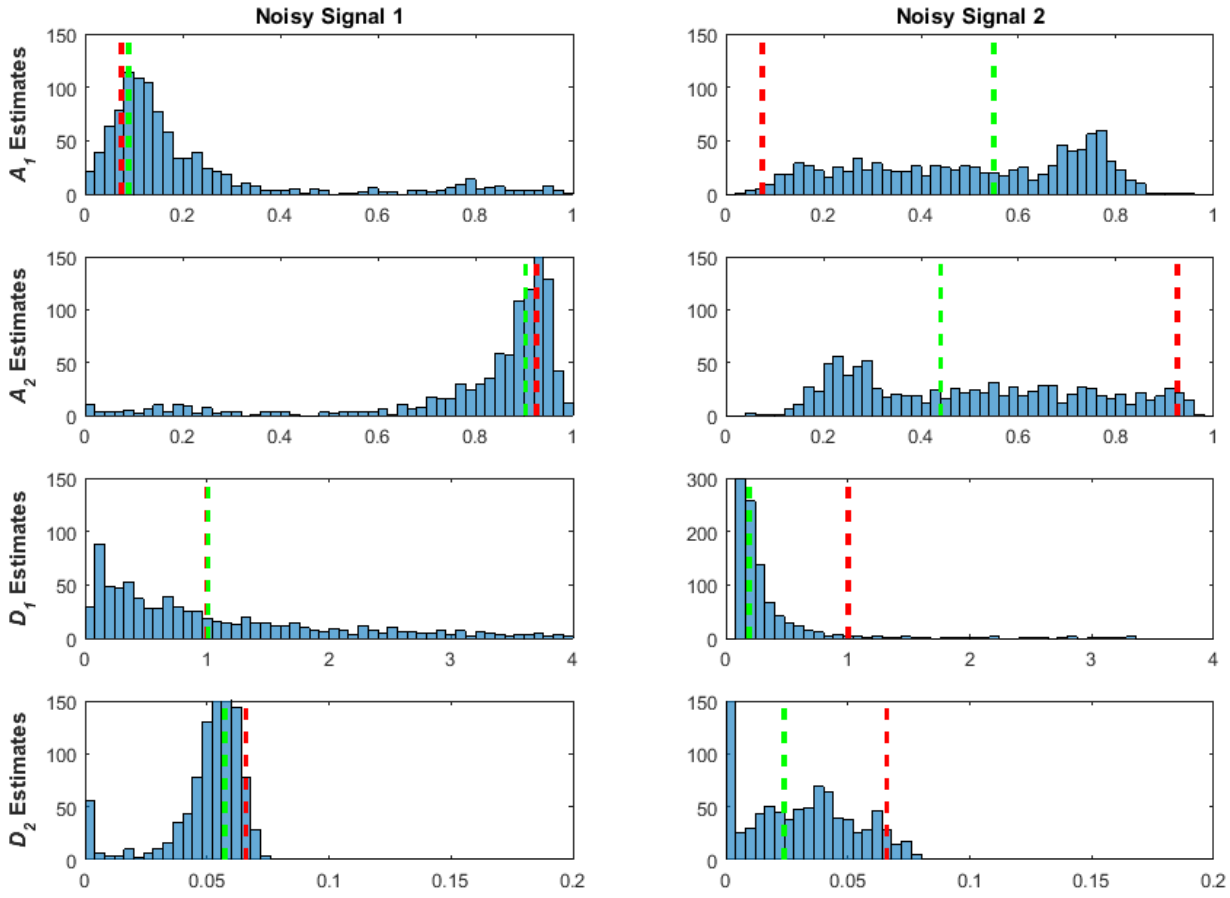

Figure 33 - Histograms of bootstrap samples from parameter estimates for two noisy regression fits from the same noise-free signal in an area of low true SF_1

True parameter values of the noise-free signal (dashed red lines on the histograms) along with the parameter estimates from the two noisy regression fits (dashed green lines) are displayed in the table at the top of the image. These distributions show that regardless of whether the parameter estimates are close to the true parameter values, future measurements of this same signal are likely to be unreliable.

Table 5 – True parameter values and regression fit estimates for Figure 34.

	A_1	A_2	D_1	D_2
True values	0.53	0.48	1.00	0.41
Noisy Signal 1 Estimates	0.51	0.51	1.10	0.37
Noisy Signal 2 Estimates	0.98	0.02	0.68	< 0.0001

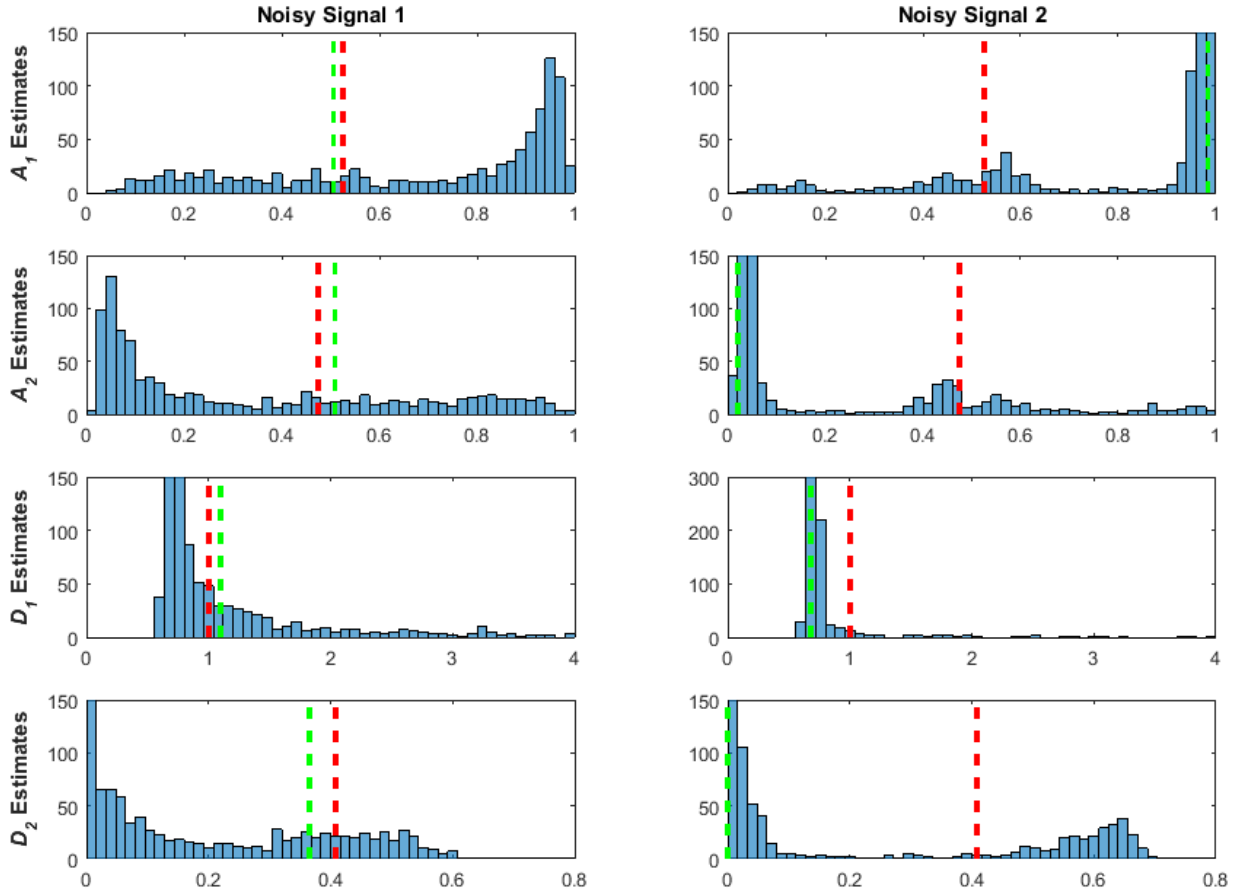

Figure 34 - Histograms of bootstrap samples from parameter estimates for two noisy regression fits from the same noise-free signal in an area of low true D_1/D_2 ratio

True parameter values of the noise-free signal (dashed red lines on the histograms) along with the parameter estimates from the two noisy regression fits (dashed green lines) are displayed in the table at the top of the image. These distributions show that regardless of whether the parameter estimates are close to the true parameter values, future measurements of this same signal are likely to be unreliable.

The bootstrap sample distributions were also used to produce more reliable 95% confidence intervals for the data. Instead of estimating from the regression fit, a 95% confidence interval was estimated by calculating the 2.5% and 97.5% percentiles of the bootstrap sample distribution. Unlike the earlier estimated confidence intervals based on the *t*-distribution, using the bootstrap samples did not assume that the errors are symmetrical about the parameter estimates, and since

the values were bounded to zero on the lower side, the lower bound interval values wouldn't be negative. The bootstrap samples in all regression fits were estimated using *no* upper bounds on the NLLS algorithm, however, to investigate the true deviation for unreliable parameter estimates. The horizontal axes for the bootstrap distributions in the previous three figures were limited to the same values as the parameter bounds for the original fit, but the bootstrap samples for the unreliable estimates had much higher values, especially for the decay parameters. For example, Figure 33 had D_1 sample estimates that were much greater than 4, with some values as high as 200. Figure 35 shows the bootstrap 95% confidence intervals compared to the 95% confidence intervals calculated via the *t*-distribution from the original regression fit.

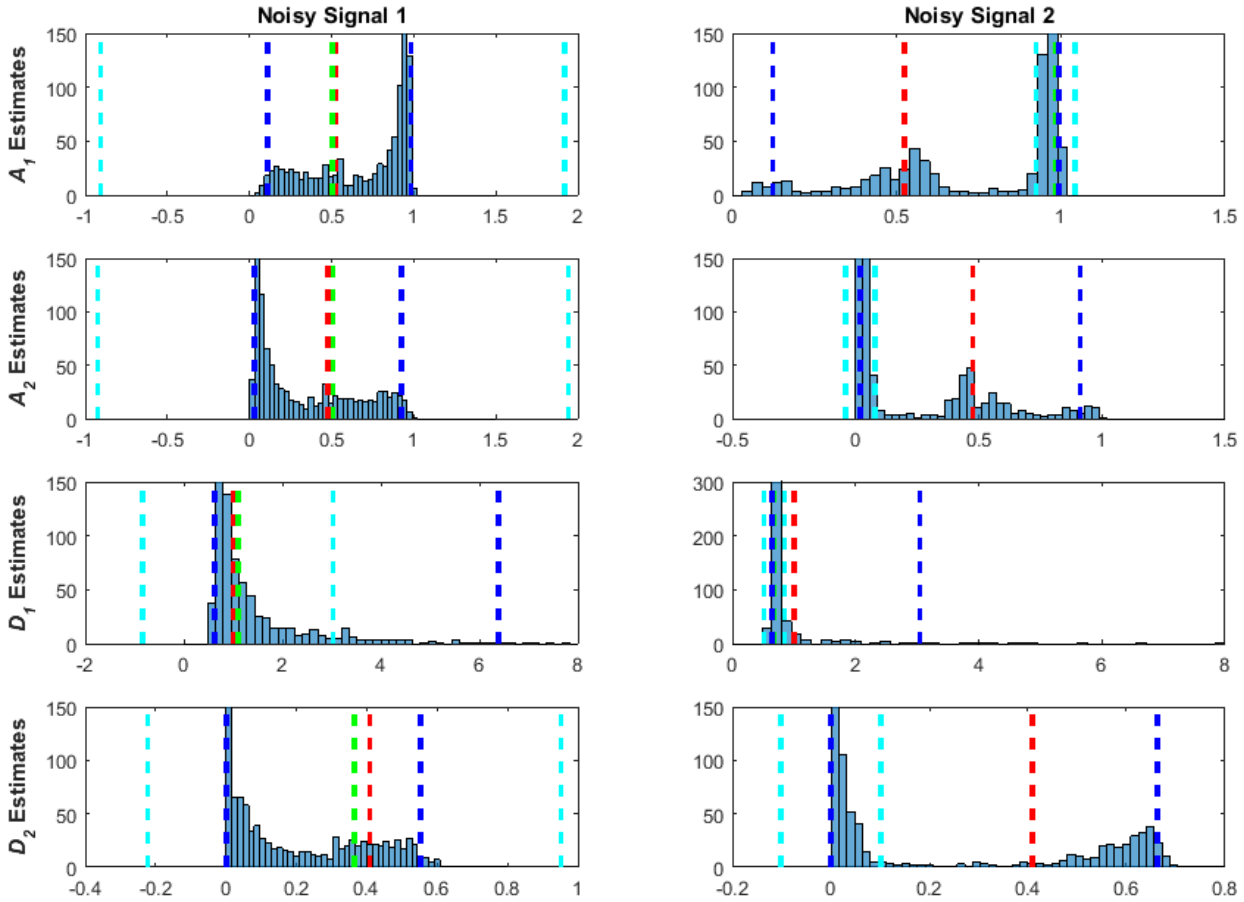

Figure 35 - Histograms of same bootstrap samples in Figure 34 with added confidence intervals

True parameter values of the noise-free signal (dashed red lines on the histograms) along with the parameter estimates from the two noisy regression fits (dashed green lines) are displayed in the table at the top of the image. Dashed cyan lines indicate the 95% *t*-distribution based confidence intervals calculated from each regression fit, and dashed blue lines indicate the 95% percentile intervals based on the bootstrap sample distribution. The bootstrap distributions better assess the range of values likely to be seen for future measurements.

Figure 35 shows that the A_1 and A_2 bootstrap confidence intervals in blue encompass the distributions much better and fell within the original bounded values of 0 and 1, while the *t*-distribution confidence intervals didn't properly encapsulate the range of possible values. The D_1 bootstrap confidence intervals encompassed the bootstrap samples that varied well above the original parameter estimate upper bound of 4 and also didn't become negative. The bootstrap confidence intervals were a better assessment of the range of possible values that the parameter estimates could have, but obviously, confidence intervals that encompass nearly the entire possible range of values for a parameter, 0 to 4 for D_1, for example, indicate that the true value could be anywhere in the possible range of values and that this particular fit is very unreliable. These confidence intervals, or any variance measure such as standard deviation, IQR, etc., produced by the bootstrap sample distributions give a significantly better estimation of the variance associated with the parameter estimates of a single regression fit.

2.3.6 Graphical Analysis

Figure 36 shows a contour plot of the RSS values across a discrete grid of points with different values of SF_1 and D_1. with each point compared to two noise-free signals, both with $A_1 = A_2 = SF_1 = 0.5$ and $D_1 = 1$, with one signal having a D_1/D_2 ratio = 2 and the other a D_1/D_2 ratio = 20. With the same colour scale set for the RSS value of each contour map, these two maps show how much flatter and wider the RSS contours are for the D_1/D_2 ratio = 2 signal than the D_1/D_2 ratio = 20 signal. The innermost contour at the ratio-of-2 signal stretches from 0.35 to 0.9 for SF_1 value, whereas the same contour for the ratio-of-20 signal is an ellipse tightly centred on the true value of 0.5. This innermost contour represents RSS values between 0 and 0.0005, and the ratio-of-2 signal has a "shallower" floor with more parameter combinations falling in that RSS range. While a standard NLLS algorithm would find the global minimum of 0 and the true parameter values for both signals in this case, the addition of noise would affect the ratio-of-2 signal much more than the ratio-of-20 signal. Put another way, if the addition of noise means that a NLLS fit of this a noisy signal measurement has an RSS of 0.0005, there is a larger range of possible parameter estimates for the ratio-of-2 signal due to the larger contour. If the noise added to the signal increases, such that the RSS from a NLLS fit is now 0.001, the increase affects the ratio-of-2 signal more than the ratio-of-20. The ratio-of-2 contour is also elongated and banana shaped, and projecting it onto the horizontal axis to create a one-dimensional distribution helps to illustrate why the parameter estimates are no longer normally distributed and/or symmetrical (see Figure 12).

If these same sum-of-squares contours are applied to a noise-free signal that is monoexponential and the D_1/D_2 ratio = 1 ($D_2 = 1$), the resulting map in Figure 37 shows that the minimum contour stretches across all possible values of SF_1. This illustrates what Equation 33 demonstrated in this chapter's introduction of collinearity – if the values of D_1 and D_2 are equal, there can be an infinite number of possible values for the amplitude coefficients.

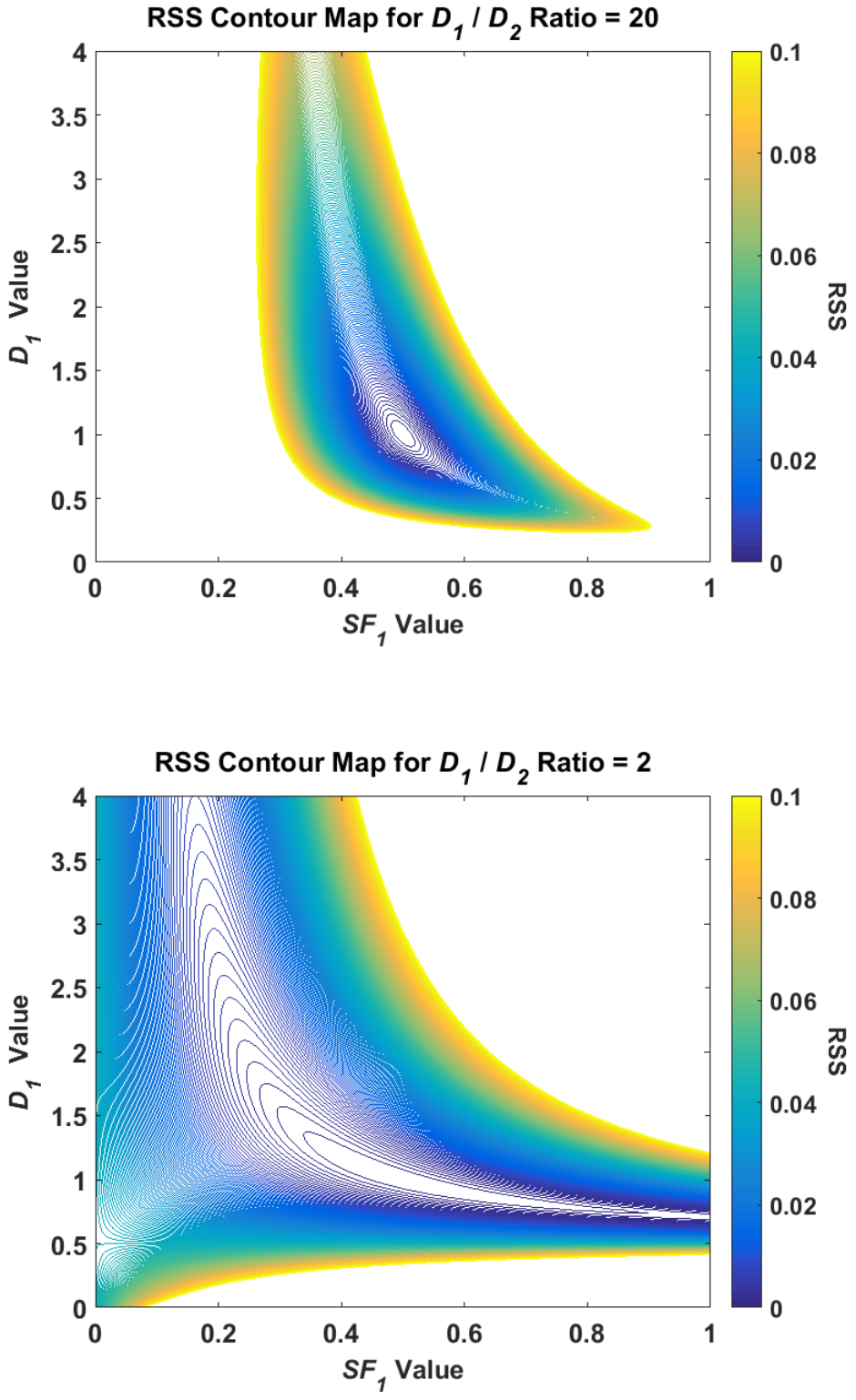

Figure 36 – Sum-of-squares contour maps of the RSS value for a D_1/D_2 ratio = 20 signal (top) and a D_1/D_2 ratio = 2 signal (bottom)

The ratio = 2 signal has a much wider range of possible estimate values than the ratio = 20 signal.

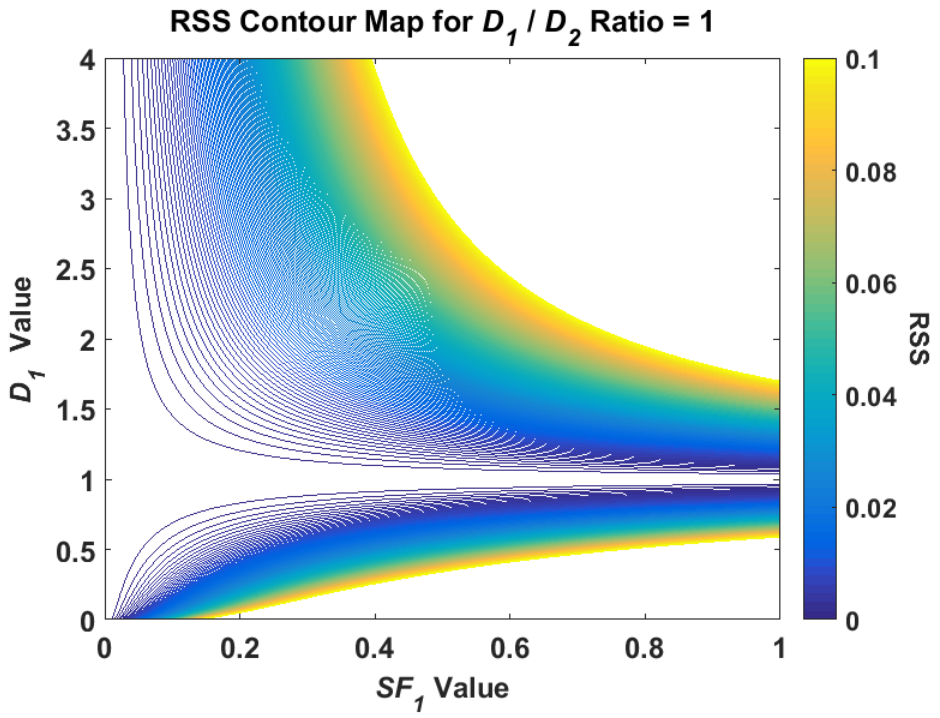

Figure 37 – Sum-of-squares contour maps of the RSS value for a D_1/D_2 ratio = 1 signal

For this signal, the value of SF_1 could be estimated at any value from 0 to 1.

The contour plots from these three signals demonstrate that the issues with the biexponential model and its large parameter errors stem from the nature of the model and not the inadequacy of the algorithm in finding the global minimum. They also illustrate what Acton meant in the chapter introduction by many combinations of parameters fitting data quite well. The results in this chapter have shown, however, that using the biexponential model is not a completely hopeless prospect, but instead the degree of hopelessness depends on the nature of the signal itself.

2.3.7 Comparison of Simulation Results to the Literature

The simulated results presented in this chapter illustrated how and why the biexponential model has uncertainty issues with its parameter estimates. While new methods such as the bootstrap analysis were presented to assist researchers in evaluating possible uncertainties in the model estimates from their data, it also might be useful to assess some of these simulated results in the context of reported literature values. The reported literature values from Figure 6 can be plotted on the SD contours of the SF1 estimates for an SNR of 25 (Figure 10, top left plot), resulting in the image in Figure 38.

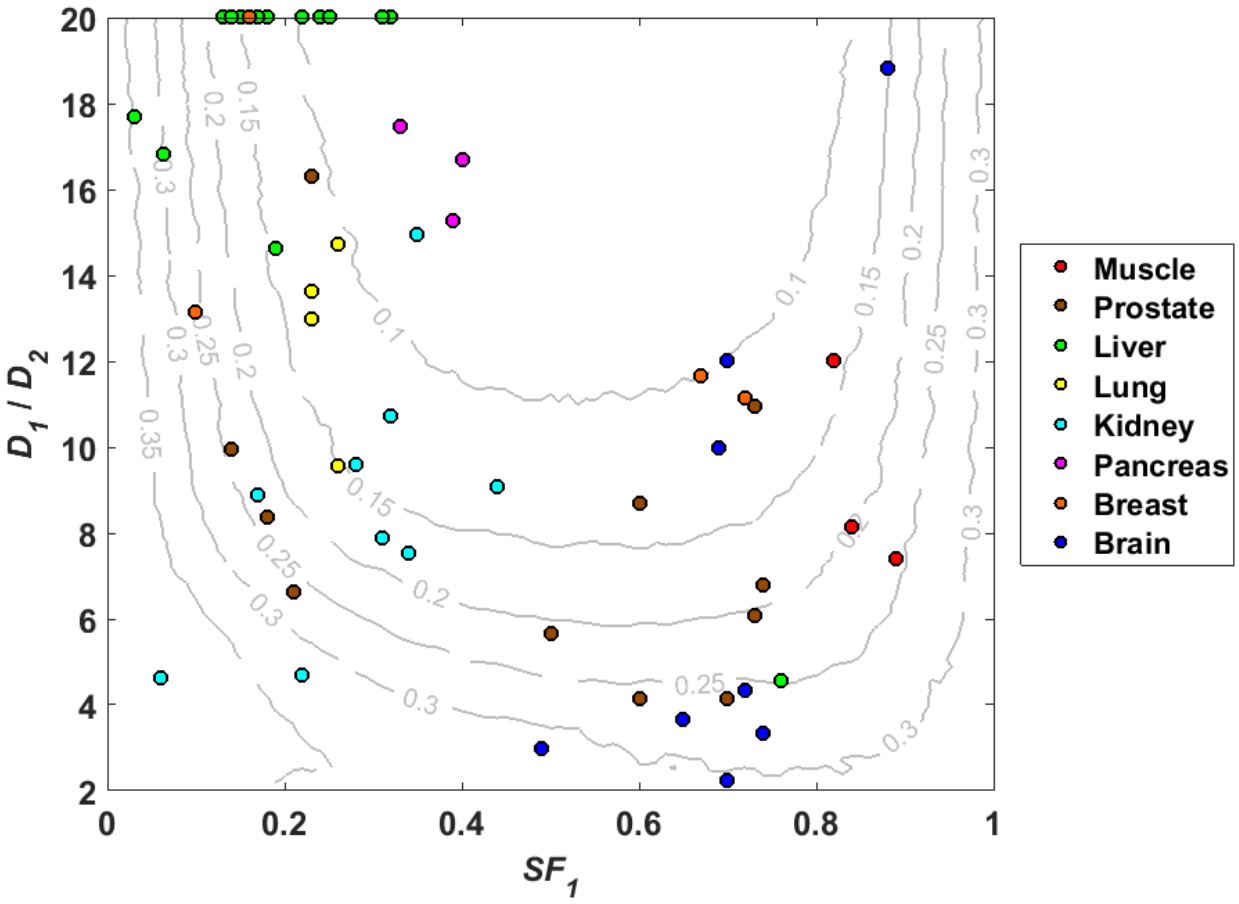

Figure 38 – Reported study parameter values (dots) from Figure 6 overlaid on the CV values (contours) for the SF_1 estimates at SNR = 25 plot (top left) in Figure 10

This figure shows that many of the reported literature values can be found in areas of increased variability in the SF_1 estimates. The reported estimates for most of the brain, prostate, and kidney studies have a D_1/D_2 ratio lower than 10, which puts them in an area of increased variability in the SF_1 estimates, indicating that it may be good to revisit these studies in the context of these results. While the muscle and breast studies have D_1/D_2 ratios in a middle range between 8 and 16, many of the studies have low signal fraction of either the slow or fast component, causing a higher variability in the SF_1 estimates. The pancreatic studies appear to have the best overall performance, while the liver studies have the highest D_1/D_2 ratios, but some of these studies have low values of SF_1. These simulations could be extended to the reported D_1/D_2 ratios for these liver studies (> 50), if desired, however, the simulation results presented in this chapter cannot directly be related to the actual studies, as the scanner acquisition criteria are most likely different and the reported literature values are the estimates, not the true values. An interesting analysis of these high D_1/D_2 ratios would be to revisit these liver studies and examine whether these high ratios are actually due to increased perfusion or are more due to the overestimation of the D_1 estimates when the fits are ill-conditioned, as was shown in this chapter. As this chapter also showed, the number and values of the diffusion weightings affected the performance of the biexponential model, as did the value of the simulated SNR, so any tissue analysis should also take this into account.

2.4 Summary of Conclusions

This chapter presented an analysis of the biexponential model and its use with NLLS regression algorithms on simulated data. The results addressed gaps in the literature specifically described in Section 2.1.5, namely:

- Both large bias and variance were found in the biexponential model parameter estimates when fitting to a test set of simulated biexponential signal measurements that reflected the range of parameters typically seen in DWI studies. For noisy measurements of some signals, the variance in the estimates was an order of magnitude beyond the variance of the simulated noise. When significant bias affected the parameter estimates, fitting of repeated samples from the same signal did not converge to the true value of the parameters. In some cases, the distribution of the parameter estimates was heavily skewed and/or bimodal, with many outlier values.
- The large bias and variance occurred in true signals that were near monoexponential, either through low signal fraction or low decay ratio. Thus, the biexponential model was found to be non-robust and the degree of uncertainty in its estimates varies based on the nature of the signal.
- Increasing the SNR of the simulated noise from 25 to 200 decreased the overall number of signals with large uncertainty in the parameter estimates. However, at signals that were effectively monoexponential, large bias and variance were still found in the parameter estimates, signifying that increased SNR does not lead to better estimates for all signals.
- There was increased *ADC* coefficient of variation when fitting the simpler monoexponential model to much of the biexponential test set, illustrating the increase in error of fitting a simpler model to a complex signal. However, it was found that the coefficient of variation in the biexponential model decay parameter estimates when fitting a monoexponential signal were much greater, demonstrating the cost of using the biexponential model that was previously unassessed.
- The Rician bias from the magnitude measurements was found to affect both the monoexponential and biexponential models. When more measurements at higher diffusion weightings were at or below the noise floor, the variance in the parameter estimates increased. However, for data where all measurements were well above the noise floor, the biexponential model still had large uncertainty in the parameter estimates where there was very low signal fraction of one component, demonstrating that the effects of ill-conditioning and algorithmic issues are much greater than the effects of Rician signal bias.
- The ill-conditioning in the parameter estimates were found to affect the diagnostic measures derived from the NLLS regression algorithm Jacobian matrix. They were found to have poor predictive performance in identifying cases where the parameter estimates had large bias or variance. When averaging over repeated samples, an increase in the Jacobian condition number was correlated with an increase in parameter estimate uncertainty.
- The effects of the parameter standard deviations derived from the Jacobian matrix led to confidence intervals that poorly estimated the possible range of values of parameter estimates with high uncertainty.

- The parametric bootstrap perturbation analysis provided confidence intervals that gave a much better assessment of the variance possible in the parameter estimates upon fitting repeated samples from the same measurement. Examining the shape of the bootstrap sample distribution also showed whether there may be bias in the estimates where repeated samples would converge to the wrong value, giving a better assessment of the reliability of the estimates from a given fit.
- The sum-of-squares contours illustrated that the problem with uncertainty in the parameter estimates is not due to possible inadequacies in the NLLS regression algorithm, but rather issues with the mathematical nature of the biexponential model itself. They also demonstrated how noise affected a signal with low decay ratio more than one with a high decay ratio.
- Finally, these simulated results were presented in the context of the reported literature studies to notify researchers what types of studies were most likely to have biexponential reliability issues.

Chapter 3
Performance of the Kurtosis Model Using Simulated DWI Data

3.1 Introduction and Background

3.1.1 Regression Fitting with the Kurtosis Model

The kurtosis model is another model commonly used for NLLS regression fitting to DWI data, with the methodology often referred to as diffusional kurtosis imaging (DKI). DKI assesses differences in biophysical tissue structure by measuring the deviation in the shape of the molecular diffusion displacement distribution from a Gaussian distribution. Equation 13 introduced the DWI kurtosis model and its individual parameters,

$$S_b = S_0 \exp\left(-bD_{app} + \frac{1}{6}b^2 D_{app}^2 K_{app}\right). \quad (13)$$

This model has an amplitude component S_0 like the monoexponential decay model, but the apparent diffusion coefficient D_{app} appears twice in the exponential term in the same relationship as the *b*-value, linearly in the first component and quadratically in the second. This second component is then multiplied by the kurtosis parameter, with a larger absolute value of the kurtosis parameter leading to a larger deviation from the single exponential decay curve. Like the biexponential model illustration in Figure 7, the effect of the kurtosis parameter on the curve deviation can best be seen with a logarithmically scaled plot, shown in Figure 39.

At higher *b*-values, the curve with a kurtosis of 2 actually curves up and since its values are based on a quadratic component for the *b* and *ADC* parameters, its term can become larger than the linear term, causing an exponential *increase* in the signal value. This does not make sense physically, however, since molecules wouldn't all disperse and then spontaneously reform to their starting positions, so the decay curve should be monotonically decreasing or at least unchanging at high *b*-values [38]. A kurtosis greater than zero means the molecular displacement distribution is more sharply peaked and has fatter tails than a normal distribution, whereas a kurtosis less than zero has the opposite effect, the distribution is more rounded and has shorter tails than the normal distribution. The effect of a negative kurtosis on the curve is shown in Figure 39 with the signal decreasing in value more rapidly than a normal distribution as the *b*-value increases. The increase or decrease of the signal curve at high *b*-values may also mean that estimates of the kurtosis parameter may also be susceptible to Rician noise bias when fitting to noisy DWI data. The lifting of the tail of the curve where kurtosis = 2 in Figure 39 is similar to the effect at high *b*-values seen in Figure 4, which would cause a positive bias in the kurtosis parameter as well as possible bias in the decay rate. The lifting of the tail end of the signal from positive kurtosis would indicate that tissue structures are causing restrictions to the diffusion, whereas negative kurtosis would indicate something more complex is happening.

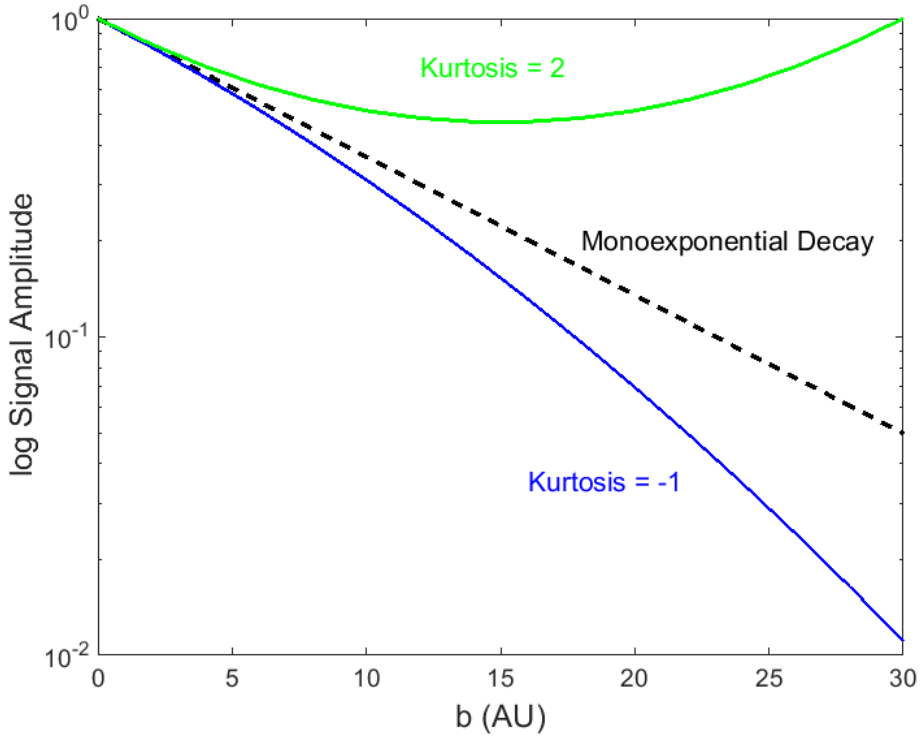

Figure 39 - Line plot of a monoexponential decay signal (black dashed) and two kurtosis signals based on Equation 13, one with K_{app} = -1 (blue) and another with K_{app} = 2 (green)

Jensen et al [36] gave specific examples where a negative kurtosis has a possible biophysical basis and also reported the existence of several negative kurtosis estimates from NLLS regression, but all negative values were rounded to zero for presentation. The authors there also report studies on phantoms filled with sucrose solution, with reported mean kurtosis values close to zero, but all kurtosis estimates are reported with a zero cut-off. Rosenkrantz et al [41] reported a few negative kurtosis values, and Latt et al [169] actually reported negative kurtosis values along with fitted kurtosis values that were much higher than 3. Negative kurtosis values have also been seen in T2* decays when measuring prostate tissue [170]. Kurtosis models are often compared to the fits from a monoexponential decay model, but if there are estimated negative kurtosis values, rounding them to zero would present the results as an overabundance of simple monoexponential decay values since a kurtosis of zero makes Equation 13 equivalent to a single exponential decay. To better assess the performance of the kurtosis model over the range of its possible estimated parameter values, analysis of negative kurtosis parameter values should be included.

Most DKI studies used the kurtosis model successfully to correlate different tissues and/or changes in the brain with differences in the kurtosis parameter estimates. Researchers have expanded the scope of kurtosis model investigations into other tissues, with a recent review paper summarizing all studies done on non-brain tissue [171]. Kurtosis model parameter estimates are also reported after combining data voxels using visual selections of regions of interest, and then reporting the mean and standard deviation of the parameter distributions, although exceptions exist where estimates are reported via distribution scatter plots or colour maps [109]. The kurtosis model is also used as a combined tensor where measurements are taken over multiple axes in three

dimensions, and the kurtosis parameter estimates are reported as a mean kurtosis calculated by averaging the different axial kurtosis estimates [38].

3.1.2 Comparison with the Biexponential Model

The biexponential and kurtosis models both contain additional parameters to assess information beyond monoexponential, Gaussian diffusion. The kurtosis model with three fitting parameters is more complex than the monoexponential decay model, but is more parsimonious than the four parameter biexponential model. As the previous chapter on the biexponential model showed, the uncertainty in the biexponential model parameter estimates was much greater than in the monoexponential model over a significant area of the biexponential true parameter space. In this case, using a more complex model to attempt to reduce model bias when fitting to complex tissue structure came at a cost of a large increase in both the bias and variance in the parameter estimates. The more parsimonious kurtosis model may have more stable parameter estimates that are less prone to error than the biexponential model, but no literature investigations into kurtosis model stability over a typical range of parameters have been made either. Kiselev et al [109] did a comparison of the biexponential model vs. the kurtosis model and showed that the kurtosis model fit nearly as well as the biexponential model for several parameter estimate cases in brain tissue. Toivonen et al [172] reported that the biexponential model was less reliable than the kurtosis model.

While the Rician noise bias can affect the kurtosis model parameters, the kurtosis model itself may also be susceptible to the same ill-conditioning and parameter identification problems that affect the biexponential model. The monoexponential model is also nested within the kurtosis model, seen in Equation 13 when K_{app} = 0. When the true kurtosis value of a signal approaches zero, the model is attempting to fit a parameter that isn't there, which is similar to the parameter identification problem that created large errors in the biexponential model parameter estimates. This can also be visually explained using Figure 39 when the true kurtosis parameter is close to zero. Small kurtosis values have a small deviation or angle from the true monoexponential decay signal, so the measurement noise will have a larger effect on the parameter estimates, and ill-conditioning may result in the NLLS regression algorithm. Equation 13 also shows that even if K_{app} is relatively large, there is an additional source of small deviation based on the values of b or D_{app}. When either b or D_{app} (or both) are small, the difference between the quadratic term $\frac{1}{6}b^2D_{app}^2K_{app}$ and the linear monoexponential term $-bD_{app}$ is smaller. If the measurements are limited to small b-values or the true value of D_{app} is small, the algorithm is attempting to assess a deviation that is small relative to the added measurement noise, which may also cause it to be ill-conditioned.

Parameter estimates were already obtained using the biexponential model on a simulated biexponential signal test set, so kurtosis model fits can be performed on this same biexponential test set for comparison. As the kurtosis model is a more parsimonious model, the parameter estimates may have lower bias and/or variance for true biexponential signals that were problematic when fitting the biexponential model. This reduction in parameter uncertainty may offset the bias from fitting the incorrect model to the known truth. Since a monoexponential model is nested within the kurtosis model, also fitting the kurtosis model to simulated monoexponential

signals can demonstrate if the kurtosis model is also susceptible to ill-conditioning when there is no kurtosis present in the true signal. Both the biexponential and monoexponential models were chosen to generate signal data, since these two models have direct biophysical basis in human tissue. As was mentioned in the introduction, the kurtosis model does not have this basis and is considered a model-free approach [173, 174], so it was not used to generate signal data here

3.1.3 Chapter Aims

This chapter will perform an assessment of the kurtosis model parameter estimates similar to the tests run on the biexponential model in Chapter 2.

The aims of this chapter were to:

- Determine the variance in the kurtosis model parameter estimates by testing the model on the same biexponential signal test set from Chapter 2.
- Examine how this variance changes as the true parameter values of the noise-free biexponential signals vary over the parameter space in Figure 6, and compare the results to the biexponential model parameter estimates.
- Determine how the variance in the kurtosis model parameters estimates compares to monoexponential model estimates by fitting both models on noisy measurements derived from signals generated using a monoexponential model.
- Examine the effects on the parameter estimates as the measurement SNR changes by adjusting the magnitude of the simulated noise added to the noise-free signals.
- Assess the effects of Rician bias on kurtosis model estimates.
- Investigate the average condition number across the fits for each noise-free signal to determine when ill-conditioning, if any, was present in a given fit.
- Investigate the results from the parametric bootstrap perturbation analysis and determine its effectiveness in detecting ill-conditioning and large variance in the kurtosis model parameter estimates.

3.2 Methods

3.2.1 Error in Kurtosis Model Parameter Estimates on Biexponential Truths

The kurtosis model was fit to the same simulated biexponential data test set in Section 2.2, using the same parameter space and diffusion weightings (0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6) to create 50,000 noise-free signals. Additionally, the 200 noisy measurements for each noise-free signal were used, but only the test sets for SNR values of 25 and 200 were used in this section. Fitting of the kurtosis model to all simulated noisy data was again performed using a NLLS algorithm (*lsqcurvefit* in MATLAB) with a trust-region-reflective optimization option. Equation 13 was used as the regression model, the amplitude S_0 estimates were bound in the algorithm to a range between 0 and 2, the D_{app} parameter was bound to a range between 0 and 4, and K_{app} bound between -1 and 2. For each noisy signal, five separate regression fits were performed using random start values. Because certain combinations of the above parameters within those bounds can cause

very large values (>10^{10}) due to the exponential *increase* of the quadratic parameter at high diffusion weightings, the random starting values were bound to a more restricted range – S_0: 0.8 – 1.2, D_{app}: 0.5 – 1.5, K_{app}: -0.2 – 0.2. For each noisy signal, the regression fit with the minimum RSS value was kept and the others discarded. To reduce the effects of the Rician bias on the signal, the bias reduction formula in Equation 24 was applied to all noisy signals before fitting, using the known standard deviation from the SNR used to create the data, plus a small random perturbation since its exact value would not be exactly obtained in a physical DWI measurement.

3.2.2 Condition Number

For all kurtosis model regression fits in this chapter, the final Jacobian matrix returned by the algorithm was also saved and used to calculate the Jacobian condition number introduced in Section 2.1.3.

3.2.3 Rician Bias and Low SNR Rejection Strategy

Testing of the effects of Rician bias was performed similar to the methodology in Section 2.2.3, i.e., for each measurement, the first diffusion weighted data point with an SNR < 2 was removed along with all measurements from higher diffusion weightings. The effects of employing this measurement strategy were compared to the full eleven diffusion weighting strategy for both the kurtosis model and compared to the results in Chapter 2 for the biexponential and monoexponential models.

3.2.4 Bootstrap Analysis

The same limited bootstrap analysis subset of the biexponential test set in Section 2.2.5 was also tested with the kurtosis model. 1000 parametric bootstrap samples were created for the kurtosis model to examine the effects of ill-conditioning, if any, on the parameter estimates.

3.2.5 Error in Kurtosis Model Parameter Estimates on Monoexponential Truths

To create simulated monoexponential data, the same eleven diffusion weightings as Section 3.2.1 were used to create 1000 noise-free signals. The basis for the noise-free signals was Equation 8 with signal amplitude S_0 equal to 1 and ADC randomly and uniformly chosen from a range of 0.05 to 1 for each signal, giving the same possible decay range as the biexponential signal test set shown in Figure 9. For each noise-free monoexponential signal, random Gaussian noise was added per Equation 26 to create 200 noisy magnitude measurement each at SNR $_{b=0}$ values of 25 and 200. NLLS regression fitting was performed using the same bounds and starting parameters as Section 3.2.1

3.3 Results

3.3.1 Variance in Kurtosis Model Estimates to Biexponential Truth

All noise-free signals were grouped in the same 100x100 bin values as Chapter 2 and the parameter estimates for the associated noisy measurements in these bins were grouped as distributions. The standard deviation of these distributions are shown as pseudocolour plots for all three kurtosis parameters, at an $SNR_{b=0}$ of 25, in Figure 40.

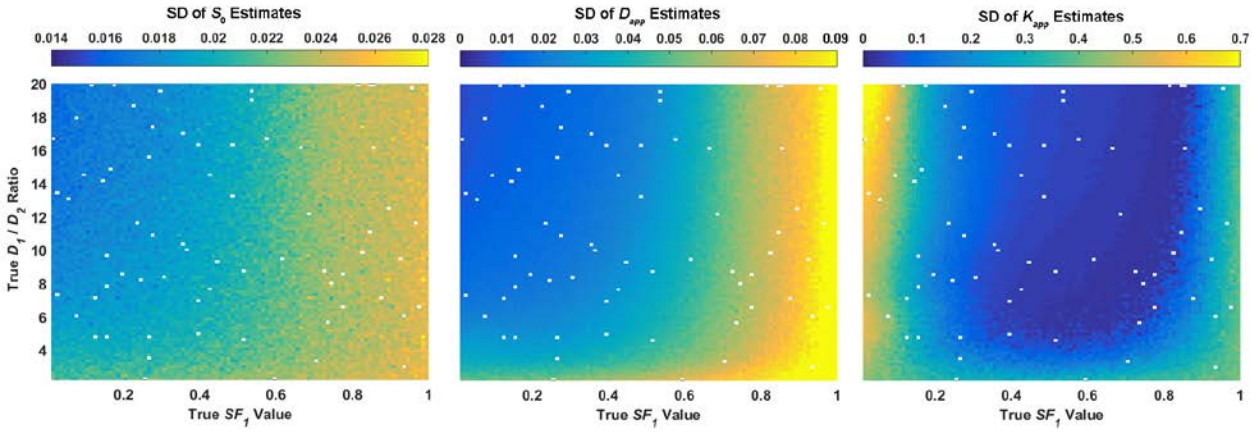

Figure 40 – Standard Deviation for all three kurtosis parameter estimates on biexponential test set at $SNR_{b=0}$ of 25

The SD of the K_{app} estimates is high on the left side of its plot which is not present on either the S_0 or D_{app} estimates.

For the data at an $SNR_{b=0}$ of 25, the SD values for the S_0 and D_{app} parameter estimates of the biexponential test set look very much the same as the deviations in their monoexponential parameter counterparts in Figure 14, albeit an increase in the D_{app} dispersion at the extreme right of the map versus the monoexponential ADC values. The SD in the kurtosis parameter (K_{app}) estimates increase with decreasing SNR on the right side of the map much like the other two parameters, however, there is a large increase on the upper-left side of the graph where the signal-averaged SNR of the biexponential test set is lowest (see Figure 19). This increase in K_{app} SD correlated with an increased condition number in these regression fits as shown in Figure 41, where the left side of the plot contains a large area where all fits in each bin have a condition number greater than 20. This plot also shows a band along the bottom and right edges where 30-50% of the fits have a condition number greater than 20, illustrating that there appears to be conditioning issues in the kurtosis model when the true signal approximated monoexponential decay. This increase in the kurtosis parameter estimate variance was also seen in the biexponential parameter estimates in Chapter 2, and like that model, this increase can still be seen when the test set $SNR_{b=0}$ is 200 as shown in Figure 42. These plots all suggest that the increased SD in the K_{app} estimates was not due to decreased signal-averaged SNR alone and that ill-conditioning also affects the kurtosis model, confirming the hypothesis posited in Section 3.1.2 that this model also has algorithmic issues when attempting to fit signals that are effectively monoexponential.

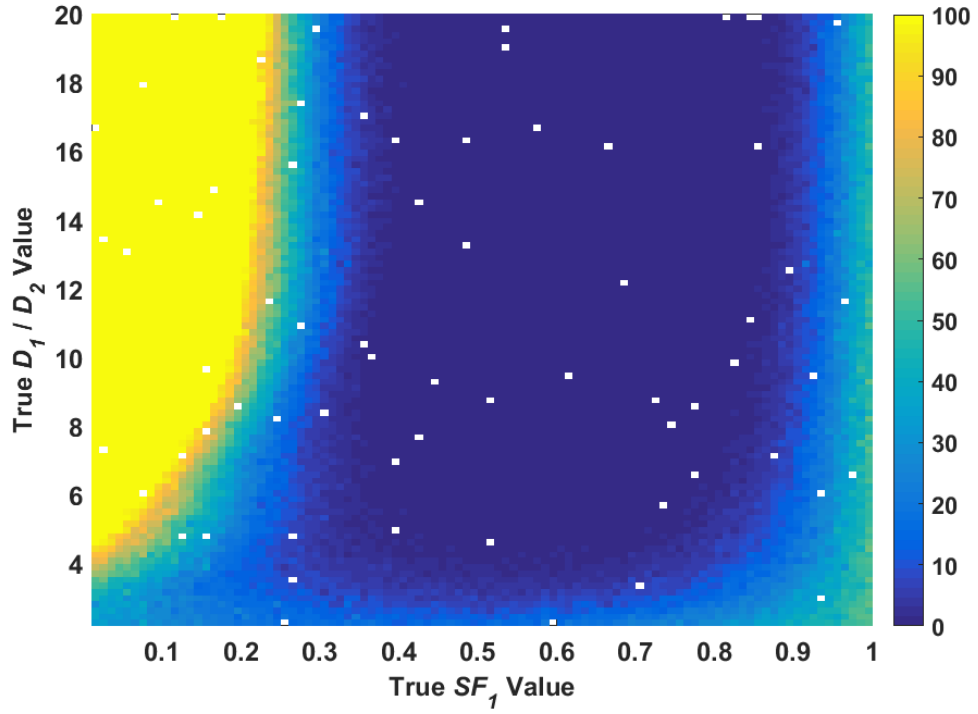

Figure 41 – Percentage of fits with condition number greater than 20 for the biexponential test set at $SNR_{b=0}$ of 25

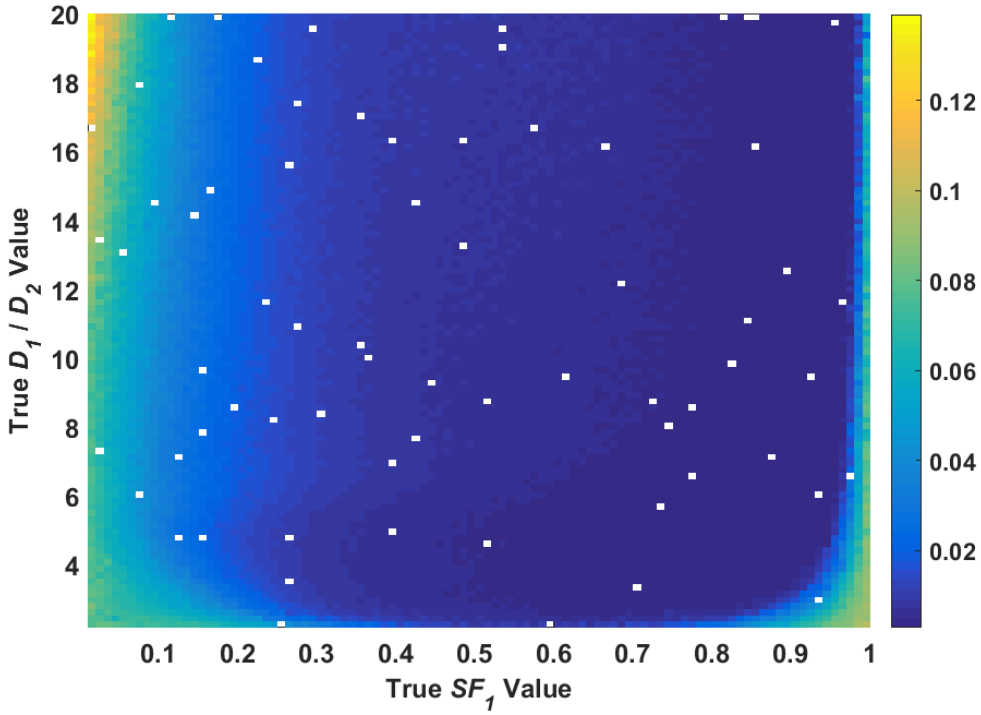

Figure 42 - Standard deviation in K_{app} parameter estimates at SNR of 200

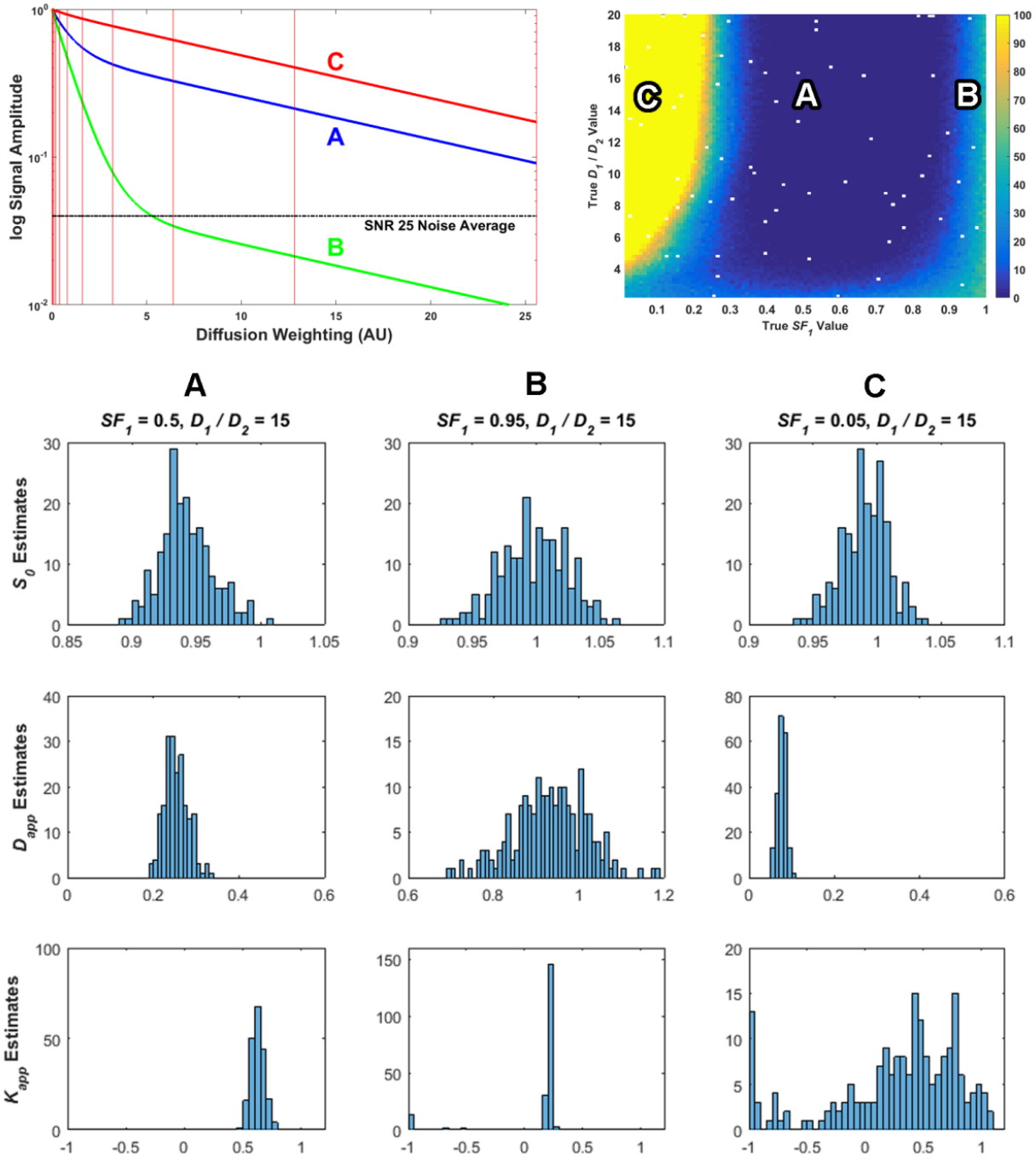

Figure 43 – Histograms of kurtosis parameter estimates for fits to noisy measurements from three noise-free signals (A-C)

The parameter values of the noise-free signals are compared to the number of fits that have condition number greater than 20 (top right, see Figure 41), and their signals also plotted (top left). Signal C has the highest condition number and uncertainty in the parameter estimates, even though it has the highest signal-averaged SNR.

As was illustrated in Chapter 2, when the fits were highly affected by ill-conditioning, the parameter estimates distributions for a given noise-free signal were widely varied and did not have a cohesive, normal-like distribution. The kurtosis model parameter estimates were examined in the same manner using three noise-free signals, each with a D_1/D_2 ratio of 15 and individual SF_1 values of 0.5, 0.95, and 0.05 respectively, with the combined parameter estimate histograms from all noisy measurement fits associated with each signal displayed in Figure 43. These histograms show that the S_0 parameter estimates have a similar variance, while the D_{app} estimates increase in variance as true SF_1 increases from 0.05 (signal C) to 0.5 (signal A) to 0.95 (signal B). This corresponds with the increase in the variance of the D_{app} estimates seen when increasing true SF_1 from left to right in the center map in Figure 40, but the distribution of the estimates are still clustered together in a distribution.

There were more severe errors in the K_{app} estimate distributions, however. At true SF_1 = 0.5 (A), where there is very little ill-conditioning indicated in the top right corner plot in Figure 43, the K_{app} estimates are clustered together in a distribution. The true SF_1 = 0.95 (B) estimates are mostly clustered in a distribution around 0.2, however there are several outliers shown at much smaller values, including around 10 values with an estimated K_{app} of -1, which was the lower bound of the parameter in the regression fits. For true SF_1 = 0.05 (C), in the area where all signals had a fit condition number greater than 20, the K_{app} estimates widely varied from -1 to 1.2, with a significant left tail in the distribution and a considerable number of fits clustered near the lower bound of -1. Thus, ill-conditioning in the kurtosis model affected the parameter estimates when fitting simulated DWI data, however large uncertainty only appeared to affect the kurtosis parameter K_{app}.

3.3.2 Bootstrap Samples

A single, noise-free biexponential signal (true SF_1 = 0.025, D_1/D_2 = 15.05) from the smaller bootstrap test set was also selected from the kurtosis model fits similar to Signal C in Figure 43. One fit was selected from the 25 noisy measurements derived from that noise-free signal, with kurtosis model parameter estimates of S_0 = 0.97, D_{app} = 0.071, K_{app} = 0.068. The 1000 bootstrap samples from that fit were grouped together and plotted as histograms in Figure 44. The histograms for the S_0 and D_{app} bootstrap samples show well-formed distributions that are Gaussian-like and centred around the original signal estimates of 0.97 and 0.071 respectively. The K_{app} bootstrap samples, however, show a widely dispersed distribution with many values found at the lower bound of -1. Confirming the results in Chapter 2 with the biexponential model, performing a perturbation analysis with the parametric bootstrap also effectively identified ill-conditioning in a model fit and high variance in the parameter estimates. Any variance measures derived from the bootstrap samples could indicate problematic fits, as could an examination of the shape of the distribution.

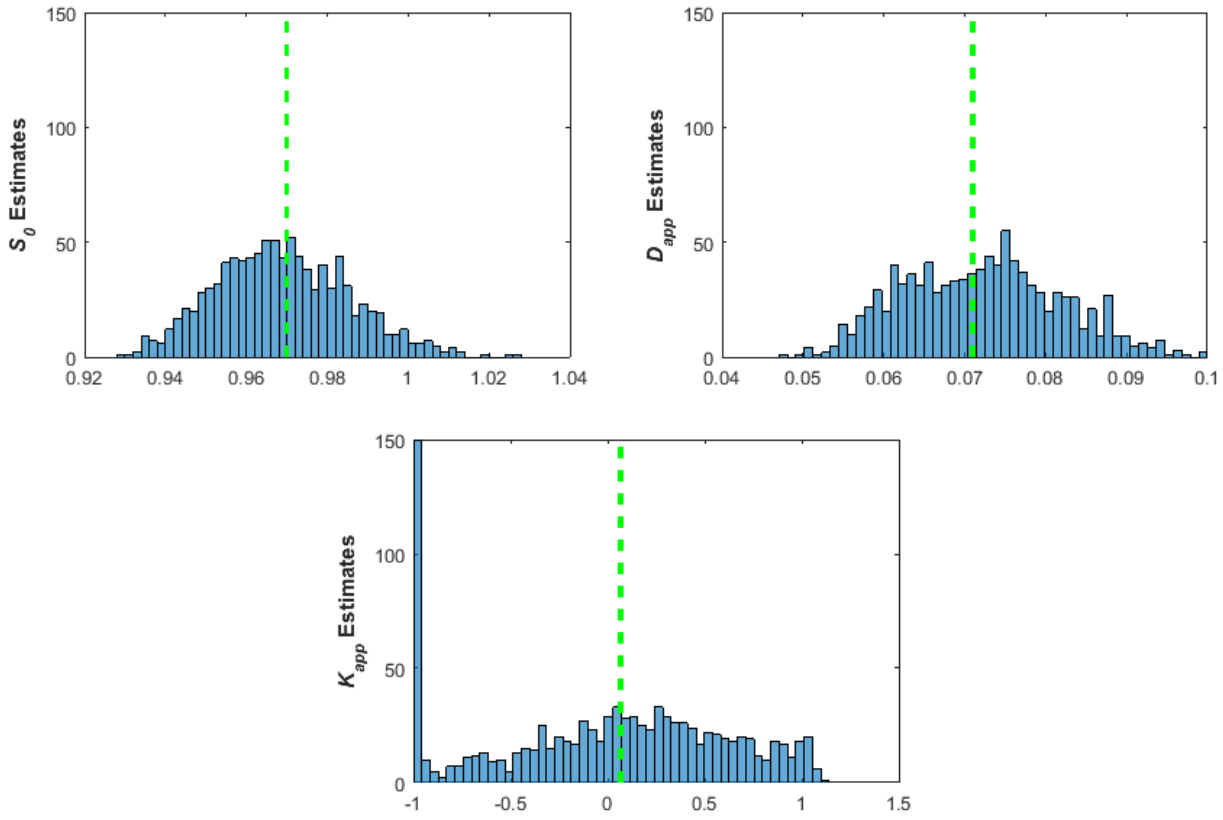

Figure 44 – Histograms of kurtosis bootstrap samples for each model parameter along with the estimated parameter value from the original fit (green dashed line)

The bootstrap distribution show that S_0 and D_{app} have increased uncertainty in their estimates, but not to the degree that K_{app} does, with many values found at the lower bound of -1.

3.3.3 Low SNR Data Rejection

The methodology of rejecting low SNR values improved the overall, signal-averaged SNR of the biexponential test set (Figure 20), but it was found that little improvement was made in the monoexponential parameter estimates and there was a significant degradation in the biexponential parameter estimates. The kurtosis model was fit to the biexponential test set using this same method of rejecting any signal measurement below 2 times the noise standard deviation ($SNR_{b=0}$ = 25). The variance in the parameter estimates from those regression fits were mapped over the parameter space of the biexponential signal test set and compared to the estimates from fitting all eleven diffusion weighted measurements. In these results, the change in variance of the S_0 and D_{app} estimates between the SNR < 2 rejection and full fit strategies was negligible (< 0.005), like the monoexponential model, but there was a slight improvement in the variance of the K_{app} estimates. Figure 45 shows a plot of the difference in SER between the SNR < 2 rejection strategy and a fit of all eleven diffusion weightings. This plot shows that in the portion of the parameter space where there were one or two measurements removed (bottom left quadrant of plot), the SER decreased using an SNR < 2 strategy. However, for the area where the most measurements are removed (right

side of plot), in the hope of improving the signal-averaged SNR, the SER actually *increased* when discarding data points. A negligible parameter estimate benefit and worse overall fitting ability suggested that a variable SNR < 2 rejection strategy did not significantly improve the results when fitting a kurtosis model on this biexponential test set.

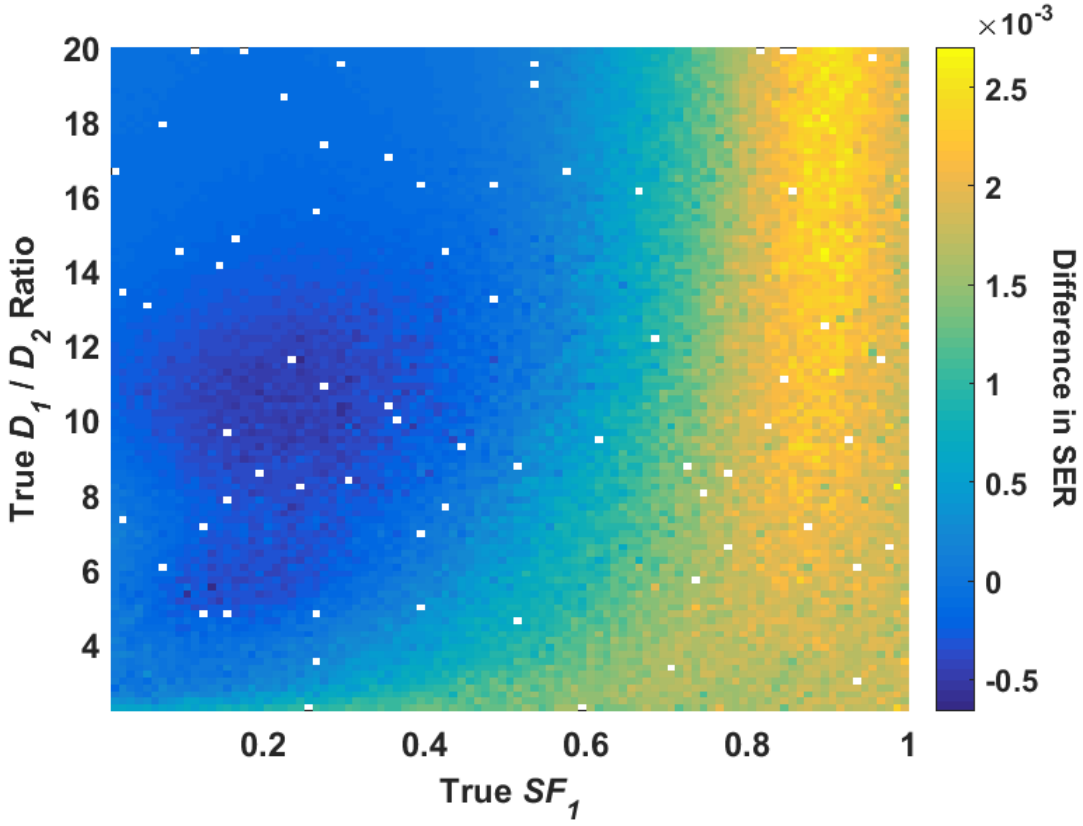

Figure 45 – Change in SER utilising SNR < 2 rejection strategy versus fitting all 11 diffusion weightings

Negative value indicates improvement in fitting ability of SNR < 2 strategy.

SER

The mean SER of all kurtosis regression fits to biexponential truth is shown in Figure 46. This plot is similar to Figure 16, where the regression fits were able to closely fit the biexponential signal when it was most like a monoexponential signal, but at high D_1/D_2 ratio and near-equal signal fraction, the regression error increased. The kurtosis fit does have more flexibility as the highest SER for the regression fits is around 0.09, compared to 0.11 SER when fitting a monoexponential model (Figure 16), showing that the flexibility of the extra term in the kurtosis model allows for an improvement of fitting on a true biexponential signal. However, this increase in SER versus the biexponential model indicated that the kurtosis model was not as flexible as the biexponential model which had an SER over the entire test set of 0.042. This showed that there was still considerable model bias when fitting a kurtosis model to typical biexponential based signals. The kurtosis model has a larger area that is relatively free of ill-conditioning compared to the biexponential model (compare Figure 24 and Figure 41), though, but there is still a significant area

of large uncertainty, so the kurtosis model is also not robust over the entire range of signal measurements derived from this biexponential test set. Furthermore, the kurtosis and biexponential model both have these problems for measurements of signals that are effectively monoexponential.

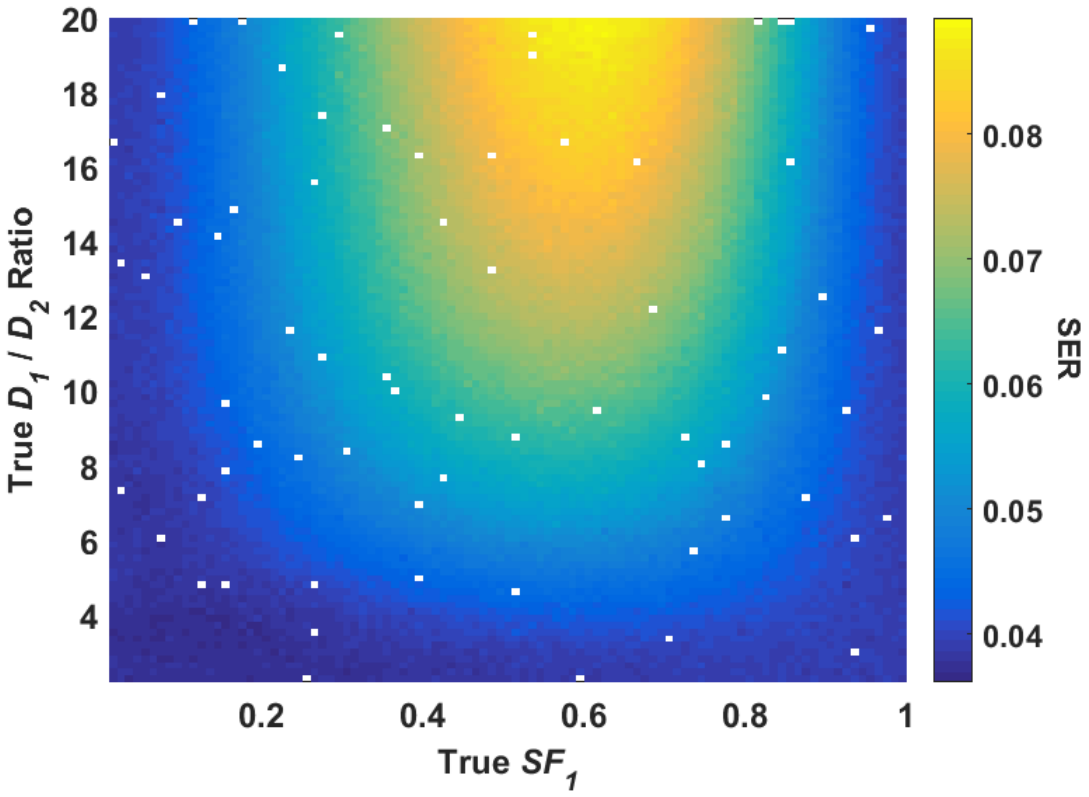

Figure 46 – Mean SER for all kurtosis fits on biexponential signal test set for $SNR_{b=0}$ = 25

The kurtosis model fits the biexponential data better than the monoexponential model when compared with Figure 16, however, it also does not fit the data as well when the signal has equal signal fraction and high D_1/D_2 ratio.

3.3.4 Fitting a Kurtosis Model to Monoexponential Truth

All noise-free monoexponential signals were sorted into an array of 95 bins based on true *ADC* value. Signals with true *ADC* close to 0.05 were similar to signal C in Figure 43 as they remained above the noise floor for all eleven diffusion weightings. Signals with true *ADC* closer to 1 were similar to signal B in Figure 43 as they decreased rapidly at lower diffusion weightings and the remaining four measurements were biased by the noise floor. The kurtosis model parameter estimates from all noisy measurements, based on the signals in each bin, were combined as a distribution with the variance calculated for each bin. Figure 47 displays histograms of the variance in these estimates along with the mean condition number, and show that as signal *ADC* increased, the standard deviation in the S_0 and D_{app} parameter estimates also increased, since more diffusion weighted measurements fell below the noise floor and the signal-averaged SNR of the

signal measurements was lower. The trend in the K_{app} parameter estimates was the opposite, however, as the standard deviation *decreased* as the signal-averaged SNR decreased. As the plot of the mean Jacobian condition number shows, this was due to the reduction of ill-conditioning in the signal. For these monoexponential measurements, the Rician bias actually had a positive effect on the K_{app} estimates, since it artificially lifted the tail of these measurements enabling the kurtosis model to fit the data with less ill-conditioning. However, even with the beneficial effect from Rician bias, the minimum standard deviation of the K_{app} estimates was around 0.4, which was quite high compared to the true value of zero, and was also high compared to the estimates returned across much of the biexponential test set shown in Figure 40.

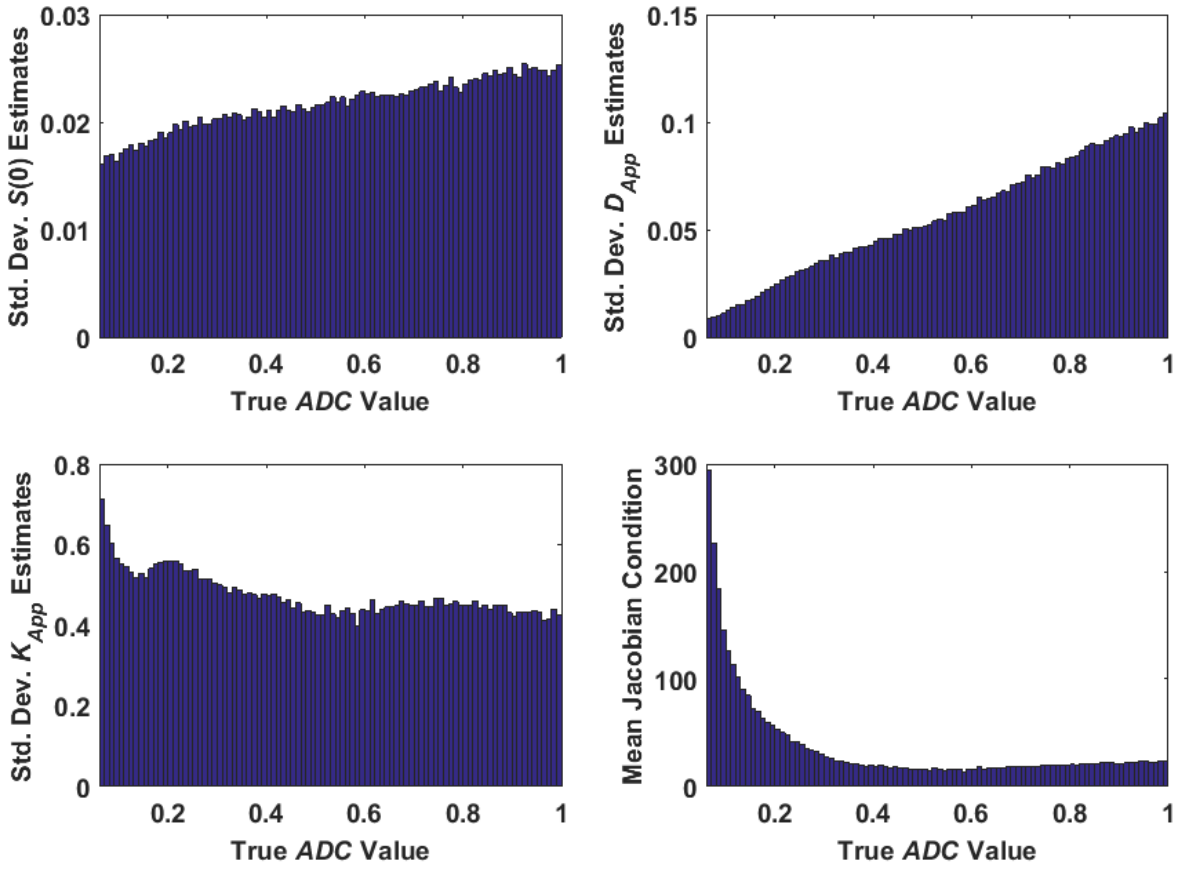

Figure 47 – Histograms of kurtosis model parameter estimates to a monoexponential test set

While the S_0 and D_{app} parameters have higher SD as the true *ADC* increases (which lowers the signal-averaged SNR), this is not the case with K_{app} which actually decreases as the SNR decreases. This is most likely due to the lower ill-conditioning in the fits, indicated by the mean Jacobian condition number plot.

3.4 Summary of Conclusions

This chapter presented an analysis of the kurtosis model and its use with NLLS regression algorithms on simulated data, specifically data generated from both the monoexponential and

biexponential models. The results addressed gaps in the literature specifically described in Section 3.1.3, namely:

- Large variance was found in the kurtosis model parameter estimates when fitting to a biexponential test set, specifically when the biexponential signal was close to being monoexponential. However, this large variance was confined to the parameter that assesses the kurtosis (K_{app}).
- These high levels of variance were found in this kurtosis parameter as the true biexponential signal was close to monoexponential, similar to the findings for all biexponential model parameter estimates in Chapter 2.
- Again like the biexponential model, while increased SNR alleviated the large variance in the estimates for some signal measurements, at effectively monoexponential signals, there was still significant variance.
- This large variance was isolated to the effects of ill-conditioning, as evidenced by large variance in the parameter estimates even at the highest levels of SNR, as well as a large condition number.
- The parametric bootstrap was also able to identify measurements with significant uncertainty due to ill-conditioning.
- Implementing a SNR < 2 rejection strategy did not have a significant improvement on the kurtosis parameter estimates and actually increased the fitting error in many cases.
- Testing the kurtosis model on a test set generated from all monoexponential signals showed that the Rician bias actually had a beneficial effect on the kurtosis parameter estimates, since it artificially induced a deviation from a straight monoexponential decay signal where the ill-conditioning was highest.

Chapter 4
Model Selection Using Information Criteria and Cross-Validation

4.1 Introduction and Background

The previous two chapters demonstrated the magnitude of the bias and variance in both the biexponential and kurtosis models parameter estimates, such that reliable inference could not be made about the underlying true signal. For both models, the uncertainty in these estimates was greatest when the simulated true signal was most like monoexponential decay. Applying a biexponential or kurtosis model to a monoexponential signal is then a *misspecifcation* of the model, since the monoexponential model is obviously the most appropriate model to apply to the data in this situation. However, when assessing a set of tissue data, the best model for a data set is unknown prior to measurement and analysis, and often this best model can change on a voxel-wise basis [37]. Additionally, the true voxel signals may have dozens of dimensions far beyond the simple structure of the DWI models presented thus far, so the best model to apply would likely not be obvious. Thus, the basis for the DWI model selection studies presented in Section 1.5 was to demonstrate to researchers what model is likely to provide the lowest MSE parameter estimates in future studies of a particular tissue or condition. If a more complex model is indicated as the best model in a study, and demonstrations of the additional information it provides are often included (e.g. [87]).

4.1.1 Model Selection Uncertainty

While the premise behind model selection is to obtain the best model that provides the lowest error between model and data *over repeated sample measurements*, little consideration is taken into account on the effects of noise on the model selection process. Burnham and Anderson's book on model selection, however, specifically notes, "…a different model (in the fixed set of models considered) may be selected as best for a different replicate data set arising from the same experiment." (Section 1.7, [62]). For example, 100 sample measurements are obtained from the same voxel, and although the biexponential model was selected as best for the first sample, the monoexponential model was selected as best for 95 of the 100 total samples. Given this information, lower MSE would be achieved over all samples by using the monoexponential model for future data acquisitions on this voxel. However, what if the biexponential model selection rate over repeated samples of a different voxel was 50%, with no clear best model in this case? Which model should be used in this scenario? Model selection criteria don't have the capability to assess future measurements, and the high cost of additional DWI acquisitions means that researchers have to do the best with the limited information they have. To gain better insight into the selection rate between two models over the long run, however, DWI model selection studies are often conducted by combining the number of times each model is selected over multiple voxels in an ROI or organ. But what would be the effects on future studies when the biexponential model is selected as best for 51% of these voxels versus 100%? Knowledge of this information would be a useful addition to the DWI literature.

The appeal of using the AIC in model selection, for example, is its ability to assess which model provides the most *information* on the data because it has the minimum K-L distance to the truth. Yet, the LS regression formula for the AIC in Equation 30 has the value of RSS in it, and as was seen in Chapter 2, the SER, derived from the RSS, changed depending on the signal-averaged SNR as well as the number of diffusion weightings. As was also shown with the biexponential signal test set in Chapter 2, while the SER for the biexponential model fits in Figure 17 stayed relatively constant over the entire test, the SER for the monoexponential model fits in Figure 16 changed depending on the true signal, and when the true signal was most like a monoexponential signal, the SER for both the biexponential and monoexponential model fits had similar values. Similar values of SER would mean the RSS values would be similar, and therefore, the two models would fit the same signal similarly. In this case, it would be expected that when comparing model fits on repeated noisy measurements samples from one signal, the selection rate of the biexponential model as best would be different than fits on repeated samples of a biexponential signal where the monoexponential model SER was much higher (i.e., one with equal SF_1 and high D_1/D_2 ratio).

What researchers need to know then is what the cost in MSE is when a model is misspecified, i.e. when a biexponential model is applied to a monoexponential signal or vice versa. In terms of the ability to fit a signal, there may appear to be no cost, but as was demonstrated in Section 2.3.2, the CV in the biexponential model parameter estimates when assessing an effectively monoexponential signal was nearly 7.5 times higher than the highest CV when assessing the monoexponential model to the biexponential signal. Chapter 2 also demonstrated that measures based on goodness-of-fit had no ability to detect unreliable parameter estimates. Therefore, relying solely on model selection methods to pick the best signal may come with a hidden cost in lower parameter estimate reliability. A study that relates the results presented thus far in this thesis on uncertainty in the parameter estimates to the effects of uncertainty in model selection would also be a useful supplement to the literature.

4.1.2 Akaike Information Criterion

The theoretical definition of a monoexponential signal uses Equation 9 as the generating model, whereas a signal using Equation 12 with SF_1 = 0.001 is theoretically a biexponential model, but the signal is effectively monoexponential. If one of these models is used to create a signal, with the addition of measurement noise, identifying which of these two models is the true basis model becomes extremely difficult. The effect of measurement noise compared to the parameter penalty and the difference in AIC values can be formulated using Equation 30 for the AIC of two arbitrary models *x* and *y*,

$$\Delta\mathrm{AIC}_{x-y} \equiv \mathrm{AIC}_x - \mathrm{AIC}_y = n\log\left(\frac{\mathrm{RSS}_x}{n}\right) - n\log\left(\frac{\mathrm{RSS}_y}{n}\right) + 2\left(k_x - k_y\right), \text{then}$$

$$\Delta\mathrm{AIC}_{x-y} = n\log\left(\frac{\mathrm{RSS}_x}{\mathrm{RSS}_y}\right) + 2\left(k_x - k_y\right). \tag{42}$$

The expectation is that the AIC is "tuned" to make the distinction between a theoretical biexponential model and a monoexponential model, but it is easily shown that this is not the case

using Equation 42, assigning model *x* as the monoexponential model and model *y* the biexponential. As a true biexponential signal approximates monoexponential decay, the biexponential and monoexponential models fit this signal equally, so RSS_x/RSS_y goes to 1 and the first term in Equation 42 becomes zero. When this happens, the value of ΔAIC is 4, with the monoexponential selected as the best model, and this bias occurs regardless of SNR. The other two model comparison combinations, kurtosis vs. monoexponential and biexponential vs. kurtosis, both have a ΔAIC of 2 in favour of the simpler model when the fits are effectively equal. Because this parameter bias exists even with no added measurement noise, the decision boundary between the monoexponential and biexponential models is different in a theoretical sense than in a statistical/information-theoretic one.

Equation 42 can also be used to demonstrate the AIC selection rates between the biexponential and monoexponential models that are equidistant (K-L) to a measurement of a true signal, with their AIC values equal and so ΔAIC between the two models is zero. If this measurement has eleven diffusion weightings, using Equation 42, the biexponential and monoexponential models will equally fit this measurement when the ratio of RSS_x (mono) to RSS_y (biexp) is approximately exp(4/11) or 1.44. This means that when the AIC values are equal, the monoexponential model fit to the measurement has a slightly larger error than the biexponential model fit, which compensates for the parameter penalty. For this measurement of an unknown signal where the ΔAIC (biexponential – monoexponential) is zero, if AIC values on these two models are compared to repeated noisy measurements of the same signal, acquired at the same SNR, it is plausible that for some measurements the biexponential model will fit better, and for others, the monoexponential model. Thus, the selection rate of the biexponential model across all samples of the signal won't be 100%, and if the ΔAIC values for these fits are distributed equally around zero from added noise, the selection rate for each model will be 50% – an equal chance.

While the AIC values of future samples can't be estimated from a single measurement with unknown truth, examining the ΔAIC value between the best model and the other(s) provides an additional measure of the strength of inference. Burnham and Anderson specifically give an estimated scale that relates the strength of inference of a model to the value of ΔAIC, with a value greater than 10 indicating that the selected model has a "substantial level of empirical support" versus the other tested model(s) (Section 2.6, [62]). When using only the lowest AIC to determine the best model, a difference in AIC values of -0.001 and -100 carries the same inferential weight on model selection. Compared to a ΔAIC value of 0.001, a ΔAIC value of 100 between the biexponential model (selected best) and the monoexponential means that future measurements of the same truth are more likely to have the biexponential model selected as best. A recent DWI study [175] demonstrated that the ΔAIC_c values were large between the combinations of four tested models (biexponential, monoexponential, stretched exponential, kurtosis) when measurement noise added to simulated signals was low, but as the noise increased, the AIC_c values of all four models converged together, so the ΔAIC_c values between models went to zero, and it was more difficult to distinguish models. This demonstrated that values of ΔAIC_c (and likely, ΔAIC) are sensitive to noise and as these values are smaller, it is more difficult to distinguish models.

4.1.3 F-test

The *F*-test is a statistical test that compares two nested models with the null hypothesis being that there is no difference between the models. The formula for comparing two nested models (e.g. biexponential and monoexponential) based on their RSS values from a given LS fit is [176]

$$F = \frac{\left(\frac{RSS_x - RSS_y}{RSS_y}\right)}{\left(\frac{k_y - k_x}{n - k_y}\right)}, \tag{43}$$

with n the number of measurements, k the number of parameters for that model, and x always set to the simpler model. The value of F is calculated for the fit of the two models on a given signal, and will have an *F*-distribution with $(k_y - k_x)$ and $(n - k_y)$ degrees of freedom. If the value of F is greater than the critical value of the *F*-distribution with these degrees of freedom at a given level of significance (usually 0.05), then the null hypothesis is rejected at that significance level. While the *p*-value, and its decree of "significance" when meeting an arbitrary criterion, is being widely criticised in the scientific literature at present (e.g. [177]), it does provide a way of assessing an error rate. One of the aims of this thesis was determination of the reliability of model selection methods, including type I, false positive errors, so it is more important to analyse what a significant result of an *F*-test means in terms of error rates. $p < 0.05$ indicates that for less than 5% of the time, the *F*-test will erroneously reject the null hypothesis that there is no significant difference in the models. In terms of DWI multimodel analysis, this would mean that given the null hypothesis is true, for 5% of measurements the biexponential model would be falsely selected as being significantly better than the monoexponential model.

This is one advantage that the *F*-test has over the AIC, since the AIC has no such error specification or indication of its selection reliability [178]. A certain amount of error in model selection methodology should be expected, with a scientific goal of minimising such error [179]. The worst case method, when making a decision between two models, is a random, 50/50 coin flip. The previous section illustrated a case where two models fit a measurement equally, and the AIC effectively gives a 50/50 chance in selecting the best model. Assessing a minimum ΔAIC value then is similar to a *p*-value, with a recent paper [180] demonstrating the similar analysis properties of the *p*-value, confidence intervals, and ΔAIC, and how a minimum ΔAIC value "breaks ties" in the selection process. This paper also demonstrates, though, that the choice of a minimum ΔAIC value is as arbitrary as the selection of significance level (e.g. 0.05) for a *p*-value. It also reviews the various scales presented in the literature, relating a reported ΔAIC value and the strength of inference that provides. This referenced Burnham and Anderson's scale presented in the previous section, upon which they wrote a spirited-yet-puzzling reply that essentially says that *p*-values are worthless and antiquated compared to the "21st century" information-theoretic properties of the AIC [181]. With these conflicting views on the *F*-test in the literature, comparing the AIC and *F*-test selection rates on DWI data would be useful to researchers, since both methods are currently used in DWI analysis.

4.1.4 Additional Selection Criteria

The corrected version of the AIC, the AIC_c, has a rule of thumb which states if the ratio of n/k is less than 40, then the AIC_c should be used [62]. For the four parameter biexponential model, this would require 160 b-value measurements for n/k to be higher than 40, which is unrealistic for an actual DWI acquisition. According to this rule, then, the AIC_c should be used when making inferences in nearly all DWI studies. The correction term of the AIC_c, however, is simply a mathematical bias component added to the calculated AIC value. For a signal with eleven measurements, calculating the additional bias term for the AIC_c found in Equation 31,

$$\frac{2k(k+1)}{n-k-1},$$

gives values for this bias term of 12 for a biexponential model, 6.67 for a kurtosis model, and 3.43 for a monoexponential model. These factors include the additional parameter for fitting the variance in least squares, so k = 5 for the biexponential model, 4 for the kurtosis, and 3 for the monoexponential. For eleven signal weightings, this AIC_c will then add an even larger penalty to more complex models than just the AIC. For the AIC example where the biexponential and monoexponential fit a monoexponential signal equally, resulting in a ΔAIC of 4, the AIC_c adds another 8.57 (12 - 3.43) to this value, giving a ΔAIC_c value of 12.57 at the theoretical boundary of the biexponential and monoexponential models. The added bias term in the AIC_c, then, is basically an additional parameter penalty favouring simpler models based on the number of measurements. As the number of measurements goes up, however, the effect of this correction factor becomes smaller, and eventually the AIC_c value converges to the AIC. The other model selection criterion that adjusts the parameter penalty based on the number of measurements is the BIC, which has an equation of [73]

$$\mathrm{BIC} = -2\log\left(L\left(\hat{\theta}\middle|y\right)\right) + k\log(n). \tag{44}$$

The first term is the same as Equation 29 with the likelihood function, which is equivalent to the $n\log(\mathrm{RSS}/n)$ term for the AIC in Equation 30. The difference between the AIC and BIC, then, is a difference in parameter penalties: $2\times k$ for AIC, $\log(n)\times k$ for the BIC. While the BIC parameter penalty increases with the number of measurements, the value of $\log(n)$ for the range of measurements in the test sets presented in this chapter is between 1.95 (seven measurements) and 2.40 (eleven measurements). For eleven diffusion weightings, then, the difference between a two and four parameter model that fit a signal equally would be 4.8 instead of 4. This would result in a selection rate similar to the AIC for most DWI models, and the BIC has a parameter penalty less extreme than the AIC_c, so its results could be inferred based on the results from the other criteria and it won't be assessed further in this chapter.

4.1.5 Information Criteria vs. Cross-Validation

The AIC and Leave-One-Out Cross-Validation (LOOCV) procedures are asymptotically equivalent in model selection [78], but the effects of limited measurement samples on selection rates of DWI models have not been widely studied. A recent DWI study used LOOCV model selection on ex vivo

data to confirm the rankings produced by the AIC and AIC_c, finding good agreement in selection of the best model between LOOCV and AIC [37]. While the AIC and LOOCV rates converge with a large number of samples, it is unknown which method performs better on a limited number of measurements. Cross-validation methods estimate the error across all measurements as opposed to an approximate penalty like the AIC, so it's possible for it to be more reliable for small samples. Conversely, if there are only seven diffusion weightings, leaving one measurement out will be a significant loss of information compared to the overall signal, causing significant errors in the selection rates. The AIC has already shown in this chapter to be mathematically biased at the theoretical boundary between a monoexponential and biexponential model, so it may be possible for LOOCV to have an improved selection rate near this boundary. The LOOCV also returns arbitrary values for each model and like the AIC, the boundary between selection of models is a point estimate with no assessment of error rate or strength of inference of the best model. The LOOCV also comes with a cost of increased computation time, and this trade-off should also be considered when comparing model selection methods.

4.1.6 Effects of Rician Bias and Number of Diffusion Weightings

It's been demonstrated previously in this thesis that having more measurements with lower SNR meant that the increased noise and Rician bias increased the uncertainty in the some of the model parameter estimates and reduced it in others. While complex phased data can be used to avoid Rician signal bias, this causes problems by inducing phase artefacts in the image, so common practice is to use magnitude data [54]. To avoid the effects of the Rician signal bias, the diffusion weightings of a signal acquisition strategy are often limited to low *b*-values to keep the SNR at each data point high, i.e. 5 or greater. A strategy of removing any measurements with an SNR less than 2 has been used in this thesis, but this changed the number of measurements used for fitting across the entire test set, giving a range of measurements from seven to eleven for the biexponential test set. Different diffusion weightings means the RSS value changes over the entire test set, which would affect the AIC (and AIC_c), as well. Additionally, the LOOCV would change based on differing the number of measurements over the test set. To maintain consistency, the model selection methods will be tested for two different numbers of diffusion weightings, seven and eleven. These weightings will be the same as the simulated biexponential and monoexponential tests used earlier, and as Figure 9 shows, using only the first seven diffusion weightings means that all signal SNR values will be greater than 5.

Using all eleven diffusion weighting will examine cases where signals have some measurements in the noise floor, as well as some signals that remain well above the noise floor. Limiting measurements to the lowest seven diffusion weightings will remove the effects of the noise floor altogether on the model selection process. However, if the signals are limited to lower weightings, there will be less divergence from a monoexponential signal, and therefore, the resolvability of the biexponential and kurtosis models using model selection methods decreases. While a biexponential model will fit a true biexponential signal better than a monoexponential model, the biexponential signal will fit much closer at higher weightings, such that the RSS value for a monoexponential model will be much larger. Limiting the signal to low weightings means the RSS values for both

models will be similar, and with the parameter penalty bias, the monoexponential model will be selected more often than with the higher weightings.

4.1.7 Chapter Aims

This chapter focuses on providing researchers with a detailed picture of how the most common selection methods are affected by measurement noise as well as varying DWI acquisition parameters. More importantly, this chapter will also focus on what the cost is when these methods, and the uncertainty in them, misspecify a model, leading to unreliable parameter estimates. This information will help researchers in avoiding these instances, and improve the reliability of their parameter estimates. To accomplish these goals, this chapter used the simulated biexponential signal data from Chapter 2 as well as the simulated monoexponential signal data from Chapter 3. The model selection rates can then be related to the uncertainty in the parameter estimates from these previous chapters to present a comprehensive data set on how model selection affects model parameter estimates.

The aims of this chapter were to:

- Determine the differences, if any, in how the AIC, AIC_c, and LOOCV each select the best model over repeated measurements from simulated signals where the truth is known.
- Examine how these selection rates of different models change as the true parameter values of the simulated signals vary.
- Assess the effects of varying the SNR of the added measurement noise to determine any changes in the selection rates among the model selection methods.
- Compare the changes, if any, to the selection rates when the number of diffusion weightings is reduced to increase the signal-averaged SNR and avoid Rician bias.
- Do a direct comparison of the AIC and *F*-test to compare their selection rates across the same measurements.
- Compare the effects of the additional bias correction factor in the AIC_c and compare its selection rates directly with the AIC.
- Investigate how the difference in AIC scores (ΔAIC) between two models compares with the selection rate of these two models on fits from a set of measurements from one signal.
- Compare the selection rates of the tested model selection methods and measures with the occurrence of significant ill-conditioning in the regression fits and large uncertainty in the parameter estimates.

4.2 Methods

4.2.1 Simulated Biexponential Signal Test Set

To compare the selection rates of the AIC, AIC_c, and LOOCV on simulated biexponential truth, Equation 12 was used as the generating model for 4900 noise-free signals, with 70 discrete SF_1 parameter values evenly spaced from 0 to 1, and 70 discrete D_2 values evenly spaced from 2 to 20.

A_0 and D_1 were set to 1 for all signals, and two test sets were created for both seven and eleven diffusion weightings (again in arbitrary units):

7: (0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6)

11: (0.025, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8, 25.6)

The eleven diffusion weightings set uses the same values as in Chapter 2, while the seven diffusion weightings set merely has the last four weightings removed. For both of these test sets, 200 noisy signals were created for each of the 4900 noise-free signals with simulated Gaussian noise added per Equation 26 at SNR values of 25 and 200, resulting in the test set matrix in Table 6.

Table 6 – Biexponential Test Set for Model Selection

	SNR 25	SNR 200
7 diffusion weightings	4900 x 200 noisy signals	4900 x 200 noisy signals
11 diffusion weightings	4900 x 200 noisy signals	4900 x 200 noisy signals

4.2.2 Simulated Monoexponential Signal Test Set

The monoexponential test set has the same 1000 parameter combinations from Section 3.2, using Equation 8 as the generating model for the noise-free signals, with signal amplitude S_0 equal to 1 and *ADC* randomly chosen from a range from 0.05 to 1 for each signal. These combinations were used to create two sets of noise-free signals with the same seven and eleven diffusion weightings as the biexponential test set above. Simulated noise was also added to the noise-free signals at SNR values of 25 and 200, resulting in the monoexponential test set matrix in Table 7.

Table 7 – Monoexponential Test Set for Model Selection

	SNR 25	SNR 200
7 diffusion weightings	1000 x 200 noisy signals	1000 x 200 noisy signals
11 diffusion weightings	1000 x 200 noisy signals	1000 x 200 noisy signals

The diffusion weightings and signal ranges match the plot shown in Figure 9.

4.2.3 Calculating AIC, AIC_c, and LOOCV from NLLS Regression Fits

Fitting the monoexponential, kurtosis, and biexponential models to the simulated data was performed using the same NLLS regression algorithm as in the previous two chapters (*lsqcurvefit* in MATLAB, trust region reflective option). The model equations used were 9 (monoexponential), 13

(kurtosis), and 32 (biexponential) with the lower and upper bounds on each model set per Table 8. Since the LOOCV required a twelve-fold increase in computation time, a single starting value combination was only used in each fit per Table 8, also. To calculate the AIC and AIC_c values for each signal fit, the returned RSS value from each model was used in both Equation 30 and Equation 31 to calculate the respective values. For the LOOCV value of each model fit of a signal measurement, each diffusion weighted measurement was left out in turn, with the parameter estimates from fitting the other diffusion weighted measurements used to predict the missing value. For all combinations in a given signal measurement, the squared differences between the original signal measurement and the predicted value were totalled, giving the total LOOCV value for each model. For each fit of the noisy measurements from each noise-free signal, the AIC, AIC_c, and LOOCV values were compared for each model, and the lowest value of each method signifying that model as best. Then, for each method, the number of times each model was selected as best was tallied over the total measurements for each signal, and a selection rate for each model calculated as a percentage.

Table 8 – Bounds and Starting Values for NLLS Regression Model Parameters

Amplitude bounds and starting values were set to either the maximum value of a given noisy signal measurement or a multiple of that value.

Parameter	Lower Bound	Upper Bound	Starting Value
A_0 *(mono)*	0	(2x) max signal	max signal
ADC	0	4	1.5
A_0 *(kurt)*	0	(2x) max signal	max signal
D_{app}	0	4	1
K_{app}	-1	2	0.5
A_1	0	max signal	(0.5x) max signal
A_2	0	max signal	(0.5x) max signal
D_1	0	4	1
D_2	0	4	1/6

4.2.4 Calculating ΔAIC and ΔAIC_c Values Between Two Models

For all measurements from each signal, the ΔAIC values for all three two-model combinations (biexponential vs. kurtosis, biexponential vs. monoexponential, kurtosis vs. monoexponential) were calculated for the biexponential signal test set with 11 diffusion weightings, and added measurement noise at an SNR of 25. For the each two-model comparison, both the mean ΔAIC and ΔAIC_c were calculated for the measurement fits for each noise-free signal and directly compared with the selection rate of each model.

4.2.5 Comparing the AIC and *F*-Test

To directly compare the AIC and *F*-Test selection results, the biexponential and monoexponential model were compared on the biexponential test set in Table 6 at an SNR of 25, and the selection rate that each model was chosen as best by both methods compared. For each measurement regression fit, the returned RSS values from both the biexponential and monoexponential model fits were used in Equation 43 with model x set to the monoexponential model and model y to the biexponential model. The value returned by Equation 43 was compared with the calculated critical value of the *F*-Distribution, which was calculated using the *finv* function in MATLAB with the numerator degrees-of-freedom (v_1 or d_1) equal to $k_y - k_x$, the denominator degrees-of-freedom (v_2 or d_2) equal to $n - k_y$, and the significance level (α) set to 0.05.

4.2.6 Diagnosing Ill-Conditioned Fits and Uncertain Parameter Estimates

The difficulty and time involved for a visual examination of the four parameter estimate distributions from the regression fits of all measurements from all 4900 signals was considerable. To make the information presented here useful for researchers, as well, algorithmic tests of ill-conditioning in the fits were needed. The previous biexponential and kurtosis chapters showed that when the estimates were uncertain, the distributions were heavily skewed and/or not well-formed. Assessment of the distribution shape was then performed on all parameter estimate distributions from the measurements of each signal in two ways. The first was to assess the skewness of the distributions by simply subtracting the mean from the median. The second way was using a standard test of normality, specifically, applying a Lilliefors normality test with a significance level (α) of 0.001. If the null hypothesis that the distribution is normal was rejected, this was taken as evidence that there was considerable uncertainty in the parameter estimates.

4.3 Results

4.3.1 Fitting Three Models to Biexponential Truth with Eleven Diffusion Weightings

SNR 25

The percentage rates that each model was selected as best, when fitting the noisy measurements generated by each noise-free signal at SNR = 25, were calculated for each selection method and displayed as individual pseudocolour plots in Figure 48. These plots show that the biexponential model is selected as the best model for most of the test set, including a considerable portion of the test set where 100% of noisy signals are selected as biexponential. However, the biexponential model was *not* selected as the best model over the entire test set, even though it was the basis model for all signals, and the decrease in the selection rate was greatest when the true signal was closest to monoexponential decay. This decrease in selection rate was inversely correlated with the increase in the bias and variance of the parameter estimates seen in Chapter 2, and hence, this can be a beneficial side effect of model selection, since the monoexponential model will be selected more often when the uncertainty in the biexponential model parameter estimates is highest.

The AIC and LOOCV have very similar selection rates of the monoexponential and biexponential models for the test set, confirming earlier findings in the literature noted in Section 4.1.5. The extra bias of the AIC_c toward simpler models was evident here, too, as it selected smaller percentages of the biexponential model as best, while selecting larger percentages of the monoexponential as best. However, the AIC and AIC_c selected the kurtosis model for approximately the same portion of the signals with a selection range between 0 and 40%.

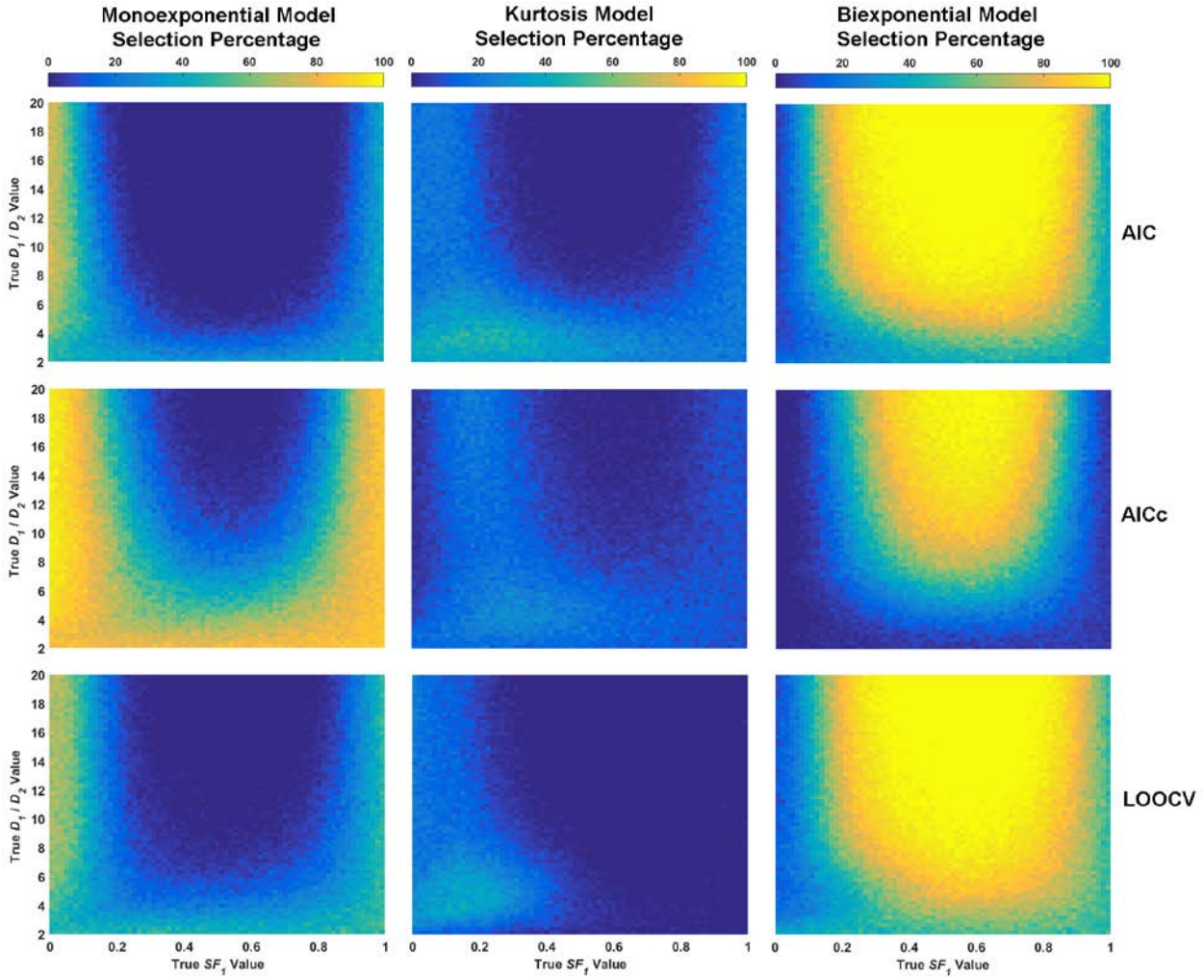

Figure 48 – Selection rate of the monoexponential, kurtosis, and biexponential models as best by the AIC, AIC_c, and LOOCV for a simulated biexponential signal test set with eleven diffusion weightings and noise added at an SNR of 25

For most of the biexponential data, the biexponential model was selected by all three selection methods as best, however, the other two models are still selected as best for some data, specifically as the true signals are closer to monoexponential.

The LOOCV selected the kurtosis model for the left half of the test set plot, but selects very few signals on the right half. After further investigation, this was found to be a complication from the

K_{app} parameter when holding out the largest diffusion weighting (25.6). As Figure 9 shows, this weighting is at the far right of the line plot while the next closest weighting is in the middle of the graph (12.8). If the largest weighting was held out, the value of K_{app} set by the first ten weightings was estimated much higher than when it was included, such that the predicted signal value at the highest diffusion weighting was much greater than the initial value (on the order of 10^4). This caused the total LOOCV value to be much greater than the biexponential or monoexponential values and it was rarely selected. This issue with the non-monotonically decreasing possibilities of the kurtosis model was also described in Chapter 3.

SNR 200

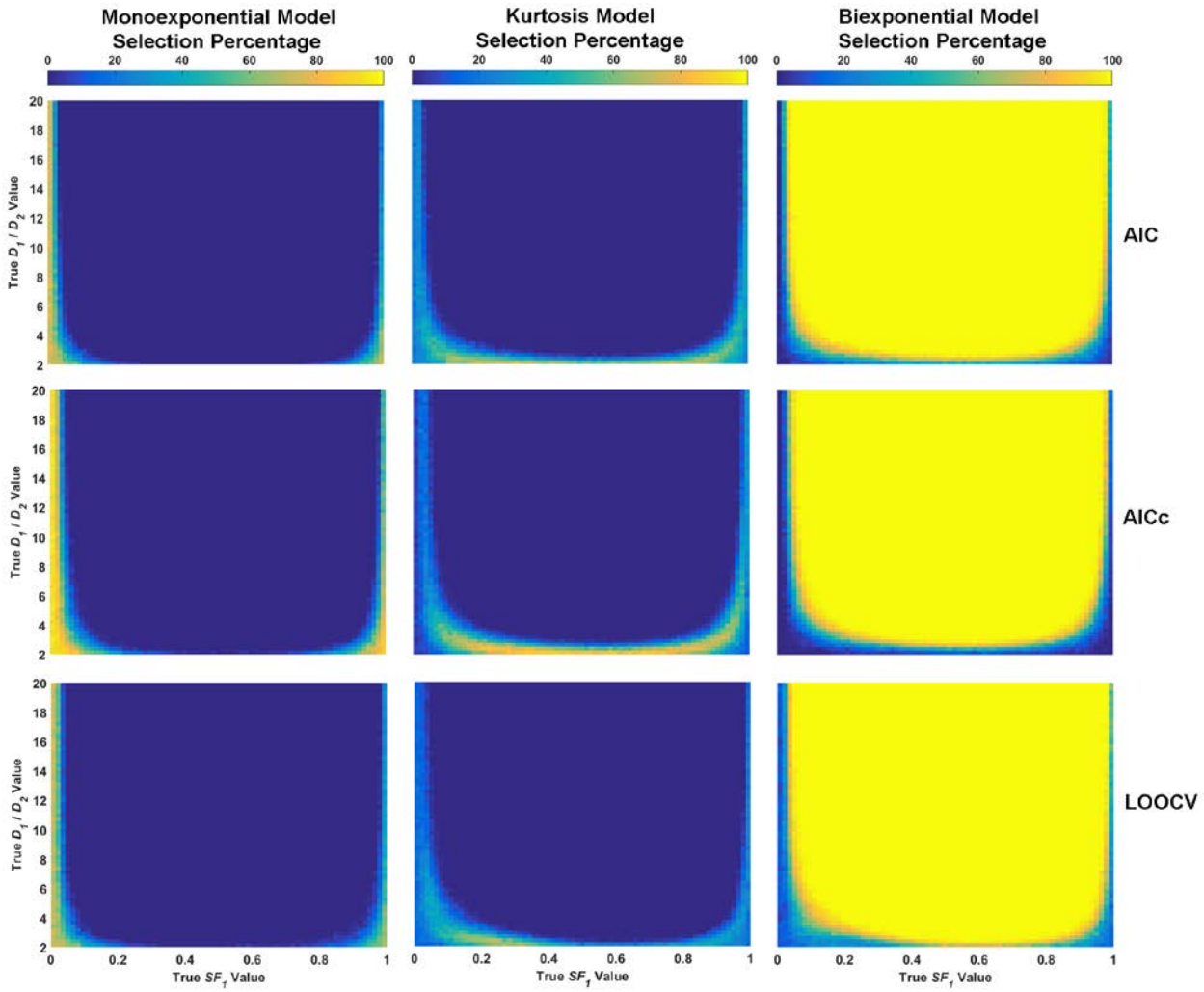

Figure 49 – Selection rate for biexponential test set with eleven diffusion weightings and noise added at an SNR of 200

Increasing the simulated SNR to 200 led to the biexponential model being selected for even more data than when the SNR was 25, however at the margins where the signal is effectively monoexponential, the other two models are still selected.

Increasing the SNR to 200 produced a distinct increase in the selection rate of the true biexponential model as best for all three selection methods as shown in Figure 49. This increase in area of the test set, where the biexponential selection rate is 100%, is similar to the area of reduced parameter error at an SNR of 200 in Figure 10. This shows that an increase in SNR not only reduced the occurrence of ill-conditioning, but also improved the reliability of the selection methods in selecting the true model. However, like Figure 48, when the true signal was effectively monoexponential, the monoexponential model was still selected as best, with the kurtosis model selected as an intermediary model for a portion of the test set. The additional bias of the AIC_c toward simpler models is still evident with a slight increase in the amount of test set that the monoexponential model was selected as best, however, the effect of this additional bias was much less than at an SNR of 25. The selection rates of the AIC and LOOCV methods are also similar here, however the LOOCV selected slightly more of the biexponential model at the extreme bottom corners of the test set plot.

4.3.2 Fitting Three Models to Biexponential Truth with Seven Diffusion Weightings

SNR 25

As shown in Figure 50, reducing the number of diffusion weightings to seven made a dramatic decrease in the reliability of the model selection methods, with the highest selection rate of the biexponential model by the AIC at 38%. The kurtosis model is selected by the AIC about 40% of the time for much of the test set, but at the extreme left of the kurtosis plot, this rate drops to zero. This was due to this portion of the test set consisting mostly of the slower decay component, such that signal remained flat for most of the diffusion weightings, with very little curvature or deviation from a monoexponential decay for the kurtosis model to fit to. The additional bias of the AIC_c obviously had too much of an effect here, since it selected the monoexponential model as best for nearly 100% of the entire test set. LOOCV produced an increase of the biexponential and kurtosis models selection rate over the AIC for this test set, suggesting that the similar performance for these two methods didn't hold for this reduction in the number of diffusion weightings.

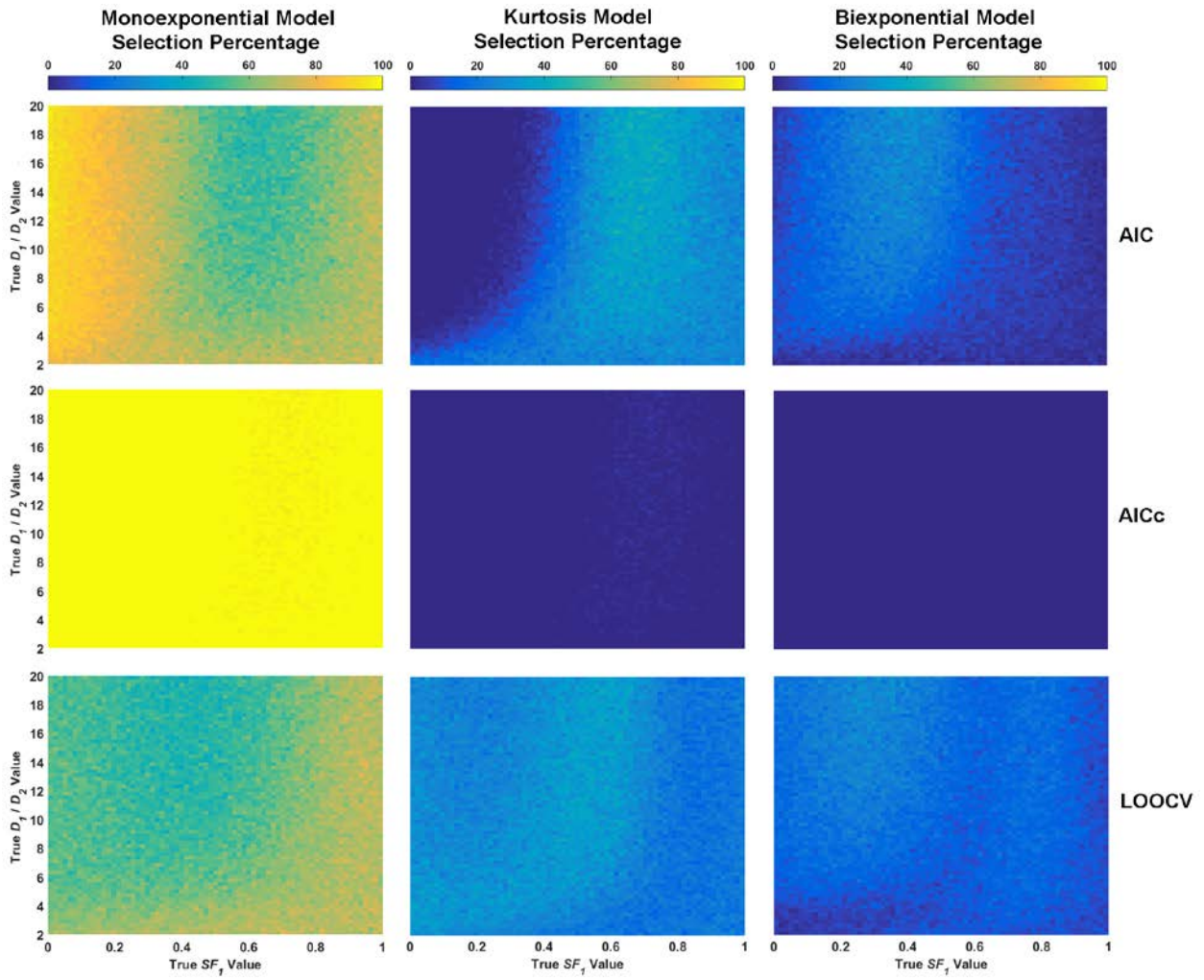

Figure 50 – Selection rate for biexponential test set with seven diffusion weightings and noise added at an SNR of 25

Reducing the number of diffusion weightings to seven led to the biexponential model rarely being selected for data generated by itself.

SNR 200

Increasing the SNR to 200 for the seven diffusion weighting biexponential test set produced different patterns for the model selection rates, as shown in Figure 51. The monoexponential model was barely selected by the AIC and LOOCV methods, except where the signal was effectively monoexponential, similar to the eleven diffusion weighting plots in Figure 49. The kurtosis model was then selected as the best model for most of the test set, except for the upper left portion of the test set, again where the signal was flat and did not decay much. For the AIC_c, this portion of the test set now had the monoexponential model selected as best, and the biexponential model was not selected at all. The LOOCV method selected the biexponential as best for much more of the test set here, suggesting the AIC approximation was not adequate in this case.

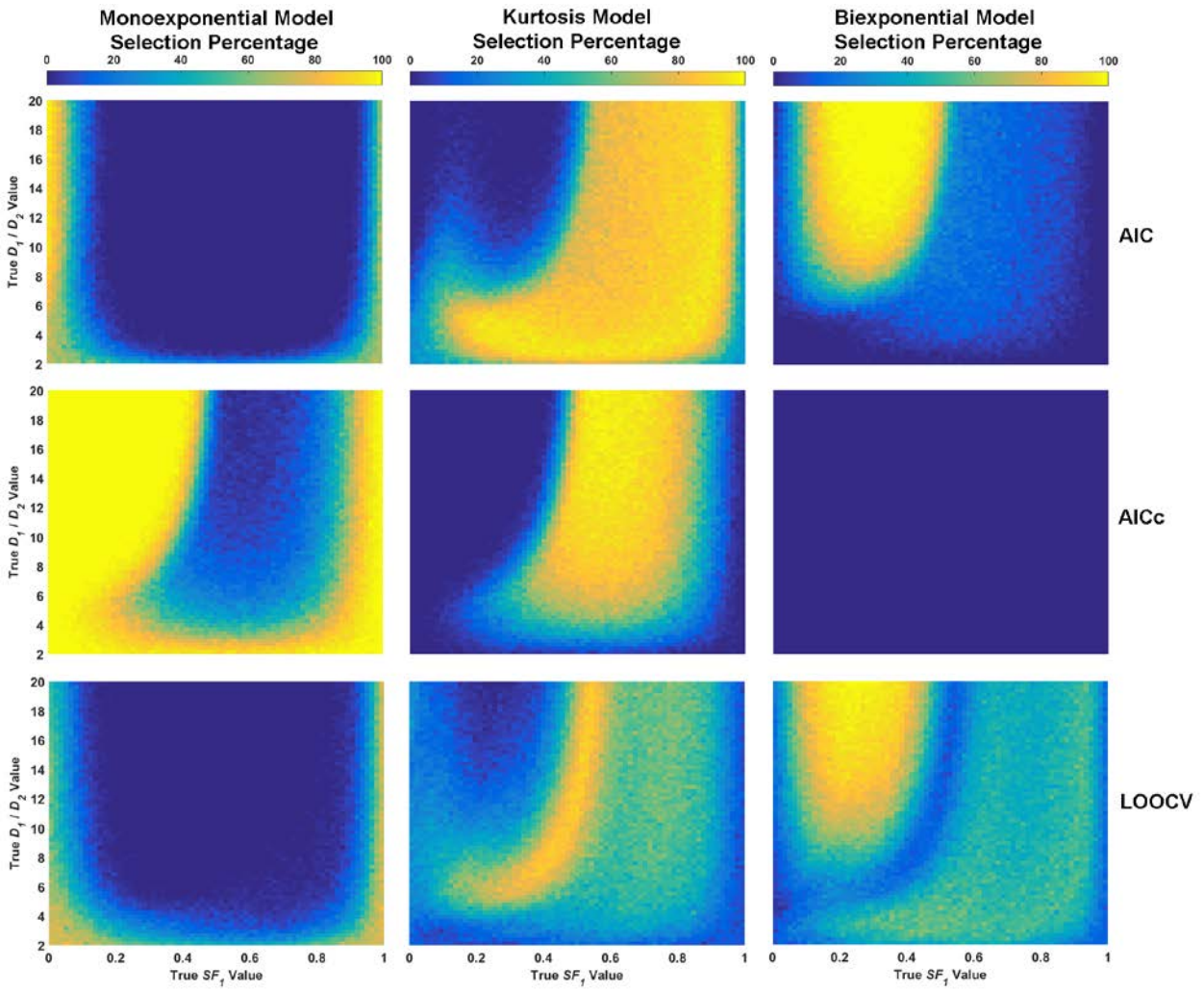

Figure 51 – Selection rate for biexponential test set with seven diffusion weightings and noise added at an SNR of 200

Increasing the SNR of 200 led to the biexponential model being selected for more data than the SNR 25 results, but the pattern of selection does not relate to whether the signals were monoexponential.

4.3.3 Fitting Three Models to Monoexponential Truth with Eleven Diffusion Weightings

SNR 25

With the monoexponential model set as the basis for the signals, the selection rates of each model were displayed as histograms for each method in Figure 52, with the true *ADC* being the only variable compared to the results. Like the biexponential test set results, the monoexponential model was not selected as the best model for all signals, with a maximum selection rate via the AIC of around 70%, which then decreased to 40% as the true *ADC* increased. This decrease in the rate that the monoexponential model is selected as best can be explained by the effects of Rician bias on the signal measurements. For low values of *ADC* around 0.05, where the signal was flat and

remained well above the noise floor, the monoexponential model selection rate was highest. Yet, the selection rate was still below 100%, indicating that there is still a considerable amount of noise affecting the data such that the other two models were selected as best for some measurements. As the *ADC* increased and the signal measurements were lifted by the effect of the noise floor, the kurtosis and biexponential models selection rates increased, since these models are able to better fit the curvature of these measurements. This was similar to the effect on the variance in the kurtosis parameter estimates in Section 3.3.4, and showed that the Rician bias does have an effect on the selection rate of all three model selection methods.

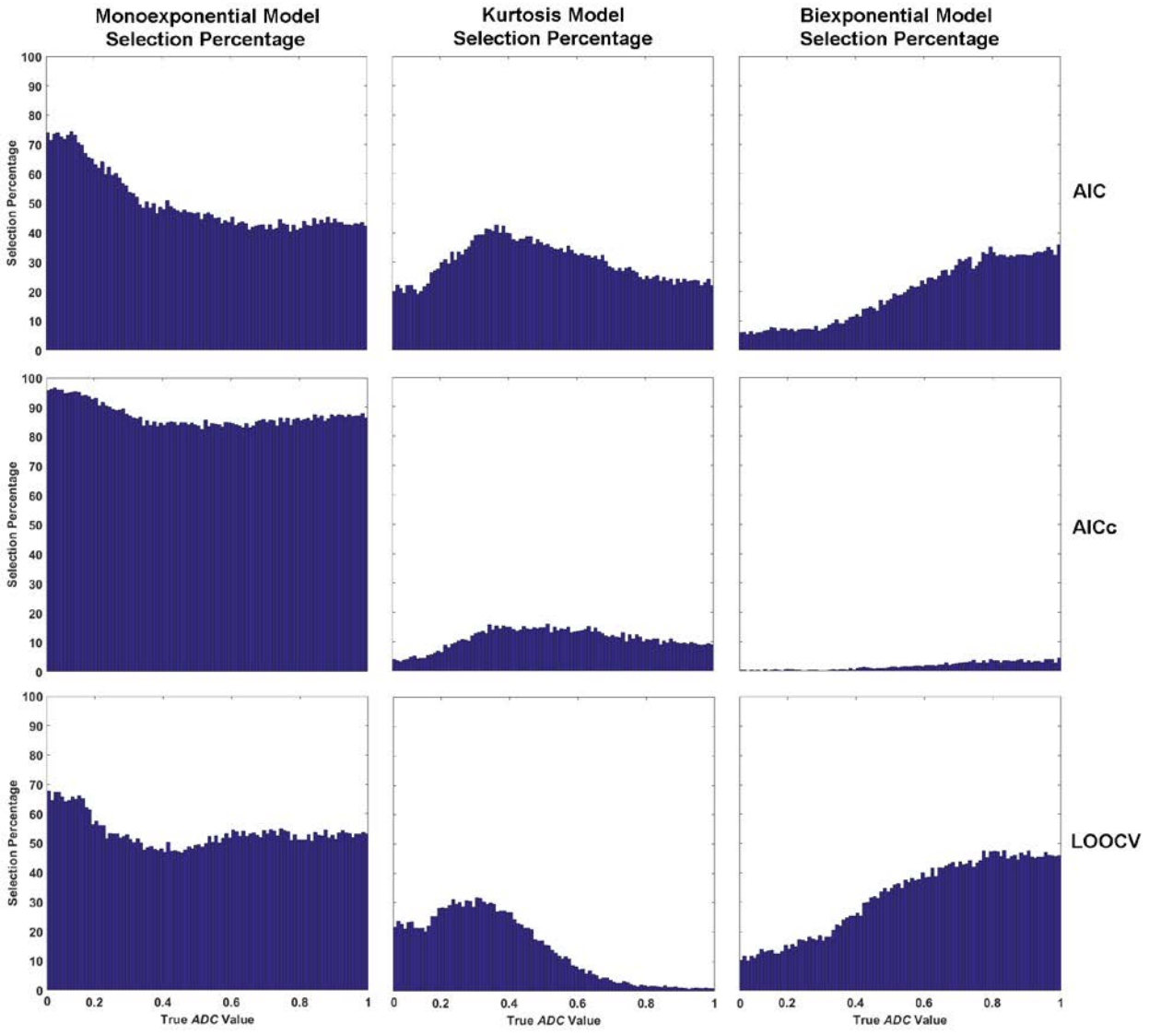

Figure 52 – Selection rate for monoexponential test set with eleven diffusion weightings and noise added at an SNR of 25

With the monoexponential model used to generate the data, the other two models are still selected for some data.

The additional bias in the AIC_c had a significant effect on the selection rates, as the simpler monoexponential model was selected correctly at least 80% of the time over the entire test set. As opposed to the biexponential test set, where the selection rate of the signal basis model decreased, for the monoexponential test set, this additional bias was beneficial as it increased the selection rate of the simpler monoexponential model. Figure 52 also shows that the AIC and LOOCV methods had similar selection rates here.

SNR 200

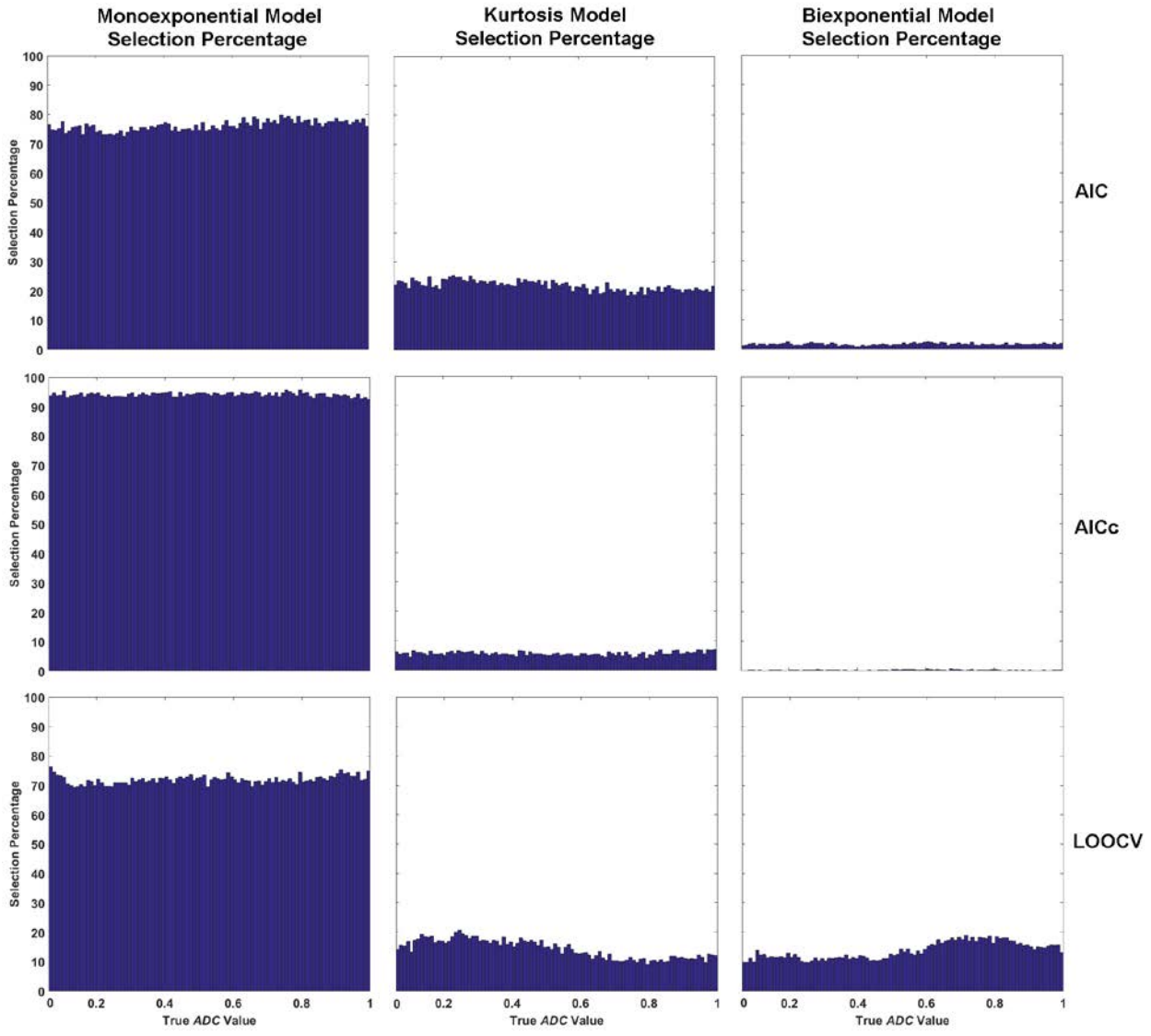

Figure 53 – Selection rate for monoexponential test set with eleven diffusion weightings and noise added at an SNR of 200

Increasing the SNR to 200 led to the monoexponential model being selected more often.

Increasing the SNR to 200 increased the AIC selection percentage of the monoexponential model to over 70% for the entire test set, and also eliminated the lifting effects of the noise floor, as seen in Figure 53. However, the kurtosis model was still incorrectly selected at a rate of over 20% of the entire test set, while the biexponential model was selected less than 5% of the time. The additional bias of the AIC_c increased the monoexponential selection rate to above 90%, with the kurtosis model selected for nearly all of the remainder of the signals. The AIC and LOOCV methods differed in selection rate of the more complex models, with the selection rate of the biexponential model increased. Compared to the divergence in selection rates for the biexponential test set, the increased selection rate of the biexponential model by the LOOCV method is incorrect in this case, suggesting that the LOOCV is somewhat biased toward the more complex models. The selection rate for the monoexponential model still does not reach 100% for these data, either, suggesting that even at an SNR of 200, the noise is still enough such that the kurtosis model fits better than the monoexponential model for many measurements.

4.3.4 Fitting Three Models to Monoexponential Truth with Seven Diffusion Weightings

SNR 25

The reduction of diffusion weightings to seven in Figure 54 shows the increase in selection of the simpler models that was also seen for the biexponential signals, due to restriction of the data to lower diffusion weightings as well as a lower overall number of weightings. The lifting tail effect, however, is again seen in the AIC where the monoexponential selection rate decreased as the *ADC* increased, suggesting that the decrease in signal-average SNR still affects the data. The AIC_c is severely biased here such that the monoexponential model was selected as best for nearly all measurements. The LOOCV selection rates differ completely than the AIC for this test set, as the monoexponential model selection rate increased as the *ADC* increased, opposite to the AIC trend, with the more complex models being selected at a higher rate overall.

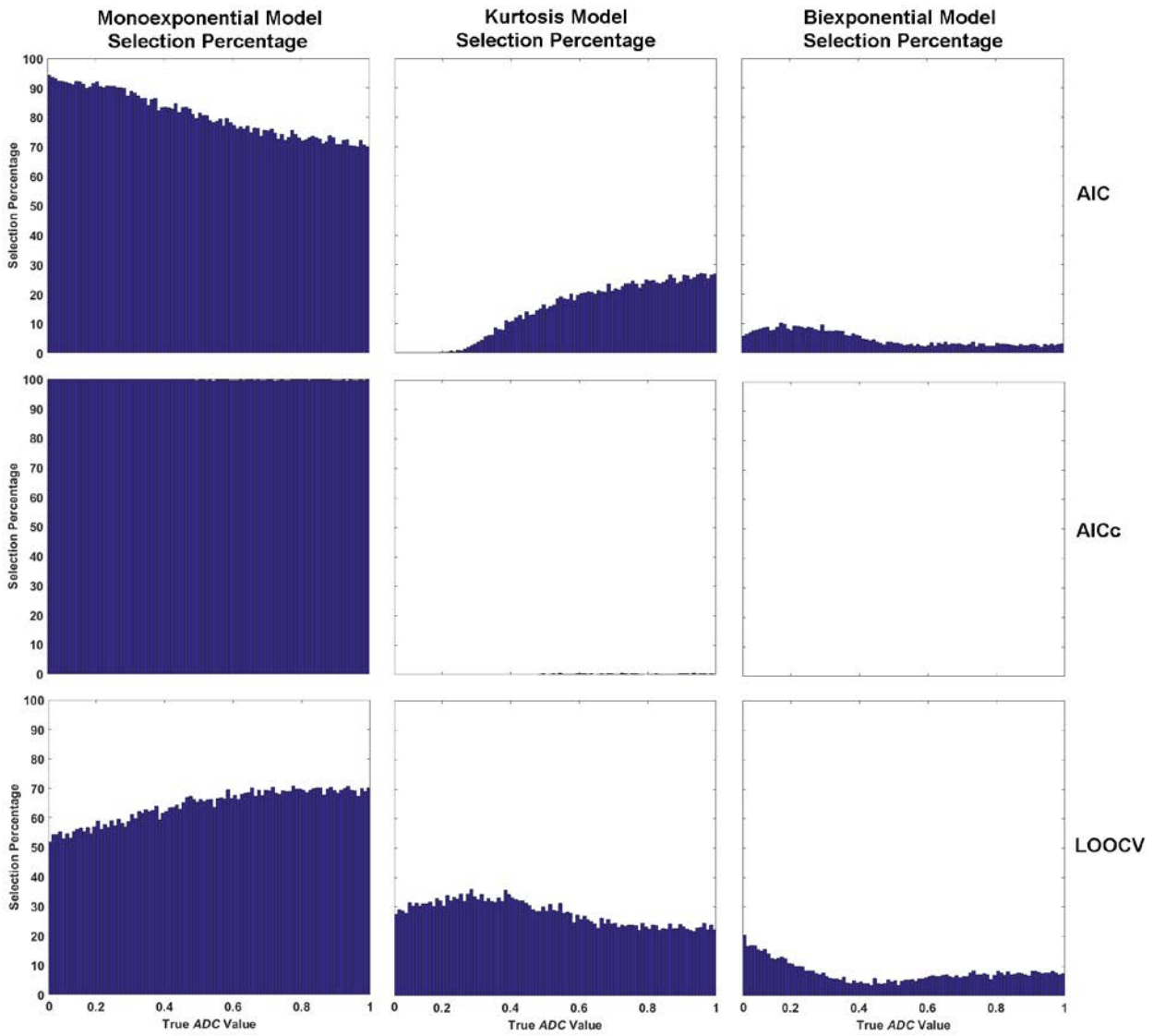

Figure 54 – Selection rate for monoexponential test set with seven diffusion weightings and noise added at an SNR of 25

Reducing the number of diffusion weightings led to a higher selection of the simpler models.

SNR 200

For an SNR of 200 and seven diffusion weightings, the monoexponential model had a selection rate of 90% for the AIC, but this rate decreased as the *ADC* increased, as shown in Figure 55. For the remainder of the test set, the AIC selected the kurtosis as the best model, aside from a small portion where the biexponential model is selected. Like the SNR 25 test set, the AIC_c selected the monoexponential model as best for the entire test set, and the selection rates between AIC and LOOCV again differed, favouring the more complex models.

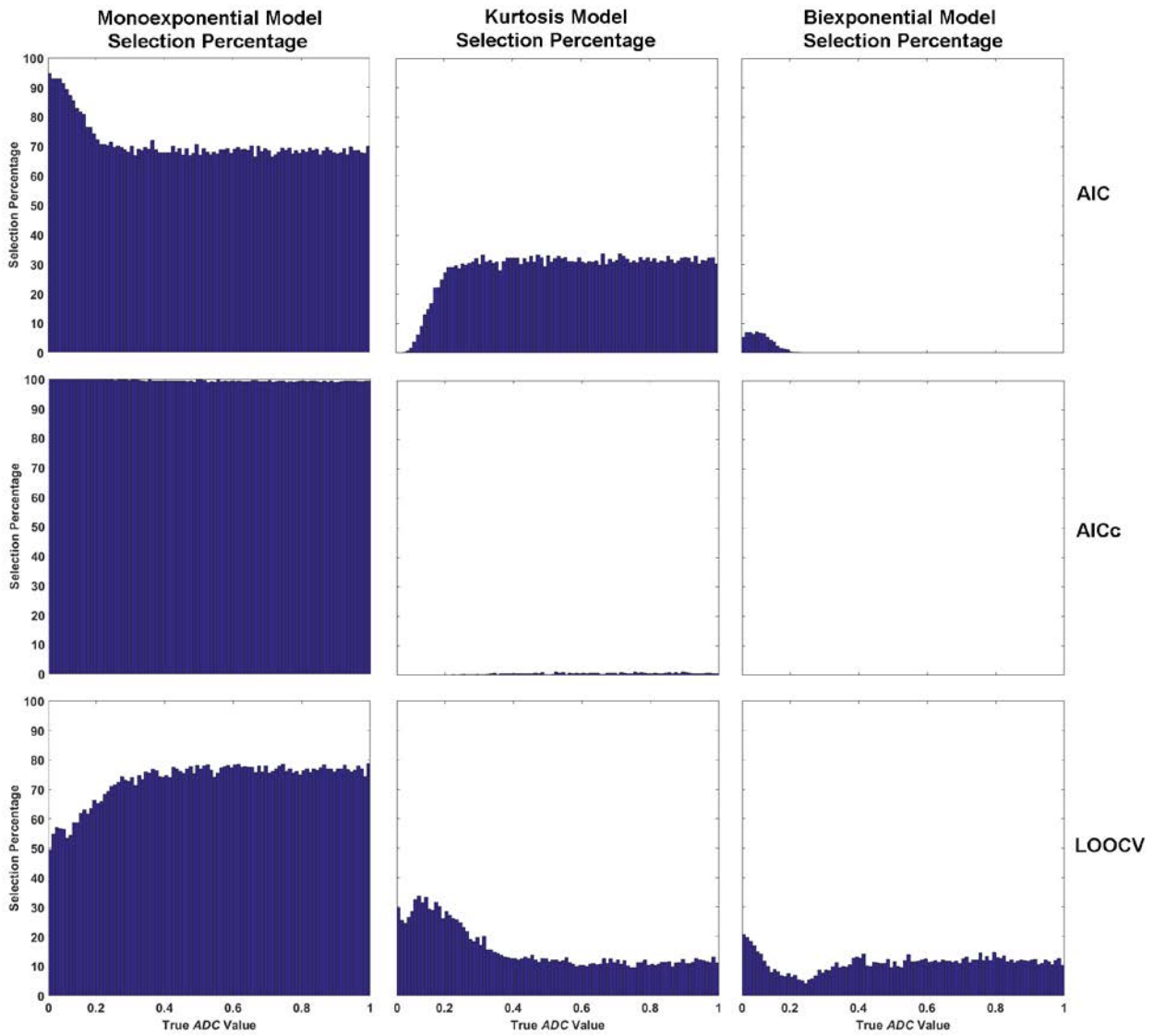

Figure 55 – Selection rate for monoexponential test set with seven diffusion weightings and noise added at an SNR of 200

4.3.5 ΔAIC and ΔAIC$_c$ Differences in Model Combinations

These mean ΔAIC values are shown in the left column of the pseudocolour plots in Figure 56 for each two-model combination, with the right column displaying the corresponding head-to-head selection rates. These plots show that where a particular model is selected 100% of the time versus the other, the mean ΔAIC value is much lower than zero. As the selection rate decreased, eventually the mean ΔAIC value reached 0, indicated by the black contour line in the left column plots. This contour location corresponds fairly well with the 50% selection rate contour in all three plots in the right column, aside from the right side of the kurtosis vs. monoexponential plots in the bottom row.

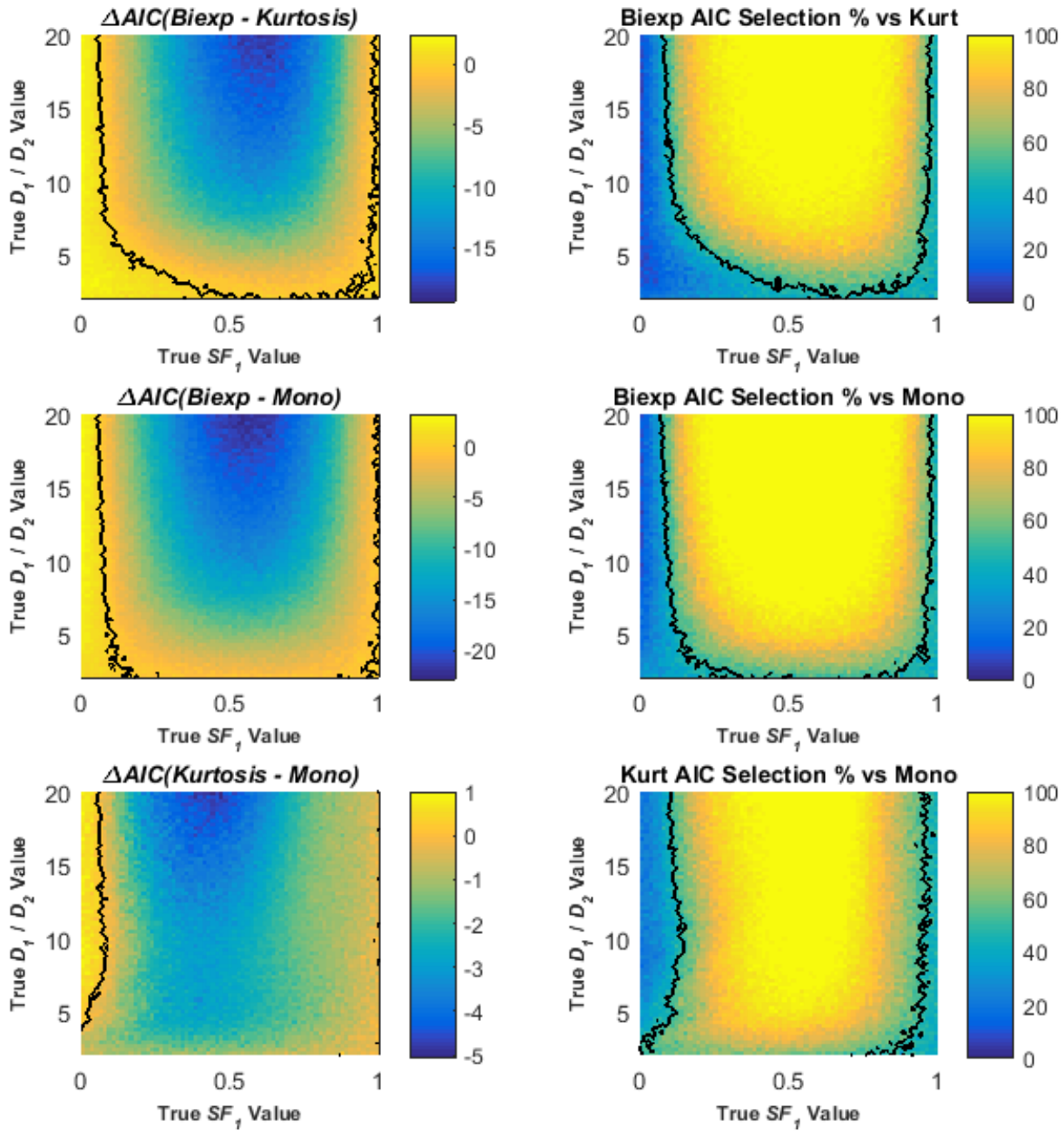

Figure 56 – Mean ΔAIC value (left column) and model selection rate of the noisy signal fits, based on each noise-free signal, for the three head-to-head model comparison pairs (rows)

The black contour line indicates where the mean ΔAIC value is zero (left column) or where the selection rate of each model is 50% (right column). This shows the signals where the ΔAIC was zero and the selection rate was 50% were at similar values.

This correlation between mean ΔAIC and selection rate is illustrated in more detail in Figure 57 showing the biexponential and monoexponential model fits to the noisy measurements from three

separate noise-free signals selected from the test set. For signal A (true SF_1 = 0.49, true D_1/D_2 ratio = 20), where the biexponential model was selected for 100% of signals, the mean ΔAIC value was -22, the standard deviation 6.3, and the ΔAIC between the two models greater than 10 in all fits. This showed that repeatedly fitting noisy measurements from one signal varies the difference in model AIC value considerably, with a difference in minimum and maximum ΔAIC values of 34. Signal B (true SF_1 = 0.087, true D_1/D_2 ratio = 19.5) had the ΔAIC distribution evenly divided about zero in count with an equal 50% selection rate, however, the maximum value is around 4 (the difference in the 2k parameter penalty between the two models), and the minimum value -17. Signal C, actually a monoexponential signal (SF_1 = 0, D_1/D_2 = 13.3), had 95% of the ΔAIC distribution greater than zero, with most values grouped near 4, however there was one signal with a ΔAIC value of -8.

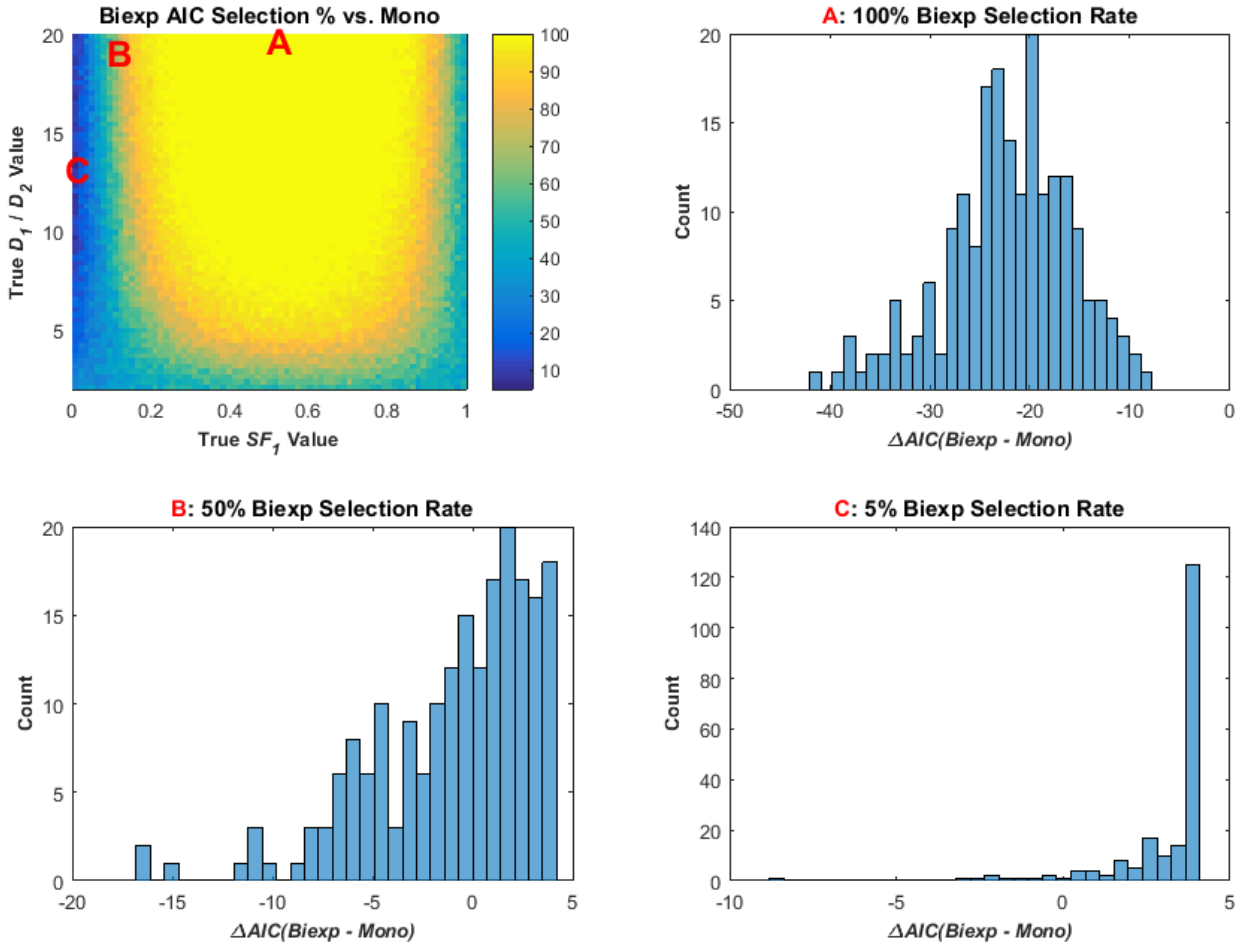

Figure 57 – Histograms of the ΔAIC values for three case studies of the model fits from three noise-free signals (top left plot) with added noise at SNR of 25

Each histogram shows the difference in AIC value between the biexponential and monoexponential model where the selection rate of the biexponential model is 100% (A), 50% (B), and 5% (C), with a higher ΔAIC value when the signals was less like a monoexponential.

These three signal examples show that just by adding noise at an SNR of 25 to one signal, the difference in AIC value can have a large deviation across the fitted noisy measurements derived from one signal. This is largely due to the deviations in the RSS values when fitting the models, as shown in Figure 58. Signal A had RSS values for the biexponential model distributed between zero and 0.04, while monoexponential RSS values vary over a much wider range with a minimum value of 0.05, leading to the clear distinction between AIC distributions. In signal B, the distribution of monoexponential RSS values is much closer to the biexponential RSS distribution, but still has a higher mean value. The effect of the AIC parameter penalty, however, means that the distribution of the AIC values overlap and are similar with a 50% selection rate for both models.

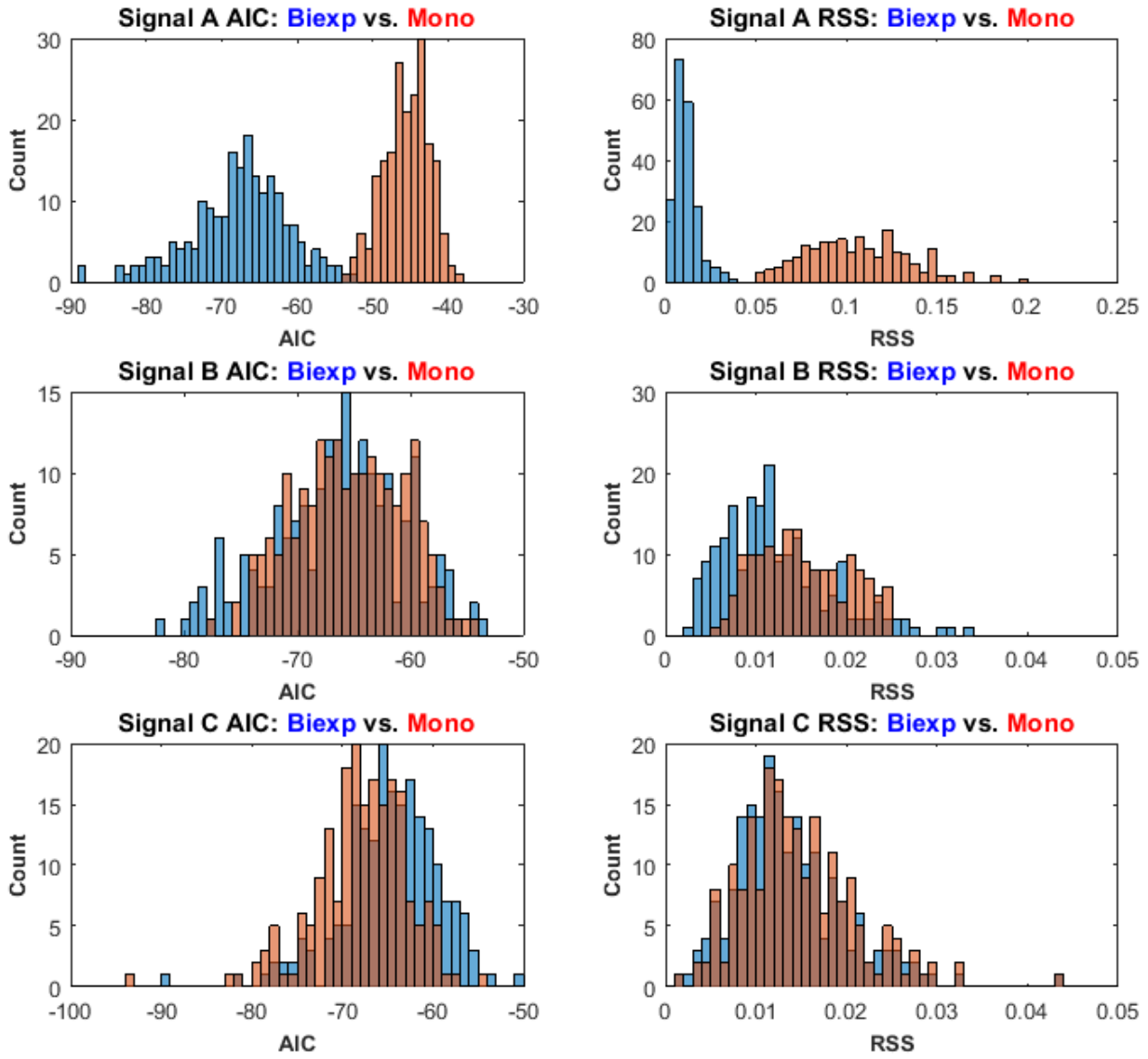

Figure 58 – Histograms of the AIC values for the biexponential (blue) and monoexponential (red) models (left column) along with the RSS distributions (right column) for the three signals in Figure 57

Finally, the RSS values for signal C show that the RSS distributions are nearly the same, since the two models fit the signals similarly. With the parameter penalty, however, the monoexponential AIC distribution has a lower mean, leading to the lower biexponential selection rate.

SNR 200

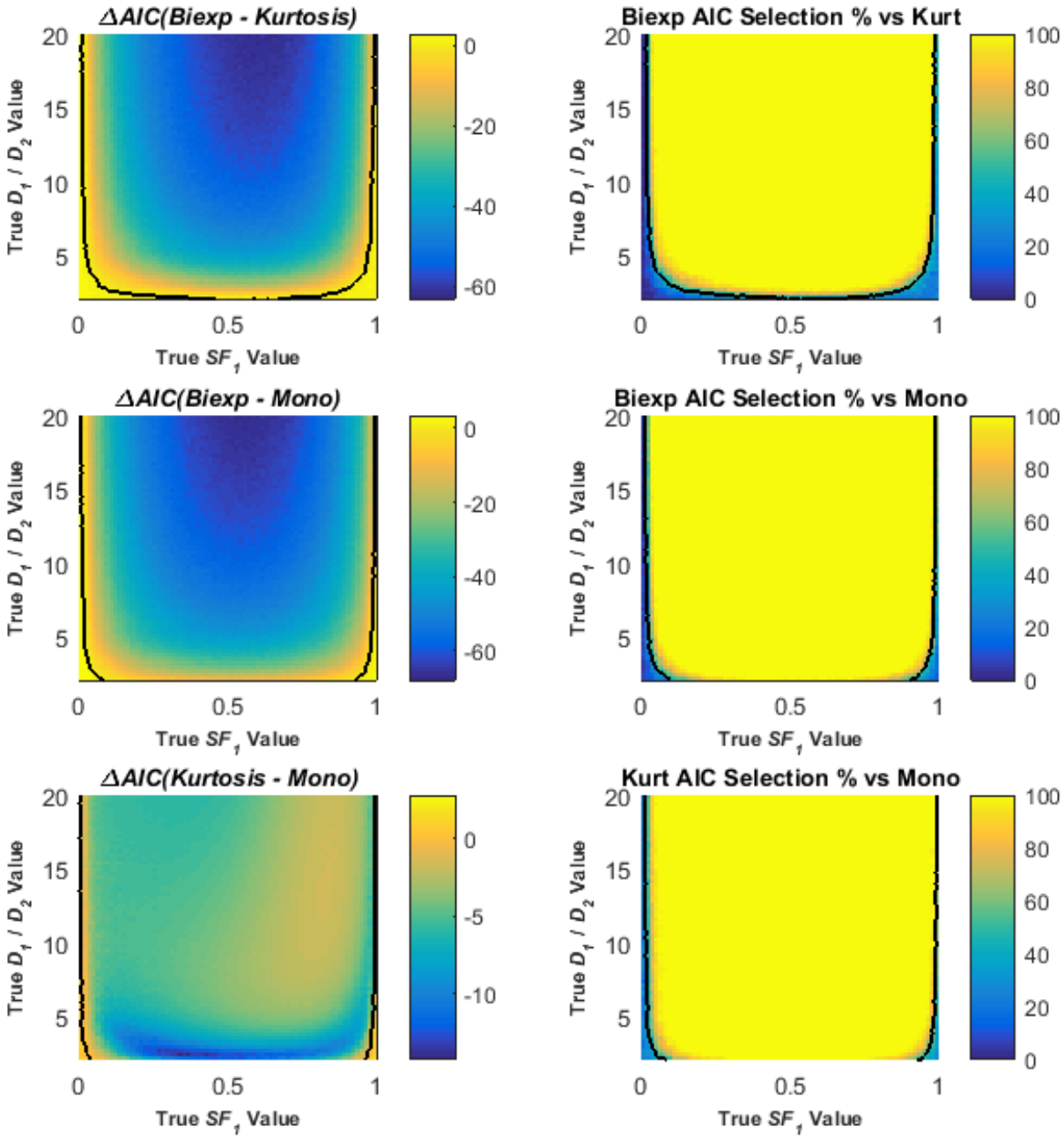

Figure 59 – Mean ΔAIC value (left column) of the noisy signal fits at an SNR of 200 versus the corresponding selection rate (right column) for the three head-to-head model comparison pairs (rows)

The black contour line indicates where the mean ΔAIC value is zero (left column) or where the selection rate of each model is 50% (right column).

Figure 59 shows the changes in the ΔAIC value when the SNR of the added noise increased to 200. For all three model pair combinations, the magnitude of the minimum ΔAIC value has greatly decreased, leading to an increased selection rate for the more complex models over most of the parameter space compared to Figure 56. A case study of three signals similar to Figure 57 was performed for the SNR 200 test set, with the results shown in Figure 60. While the selection rate of the biexponential model versus the monoexponential model increased over much of the parameter space, the spread in the ΔAIC values was largely the same. For example, signal A in Figure 57 had a difference in minimum and maximum ΔAIC values of 34, and in Figure 60 (same noise-free signal), this difference was 35.

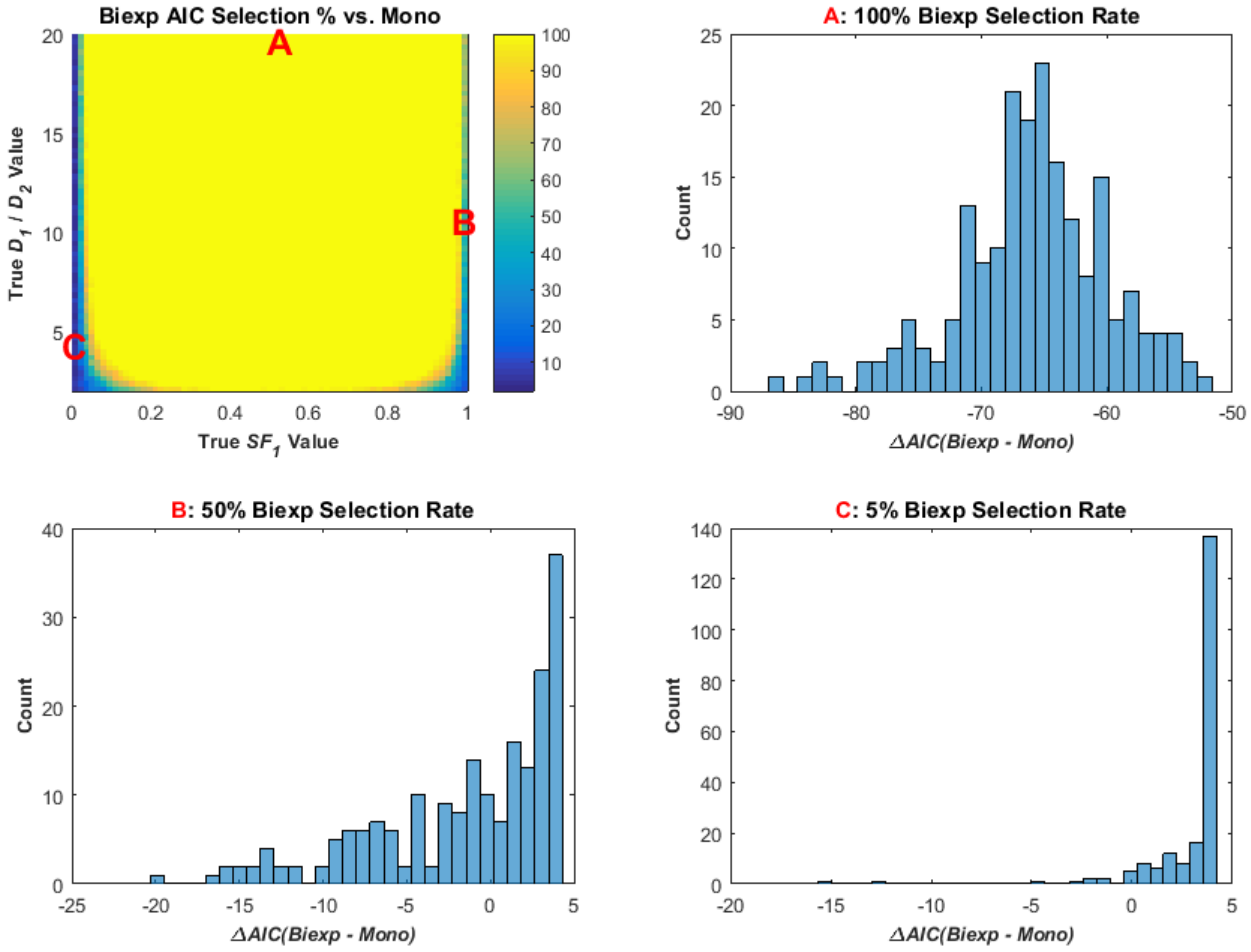

Figure 60 – Histograms of the ΔAIC values for three case studies of the model fits from three noise-free signals (top left plot) with added noise at SNR of 200

Each histogram shows the difference in AIC value between the biexponential and monoexponential model where the selection rate of the biexponential model is 100% (A), 50% (B), and 5% (C).

While the overall variance in RSS value for the SNR 200 signals decreases for both the monoexponential and biexponential model fits (see Figure 61 compared to Figure 58), Equation 42 is based off of the log of the ratio of the two RSS values, so the variation in AIC values remains at a similar value. With the biexponential and monoexponential models, this consistency of the ΔAIC value shows that it could be used as a measure to infer that the biexponential model was indeed better for a given fit, regardless of SNR. For example, only using the biexponential model when the difference in AIC to the monoexponential model is -10 or more increases the chance that the biexponential model is indeed better as opposed to strictly basing selection on whether the AIC is simply lower. This finding confirmed the literature studies introduced in Section 4.1.3 that a higher ΔAIC value correlates with the strength of inference between models, although here it was specifically demonstrated as a higher model selection rate over repeated samples.

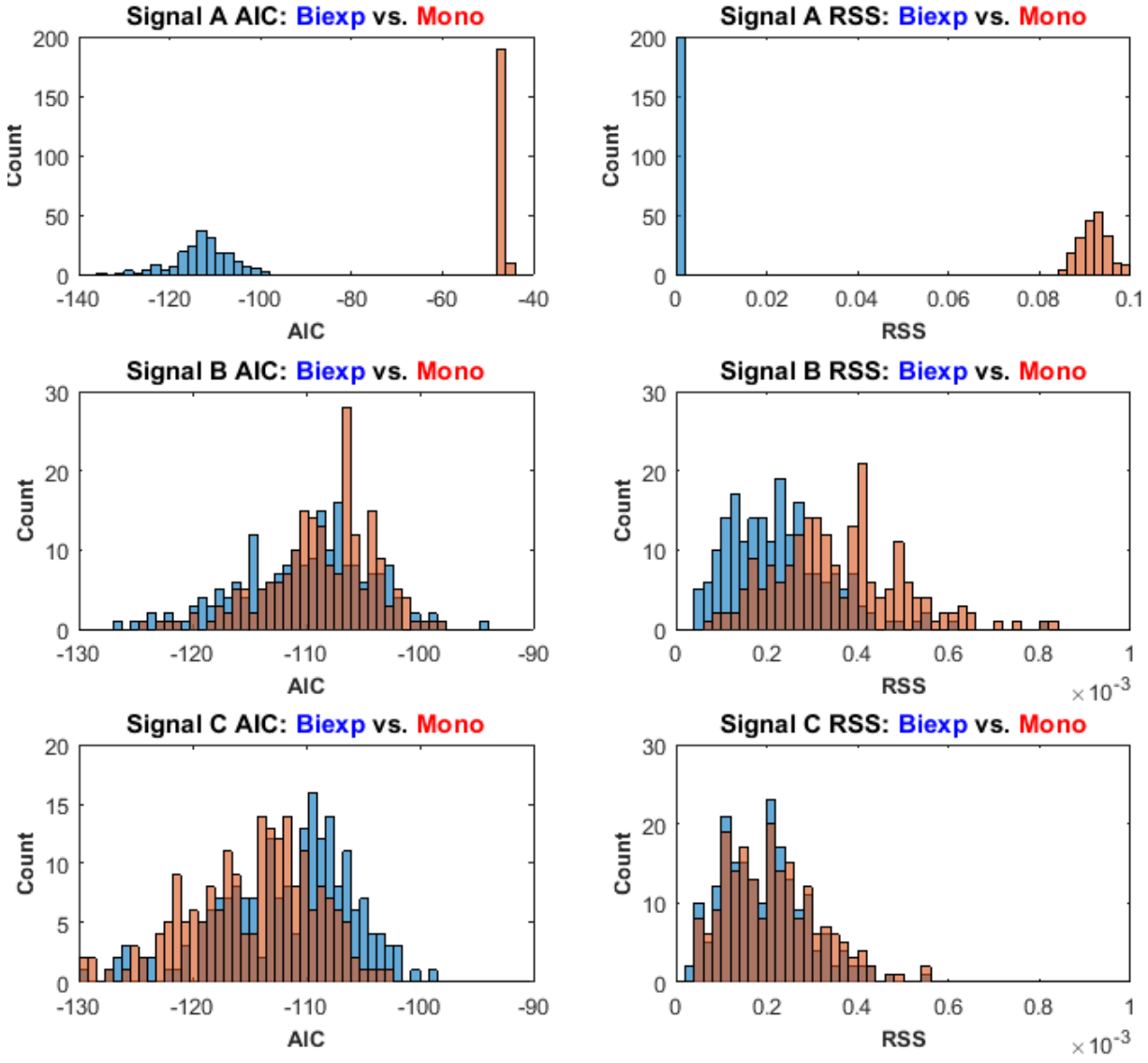

Figure 61 – Histograms of the AIC values for the biexponential (blue) and monoexponential (red) models (left column) along with the RSS distributions (right column) for the three signals in Figure 60

ΔAIC$_c$ – SNR 25

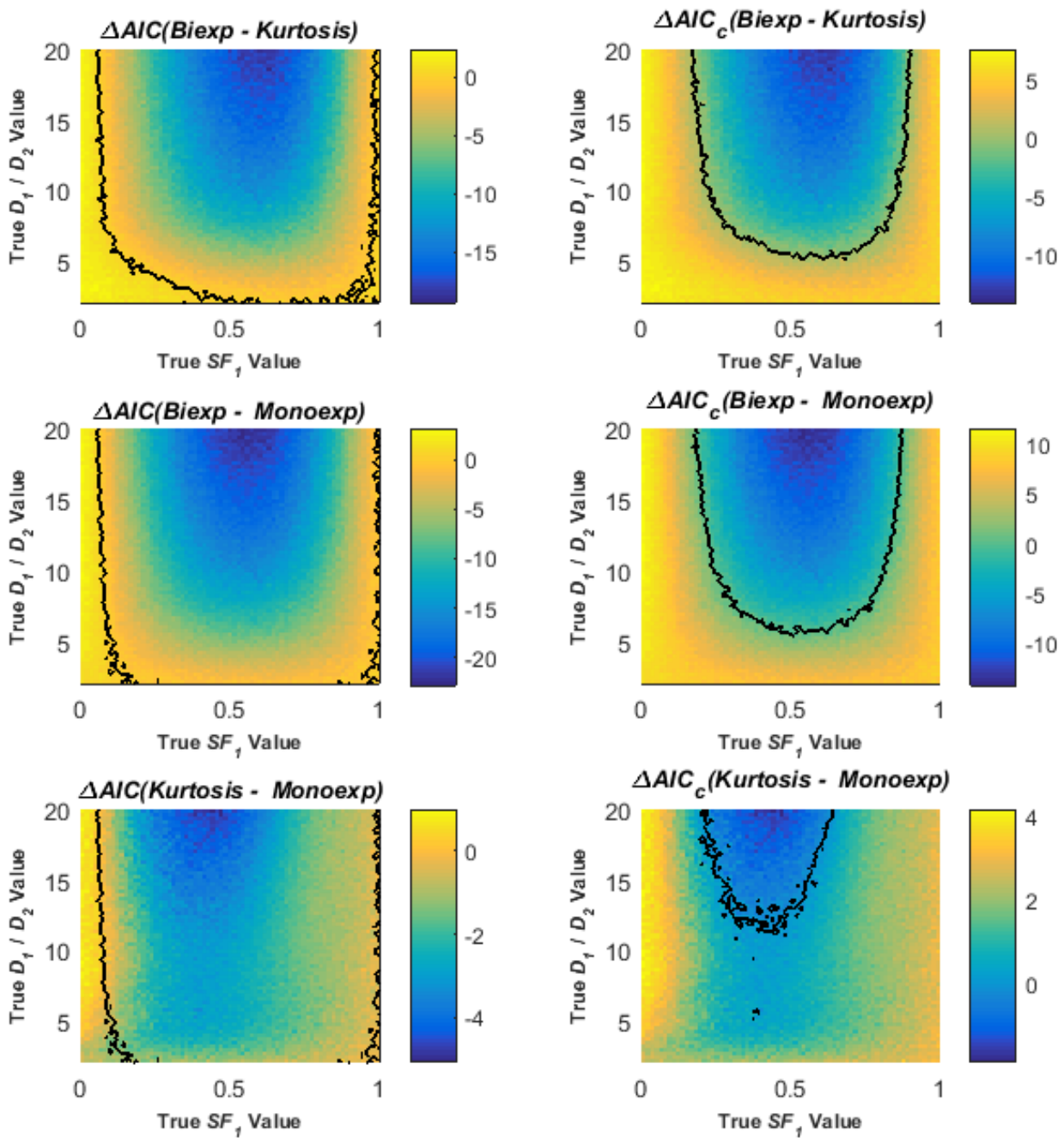

Figure 62 – Mean ΔAIC value (left column) of the noisy signal fits at an SNR of 25 versus the mean ΔAIC$_c$ value (right column) for the three head-to-head model comparison pairs (rows)

The black contour line indicates where the mean ΔAIC or ΔAIC$_c$ value is zero. Note, that the colour plots are the same column-wise, but each colourbar has a different scale.

An examination of the difference in AIC_c values shows the effect of the selection rate is merely the added parameter penalty. The values of ΔAIC_c are shown in the right column of Figure 62, and are plotted against the ΔAIC values at SNR of 25 from the left column of Figure 56. If the scales of the pseudocolor plots are adjusted based on the difference in the additional AIC_c parameter penalty, the colour distribution in the plots in each row are the same. However, the decision line of the mean $\Delta AIC_c = 0$ now covered a smaller portion of the parameter space, and hence this bias of the AIC_c toward simpler models was reflected in the selection rates in the centre row of Figure 48.

4.3.6 *F*-Test

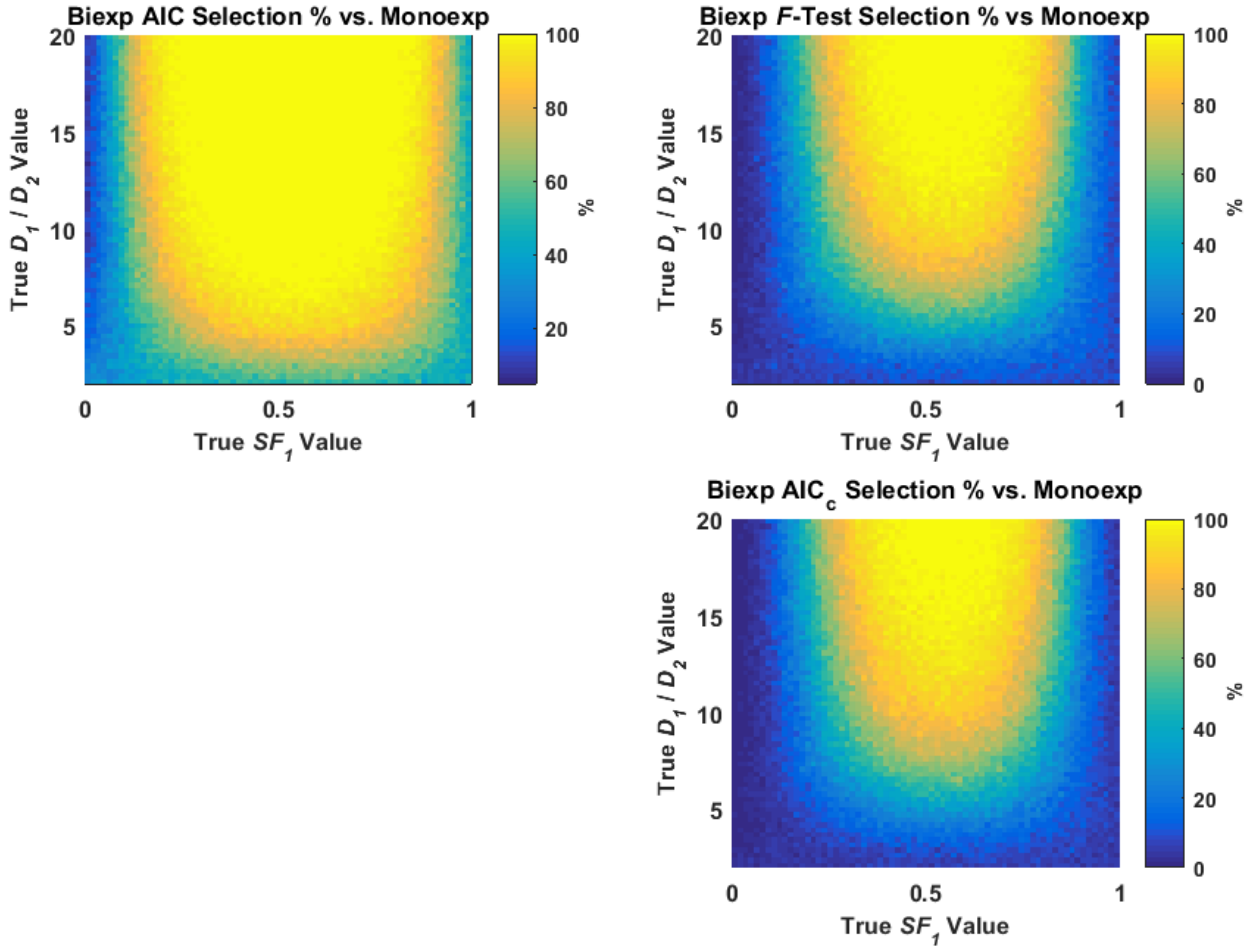

Figure 63 – Selection rate of the biexponential model vs. the monoexponential model for the biexponential test set (11 weightings, SNR = 25) by the AIC (upper left), *F*-Test (α = 0.05, upper right), and AIC_c (bottom right)

Note that the *F*-Test results are more similar to the AIC_c than the AIC.

A comparison of AIC, AIC_c and *F*-Test selection rates, when comparing the biexponential model vs. the monoexponential on the biexponential test set (11 diffusion weightings, SNR 25), is shown in Figure 63. This figure shows that the *F*-Test, with the significance level (α) set to 0.05, had similar

selection rates across the biexponential parameter space to the AIC_c, with both biased toward the simpler, monoexponential model than the AIC.

4.3.7 Ill-Conditioning and Normality in Parameter Estimates

Assessment of the skewness of the biexponential model parameter distributions, where the mean of the distribution was subtracted from the median and performed on all measurements of each noisy signal are plotted in Figure 64. These plots show that the skewness was indeed minimised in the areas where parameter uncertainty was lowest at equal signal fraction and maximum D_1/D_2 ratio. This measure also detects the difference in left- and right-handed skewness in the amplitude parameters, but may not be a useful measure, since a considerable amount of the D_1 distributions had a high amount of skewness.

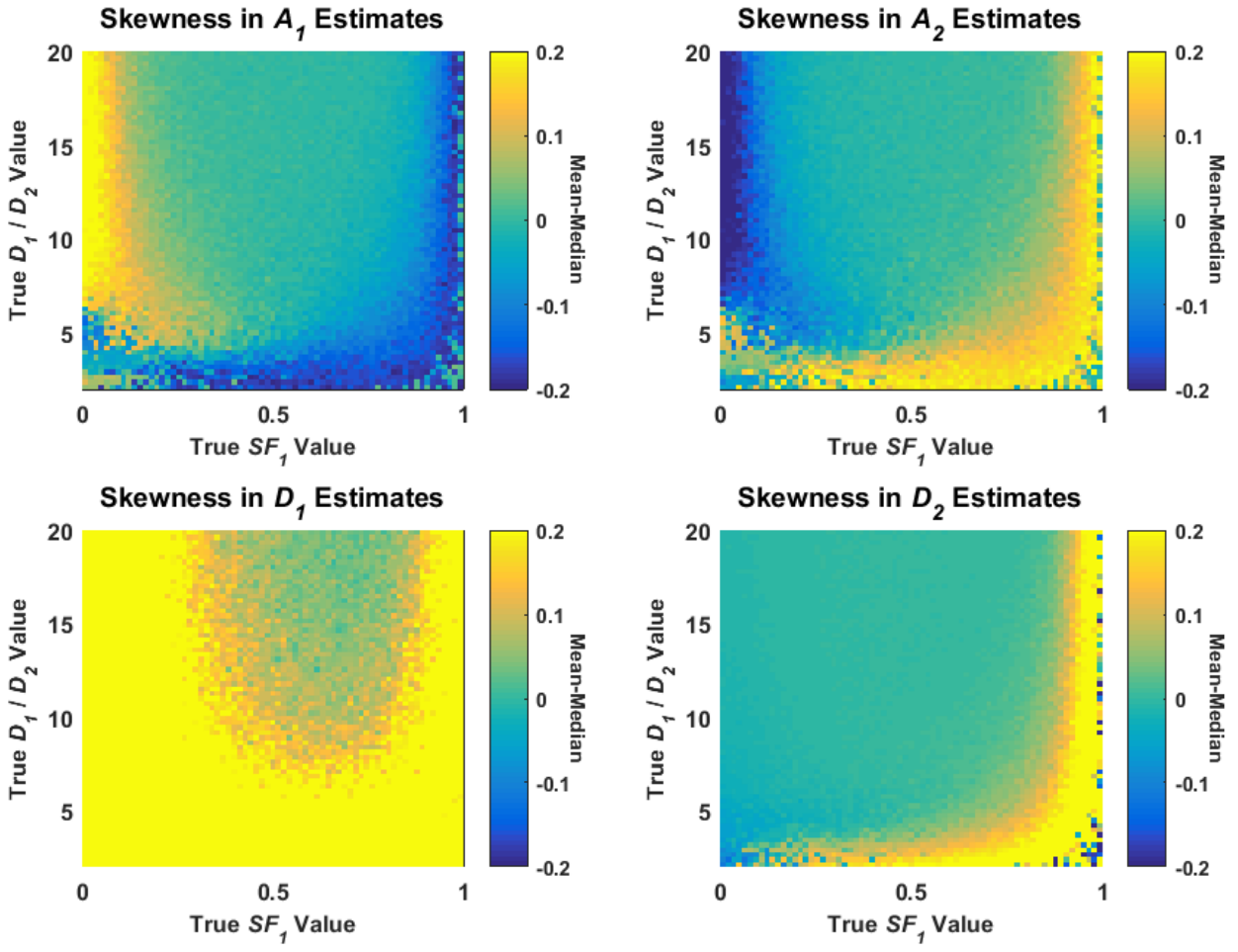

Figure 64 – Skewness (mean – median) for the parameter estimate distributions of the noisy signal fits for each noise-free signal at an SNR of 25

The skewness is higher when the signal is more monoexponential, which is also where the fitting is more likely to be ill-conditioned.

The parameter distributions were also compared with the Lilliefors normality test (α = 0.001) with the results of this test for each biexponential model parameter shown as pseudocolour plots in Figure 65. These plots also show that the estimate distributions were normal more often where parameter uncertainty was lowest at equal signal fraction and maximum D_1/D_2 ratio. These test results were noisy, even in the areas of lowest parameter uncertainty, however, with very few D_1 estimates passing the normality test. Thus, to not completely rule out all parameter estimates, a logical "OR" combination of the four parameter estimates was done, where the combined test was set to pass if *any* of the parameter estimates passed the normality test. This gives the plot seen in Figure 66.

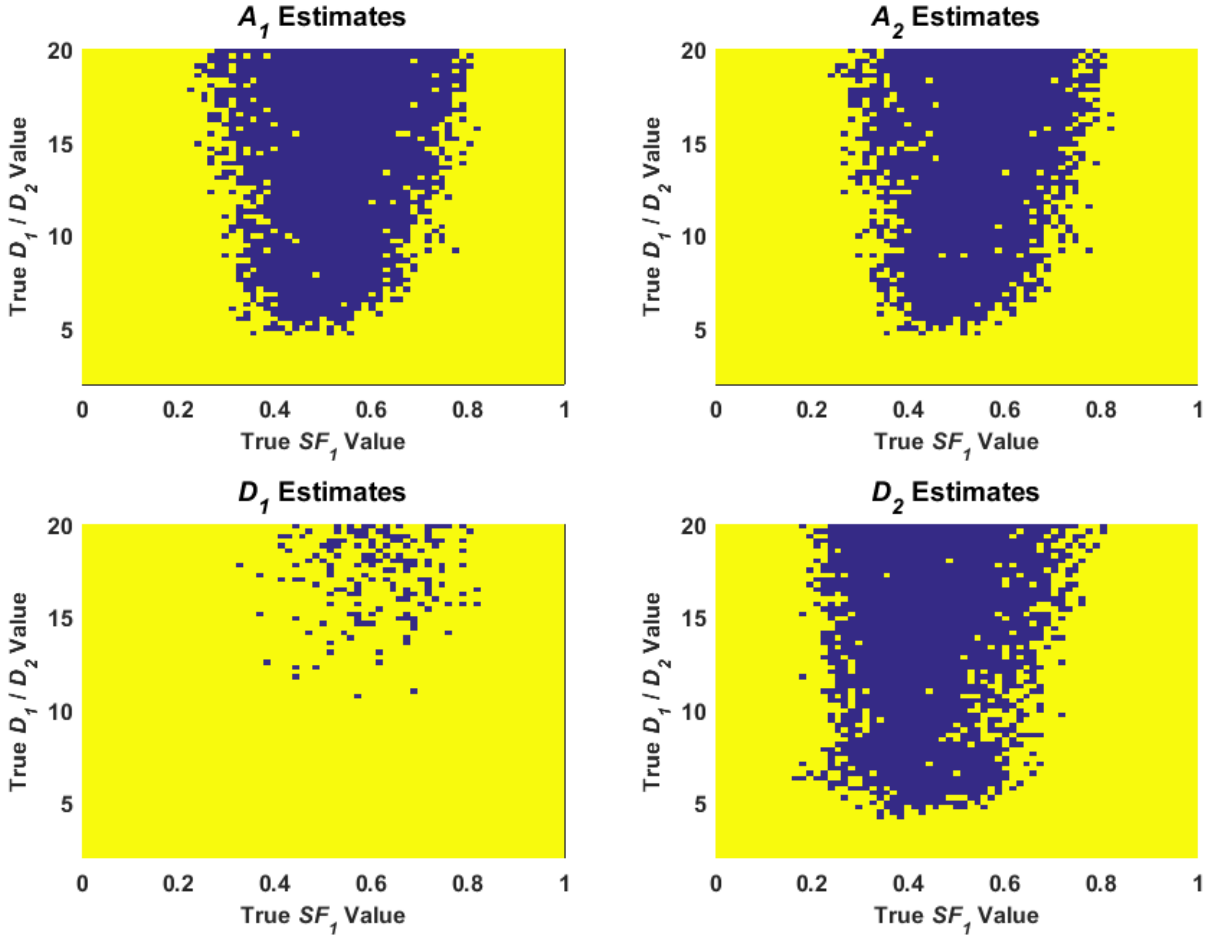

Figure 65 – Lilliefors normality test (α = 0.001) of the noisy signal parameter estimate distributions. Blue dots are true signals where the parameter distributions are normal, yellow dots are non-normal

This test also is likely to fail when the signal is more like a monoexponential decay.

Many automated tests of normality have very tight constraints, causing many distributions to fail, and different normality tests will have different assessments and selection rates [182, 183]. While the combined test could still be erratic in its selections, it did provide a decision boundary between the areas of low parameter errors and high, ill-conditioned parameter errors. Thus, this test could

be used with the bootstrap parameter estimates to provide a decision on whether to use the biexponential model parameter estimates from a single fit.

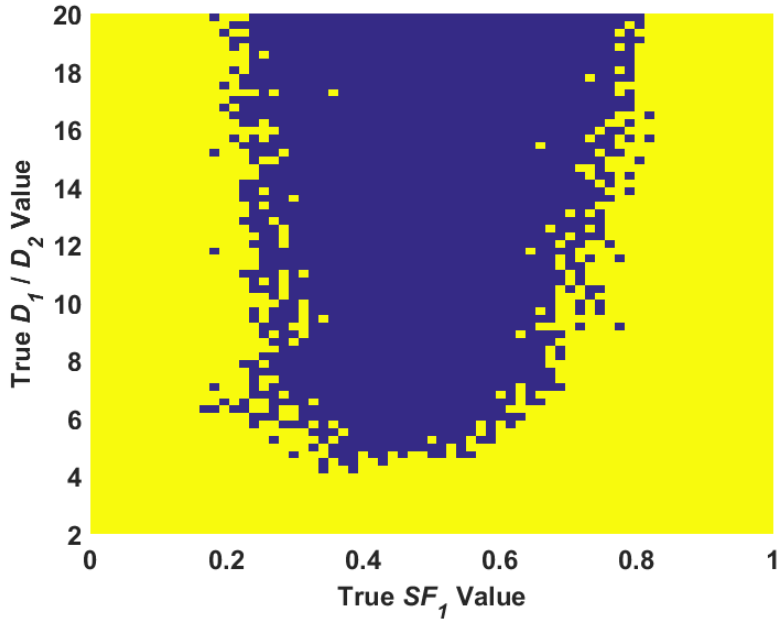

Figure 66 – Combined results of the individual parameter normality tests in Figure 65 where the test is set to pass if any of the parameter estimate tests passed

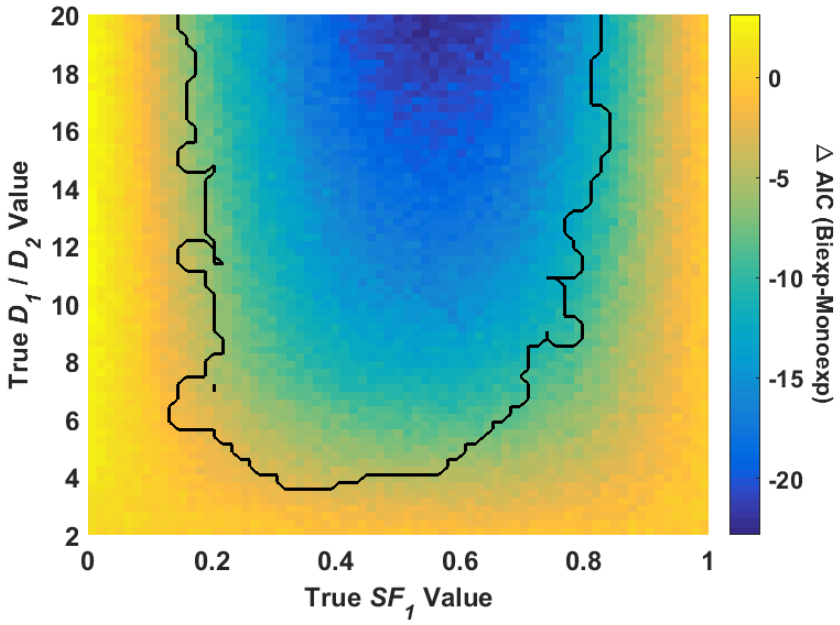

Figure 67 – Pseudocolour plot of mean ΔAIC between the biexponential and monoexponential model fits, with overlaid contour of the test boundary between normal and non-normal parameter estimate distributions from Figure 66

If this test indicates non-normal parameter estimates, there is a considerable chance that there will be large uncertainty in the parameter estimates, since this is an indication that the parameter estimates may be unreliable. If the decision boundary from the combined test in Figure 65 is overlaid as a contour on a plot of the ΔAIC value between the biexponential and monoexponential model, the value of ΔAIC was made clearer as shown in Figure 66. If a minimum ΔAIC value of say -10 was chosen for a decision boundary on whether to keep the biexponential model parameter estimates, then *on average* there was a lower chance of the parameter estimates being unreliable. This could be useful if bootstrap samples from fits are not able to be calculated to check normality due to the large increase in computation time. In this case using a specified minimum difference in ΔAIC will reduce the overall bias and variance of the biexponential model estimates.

Difference in Parameter Estimates when AIC Selects Two Models Equally

To illustrate the difference in parameter estimates made when relying solely on the AIC to select a model, one noisy measurement of signal B in Figure 57 was chosen for analysis. When fitting the biexponential and monoexponential model to the signal measurement, the difference in AIC values was 0.1, with the biexponential model selected as best. For both the biexponential and monoexponential fits, a perturbation analysis was performed for both models with 500 parametric bootstrap samples for each, in order to calculate a confidence interval for each parameter and assess the distributions for normality. The parameter estimates for both models, along with their calculated 95% bootstrap confidence intervals are listed in Table 9.

Table 9 – True Value, Parameter Estimates, and Bootstrap Confidence Intervals from Monoexponential and Biexponential Fits with near-equal AIC values.

	True Value	Original Fit Estimate	Confidence Interval
Biexponential A_1	0.09	0.10	(0.038, 0.82)
A_2	0.91	0.89	(0.23, 0.95)
D_1	1	1.17	(0.093, 34.7)
D_2	0.05	0.05	(0.0, 0.06)
Monoexponential S_0	N/A	0.96	(0.92, 0.99)
ADC	N/A	0.06	(0.05, 0.07)

The table shows that the parameter estimates of the original fit here were very close to their true values. However, this is just from one fit, and the 95% confidence intervals indicate that the A_1, A_2, and D_1 parameters could assume a very wide range of values over repeated measurements of the same signal. The confidence intervals for the monoexponential model are much smaller and evenly distributed around the original estimates. This specific example illustrates why caution must be used when relying on the AIC to choose the best model, and that when comparing the biexponential

and monoexponential models, signals with equal AIC values can have significantly worse parameter estimates in the biexponential model when assessing multiple measurements.

4.4 Conclusions

This chapter analysed three model selection methods on simulated data to examine the effects of noise and acquisition parameters on their selection rates, specifically:

- When comparing the biexponential, kurtosis, and monoexponential models, the selection rate that each model was selected as best by the AIC, AIC_c, and LOOCV selection methods varied over repeated noisy measurements of the same simulated signal.
- The rate that each model was selected by all three methods also changed as the parameter values for the true simulated signals varied, whether the signal basis model was biexponential or monoexponential.
- When the SNR increased to 200, the biexponential model selected more of the biexponential signals as the best model, however, when the signals were effectively monoexponential, the kurtosis and monoexponential models were still selected for many of the signal measurements.
- Removing the highest diffusion weightings to remove the effects of Rician bias from the measurements led to a higher selection rate of the simpler monoexponential model. When Rician bias affected the measurements when the highest diffusion weightings were included, the noise floor effects lifted the tail of the signals and the kurtosis and biexponential model were selected more often when the monoexponential model was the signal generating model.
- When comparing the biexponential and monoexponential models over all signals in a biexponential test set, the rate that the *F*-test selected the biexponential model as best was very similar to the AIC_c as opposed to the AIC.
- When comparing the AIC and AIC_c directly, it was shown that the additional correction term in the AIC_c merely operates as a source of additional bias favouring simpler models. For a biexponential test set with limited diffusion weightings, the AIC_c selected the monoexponential model as best for all signals, so it is possible that the AIC_c can be biased too much in specific examples.
- An increase in the value of ΔAIC between two models was associated with a higher selection rate of the model initially chosen as best, when repeated measurements from the same signal were tested.
- When selecting between the biexponential and monoexponential model, a higher minimum value of ΔAIC for a given fit where the biexponential model was selected as best was associated with a lower average chance that ill-conditioning would affect the biexponential model parameter estimates.
- When comparing the biexponential and monoexponential models on a given measurement, and the AIC values for both models were similar, there was considerable ill-conditioning in the biexponential parameter estimates and therefore, relying solely on model selection

methods to choose whether to use the parameter estimates can lead to much larger error in future repeated measurements.

Chapter 5
Parameter Estimation and Model Selection of Actual DWI Data

This chapter references the author's conference publication, "Biexponential modelling of diffusion in stroma and epithelium of prostate tissue", Ned Charles, Gary Cowin, Nyoman Kurniawan, Roger Bourne, Joint Annual Meeting ISMRM-ESMRMB, 2014 [184]

The previous chapters demonstrated large uncertainties in both the biexponential and kurtosis model estimates on simulated data, along with expanded methods on how to identify these issues. Additionally, a comparison of model selection methods was performed on this simulated data, demonstrating that even if a model was selected as best by one of these methods, there could still be unreliable parameter estimates. This chapter compares these simulated data results to fitting of actual DWI data from a prostate tissue sample scan performed ex vivo.

5.1 Introduction and Background

5.1.1 DWI Analysis of Prostate Tissue

As of 2015, prostate cancer accounts for one-fourth of all new cancer diagnoses in American men and is their second leading cause of cancer deaths [185]. The number of men in New South Wales, Australia living with prostate cancer is expected to rise by 59 – 73% from 2007 to 2017, with an estimated 60,000 men to be affected then [186]. The diagnosis of prostate cancer using MRI has improved greatly since the 1980's, and the current PI-RADS standard includes DWI as part of a multiparametric assessment strategy [31]. For a DWI acquisition in vivo of the prostate, the initial PI-RADS recommended a maximum *b*-value up to 800 – 1000 s/mm^2, which avoids the increased noise of the Rician signal bias [32]. While measurements of the *ADC* value using the monoexponential model should stick to these lower *b*-values, the newest version 2 standard now recommends the use of higher *b*-values between 1400 – 2000 s/mm^2, as long as the SNR is high enough. The addition of higher *b*-values means that clinicians may be investigating more complex models such as the biexponential and kurtosis models.

The author's conference publication referenced above used even higher *b*-values, up to 4.65 ms/μm^2 (4650 s/mm^2), with an estimated SNR at the lowest *b*-value (0.335 ms/μm^2) equal to 40. However, this scan was performed ex vivo on tissue, which removes the noisy effects of patient movement and perfusing blood from the data. The study was performed on a 16.4 T MRI scanner, and the high magnetic field strength and small bore allowed for DWI voxel sizes of 80 μm (length of all three sides), and T2* voxel sizes of 40 μm. The purpose of the research study was to use this high scan resolution to individually investigate the properties of the three main types of prostate cellular tissue: epithelium, stroma, and lumen (duct). In vivo studies, for example, typically have voxel sizes around 1-2 mm, and contain heterogeneous mixtures of all three tissue types. The study was performed on three, 3 mm-diameter tissue core samples from three separate prostates analysed after radical prostatectomy, with a 40 μm T2* image of a middle slice through the cores shown in Figure 68. From this T2* image, five ROI were manually drawn on the image, representing stroma (S1 and S2), normal, epithelium-rich glands (E1 and E2), and one region (C1)

likely to be low-grade cancer based on macroscopic tissue features and previous patient biopsy results (though not confirmed by histopathology analysis, which was unable to be performed).

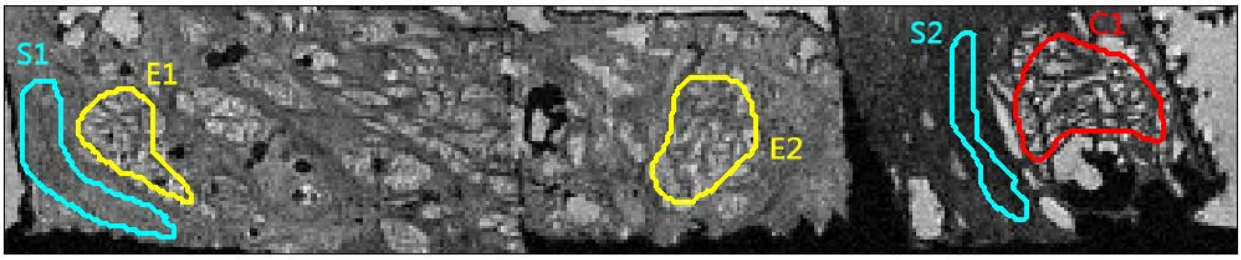

Figure 68 – 40 µm T2* image of three 3 mm diameter prostate tissue samples with five manually selected ROI illustrated

Reproduced from [184] for this thesis.

Biexponential Model Fitting

The five ROI from the T2* image were rescaled to the 80 µm DWI image, and all voxels within each ROI were fitted with a biexponential model using NLLS regression. The 80 µm voxel size used in this study (each voxel containing ~200 cells) is eight times smaller than that used for a previous investigation that applied a biexponential model to DWI signal attenuation in fixed prostate tissue [187]. The results from this 80 µm biexponential study showed that both D_1 and SF_1 were lower in the epithelium-rich regions (E1, E2, C1) than in stromal regions (S1, S2). These results should also lead to a lower ADC in epithelial tissue than in stroma when examining the monoexponential model, although this test was not performed on the study. Additionally, the lowest value of D_2 was found in the epithelial tissue in C1, suggesting a more restrictive diffusion environment than the other epithelial ROI. The difference in parameter estimates were analysed and presented by grouping the parameter estimates for the voxel fits in each ROI and reporting a mean and standard deviation of these distributions.

5.1.2 Chapter Aims

Given the issues with the biexponential model and its uncertainties presented thus far in this thesis, this data set was chosen for a study of the biexponential model fits on actual tissue. This included an expanded analysis with some of the measures introduced earlier in this thesis that indicated large uncertainty in the parameter estimates. Reviewing this earlier study established whether there were also large uncertainties in the biexponential parameter estimates when examining real data, determined the magnitude of the variance in these estimates, and the number of voxels that were affected. It also determined whether large uncertainties in the parameter estimates could be alleviated through testing the bootstrap sample distributions for non-normality, as well as choosing voxel fits with a minimum difference in AIC score from the monoexponential model. This analysis examined the kurtosis model in a similar fashion, and also examined monoexponential model estimates to compare to the two more complex models.

The aims of this chapter were to:

- Investigate the uncertainty in the biexponential, kurtosis, and monoexponential model parameter estimates of the data by examining both parameter plots and histograms of the regression fits for outliers.
- Examine the bootstrap confidence intervals for all fits and determine whether there were large discrepancies and ill-conditioning present in the parameter estimates.
- Use the AIC to select the best model in all voxels and determine the percentage that each model was selected as best.
- Perform a normality test on the bootstrap sample distributions for all biexponential and kurtosis model fits and determine whether excluding non-normal fits from the analysis improved the uncertainty in the parameter estimates.
- Individually compare the biexponential and kurtosis models to the monoexponential model, exclude voxel fits under a minimum difference in AIC score, and determine whether this exclusion improved the uncertainty in the parameter estimates.

5.2 Methods

5.2.1 Data Acquisition

Three, 3 mm-diameter core samples were obtained from three separate prostates after radical prostatectomy with patient consent. These samples were initially fixed in formalin for 24 hours, then immersed in 0.2% v/v Magnevist for over 48 hours. The samples were set in a fixed position and imaged on a 16.4 T Bruker AV 700 microimaging system (5 mm birdcage coil, Micro5 gradient set) using a 3D spin echo DTI sequence with TE/TR = 28/500 ms and δ/Δ = 2/20 ms. 80 μm isotropic voxels were acquired in six gradient directions with *b*-values of 0.50, 0.90, 1.42, 2.06, 2.78, 3.59, 4.65 ms/μm^2, along with two reference images at an effective *b*-value of 0.335 ms/μm^2. All scans were performed at room temperature (22° C) and typical voxel $SNR_{b=0.335}$ = 40. In the original conference publication study, a diffusion tensor was calculated for each *b*-value in the 80 μm data set and the mean diffusivity used to calculate normalized signal intensity at each *b*-value, independent of gradient direction. In this thesis study, the mean diffusivity was not calculated and instead, biexponential model fitting was performed using the data from one axis, combining the mean signal from the two reference images (b = 0.335 ms/μm^2) and the remaining seven *b*-value acquisitions for a total of eight diffusion weighted measurements for each voxel.

5.2.2 NLLS Regression Fitting

The monoexponential, kurtosis, and biexponential models were each fit to the data using the same NLLS regression algorithm as the previous three chapters (*lsqcurvefit* in MATLAB, trust region reflective option). The model equations used were 9 (monoexponential), 13 (kurtosis), and 32 (biexponential) with the starting values and lower and upper bounds on each model set per Table 8. The SD of the background noise for this acquisition was calculated over a few thousand voxels using the difference in value between the two reference images. Each voxel measurement was also

compared to the noise level and if the measured magnitude of the lowest *b*-value was not at least three times the noise standard deviation, the models were not fit.

Table 10 – Bounds and Starting Values for NLLS Regression Model Parameters

Amplitude bounds and starting values were set to either the maximum value of a given noisy signal measurement or a multiple of that value. Values for ADC, D_{app}, D_1 and D_2 are in ($\mu m^2/ms$) and starting value of 2.1 chosen since it's the diffusion coefficient of free water at 22° C.

Parameter	Lower Bound	Upper Bound	Starting Value
A_0 *(mono)*	0	(2x) max signal	max signal
ADC	0	10	2.1
A_0 *(kurt)*	0	(2x) max signal	max signal
D_{app}	0	10	2.1
K_{app}	-1	2	0.2
A_1	0	(1.2x) max signal	(0.5x) max signal
A_2	0	(1.2x) max signal	(0.5x) max signal
D_1	0	10	2.1
D_2	0	10	2.1/6

5.2.3 Parametric Bootstrap, Confidence Intervals, and Normality Testing

For all voxel regression fits with the monoexponential, kurtosis, and biexponential models, 200 parametric bootstrap samples were created for each model fit, by resampling the residuals and fitting each sample fit with that same model. For all bootstrap fits on all three models, all upper bounds in the regression fitting were removed, and only the lower bound of K_{app} was removed with the rest left at zero. A 95% confidence interval was determined for each resulting bootstrap parameter estimate distribution from each model by calculating the 2.5 and 97.5% percentile values of the distribution. A Lilliefors normality test with a significance level (α) of 0.001 was also applied to each bootstrap parameter distribution. For the biexponential model, if any of the four parameter distributions passed the normality test (did not reject the null hypothesis), then that voxel passed the normality test. For the kurtosis model, the normality test was applied to the K_{app} distributions only, based on the simulation results in Chapter 3.

5.2.4 AIC and Model Selection

The AIC value was calculated for the monoexponential, kurtosis, and biexponential fits for all voxels. The model with the lowest AIC value for each voxel was selected to be the best model for

that voxel. The ΔAIC values were also calculated for both the biexponential-monoexponential and kurtosis-monoexponential model combinations.

5.3 Results

5.3.1 SNR

For reference when analysing the parameter estimates of the tissue data, both the SNR at the lowest *b*-value ($SNR_{b=0}$) and the signal-averaged SNR demonstrated in Chapter 2 were calculated for all voxels. The noise SD was calculated by subtracting the two reference images from each other and taking the SD of a selected group of voxels, with a resulting value of 350, which can be used to reference the amplitude parameter values later on in this chapter. Figure 69 shows the maps for $SNR_{b=0}$ and signal-averaged SNR.

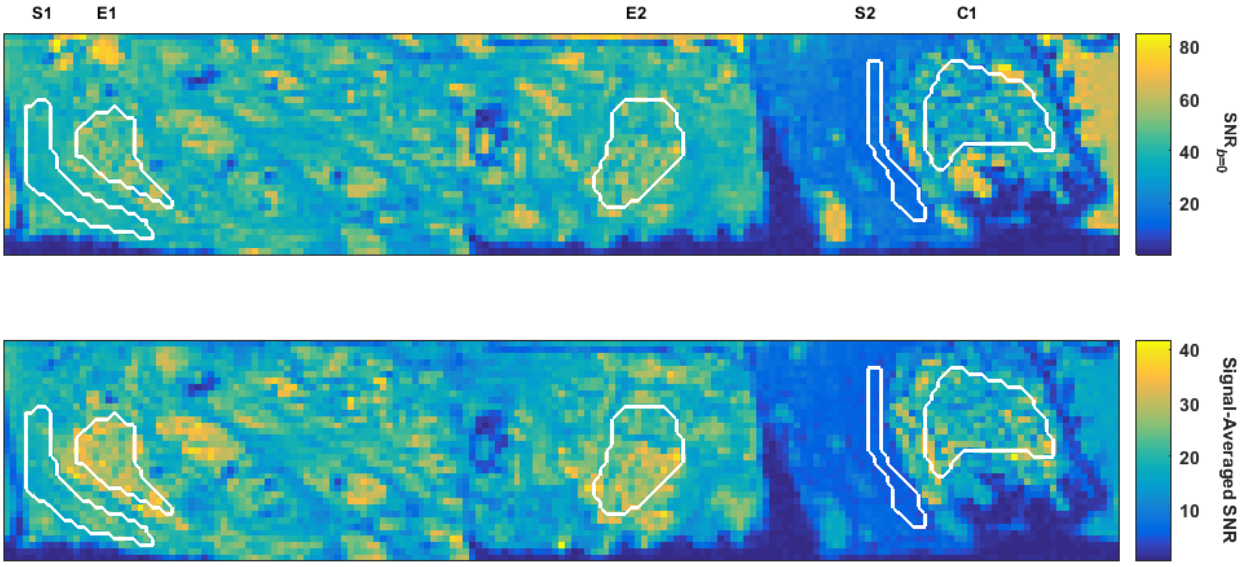

Figure 69 – $SNR_{b=0}$ and signal-averaged SNR of slice voxels

White contours indicate selected ROI from Figure 68.

For the voxels in all five ROI, the mean of the lowest *b*-value SNR and signal-averaged SNR were both calculated and are given in Table 11.

Table 11 – Mean $SNR_{b=0}$ and mean signal-averaged SNR for all voxels (*n* = number) in the five ROI shown in Figure 68

	S1 (*n*=156)	**S2** (*n*=89)	**E1** (*n*=123)	**E2** (*n*=189)	**C1** (*n*=248)
$SNR_{b=0}$	39	15	47	45	35
Signal-Averaged SNR	20	6	28	25	20

This table shows that the epithelial regions (E1 and E2) both had a higher mean $SNR_{b=0}$ and signal-averaged SNR that the other regions. Additionally, the stromal region S2 had both mean SNR values considerably lower than the other regions, and a signal-averaged SNR of 6 meant that many of the signals were affected by Rician bias.

5.3.2 Model Parameter Estimates

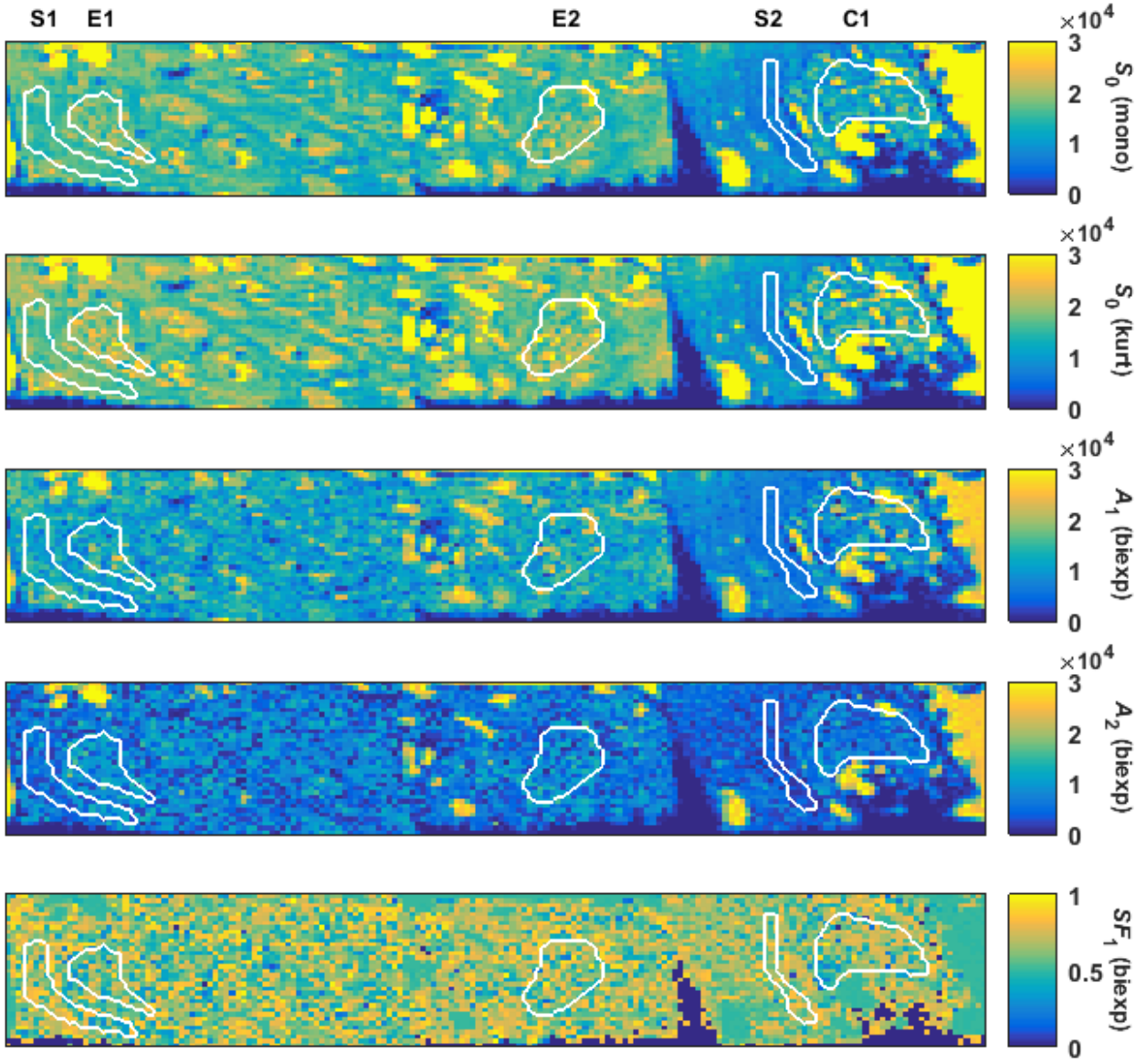

Figure 70 – Amplitude-related parameter pseudocolour plots from the monoexponential, kurtosis, and biexponential model fits

Zero-valued parameters near bottom indicate where no fits were performed due to low signal intensity. Note the increased noise in the biexponential parameter estimates (A_1, A_2, and SF_1)

The parameter estimates for the voxels fits were grouped together by similar parameter type, i.e. S_0, A_1, A_2, and SF_1 for the amplitude-related parameters, and ADC, D_{app}, D_1, D_2, and K_{app} for the decay-related parameters, with the parameter pseudocolour plots displayed in Figure 70 and Figure 71.

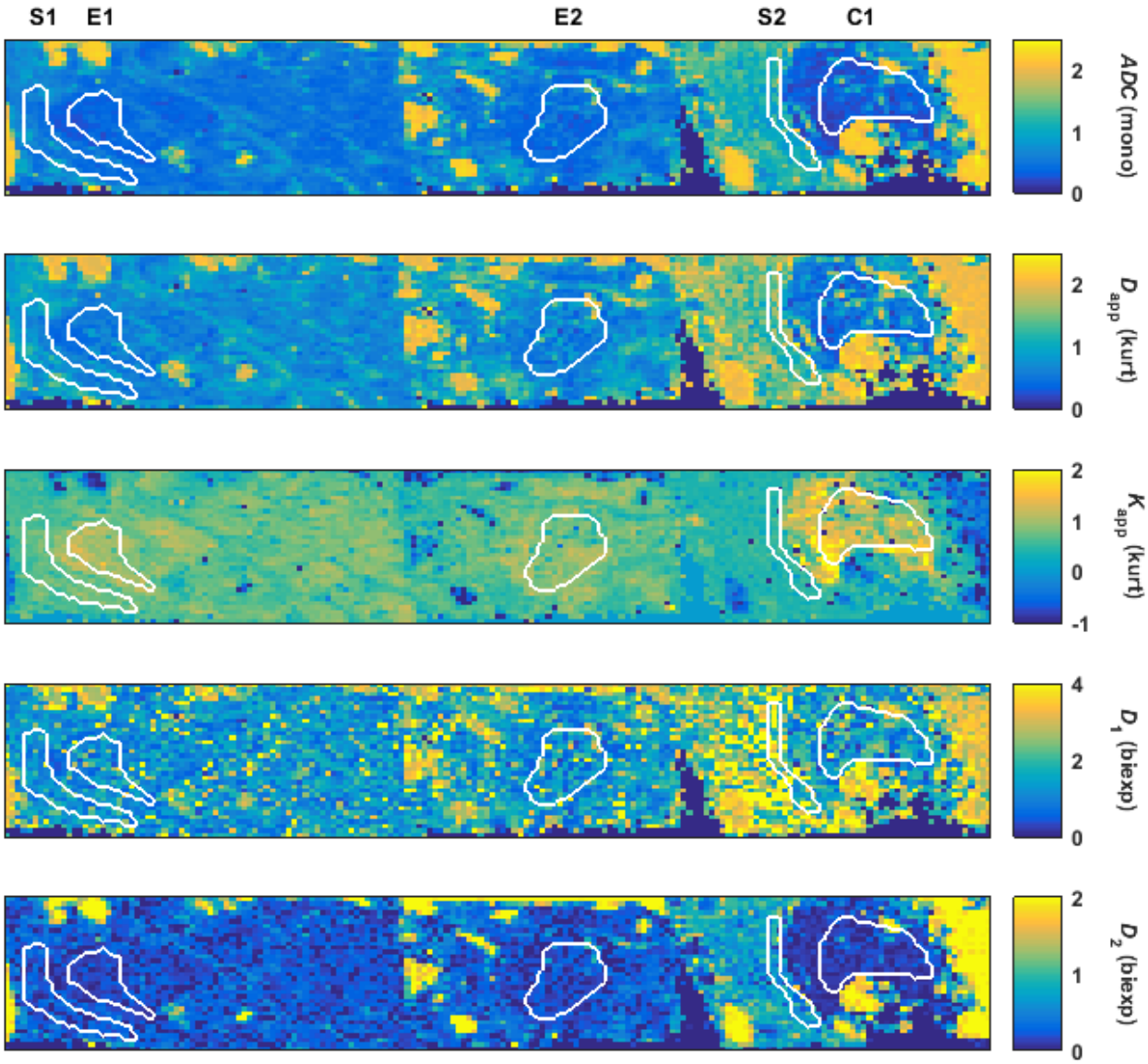

Figure 71 – Decay-related parameter pseudocolour plots from the monoexponential, kurtosis, and biexponential model fits

ADC, D_{app}, D_1, and D_2 are in units of $\mu m^2/ms$. Note that the biexponential parameter estimates (D_1 and D_2) are noisiest, and the kurtosis parameter (K_{app}) has values of -1 scattered across the data.

The S_0 maps for the monoexponential and kurtosis models showed similar signal magnitudes across all three tissue cores, with the glandular structure across most of the tissue visible, and the high signal, liquid-filled ducts easily visible. In the biexponential amplitude maps, A_1 and A_2, the ducts

were still visible at high signal value, but there was a considerable amount of noise across both parameter plots, which caused a loss of visibility of the tissue structure. This noise is even more pronounced in the SF_1 map, which lost much of the visible structure in the tissue. In the diffusion rate parameter plots, the high-signal ducts also had the highest values of ADC, D_{app}, D_1, and D_2, and these values were all around 2.1 μm^2/ms, suggesting the presence of free water. The value of D_{app} from the kurtosis model was slightly elevated in value versus the ADC across much of the tissue, especially in the right one-third of the plot. The stromal ROI had the highest mean values of ADC compared to the epithelial ROI, suggesting that this tissue had less diffusion restrictions. Distributed randomly across the parameter plot of K_{app} were several voxels where the value was effectively -1. This was a high indication of ill-conditioning present in these fits, since this phenomenon was also demonstrated in simulated data in Chapter 3. There were also indications of ill-conditioned fits for the biexponential model, as there were many values of D_1 that were at 4 μm²/ms or higher, which was well above the free diffusion coefficient of 2.1 (at 22° C). Since this was ex vivo tissue, there were no perfusion effects present, so freely diffusing water should have been the maximum decay rate seen. The values of D_2 in C1 were among the lowest ones in all tissue measurements here, but this was not noticeable in the D_1 plots. Again, the biexponential decay parameter plots were considerably noisier than the monoexponential or kurtosis decay parameters, especially for D_1, which had lost considerable structural detail.

Table 12 – Mean ± SD of the five ROI voxels for all model fit parameters

	Region / Parameter	**S1** (n=156)	**S2** (n=89)	**E1** (n=123)	**E2** (n=189)	**C1** (n=248)
Mono	S_0	15800 ± 1700	7100 ± 1200	17700 ± 2700	17900 ± 4400	13600 ± 6500
	ADC	0.55 ± 0.16	1.10 ± 0.29	0.36 ± 0.08	0.48 ± 0.26	0.43 ± 0.35
Kurtosis	S_0	17500 ± 2200	8000 ± 1400	19700 ± 3600	19900 ± 5300	15600 ± 7500
	D_{app}	0.77 ± 0.23	1.41 ± 0.32	0.56 ± 0.15	0.70 ± 0.34	0.69 ± 0.45
	K_{app}	0.71 ± 0.12	0.40 ± 0.33	1.00 ± 0.17	0.80 ± 0.25	1.06 ± 0.54
Biexponential	A_1	11700 ± 3400	5900 ± 1000	12500 ± 4000	13300 ± 4800	10600 ± 6000
	A_2	6300 ± 3100	3300 ± 1700	8600 ± 3500	8200 ± 4000	6500 ± 3800
	SF_1	0.65 ± 0.17	0.66 ± 0.12	0.59 ± 0.16	0.62 ± 0.17	0.60 ± 0.17
	D_1 (μm^2/ms)	1.28 ± 0.53	2.90 ± 1.65	1.28 ± 0.79	1.49 ± 1.08	1.51 ± 0.91
	D_2 (μm^2/ms)	0.23 ± 0.13	0.67 ± 0.38	0.14 ± 0.09	0.21 ± 0.20	0.17 ± 0.27

Table 12 displays the common way of examining the five ROI indicated in the parameter estimate plots by reporting a mean and standard deviation of the voxel estimates. For the monoexponential parameter estimates, the mean *ADC* values were lower in the epithelial regions (E1, E2, C1) than the stromal regions (S1, S2), which was also true for D_{app} and D_2. These D_2 findings agreed with the original conference paper findings in Section 5.1.1, but in the current study, the SF_1 values were very close in all regions and the D_1 values for S1 were actually lower than two of the epithelial regions. In addition to lower D_{app} values in the epithelial regions, the kurtosis parameter K_{app} was also higher in the epithelial regions than the stromal regions, showing that the mean signal value in these regions decayed slower, suggesting more restrictions to diffusion. The original conference paper reported a lower mean D_2 value in the C1 epithelium versus the other epithelial ROI, but Table 12 showed D_2 lower in E1 than C1, although the SD of the C1 voxels were much higher.

Histograms of Parameter Estimates

The histograms of the parameter estimates for all ROI are shown for all three models in Figure 72 (monoexponential), Figure 73 (kurtosis), and Figure 74 (biexponential).

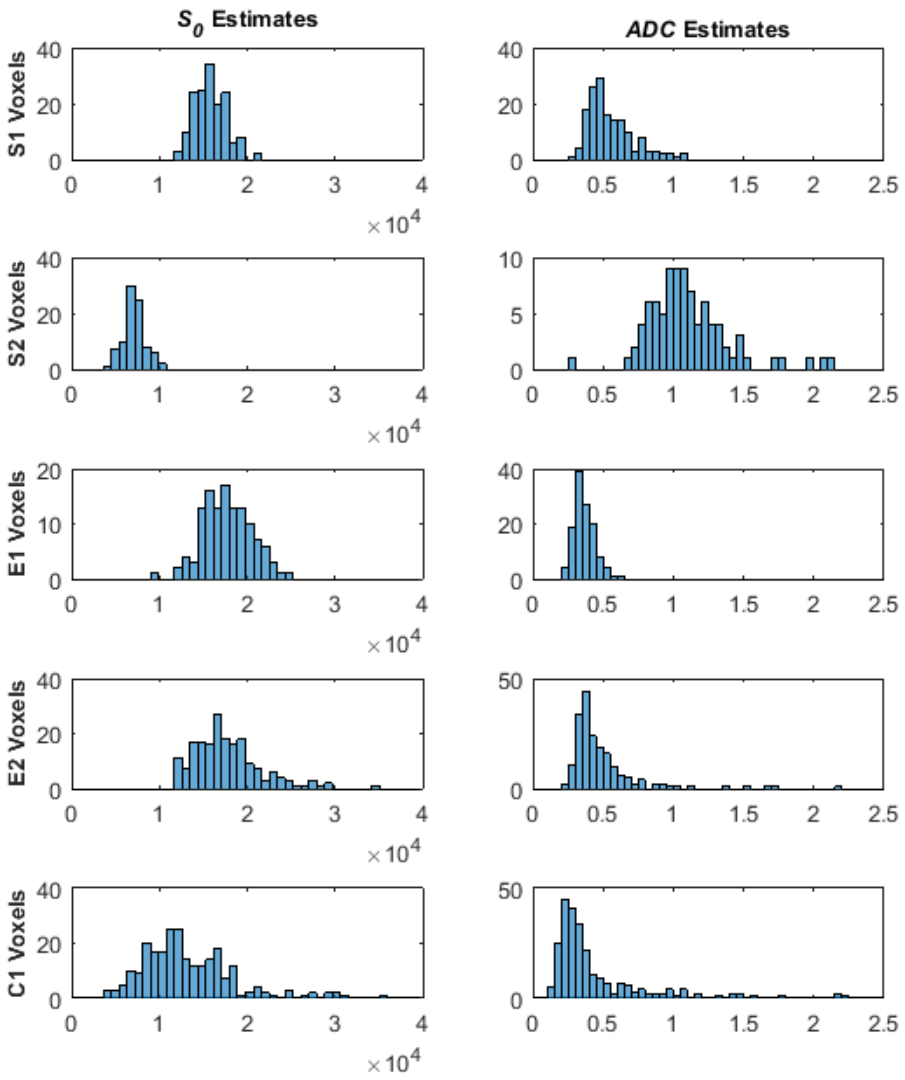

Figure 72 – Histograms of monoexponential parameter estimates for each ROI

ADC is in units of $\mu m^2/ms$. Note that the highest *ADC* values are around the free diffusion coefficient of 2.1.

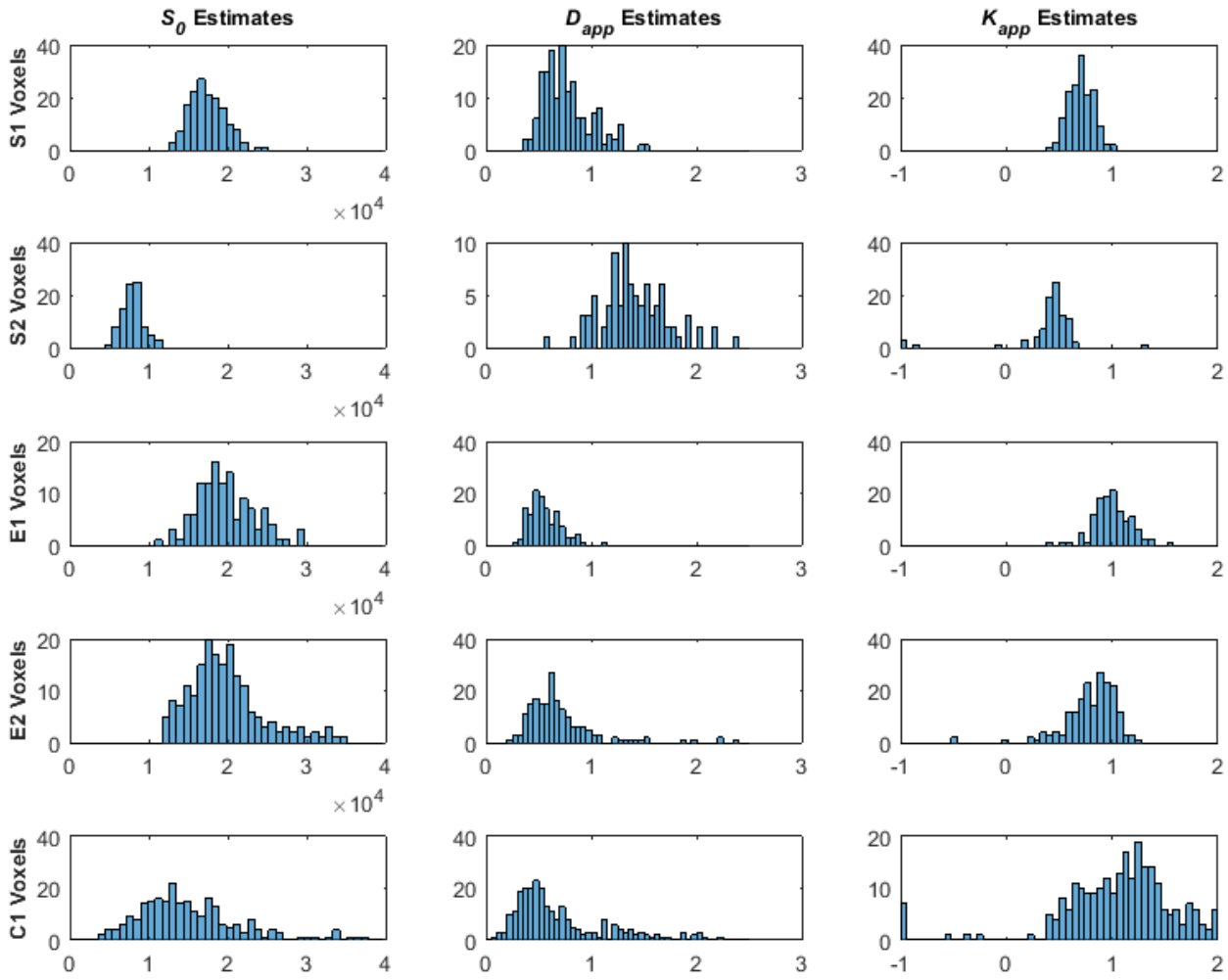

Figure 73 – Histograms of kurtosis parameter estimates for each ROI. D_{app} is in units of µm^2/ms

Note the increased uncertainty in the K_{app} estimates, along with several values found near -1. These phenomena were seen using the simulated data in Chapter 3, Figure 43.

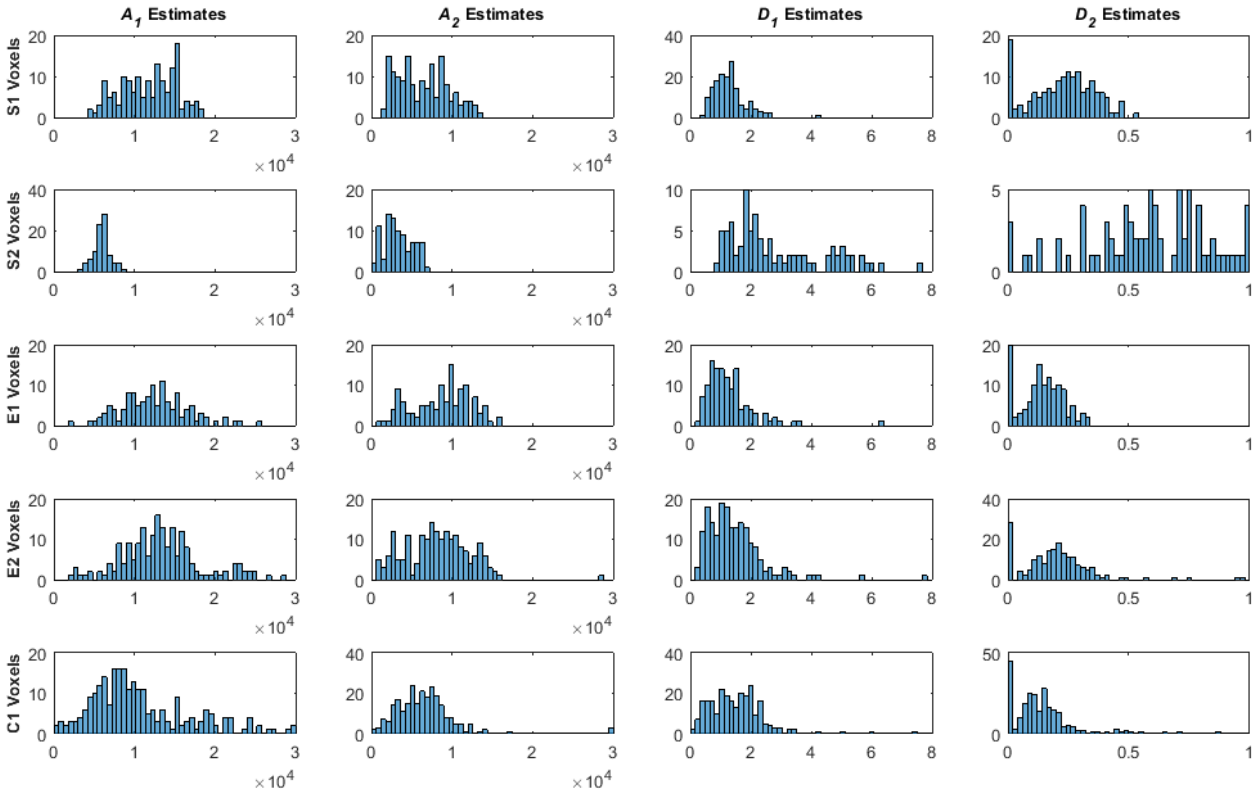

Figure 74 – Histograms of biexponential parameter estimates for each ROI

D_1 and D_2 are in units of μm^2/ms. These increases in uncertainty and outlier values of these parameter estimates are similar to the simulated histograms seen in Chapter 2, Figure 12. Note that many of the D_1 values are above the free diffusion limit of 2.1, indicating that these data do not reflect real phenomena.

The *ADC* estimate distributions in Figure 72 show decidedly non-normal distributions for S1, S2, E2, and C1, with several values found outside of the main distribution grouping. Since the monoexponential model was not known to have ill-conditioning problems, and these values were close to or below the free diffusion coefficient of 2.1, these values were most likely due to heterogeneity of the tissue structure in the ROI. The *ADC* distributions also showed that the majority of the distributions in the epithelial regions were lower than the stromal regions, reflected in the decreased mean values. However, the majority of the C1 distribution is seen to be lower than both the E1 and E2 distributions, and its higher mean value than S1 was due to the number of outlier values found well above the main distribution. In this case, the median values demonstrated this phenomenon better, which were 0.50 (S1), 1.1 (S2), 0.35 (E1), 0.40 (E2), and 0.30 (C1), indicating that the voxels in C1 were indeed lower for *ADC*. These median values, the mean values in Table 12, and the distributions in Figure 72, also illustrate that the S2 region has a much higher *ADC*, indicative of less restrictions to diffusion and a higher water content.

The D_{app} values for the kurtosis model in Figure 73 show a similar situation where C1 had a higher mean value than S1 per Table 12, but a lower median value 0.52 (C1) vs. 0.54 (E1). The kurtosis parameters also displayed signs of ill-conditioned fits, as both S2 and C1 had values of K_{app} at its

lower regression bound of -1, a phenomenon seen in Chapter 3 with simulated data, specifically Figure 43. The C1 K_{app} estimates also had a larger variance, but many more values at higher values near the upper regression bound of 2, giving it a higher mean value than E1. In Figure 74, the estimates of A_1 and A_2 for the biexponential model varied more than the S_0 values in the monoexponential and kurtosis models, but this could be attributed to ROI signal heterogeneity. More extreme outliers were seen in the D_1 estimates, where there were many estimates reported above 2.1 in the distributions, especially in region S1. Likewise, there were many D_2 estimates found at very small values near the lower bound of zero in all five ROI. These estimates were very similar to the effects seen in the biexponential chapter, specifically Figure 11 or Figure 12, and showed that the demonstrated effects of ill-conditioning on simulated data parameter estimates, presented previously in this thesis, were also present in actual tissue data.

5.3.3 Bootstrap Confidence Intervals

Confidence intervals from the parametric bootstrap distributions for all voxels from all three model fits were calculated for each ROI, and the effects of ill-conditioning on the estimates and their confidence intervals can be seen by examining two of the ROI, S1 and C1. Figure 75 shows the confidence intervals for the monoexponential parameter estimates of S1, illustrating that for nearly all of the fits, the intervals were symmetric around the estimates and their ranges all contained realistic values for both parameters. The intervals grew in size when the estimated values of either parameter increased, a phenomenon better illustrated in Figure 76 for the C1 estimates. There also seems to be more heterogeneity in the C1 voxels than S1, with more values found at higher *ADC* values. The kurtosis intervals for S1, seen in Figure 77, showed similar interval ranges and parameter estimates for S_0 and D_{app}, however the K_{app} estimates had a few intervals where the range was larger and the lower bounds had negative values. These K_{app} interval effects were even more pronounced for C1 in Figure 78, with several intervals having lower bounds below -3, with some values as low as -35,000, a highly unrealistic value indicative of extreme ill-conditioning.

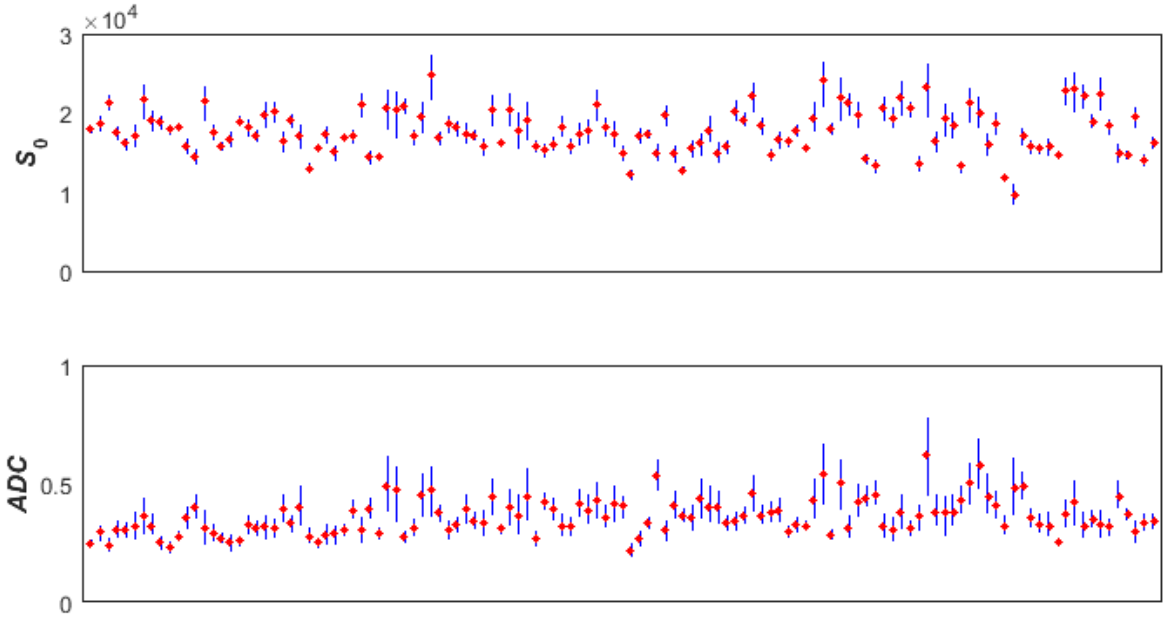

Figure 75 – Monoexponential parameter estimates (red points) plus bootstrap confidence intervals (blue lines) for voxels in E1 ROI

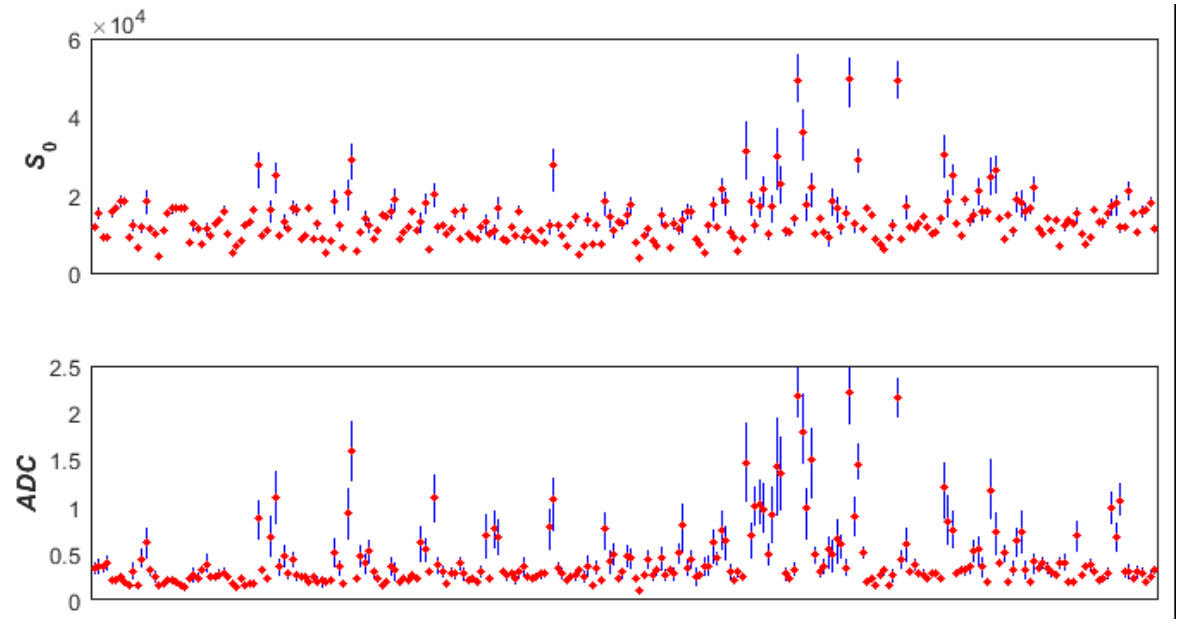

Figure 76 – Monoexponential parameter estimates (red points) plus bootstrap confidence intervals (blue lines) for voxels in C1 ROI

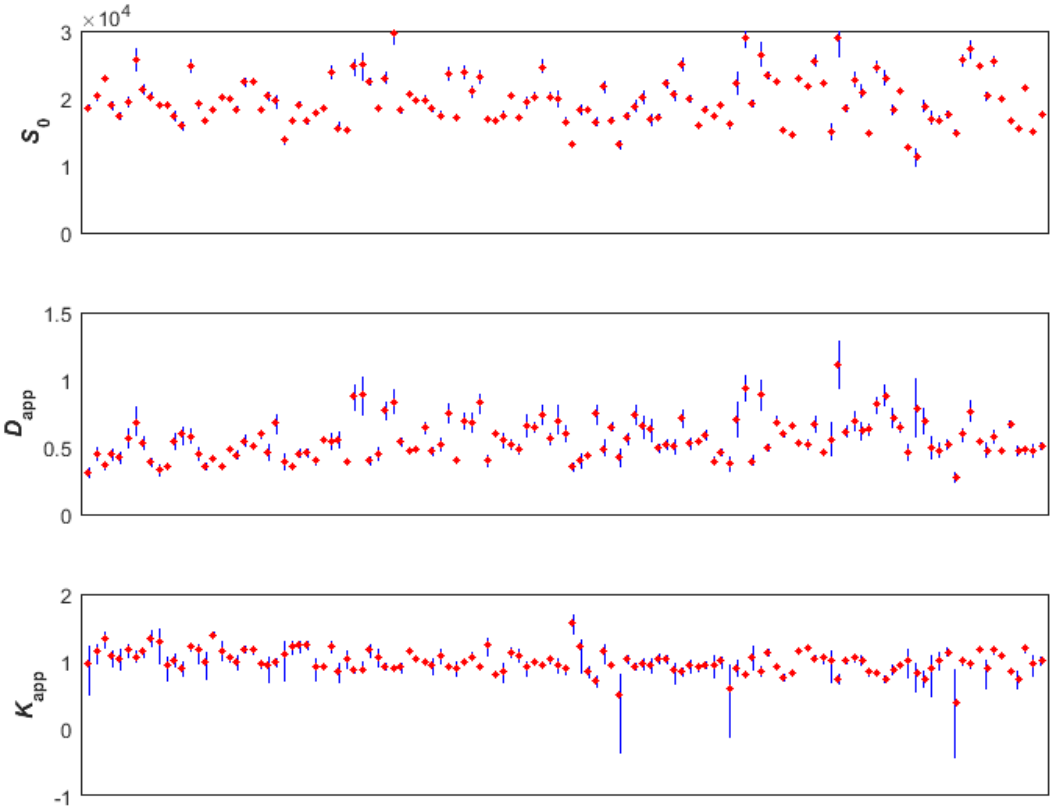

Figure 77 – Kurtosis parameter estimates (red points) plus bootstrap confidence intervals (blue lines) for voxels in E1 ROI

Note that some of the K_{app} intervals stretch into negative values.

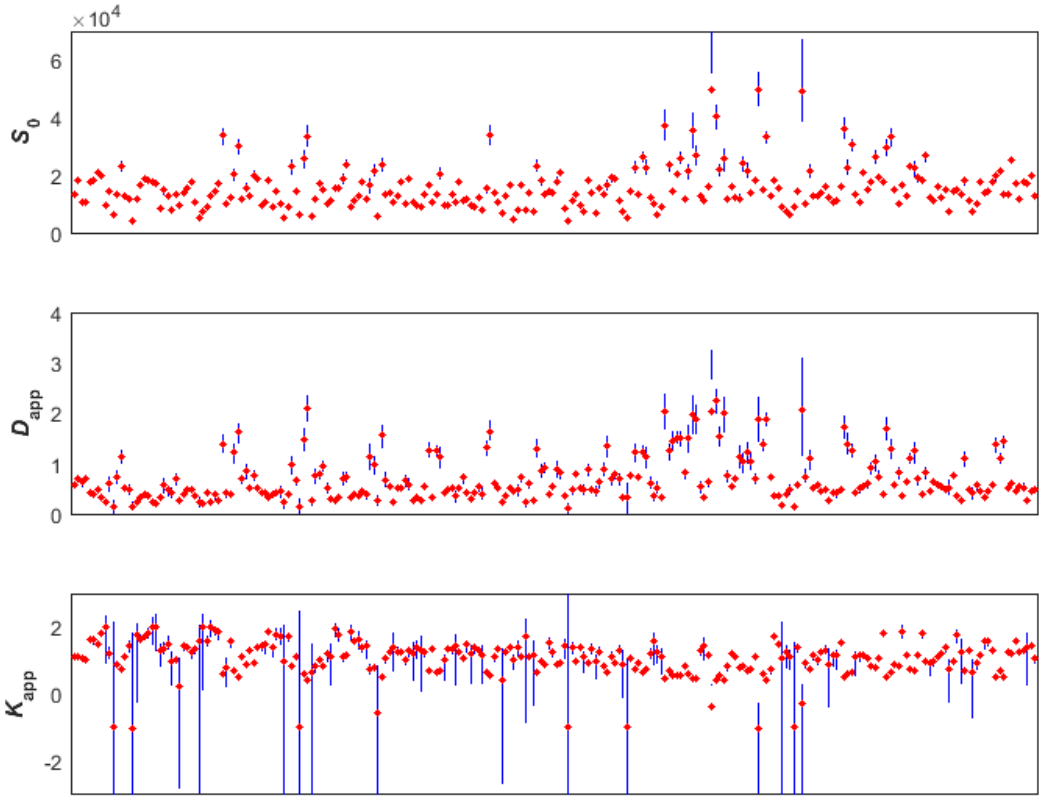

Figure 78 – Kurtosis parameter estimates (red points) plus bootstrap confidence intervals (blue lines) for voxels in C1 ROI

There are more K_{app} intervals found at negative values, indicating there is likely to be more ill-conditioned fits in this ROI.

The confidence intervals for the biexponential parameter estimates shown in Figure 79 for S1 were asymmetric for a considerable number of all four parameter fits. Both A_1 and D_1 had a few intervals where the upper and/or lower bounds were outside the displayed scale, and while a few of these intervals had unrealistic initial estimated values greater than 3, a few of these values occurred at estimates of 2 or less, illustrating that the initial estimates may have had realistic values, but repeated fits of this same voxel would be likely to have estimates that widely vary. The slow decay component parameters (A_2 and D_2) had intervals with a smaller variance around the original estimates. The intervals for C1 in Figure 80, however, had many values that go off of the displayed scale for all four parameters, with several upper bounds for D_1 over 150, with a highest value of 335. Like the kurtosis model intervals, these values were highly unrealistic and indicative of ill-conditioning in the algorithm. While, the D_1 intervals in C1 had upper values in the hundreds, the highest outlier value of the original fit estimates in Figure 74 was less than 8. Part of this was due to the regression upper bound of 10, where this bound was removed for the bootstrap estimates.

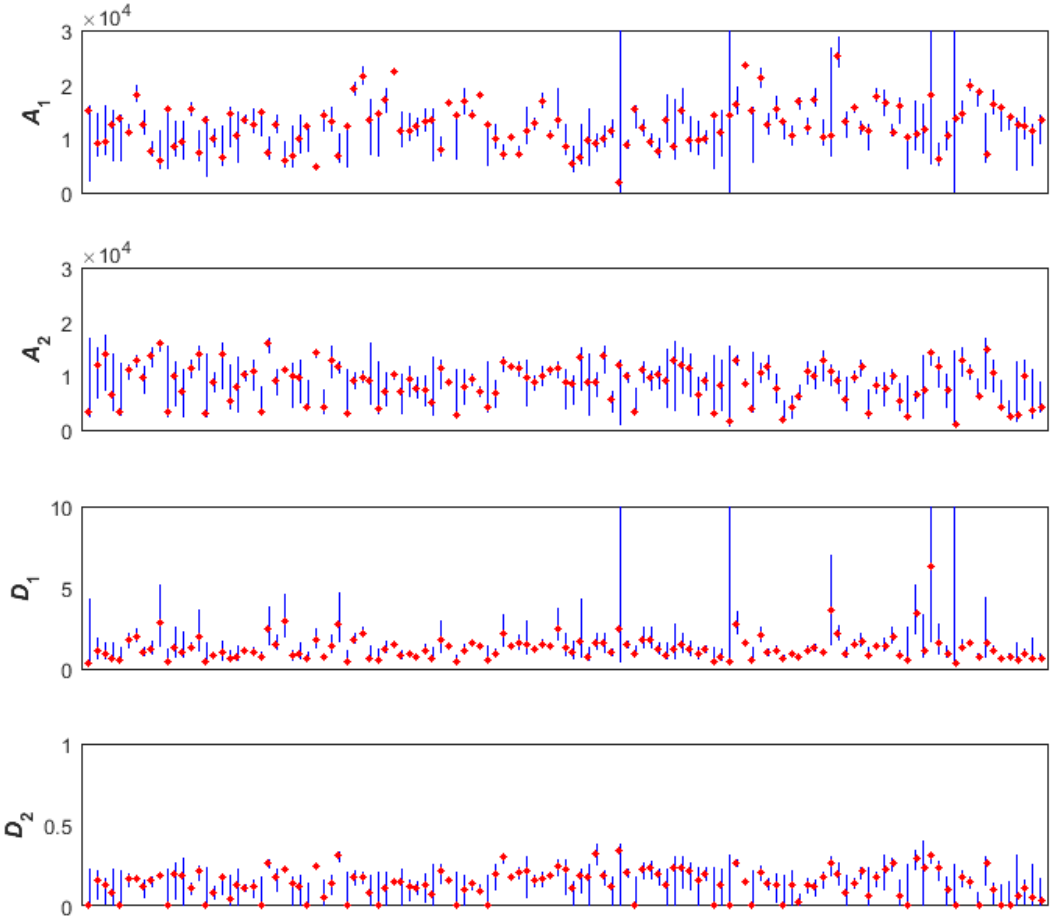

Figure 79 – Biexponential parameter estimates (red points) plus bootstrap confidence intervals (blue lines) for voxels in E1 ROI

Many of the D_1 intervals reach values of 10 or more.

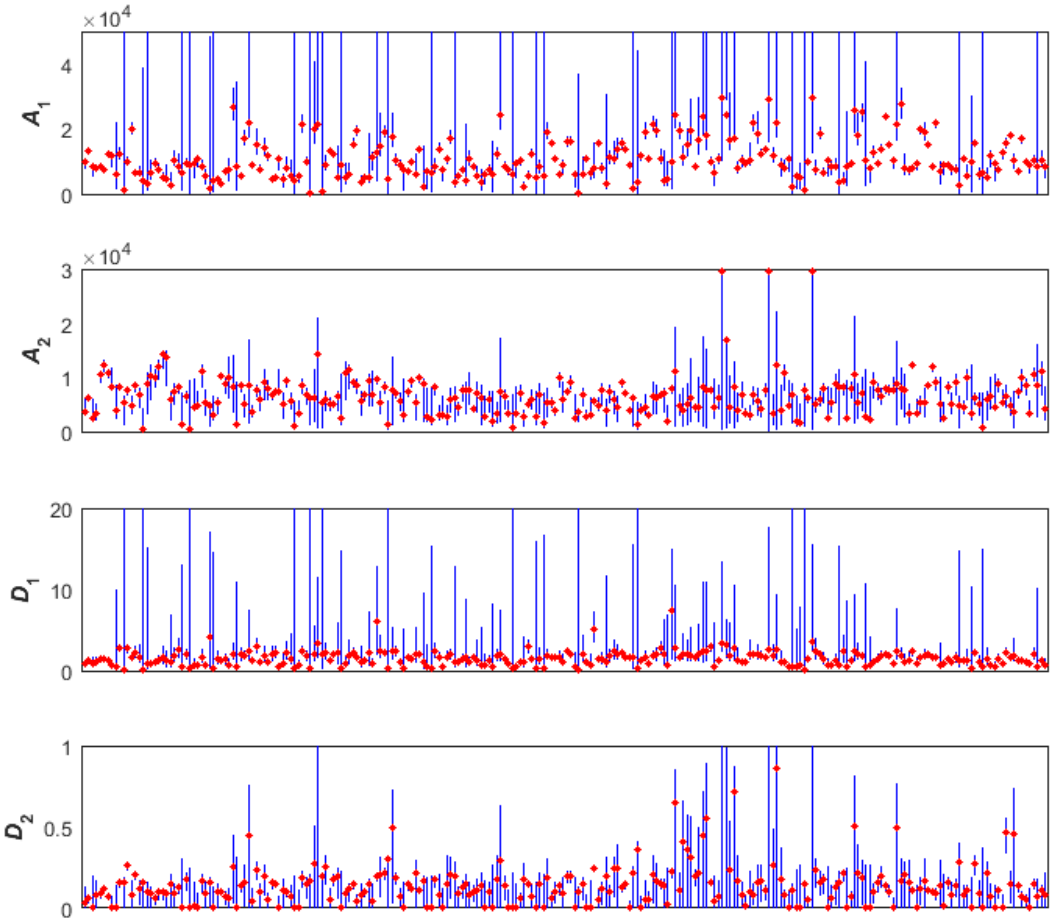

Figure 80 – Biexponential parameter estimates (red points) plus bootstrap confidence intervals (blue lines) for voxels in C1 ROI

The large interval values here show more ill-conditioned fits in this ROI were likely.

An additional illustration of why these bootstrap confidence intervals are much higher than the original estimates is shown in Figure 81. The original signal in red, with a D_1 estimate of 7.4, has a small amount of curvature, whereas the bootstrapped signal in blue happened to have resampled residuals such that the resulting signal was straighter and more monoexponential than the original signal. This resulted in the signal fit to be more ill-conditioned, with a new D_1 estimate of 46.6, and demonstrated the effectiveness of bootstrap resampling, since signals that were closer to being monoexponential had a higher likelihood that these large bootstrap parameter values will manifest themselves.

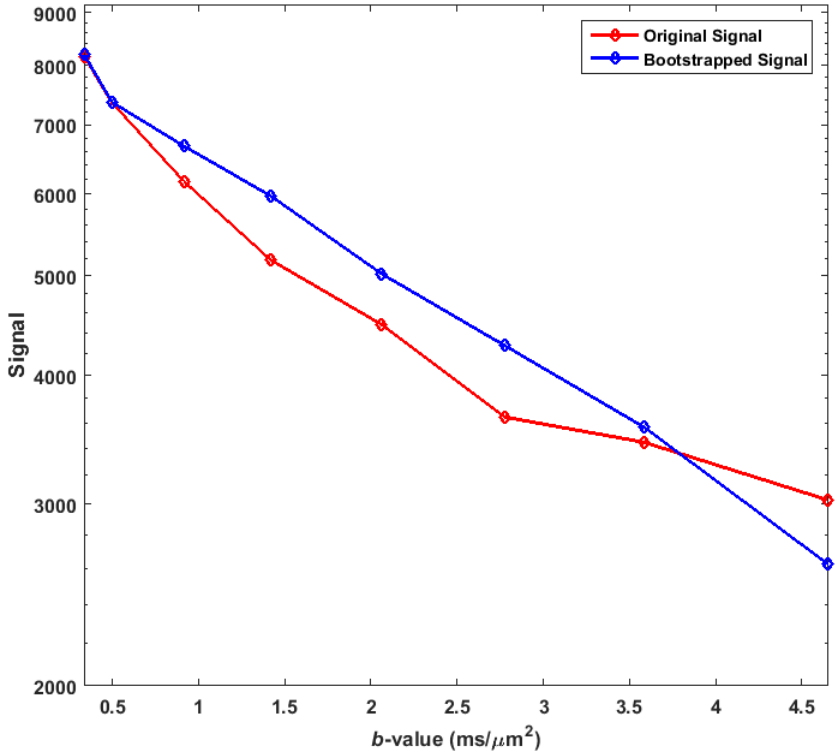

Figure 81 – Semilog plot of a selected voxel from the C1 ROI showing the original signal (red) plus a signal created via bootstrap resampling (blue)

The bootstrapped signal is straighter and closer to a monoexponential signal, leading to an estimated D_1 value of 46.6, versus the estimated D_1 value of 7.4 for the original signal. However, both signals are higher than the free diffusion limit of 2.1, indicating that ill-conditioning affected both fits.

5.3.4 Normality Testing of Bootstrap Estimates

The results from Section 4.3.7 showed that using a normality test on the bootstrap parameter distributions was an effective method to diagnose when a given biexponential fit is likely to be ill-conditioned with high uncertainty in the parameter estimates. When the normality test was applied in this study to the biexponential model fits of each ROI, the number of voxels that failed to have one normal bootstrap distribution was considerable, with a percentage of failing voxels in each ROI of 43% (S1), 73% (S2), 31% (E1), 38% (E2), and 39% (C1). As Chapter 4 showed, this indicates that the underlying tissue in these voxels produced a signal that was effectively monoexponential given the SNR and b-values. There were a considerable number of D_1 estimates higher than 2.5 for S2 in Figure 74, as well as a much larger variance in the D_2 estimates, and this is probably the reason for the larger number of voxels that failed for that ROI. After removing the voxels that failed the normality test from each ROI group, the mean and standard deviation were again tested on the ROI estimate distributions with the results given in Table 13. These results show several changes in the mean values of the various parameters in the ROI, specifically, the mean value of D_2 for region C1 is now lower than S1. More importantly, compared to the results in Table 12, every ROI/parameter distribution has a lower standard deviation apart from the A_1

estimates for E1. The improvement in estimate deviation is largely due to the elimination of outlier measurements in the distributions, as shown in Figure 83. While a large number of D_1 estimates above 2.5 remained in the distributions, especially in S2, the amount of low D_2 estimates grouped near zero all but disappeared when compared to the original estimates in Figure 74.

Table 13 – Mean ± SD of biexponential fit parameters after removing voxels that failed normality testing

Region / Parameter	**S1** (n=89)	**S2** (n=24)	**E1** (n=85)	**E2** (n=117)	**C1** (n=152)
A_1	12000± 3200	6600 ± 900	12300 ± 4400	14000 ± 4700	11900 ± 5300
A_2	7000 ± 2600	3100 ± 1100	10100 ± 2400	9300 ± 2600	7000 ± 2400
SF_1	0.63 ± 0.14	0.69 ± 0.07	0.54 ± 0.11	0.59 ± 0.11	0.61 ± 0.14
D_1 ($\mu m^2/ms$)	1.46 ± 0.50	2.47 ± 0.71	1.50 ± 0.77	1.67 ± 0.85	1.77 ± 0.65
D_2 ($\mu m^2/ms$)	0.27 ± 0.10	0.53 ± 0.20	0.18 ± 0.06	0.23 ± 0.10	0.15 ± 0.11

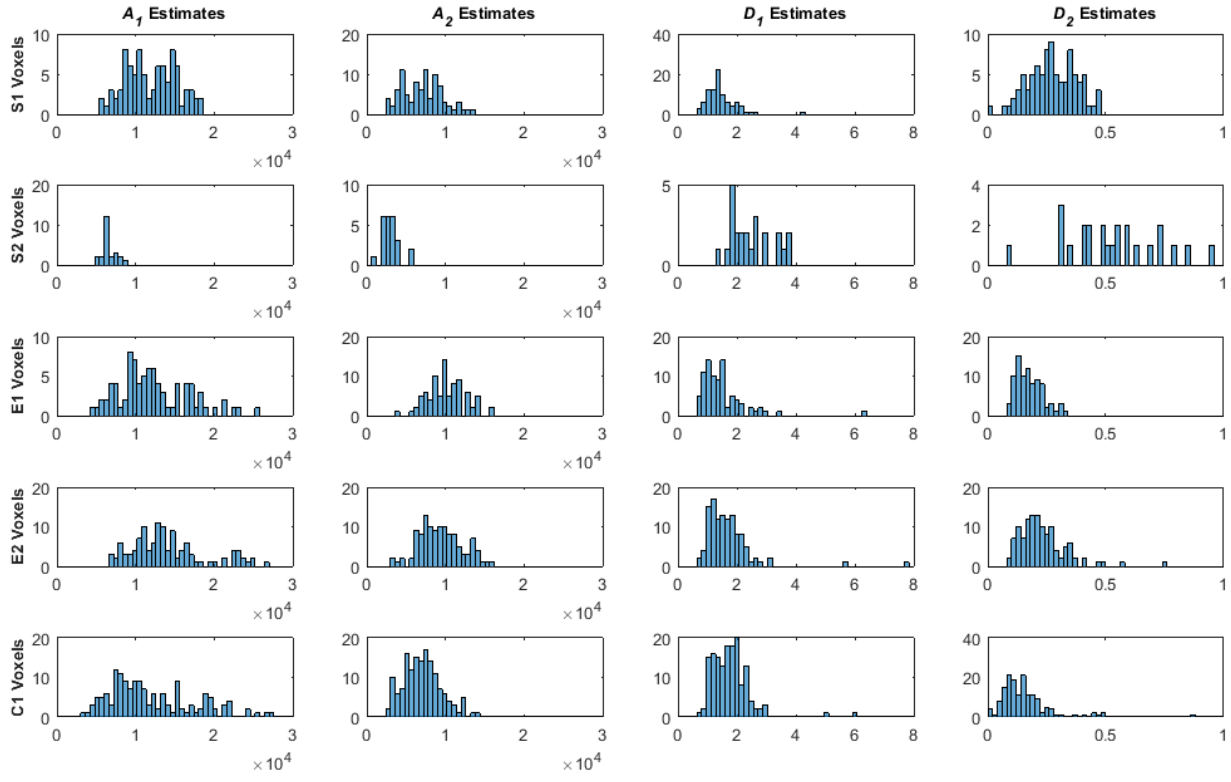

Figure 82 – Histograms of biexponential parameter estimates for each ROI with voxels that failed normality testing removed

D_1 and D_2 are in units of $\mu m^2/ms$. Many of the outlier values in Figure 74 have now been removed.

Removing voxels that failed normality testing also considerably reduced the number of fits with large confidence intervals, as shown in Figure 83. When compared to Figure 80, all D_1 intervals that were previously off the scale are gone, with only a few of these intervals found in the other parameters.

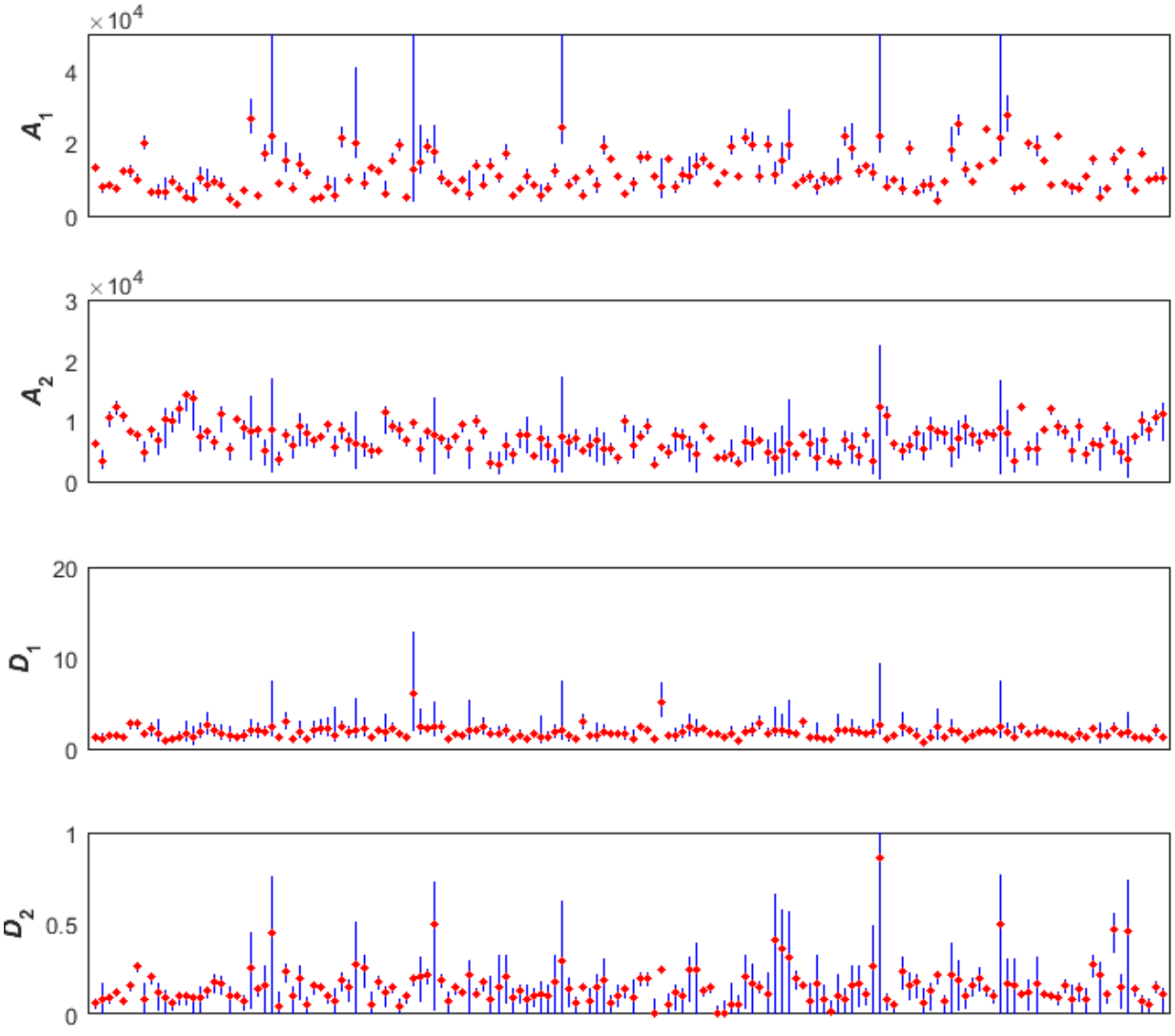

Figure 83 – Biexponential parameter estimates (red points) plus bootstrap confidence intervals (blue lines) for voxels in C1 ROI after removing voxels that failed normality testing

There is a reduction in the large interval values seen in Figure 80.

Kurtosis Model Estimates

Table 14 – Mean ± SD of kurtosis fit parameters after removing voxels that failed normality testing

Region / Parameter	**S1** (*n*=108)	**S2** (*n*=38)	**E1** (*n*=90)	**E2** (*n*=142)	**C1** (*n*=185)
S_0	17800 ± 2100	8200 ± 1400	20100 ± 3800	20500 ± 5100	16600 ± 6100
D_{app}	0.79 ± 0.24	1.55 ± 0.35	0.58 ± 0.16	0.73 ± 0.34	0.74 ± 0.41
K_{app}	0.71 ± 0.12	0.52 ± 0.15	1.00 ± 0.16	0.82 ± 0.20	1.14 ± 0.39

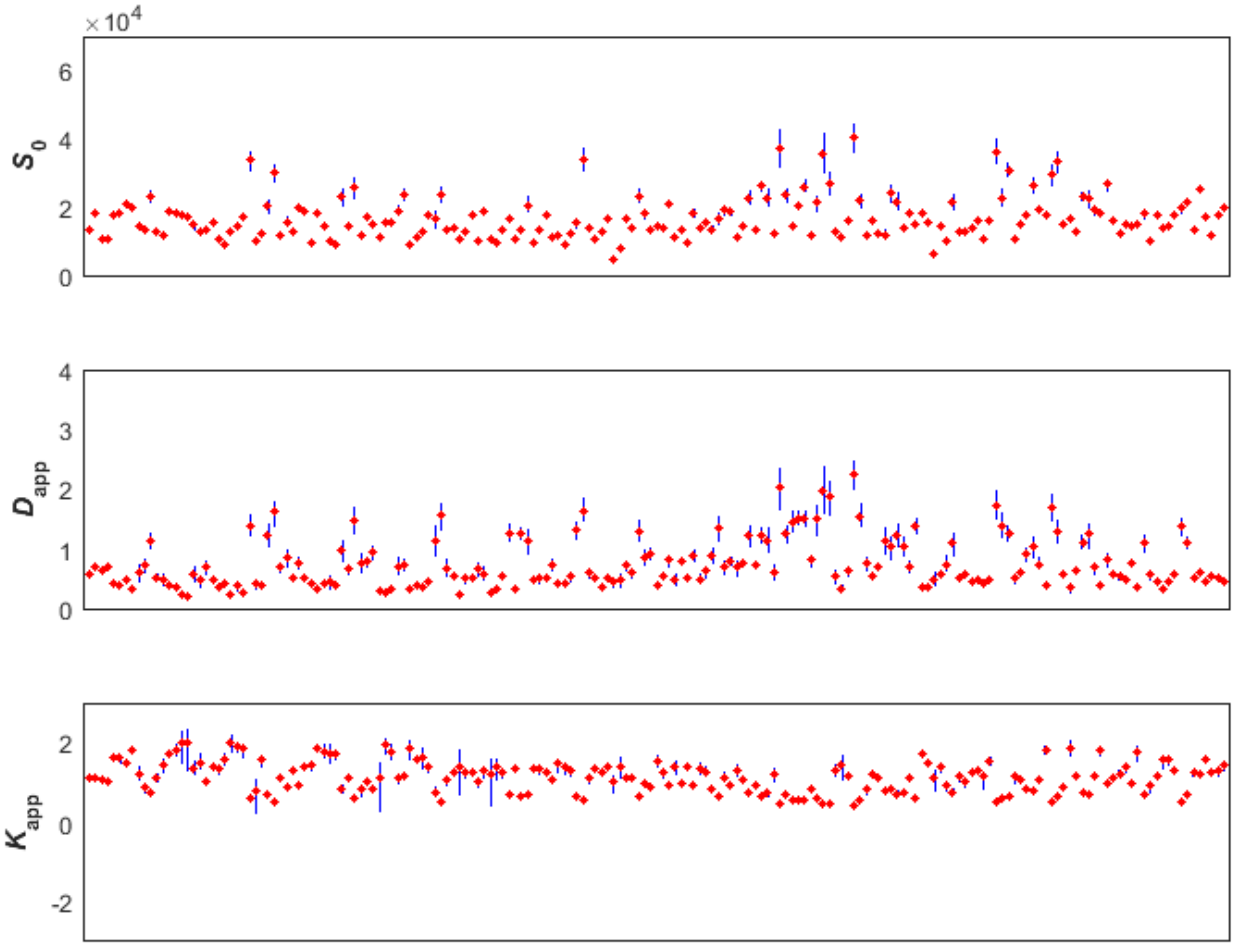

Figure 84 – Kurtosis parameter estimates (red points) plus bootstrap confidence intervals (blue lines) for voxels in C1 ROI after removing voxels that failed normality testing

Applying the normality test to the K_{app} bootstrap distribution also led to a reduction in the number of voxels in each ROI, specifically, a percentage of failing voxels in each ROI of 31% (S1), 73% (S2),

27% (E1), 25% (E2), and 25% (C1). After removing these voxels from each ROI, the mean and SD values of the S_0 and D_{app} remained close to the same, as shown in Table 12, whereas the SD in the K_{app} estimates were reduced for all ROI but S1, where it remained the same.

5.3.5 Model Selection with AIC

The results of the AIC calculations for all slice voxels in the study are shown in Figure 85, and after excluding voxels where no fits were performed, for this slice, the best model selection rates by the AIC were 6% (monoexponential), 43% (kurtosis), and 51% (biexponential).

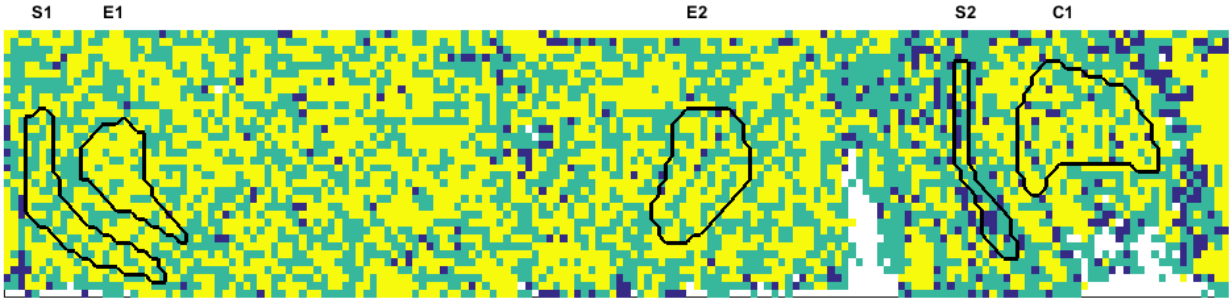

Figure 85 – Selected best model via lowest AIC value for all voxels with yellow = biexponential, green = kurtosis, and blue = monoexponential

White voxels had no fits performed due to low signal.

Of the five ROI in this study, while most voxels have the kurtosis or biexponential model chosen as best, there is no model selected as best for all voxels in a specific ROI.

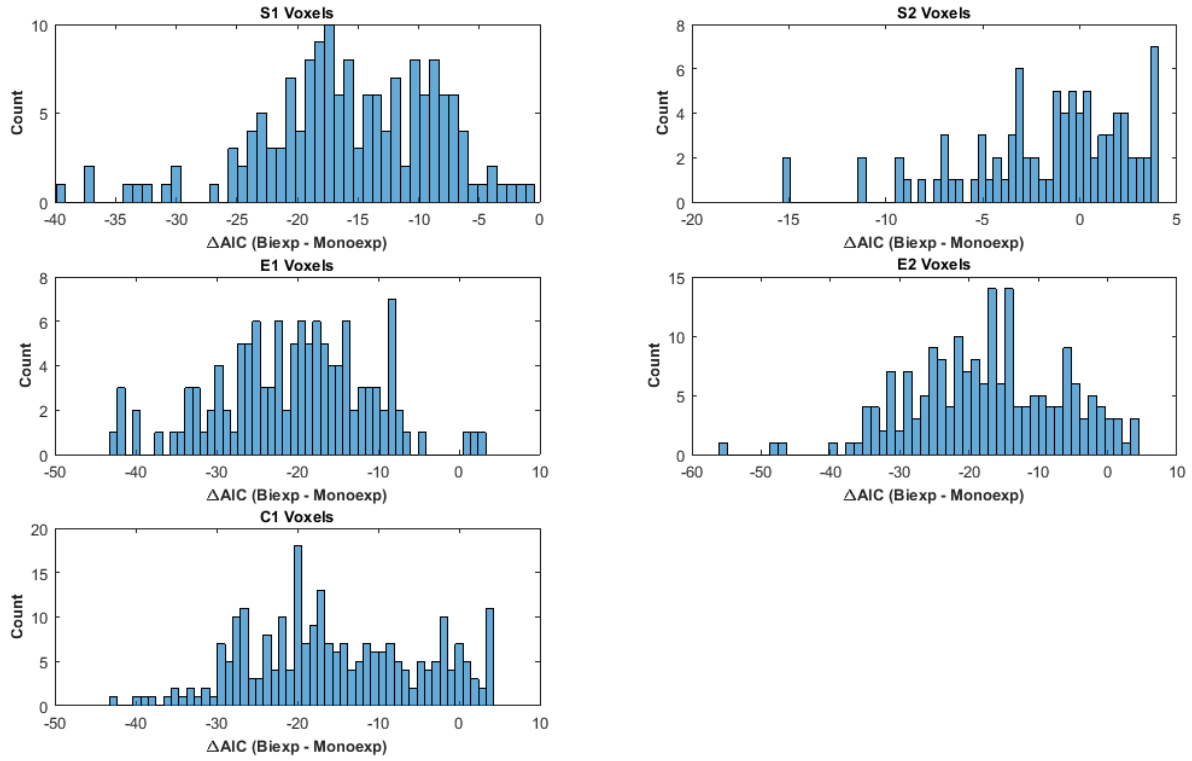

Figure 86 – ΔAIC values between the biexponential and monoexponential models for all ROI voxels

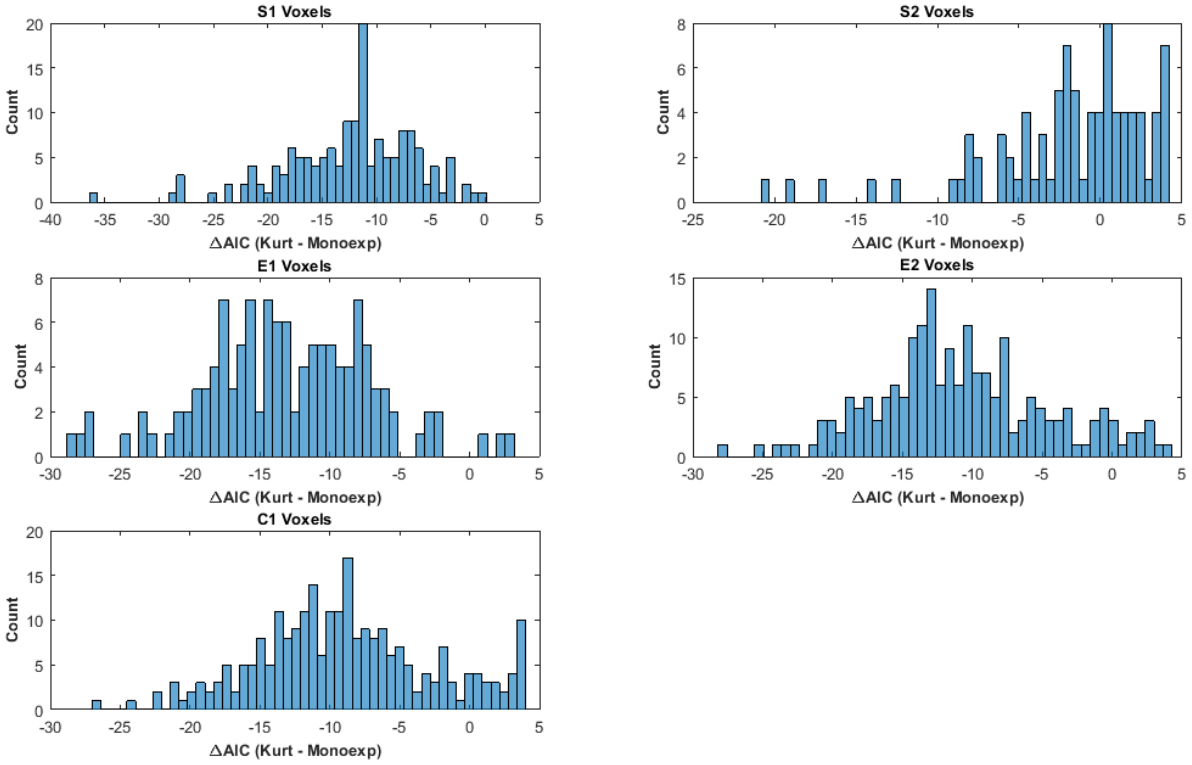

Figure 87 – ΔAIC values between the kurtosis and monoexponential models for all ROI voxels

The model selection rates were broken down further for each ROI by comparing the ΔAIC values of the biexponential model (Figure 86) and the kurtosis model (Figure 87) versus the monoexponential model. These histograms show that for most voxels in all ROI, either the biexponential or kurtosis model would be selected over the monoexponential model. However, region S2 appears to have many more voxels where the monoexponential model is selected as best, and very few voxels where the ΔAIC value between either complex model and the monoexponential model is greater than -10. As was demonstrated in Section 4.3.7, though, having the AIC select the biexponential or kurtosis models as best did not translate into better accuracy of their parameters. Section 4.3.5 established that a more negative ΔAIC value between either the biexponential or kurtosis model and the monoexponential model meant that particular signal was less like a monoexponential signal and, therefore, less likely to have ill-conditioned estimates. These ΔAIC values were used to eliminate any voxels from the ROI where the difference between the biexponential or kurtosis model and the monoexponential model was greater than -10.

Biexponential Model Estimates

Eliminating voxels with ΔAIC between the biexponential and monoexponential fits led to a reduction in the number of voxels for all ROI, specifically, 26% (S1), 96% (S2), 13% (E1), 26% (E2), and 33% (C1). This eliminated nearly all of the voxels in S2, leaving only four voxels for measurement. The mean and SD of the parameter estimates for the voxels left in each ROI are given in Table 15. Like the effect that removing non-normal bootstrap distributions had in reducing the SD of all ROI parameter estimates in Table 13, removing voxels with a ΔAIC > -10 reduced the SD of all ROI parameter estimate distributions other than the A_1 estimates for E1, and the D_2 estimates for S1. As the parameter estimate distributions in Figure 88 show, many of the outlier measurements for the D_1 estimates have been removed in all ROI, however several of the D_2 estimates near zero still remain. Removing these voxels all significantly reduced the number of C1 estimates that had large confidence intervals in Figure 89, even more so than the normality testing results did in Figure 83.

Table 15 – Mean ± SD of biexponential fit parameters after voxels with ΔAIC > -10 (see Figure 86) were removed

Region / Parameter	**S1** (*n*=116)	**S2** (*n*=4)	**E1** (*n*=107)	**E2** (*n*=139)	**C1** (*n*=167)
A_1	12300 ± 3200	5500 ± 1000	12500 ± 4100	13700 ± 4500	11400 ± 5100
A_2	6300 ± 2900	2000 ± 1600	9300 ± 3100	8600 ± 3100	7000 ± 3000
SF_1	0.66 ± 0.15	0.73 ± 0.03	0.58 ± 0.14	0.62 ± 0.14	0.61 ± 0.14
D_1 (μm^2/ms)	1.29 ± 0.46	1.79 ± 0.69	1.31 ± 0.62	1.45 ± 0.58	1.59 ± 0.63
D_2 (μm^2/ms)	0.22 ± 0.13	0.28 ± 0.21	0.14 ± 0.08	0.19 ± 0.10	0.14 ± 0.21

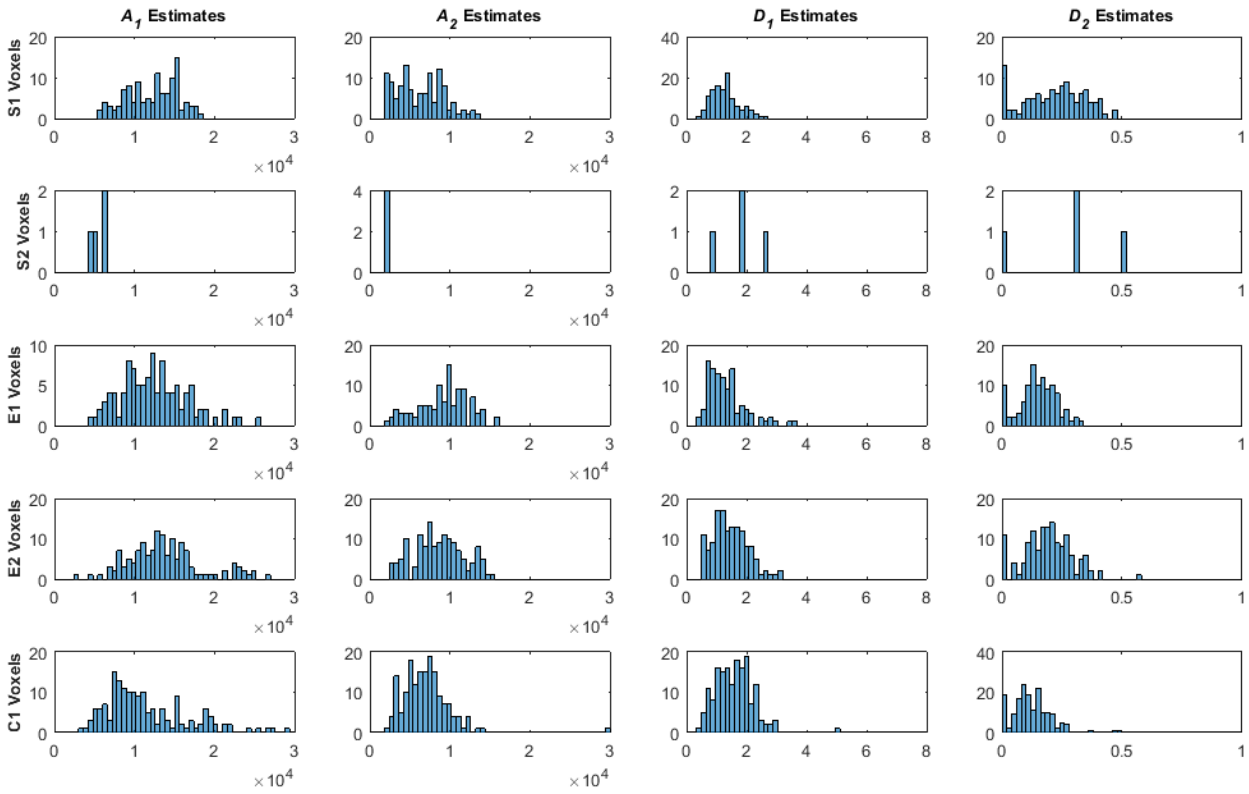

Figure 88 – Histograms of biexponential parameter estimates for each ROI after voxels with ΔAIC > -10 (see Figure 86) were removed. D_1 and D_2 are in units of μm^2/ms

This method also removed many of the outlier values from Figure 74.

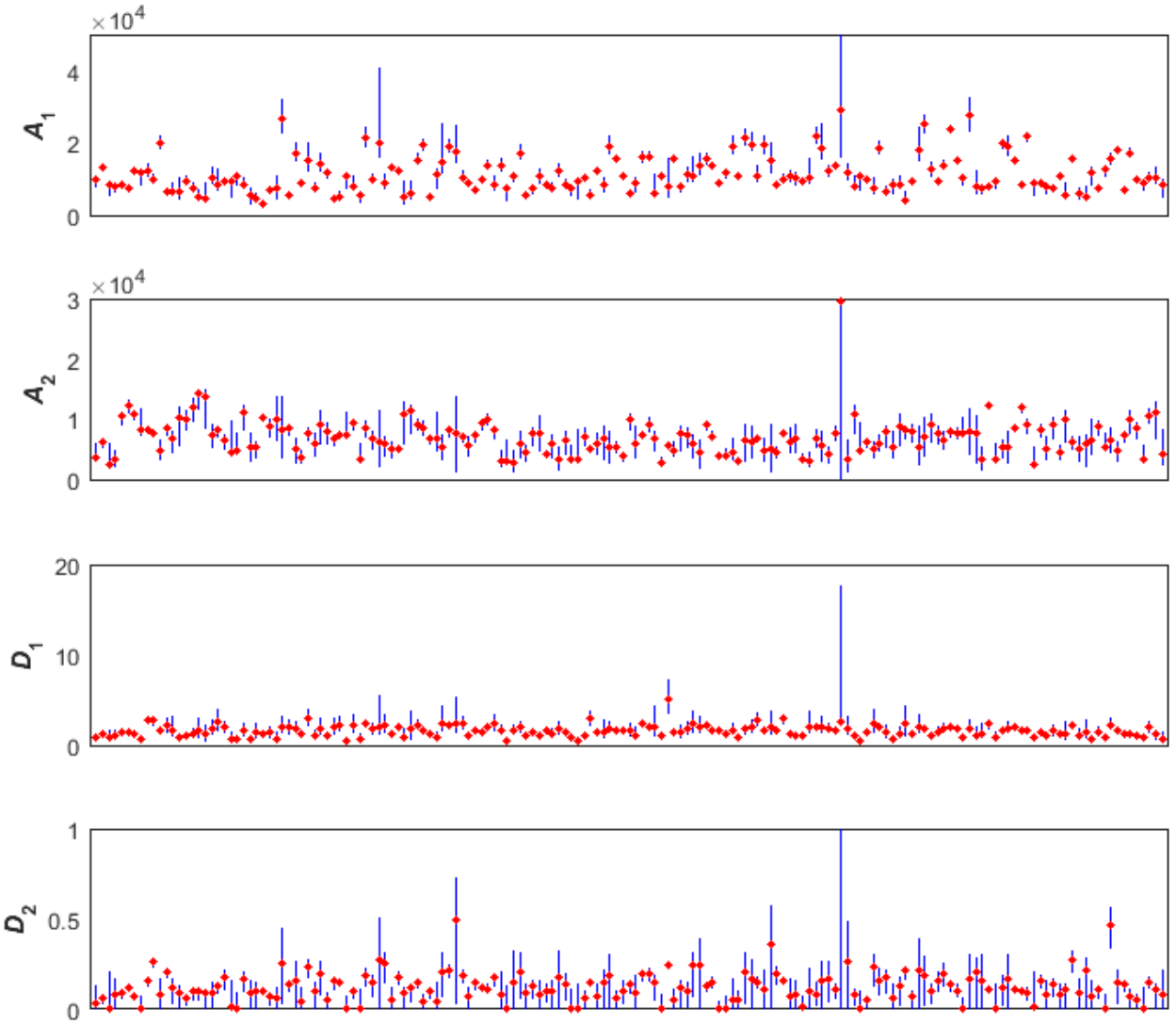

Figure 89 – Biexponential parameter estimates (red points) plus bootstrap confidence intervals (blue lines) for voxels in C1 ROI after voxels with ΔAIC > -10 (see Figure 86) were removed

Kurtosis Model Estimates

Eliminating voxels with ΔAIC between the kurtosis and monoexponential fits greater than -10 also led to a reduction in the number of voxels for all ROI, specifically, 31% (S1), 94% (S2), 27% (E1), 40% (E2), and 54% (C1). This led to an overall reduction in SD of the estimates for many of the ROI/parameter combinations, as shown in Table 16, however, a few ROI estimates had the same or higher SD values, specifically, S1 and S2. Eliminating these voxels also had the effect of eliminating the K_{app} outliers in the estimates apart from one -1 value in C1 as seen in Figure 90. As the C1 estimates with added bootstrap intervals show in Figure 91, all but one of the large K_{app} intervals seen in Figure 78 were eliminated.

Table 16 – Mean ± SD of kurtosis fit parameters after voxels with ΔAIC > -10 (see Figure 87) were removed

Region / Parameter	**S1** (*n*=103)	**S2** (*n*=5)	**E1** (*n*=84)	**E2** (*n*=114)	**C1** (*n*=114)
S_0	18000 + 2200	7900 + 1300	20500 + 3400	21200 + 4400	17500 + 6200
D_{app}	0.81 + 0.23	1.42 + 0.58	0.58 + 0.14	0.73 + 0.26	0.73 + 0.38
K_{app}	0.72 + 0.12	0.67 + 0.37	1.03 + 0.15	0.89+ 0.16	1.14 + 0.42

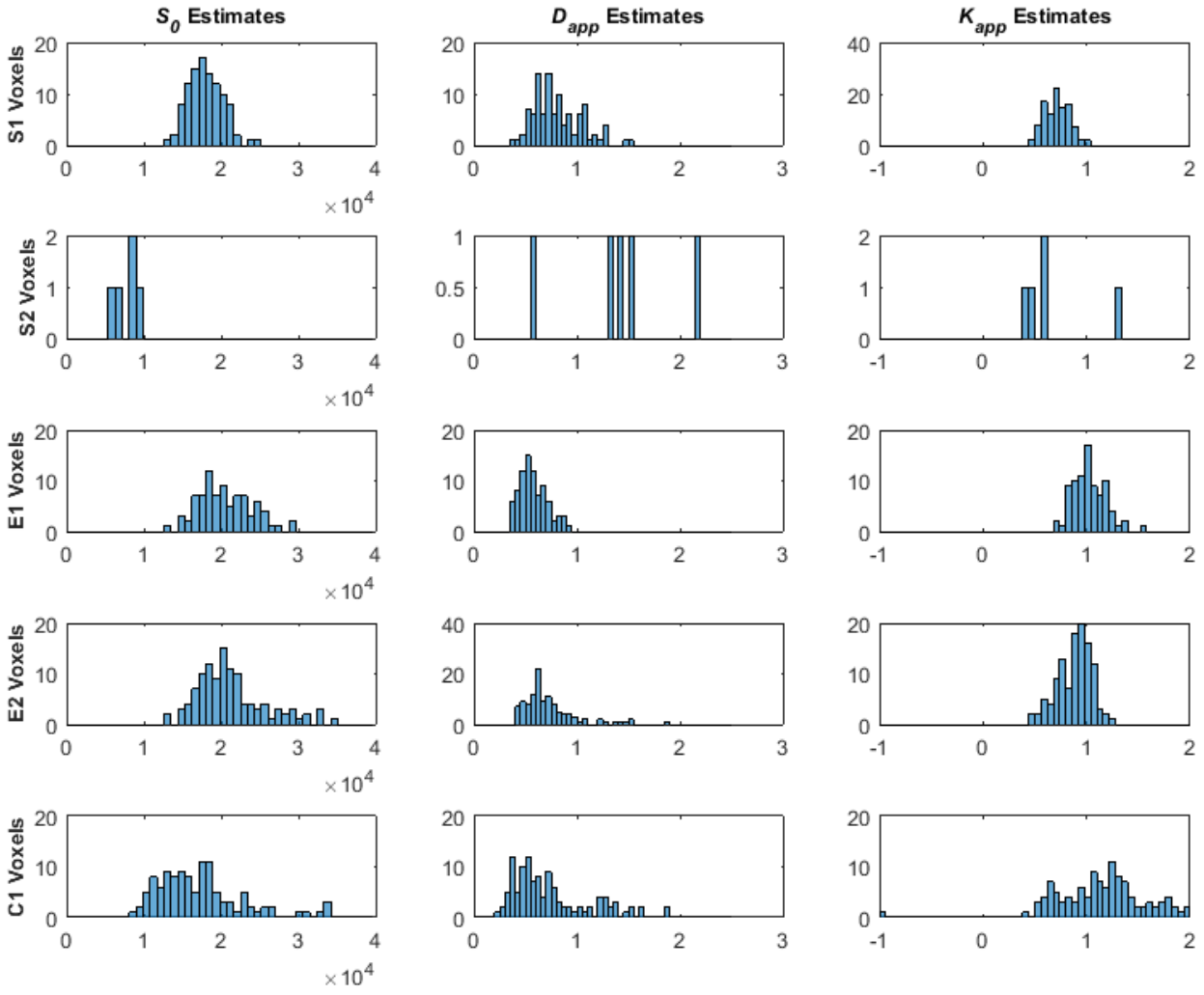

Figure 90 – Histograms of biexponential parameter estimates for each ROI after voxels with ΔAIC > -10 (see Figure 87) were removed. D_{app} is in units of μm²/ms

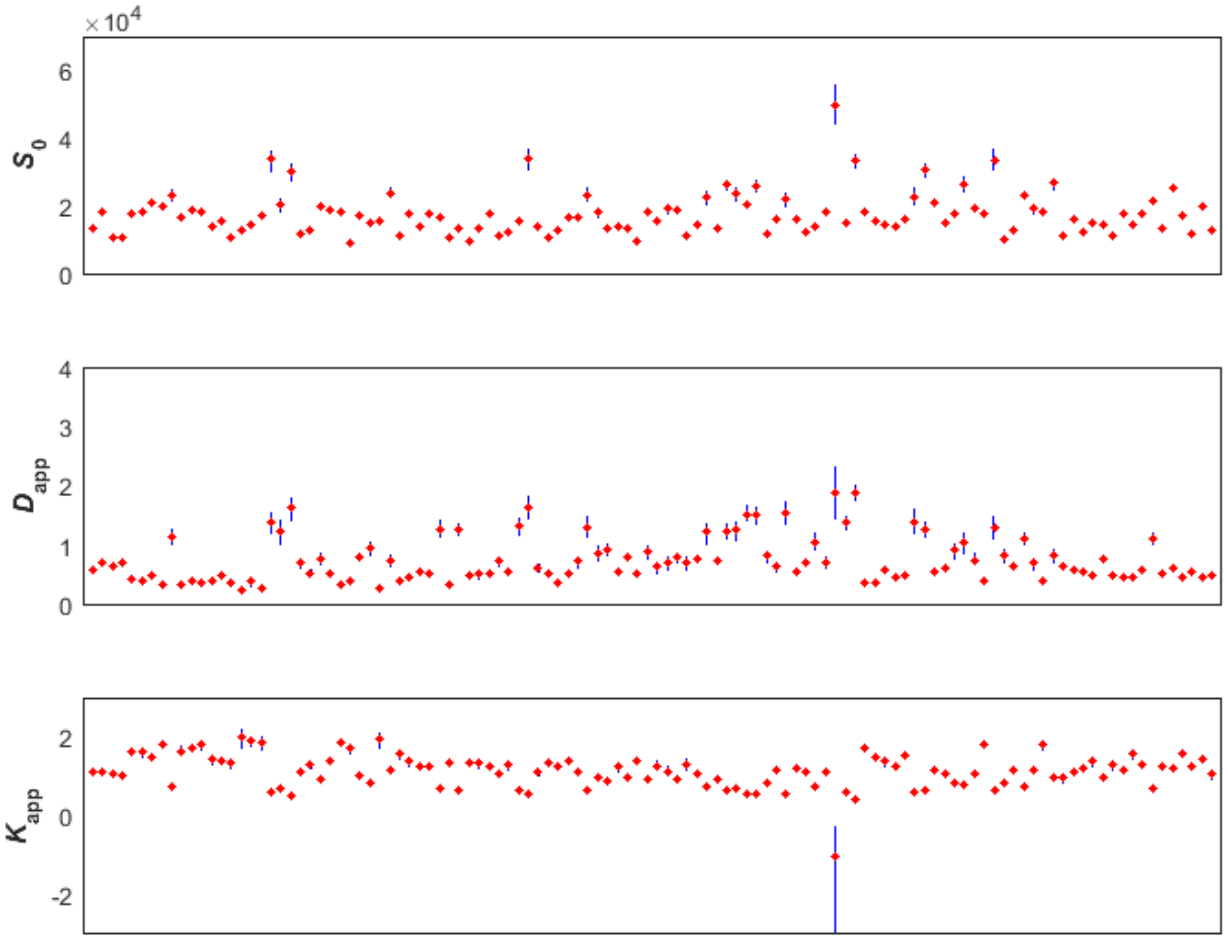

Figure 91 – Kurtosis parameter estimates (red points) plus bootstrap confidence intervals (blue lines) for voxels in C1 ROI after voxels with ΔAIC > -10 (see Figure 87) were removed

5.4 Conclusions

This chapter demonstrated that many of the phenomena discussed previously in this thesis on simulated data were also found in actual data from a DWI acquisition, specifically:

- The presence of large outlier values were found in the biexponential model parameter estimates, specifically, values for the fast decay component that were higher than the rate of free diffusion and very small estimates that were effectively zero in the slow decay component. In the kurtosis model, there were also highly negative estimates for the kurtosis parameter that did not represent realistic measurements. The phenomena seen in these model fits on real data were similar to the ill-conditioned, simulated estimates presented previously in this thesis.
- The bootstrap confidence intervals calculated from both the biexponential and kurtosis model fits showed a large number of fits with interval ranges that were orders of magnitude above and/or below the original estimate value. Many of these intervals had asymmetric,

non-normal distributions, and these large, irregular distributions were also seen in the simulated fits in Chapter 2 and Chapter 3.

- The AIC calculations for the voxels in this tissue study showed that either the biexponential or kurtosis model was selected as the model with the highest information content for most of the voxels. However, no clear model for all voxels in the individual ROI was superior.
- When applying a normality test to the bootstrap parameter estimates for both the kurtosis and biexponential model fits, there were many voxels that failed this test, which was shown to be indicative of ill-conditioned fits on simulated data in Chapter 4. When these voxels were removed from the test set and the parameter estimate distributions re-examined, a decrease in standard deviation was seen for nearly all parameters in all ROI. This was largely due to the elimination of outlier measurements in the estimate distributions, which could also be seen by the removal of nearly all voxel estimates with very large confidence intervals.
- When comparing the ΔAIC values of the biexponential and kurtosis models to the monoexponential model, the fits of all studied ROI had a wide range of ΔAIC values. When eliminating voxels fits where the ΔAIC value was greater than -10 and the more complex model was favoured, a decrease in standard deviation was also seen for nearly all parameters and ROI. As Chapter 4 showed, these voxels would be from signals that were less likely to be monoexponential, and removing them also led to a significant decrease in outliers. This showed that the ΔAIC value could also be used to improve the results of model fitting with either the biexponential or kurtosis models.
- This tissue study showed that ill-conditioned estimates were found in both stromal and epithelial regions. One stromal region (S2) had much higher *ADC* parameter values that the other regions, suggesting a higher water content in these voxels. This region also produced the largest number of D_1 estimates above the free rate of diffusion, the normality test eliminated 75% of the voxels, and the use of a minimum ΔAIC value eliminated nearly all voxels, demonstrating that the diffusion in this region was close to monoexponential, resulting in many unreliable estimates when used with the biexponential and kurtosis models. Estimates from the epithelial tissue from region C1, with glandular structure similar to that found in cancerous tissue, had the lowest ADC estimates, but the confidence intervals revealed a considerable number of voxels with highly uncertain estimates, also suggesting diffusion in these voxels that was close to monoexponential. These results demonstrated that using the biexponential or kurtosis models to assess either of these types of prostate tissue should be done so with caution, using the expanded analysis tools presented in this thesis to provide additional information.

Chapter 6
Implications of Results

6.1 The Precariousness of the Biexponential Model

A major objective of this thesis was to provide researchers with detailed information on the uncertainty and reliability of parameter estimates when using the biexponential model to fit DWI data. Chapter 2 demonstrated there were significant issues when using the biexponential model in NLLS regression fitting of simulated DWI data using parameter values similar to what has been reported in the literature. These issues were shown to be high bias and/or variance in the parameter estimates, the magnitude(s) of which greatly increased as the true signal was effectively monoexponential, and demonstrated that the biexponential model does not perform consistently and reliably across the possible parameter space. Not only did the variance of the parameter estimates increase for many tested signal measurements, but many of the estimates were outliers that were orders of magnitude higher or lower than their mean values, and had values that were physically unrealistic. While increasing the SNR improved the uncertainty in the estimates for much of the simulated data, even at a high SNR of 200, signals that were effectively monoexponential had parameter estimates that were highly unreliable. Thus, regardless of SNR, a monoexponential decay signal acts like an asymptote to the biexponential model, and fitting a biexponential model to a *true* monoexponential signal is an ill-posed problem resulting in unreliable parameter estimates.

The implication of these results is that using the biexponential model with NLLS regression fitting needs considerable reliability studies with actual tissue data before it can be deemed a replacement for the current monoexponential *ADC* model. This statement is made under the assumption that for a given DWI tissue study, there will be voxels that contain only freely diffusing water or structures that produce only non-restricted diffusion of a single, homogenous, effective diffusion coefficient. While there may be the existence of tissues that only produce non-Gaussian diffusion, the knowledge of whether non-Gaussian diffusion is present in a given voxel is unknown prior to a proposed study. Even if these voxels could be identified, the results presented in this thesis showed that the occurrence of highly uncertain estimates in the biexponential model were seen in signals that were *close* to monoexponential and that this closeness depended on SNR. The results also demonstrated that how a signal *fit* the data was independent of whether the parameter estimates were unreliable. Visual identification that a plotted signal measurement is non-monoexponential, and is better fitted by a biexponential model, is no guarantee of reliable parameter estimates. Additionally, obtaining valid, reliable biexponential parameter estimates from *one fit* of a voxel is no guarantee that this voxel may return similar estimates when fitting further sample measurements. This either requires repeated sampling from one voxel, or instead, obtaining bootstrap samples to better estimate what values are likely to be obtained over repeated measurements. The parametric bootstrap analysis introduced in this thesis was indeed shown to be effective at identifying parameter estimates with high uncertainty.

While writing the conclusion section of this thesis, a published paper by Merisaari et al was released that demonstrated both large variance *and* bias using simulated signals, and that these high uncertainties manifested themselves when either the D_1/D_2 ratio or SF_1 was at its lowest [188]. The authors also demonstrated how parameter uncertainty varied as the SNR was changed, and that a segmented fitting approach was not necessarily better. The paper also implemented bootstrap confidence intervals and reported extremely high median upper bound for the confidence intervals for D_1 estimates as well lower bounds of effectively zero for D_2 estimates. These results were all in agreement with the results presented in this thesis, demonstrating the reproducibility of the biexponential model uncertainty issues presented here with another simulated study. However, the authors there attribute the overestimated bias in the biexponential parameter values to the effects of the Rician signal bias. In this thesis, these issues were identified as problems arising from ill-conditioning issues in the fitting algorithm, since the effects of Rician signal bias were not enough to cause extreme uncertainty in the parameter estimates. Regardless of these differences, that study also reinforces the importance of repeatability studies of model parameter estimates, robustness of model to noise, and clinical usefulness of the results, and is a good complement to the findings presented in this thesis.

As the results in this thesis also demonstrated, the mathematical structure of the biexponential model meant that for some signal measurements, there were many biexponential parameter combinations that fit a signal closely. This property, as well as the general susceptibility of a model with more parameters to overfitting, meant that the biexponential model was more susceptible to noise than the monoexponential model. This was exemplified by the parameter plots in Figure 70 and Figure 71, where the biexponential parameters were considerably noisier than either the monoexponential or kurtosis parameters, resulting in a loss of visible tissue structure. This is a possible reason why most biexponential/IVIM analyses in the literature involve grouping together voxels into ROI, since combining voxels averages out the inter-voxel variation in the parameter estimates over the entire group. These noise issues of the biexponential model were highlighted in a recent review published on DWI analysis outside the brain [189]. In that paper, while the IVIM model there was indicated as a future direction for DWI use in the clinic, the increased variance of the perfusion decay coefficient estimates was specifically noted as a hindrance to its use. A relevant quote from the paper states, "While most published repeatability studies focus on estimates from small regions or whole organs, for the approach to be useful in general reliable voxel-wise estimation is important so that parameter maps may be computed for radiological assessment. The variability reported for regions or organs *underestimate the errors* with voxel-wise estimation." (Emphasis mine) Another quote states, "Data showing superiority of IVIM parameters over ADC for tissue characterization is limited, and more evidence is needed. IVIM parameter reproducibility and the role of IVIM parameters in treatment response need also to be better defined."

In the analysis in Chapter 5 using real tissue data, problematic voxels where the signal measurements were effectively monoexponential were removed, since these extreme values considerably affected the ROI statistical parameters. These ROI were made up of nearly a hundred or more voxels, so the loss of 30 or 40 of these voxels did not produce a drastic effect. However, in an in vivo analysis of prostate, for example, every voxel is crucial since a possible tumour may only be 2 or 3 voxels wide. In this case, the biexponential/IVIM model may be best suited as

supplementary information to the *ADC*, much like DWI is recommended as part of a multiparametric analysis in the PI-RADS standard. The biexponential parameter estimates can be used in addition to the *ADC* value, and if the biexponential estimates are unreliable, that at least provides additional information, namely, this voxel is likely exhibiting monoexponential decay. Leaving these unreliable voxels present in a statistical analysis could significantly skew the results of any categorical comparisons, so extra caution should be used with any statistical testing.

6.2 The Kurtosis Model – Also Not Ready to Replace the *ADC*

This thesis also analysed the uncertainty in the kurtosis model parameter estimates for the same simulated data as the biexponential model, and while the instability in the model was not as severe as the biexponential model, when assessing signals that were effectively monoexponential, the variance in the kurtosis parameter estimates increased considerably, with many outlier values found at extremely large negative values. While the parameter maps for the signal amplitude and decay rate in in Figure 70 and Figure 71 showed the ability to resolve similar tissue structure to the monoexponential model estimates, there was large variance and the presence of negative outlier values in the kurtosis parameter estimates. Compared to the biexponential model, the kurtosis model parameters may then give an *average* lower variance, but it can also produce highly uncertain parameter estimates like the biexponential model, although they appear to be confined to one parameter. Due to this same possibility of unreliable estimation, the kurtosis model would also appear to be not ready for voxel-wise clinical assessment of non-Gaussian diffusion using NLLS regression, and also needs more repeatability studies to be a replacement for the monoexponential model when assessing DWI data. Although the kurtosis does not have a biophysical basis, it would be interesting to see what sort of bias results in the parameters if the kurtosis model was used to generate the true signal data.

6.3 Model Selection and the Effects of Misspecification

Another major objective of this thesis was to determine the reliability of common model selection methods when selecting from the monoexponential, biexponential, and kurtosis models on a set of simulated DWI data. When comparing the AIC, AIC_c, and the LOOCV methods on simulated data in Chapter 4, while varying the simulated SNR and acquisition parameters, the rates that these methods selected the tested models also varied significantly. Additionally, if the biexponential model was the true model that generated the signals, after adding simulated measurement noise, the biexponential model was not selected as best for all signals. This also happened when using the monoexponential model as the true model, and also demonstrated that no one particular method was superior in reliably selecting the true model, but instead certain methods were biased in favour of simpler models and others in favour of complex models. For example, when testing data simulated from true biexponential signals, the AIC selected the biexponential model as best more often than the monoexponential model when compared to the AIC_c. However, when testing a data set generated from true monoexponential signals, the AIC_c selected the monoexponential model more often than the biexponential model when compared to the AIC. The AIC and LOOCV methods were shown to select similarly across most of the data, which confirmed earlier studies, and

indicated that when assessing the DWI measurements and models found in this thesis, spending the extra computational time to calculate LOOCV information is probably unnecessary.

Another important conclusion from these results is that across repeated measurements, the AIC, and its derived measure AIC_c, can select three different models as best for the same signal. This was shown to happen when the true generating signal was effectively monoexponential, however, this does demonstrate that the AIC has limitations with this set of models. Due to the overlapping nature of the biexponential and kurtosis models with the monoexponential model, for some signals, the AIC is effectively a random "spin of the wheel". This was shown specifically when comparing the biexponential and monoexponential models in Figure 57, where depending on how the noise affected the sample measurement, there was a 50/50 chance of selecting one of the models. This illustrates the inherent problem in using the AIC selection results from *one* sample measurement of *one* voxel – a researcher doesn't know whether this AIC best model selection came from a signal where that model would get chosen for all repeated samples, or whether the AIC is randomly selecting from the tested models. Hence, the magnitude value of ΔAIC between two models was demonstrated to be important, but this value of ΔAIC also varied over repeated individual samples. For example, a ΔAIC of 1 between two models could have come from a signal where the model would be selected for 100% of all samples or one where the selection rate would be 50%. However, the higher the ΔAIC value of a given measurement, the higher the selection rate of this model *on average* versus the other tested model across repeated measurements.

Perhaps it would help us to be reminded that the AIC is an *asymptotic measure*, and that its results apply to an assessment of many repeated sample measurements. The theoretical basis for the AIC and its roots in information theory give it a solid foundation as an objective judicator, and while its basis may be sound, as statistical consultant John D. Cook points out on his blog, "…the choice of models to compare and the choice of information criterion are not as objective, so there can be an inflated impression of objectivity [190]." Another conclusion from this thesis, then, was that various information criteria (AIC, AIC_c, BIC, etc.) all selected a specific model at different rates over repeated samples of the same signal versus the other tested models. Newer model selection criteria such as the Mixture Regression Criterion [191, 192] might possibly improve these selection rates. While an aim of Chapter 4 was to establish whether a given selection method was more reliable than the others, the more important result was establishing that these methods merely select differently. Researchers, then, must be aware that their choice of selection method is *subjective*, that another method or criterion can return different results for the same data set with recent evidence of this reported in the DWI literature [37]. Chapter 4 also demonstrated that measurement settings also affect the model selection results, for example, limiting the number of diffusion weightings to avoid Rician signal bias had the side effect of choosing simpler models more often. Not only was this due to limiting the signals to lower weightings where the divergence from non-Gaussian, monoexponential decay was harder to detect, but with less signal measurements, the difference in RSS values was lower compared to the value of the parameter penalty. Thus, for a given acquisition of DWI data, the selection of biexponential model on a set of data could be increased if the number of *b*-values was increased to 20, 30, or more. In this case, what *is* the best model? What *is* the optimal number of *b*-values? Perhaps the conclusion here is that there appears to be no such thing as an objective selection of best.

Objectivity is usually not what researchers are interested in anyway. Rather, the goal is to find and use the model with the best performance across future measurements. Best performance, then, requires model parameter estimates that are consistent and will converge toward their expected values. In the history of DWI analysis, the monoexponential model was the foundational model because it has a theoretical basis for it, and for some tissue measurements, the model may get "lucky" and assess a voxel where the diffusion is Gaussian, and so it is measuring the actual diffusivity. For non-Gaussian diffusion measurements, this model is no longer valid. If there are two distinct diffusing compartments for a given voxel, then obviously the biexponential model is the valid model. However, when this biexponential model is applied to monoexponential, Gaussian diffusion, there are now two redundant parameters, and so the model is invalid, and the parameter estimates in this case do not represent anything real. For most tissue measurements, however, the true basis model is infinite and so all models aren't really "valid" but are approximations, and so the goal instead is to find the model with the best average prediction over all future data. Assuming that this data will fluctuate across all possibilities, when assigning a model to assess data, two things will happen. A simpler model will be applied to complex data, or a complex model will be applied to simpler data. As Chapter 4 demonstrated, when using the biexponential or kurtosis model, if the AIC selects the biexponential and monoexponential models as equal for a given measurement, the variance in the biexponential model diffusion coefficient estimates was an order of magnitude higher than the monoexponential estimates.

This demonstration established that a model selection method declaring a model as best did not automatically translate into better accuracy and/or precision of the parameters. Therefore, extra caution must be taken when using model selection methods to make inference about complex models, since their tendency to overfit leads to a lack of generalization. While the simple monoexponential *ADC* value does not accurately describe underlying restrictions to tissue, giving considerable bias to its diffusion coefficient estimates, its simple structure makes it more robust to noise and algorithmic issues than the biexponential or kurtosis models. One reason that the *ADC* has made it to clinical use is its ability to be used effectively on a wide variety of measurements with various acquisition protocols. In their current state, and with the current state of NLLS regression, the biexponential and kurtosis models don't appear to have this capability. Improved algorithmic methods such as NLLS fitting with regularization [98] or mixed effect model [193] have shown some improvements on the uncertainty in the biexponential model and are examples of direction for future analysis. As this thesis showed, the current algorithms can be used with model selection criteria, specifically ΔAIC, to reduce the occurrence of these extreme parameter estimates. To be accepted into clinical use, further research must be performed on these models to attempt to reduce these ill-conditioning issues, and this will require study specifically on how to tame these errors in the worst-case scenarios.

6.4 Replication, Replication, Replication…

In 2016, one of the most widely-discussed, pressing topics in the current statistical literature has been that of a "statistical crisis in science", particularly with respect to replication of existing studies. The focal point has been mainly in the field of social psychology with one example being a large study released last year that replicated the results of 100 original experiments that had been

reported in top-ranking psychology journals [194]. The results of this replication showed that compared to 97% of the original results reported as statistically significant, only 36% of the replicated study results were statistically significant. The main object of criticism over the past few years of this replication crisis is the *p*-value, with one journal going to so far as to outright ban Null Hypothesis Significance Testing (NHST) procedures and their resulting criteria (*p*-values, *t*-values, *F*-values, etc.) [195]. Recently, the American Statistical Association released a formal statement in March 2016 that clarified the principles underlying the proper use and implementation of the *p*-value [196]. Regardless of whether researchers heed these guidelines or not, a journal paper probably has a better chance of getting accepted with a statistical test proclaiming "significant" or a model criterion proclaiming "best".

This sentiment is echoed in one of the commentaries added as supplemental information to the ASA statement [196] written by Andrew Gelman, in which he states, "…we tend to take the 'dataset' and even the statistical model as given, reducing statistics to a mathematical or computational problem of inference and encouraging students and practitioners to think of their data as given. Even when we discuss the design of surveys and experiments, we typically focus on the choice of sample size, not on the importance of valid and reliable measurements. The result is often an attitude that any measurement will do, and a blind quest for statistical significance." He later says, "…it seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an "uncertainty laundering" that begins with data and concludes with success as measured by statistical significance…This is what is expected—demanded—of subject-matter journals. Just try publishing a result with $p = 0.20$." A related article by Gelman and his co-author Loken [197] state that *p*-values are based on what would have happened under other data sets, and term these possible choices in the process of setting up and analysing a scientific study "The Garden of Forking Paths".

The topic of replication was addressed in much of this thesis by showing how results changed when assessing repeated simulated samples. These studies were specifically presented this way to inform researchers of problems that need to be addressed in order to establish the reliability of these new DWI models and methods before implementing them in clinical research. This thesis demonstrated that for a given measurement, the SNR, number and locations of the *b*-values, removal of outlier data, choice of ROI, choice of model, and choice of model selection criteria all affected results. Thus, extra care should be taken when drawing conclusions from statistical analysis and attempting to characterize the diffusion processes occurring in complex tissue.This is especially true when *p*-values are involved, for example, a multimodel analysis where several models, each with multiple parameters, are applied to a set of data has a high likelihood of achieving a significant difference in one parameter value with $p < 0.05$ *by chance alone*. Consistency in statistical significance in DWI analysis was called for by Jones and Cercignani in pitfall #23 of their 2010 paper, "Twenty-five Pitfalls in the Analysis of Diffusion MRI Data" [198], in which they warn against this multiple comparisons problem and the high probability of type I, false positive errors when researchers report *p*-values and do not correct for this issue.

This thesis is not a criticism of individual studies or researchers, nor does intend to downplay the importance of researchers conducting exploratory model analyses for DWI data. Rather, it is a plea

for more detailed model checking and validation which seems to be at odds with the current research environment. While the DWI community needs more reliable exploratory studies to lay solid foundations for future research, the current motto of "publish or perish", however, demands as many novel papers as possible. The second study of a given tissue/model combination probably doesn't get published, so replication is often not rewarded. Thus, the decree of significance is more important than detailed error analysis and control, and more and more scientific studies are published with the focus on quantity at the cost of quality [199]. This loss of quality has not gone unnoticed, for example, a recent review of the literature determined that 50% of previous pre-clinical, life sciences research was irreproducible, with approximately $28 billion spent on this irreproducible research in the United States alone [200]. Figures like this are part of the reason that in 2016, the National Institute of Health added new guidelines for scientific rigour and transparency in data and statistical analysis for all future grant applications [201].

As demonstrated in this thesis, simply fitting a biexponential or kurtosis model to one voxel measurement via NLLS regression and examining the parameter estimates will almost never lead a researcher to think there is anything wrong with a given measurement and model fit, as this cannot be assessed by a single point estimate. While the demonstrated parametric bootstrap can assess whether a given fit is likely to produce unreliable estimates in future measurements, this needs to be demonstrated with empirical studies on actual data. Likewise, empirical studies comparing model selection methods on repeated measurements of a given volume of interest needs to be assessed as well. A study examining the differences between how two different observers reported results has been performed on the biexponential model [202], but these observers tested the same data set. The previously introduced study by Merisaari et al [188] compared results from two *separate* examinations for eighty-one patients. Additionally, a recent study investigated differences in *ADC* measurements across multiple MRI scanners using a phantom [203]. More of these types of repeatability and reproducibility studies are needed in the DWI literature, especially with the biexponential and kurtosis models, with several examples now being seen [114, 121, 204-207]. The information provided in this thesis on these models, along with the techniques provided to improve their estimates, was intended to assist researchers in understanding the uncertainty involved not only in their data but also in their models. Such understanding will improve the reliability of DWI studies and help avoid any future DWI replication crisis as well.

References

1. Stevenson A. Oxford dictionary of English: Oxford University Press, USA; 2010.
2. Upton G, Cook I. A dictionary of statistics 3e: Oxford university press; 2014.
3. Lauterbur PC. Image formation by induced local interactions: examples employing nuclear magnetic resonance. Nature. 1973;242(5394):190-1.
4. McRobbie DW, Moore EA, Graves MJ, Prince MR. MRI from Picture to Proton: Cambridge university press; 2006.
5. Jones DK. Diffusion MRI: Theory, methods, and applications: Oxford University Press; 2010.
6. Callaghan PT. Translational dynamics and magnetic resonance: principles of pulsed gradient spin echo NMR: Oxford University Press; 2011.
7. Price WS. NMR studies of translational motion: principles and applications: Cambridge University Press; 2009.
8. Holz M, Heil SR, Sacco A. Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate 1H NMR PFG measurements. Physical Chemistry Chemical Physics. 2000;2(20):4740-2.
9. Havlin S, Ben-Avraham D. Diffusion in disordered media. Advances in physics. 2002;51(1):187-292.
10. Ben-Avraham D, Havlin S. Diffusion and reactions in fractals and disordered systems: Cambridge University Press Cambridge; 2000.
11. Ozarslan E, Basser PJ, Shepherd TM, Thelwall PE, Vemuri BC, Blackband SJ. Observation of anomalous diffusion in excised tissue by characterizing the diffusion-time dependence of the MR signal. J Magn Reson. 2006;183(2):315-23. doi: 10.1016/j.jmr.2006.08.009. PubMed PMID: 16962801.
12. Callaghan PT, Coy A, MacGowan D, Packer KJ, Zelaya FO. Diffraction-like effects in NMR diffusion studies of fluids in porous solids. 1991.
13. Beaulieu C. The basis of anisotropic water diffusion in the nervous system–a technical review. NMR in biomedicine. 2002;15(7-8):435-55.
14. Grebenkov DS. NMR survey of reflected Brownian motion. Reviews of Modern Physics. 2007;79(3):1077.
15. Stejskal EO, Tanner JE. Spin Diffusion Measurements: Spin Echoes in the Presence of a Time-Dependent Field Gradient. The Journal of Chemical Physics. 1965;42(1):288. doi: 10.1063/1.1695690.
16. Le Bihan D, Breton E. Imagerie de diffusion in-vivo par resonance magnetique nucleaire. Comptes-Rendus de l'Académie des Sciences. 1985;93(5):27-34.
17. Woessner DE. NMR spin-echo self-diffusion measurements on fluids undergoing restricted diffusion. The Journal of Physical Chemistry. 1963;67(6):1365-7.
18. Le Bihan D, Turner R, Douek P, Patronas N. Diffusion MR imaging: clinical applications. AJR American journal of roentgenology. 1992;159(3):591-9.
19. Le Bihan D. Looking into the functional architecture of the brain with diffusion MRI. Nature Reviews Neuroscience. 2003;4(6):469-80.
20. Le Bihan D, Johansen-Berg H. Diffusion MRI at 25: exploring brain tissue structure and function. NeuroImage. 2012;61(2):324-41.
21. Shenton M, Hamoda H, Schneiderman J, Bouix S, Pasternak O, Rathi Y, et al. A review of magnetic resonance imaging and diffusion tensor imaging findings in mild traumatic brain injury. Brain imaging and behavior. 2012;6(2):137-92.
22. Jauch EC, Saver JL, Adams HP, Bruno A, Demaerschalk BM, Khatri P, et al. Guidelines for the early management of patients with acute ischemic stroke a guideline for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2013;44(3):870-947.

23. Schellinger P, Bryan R, Caplan L, Detre J, Edelman R, Jaigobin C, et al. Evidence-based guideline: The role of diffusion and perfusion MRI for the diagnosis of acute ischemic stroke Report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology. Neurology. 2010;75(2):177-85.
24. Koh DM, Collins DJ. Diffusion-weighted MRI in the body: applications and challenges in oncology. AJR American journal of roentgenology. 2007;188(6):1622-35. Epub 2007/05/23. doi: 10.2214/AJR.06.1403. PubMed PMID: 17515386.
25. Khoo MM, Tyler PA, Saifuddin A, Padhani AR. Diffusion-weighted imaging (DWI) in musculoskeletal MRI: a critical review. Skeletal radiology. 2011;40(6):665-81.
26. Koh DM, Blackledge M, Padhani AR, Takahara T, Kwee TC, Leach MO, et al. Whole-body diffusion-weighted MRI: tips, tricks, and pitfalls. AJR American journal of roentgenology. 2012;199(2):252-62. doi: 10.2214/AJR.11.7866. PubMed PMID: 22826385.
27. Padhani AR, Makris A, Gall P, Collins DJ, Tunariu N, Bono JS. Therapy monitoring of skeletal metastases with whole-body diffusion MRI. Journal of Magnetic Resonance Imaging. 2014;39(5):1049-78.
28. Bollineni V, Kramer G, Liu Y, Melidis C. A literature review of the association between diffusion-weighted MRI derived apparent diffusion coefficient and tumour aggressiveness in pelvic cancer. Cancer treatment reviews. 2015.
29. Türkbey B, Aras Ö, Karabulut N, Turgut AT, Akpinar E, Alibek S, et al. Diffusion-weighted MRI for detecting and monitoring cancer: a review of current applications in body imaging. Diagn Interv Radiol. 2012;18(1):46-59.
30. Padhani AR, Koh D-M, Collins DJ. Whole-body diffusion-weighted MR imaging in cancer: current status and research directions. Radiology. 2011;261(3):700-18.
31. American College of Radiology. Available from: http://www.acr.org/Quality-Safety/Resources/PIRADS.
32. Barrett T, Turkbey B, Choyke PL. PI-RADS version 2: what you need to know. Clinical radiology. 2015;70(11):1165-76.
33. Grebenkov DS. Use, misuse, and abuse of apparent diffusion coefficients. Concepts in Magnetic Resonance Part A. 2010;36(1):24-35.
34. Le Bihan D, Breton E, Lallemand D, Aubin ML, Vignaud J, Laval-Jeantet M. Separation of diffusion and perfusion in intravoxel incoherent motion MR imaging. Radiology. 1988;168(2):497-505. doi: doi:10.1148/radiology.168.2.3393671. PubMed PMID: 3393671.
35. Davison AC. Statistical models: Cambridge University Press; 2003.
36. Jensen JH, Helpern JA, Ramani A, Lu H, Kaczynski K. Diffusional kurtosis imaging: the quantification of non-gaussian water diffusion by means of magnetic resonance imaging. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2005;53(6):1432-40. doi: 10.1002/mrm.20508. PubMed PMID: 15906300.
37. Bourne RM, Panagiotaki E, Bongers A, Sved P, Watson G, Alexander DC. Information theoretic ranking of four models of diffusion attenuation in fresh and fixed prostate tissue ex vivo. Magnet Reson Med. 2014;72(5):1418-26.
38. Jensen JH, Helpern JA. MRI quantification of non-Gaussian water diffusion by kurtosis analysis. NMR in biomedicine. 2010;23(7):698-710. doi: 10.1002/nbm.1518. PubMed PMID: 20632416; PubMed Central PMCID: PMC2997680.
39. Van Cauter S, Veraart J, Sijbers J, Peeters RR, Himmelreich U, De Keyzer F, et al. Gliomas: diffusion kurtosis MR imaging in grading. Radiology. 2012;263(2):492-501.

40. Huang Y, Chen X, Zhang Z, Yan L, Pan D, Liang C, et al. MRI quantification of non-Gaussian water diffusion in normal human kidney: a diffusional kurtosis imaging study. NMR in biomedicine. 2015;28(2):154-61.
41. Rosenkrantz AB, Sigmund EE, Johnson G, Babb JS, Mussi TC, Melamed J, et al. Prostate cancer: feasibility and preliminary experience of a diffusional kurtosis model for detection and assessment of aggressiveness of peripheral zone cancer. Radiology. 2012;264(1):126-35. doi: 10.1148/radiol.12112290. PubMed PMID: 22550312.
42. Tamura C, Shinmoto H, Soga S, Okamura T, Sato H, Okuaki T, et al. Diffusion kurtosis imaging study of prostate cancer: preliminary findings. Journal of Magnetic Resonance Imaging. 2014;40(3):723-9.
43. Bennett KM, Schmainda KM, Rowe DB, Lu H, Hyde JS. Characterization of continuously distributed cortical water diffusion rates with a stretched-exponential model. Magnet Reson Med. 2003;50(4):727-34.
44. Hall MG, Barrick TR. From diffusion-weighted MRI to anomalous diffusion imaging. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2008;59(3):447-55. doi: 10.1002/mrm.21453. PubMed PMID: 18224695.
45. Anderson SW, Barry B, Soto J, Ozonoff A, O'Brien M, Jara H. Characterizing non-gaussian, high b-value diffusion in liver fibrosis: Stretched exponential and diffusional kurtosis modeling. Journal of Magnetic Resonance Imaging. 2014;39(4):827-34.
46. Basser PJ, Mattiello J, LeBihan D. MR diffusion tensor spectroscopy and imaging. Biophysical journal. 1994;66(1):259.
47. Hall MG, Bongers A, Sved P, Watson G, Bourne RM. Assessment of non-Gaussian diffusion with singly and doubly stretched biexponential models of diffusion-weighted MRI (DWI) signal attenuation in prostate tissue. NMR in biomedicine. 2015;28(4):486-95.
48. Panagiotaki E, Walker-Samuel S, Siow B, Johnson SP, Rajkumar V, Pedley RB, et al. Noninvasive quantification of solid tumor microstructure using VERDICT MRI. Cancer research. 2014;74(7):1902-12.
49. Panagiotaki E, Schneider T, Siow B, Hall MG, Lythgoe MF, Alexander DC. Compartment models of the diffusion MR signal in brain white matter: a taxonomy and comparison. NeuroImage. 2012;59(3):2241-54. doi: 10.1016/j.neuroimage.2011.09.081. PubMed PMID: 22001791.
50. Wasserman L. All of statistics: Springer Science & Business Media; 2011.
51. Fahrmeir L, Kneib T, Lang S. Regression: Springer; 2007.
52. Gelman A, Hill J. Data analysis using regression and multilevel/hierarchical models: Cambridge University Press; 2006.
53. Björck A. Numerical methods for least squares problems: Siam; 1996.
54. Gudbjartsson H, Patz S. The Rician distribution of noisy MRI data. Magnet Reson Med. 1995;34(6):910-4.
55. Sijbers J, Den Dekker A. Maximum likelihood estimation of signal amplitude and noise variance from MR data. Magnet Reson Med. 2004;51(3):586-94.
56. Walker-Samuel S, Orton M, McPhail LD, Robinson SP. Robust estimation of the apparent diffusion coefficient (ADC) in heterogeneous solid tumors. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2009;62(2):420-9. Epub 2009/04/09. doi: 10.1002/mrm.22014. PubMed PMID: 19353661.
57. Henkelman RM. Measurement of signal intensities in the presence of noise in MR images. Medical physics. 1985;12(2):232. doi: 10.1118/1.595711.
58. Cárdenas-Blanco A, Tejos C, Irarrazaval P, Cameron I. Noise in magnitude magnetic resonance images. Concepts Magn Reson Part A Bridg Educ Res. 2008;32(6):409-16.

59. Sijbers J, den Dekker AJ, Van Audekerke J, Verhoye M, Van Dyck D. Estimation of the Noise in Magnitude MR Images. Journal of magnetic resonance imaging : JMRI. 1998;16(1):87-90. doi: http://dx.doi.org/10.1016/S0730-725X(97)00199-9.
60. den Dekker AJ, Sijbers J. Data distributions in magnetic resonance images: a review. Physica medica : PM : an international journal devoted to the applications of physics to medicine and biology : official journal of the Italian Association of Biomedical Physics. 2014;30(7):725-41. doi: 10.1016/j.ejmp.2014.05.002. PubMed PMID: 25059432.
61. Narsky I, Porter FC. Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning: John Wiley & Sons; 2013.
62. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretic approach: Springer Science & Business Media; 2002.
63. Claeskens G, Hjort NL. Model selection and model averaging: Cambridge University Press Cambridge; 2008.
64. Alexander D, Barker G, Arridge S. Detection and modeling of non-Gaussian apparent diffusion coefficient profiles in human brain data. Magnet Reson Med. 2002;48(2):331-40.
65. Wittsack HJ, Lanzman RS, Mathys C, Janssen H, Modder U, Blondin D. Statistical evaluation of diffusion-weighted imaging of the human kidney. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2010;64(2):616-22. Epub 2010/07/29. doi: 10.1002/mrm.22436. PubMed PMID: 20665805.
66. Veraart J, Poot DH, Van Hecke W, Blockx I, Van der Linden A, Verhoye M, et al. More accurate estimation of diffusion tensor parameters using diffusion Kurtosis imaging. Magnet Reson Med. 2011;65(1):138-45.
67. Kristoffersen A. Statistical assessment of non-Gaussian diffusion models. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2011;66(6):1639-48. Epub 2011/04/28. doi: 10.1002/mrm.22960. PubMed PMID: 21523826.
68. Akaike H, editor Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory; 1973; Tsahkadsor, Armenia, USSR.
69. Kullback S, Leibler RA. On information and sufficiency. The annals of mathematical statistics. 1951:79-86.
70. Burnham KP, Anderson DR, Huyvaert KP. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behavioral Ecology and Sociobiology. 2010;65(1):23-35. doi: 10.1007/s00265-010-1029-6.
71. Panagiotaki E, Chan RW, Dikaios N, Ahmed HU, O'Callaghan J, Freeman A, et al. Microstructural characterization of normal and malignant human prostate tissue with vascular, extracellular, and restricted diffusion for cytometry in tumours magnetic resonance imaging. Investigative radiology. 2015;50(4):218-27.
72. Freiman M, Perez-Rossello JM, Callahan MJ, Bittman M, Mulkern RV, Bousvaros A, et al. Characterization of fast and slow diffusion from diffusion-weighted MRI of pediatric Crohn's disease. Journal of Magnetic Resonance Imaging. 2013;37(1):156-63.
73. Ferizi U, Schneider T, Panagiotaki E, Nedjati-Gilani G, Zhang H, Wheeler-Kingshott CA, et al. A ranking of diffusion MRI compartment models with in vivo human brain data. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2014;72(6):1785-92. doi: 10.1002/mrm.25080. PubMed PMID: 24347370; PubMed Central PMCID: PMC4278549.
74. Burnham KP, Anderson DR. Multimodel inference understanding AIC and BIC in model selection. Sociological methods & research. 2004;33(2):261-304.

75. Kuha J. AIC and BIC comparisons of assumptions and performance. Sociological Methods & Research. 2004;33(2):188-229.
76. Hastie T, Tibshirani R, Friedman J, Hastie T, Friedman J, Tibshirani R. The elements of statistical learning: Springer; 2009.
77. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics surveys. 2010;4:40-79.
78. Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society Series B (Methodological). 1977:44-7.
79. Charlton R, Landau S, Schiavone F, Barrick T, Clark C, Markus H, et al. A structural equation modeling investigation of age-related variance in executive function and DTI measured white matter damage. Neurobiology of aging. 2008;29(10):1547-55.
80. Feis D-L, Brodersen KH, von Cramon DY, Luders E, Tittgemeyer M. Decoding gender dimorphism of the human brain using multimodal anatomical and diffusion MRI data. NeuroImage. 2013;70:250-7.
81. Santis S, Assaf Y, Evans C, Jones D. Improved precision in CHARMED assessment of white matter through sampling scheme optimization and model parsimony testing. Magnet Reson Med. 2014;71(2):661-71.
82. Cox DR. Principles of statistical inference: Cambridge University Press; 2006.
83. Bourne RM. The trouble with apparent diffusion coefficient papers. Journal of medical radiation sciences. 2015;62(2):89-91.
84. Chatfield C. Model Uncertainty, Data Mining and Statistical Inference. Journal of the Royal Statistical Society Series A (Statistics in Society). 1995;158(3):419-66. doi: 10.2307/2983440.
85. Hjorth JU. Computer intensive statistical methods: Validation, model selection, and bootstrap: CRC Press; 1993.
86. Le Bihan D. Intravoxel Incoherent Motion Perfusion MR Imaging: A Wake-Up Call 1. Radiology. 2008;249(3):748-52.
87. Luciani A, Vignaud A, Cavet M, Tran Van Nhieu J, Mallat A, Ruel L, et al. Liver cirrhosis: intravoxel incoherent motion MR imaging—Pilot study 1. Radiology. 2008;249(3):891-9.
88. Sigmund EE, Cho GY, Kim S, Finn M, Moccaldi M, Jensen JH, et al. Intravoxel incoherent motion imaging of tumor microenvironment in locally advanced breast cancer. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2011;65(5):1437-47. Epub 2011/02/03. doi: 10.1002/mrm.22740. PubMed PMID: 21287591.
89. Thelwall PE, Grant SC, Stanisz GJ, Blackband SJ. Human erythrocyte ghosts: exploring the origins of multiexponential water diffusion in a model biological tissue with magnetic resonance. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2002;48(4):649-57. Epub 2002/09/28. doi: 10.1002/mrm.10270. PubMed PMID: 12353282.
90. Thelwall PE, Shepherd TM, Stanisz GJ, Blackband SJ. Effects of temperature and aldehyde fixation on tissue water diffusion properties, studied in an erythrocyte ghost tissue model. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2006;56(2):282-9. Epub 2006/07/15. doi: 10.1002/mrm.20962. PubMed PMID: 16841346.
91. Grant SC, Buckley DL, Gibbs S, Webb AG, Blackband SJ. MR microscopy of multicomponent diffusion in single neurons. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2001;46(6):1107-12. Epub 2001/12/18. PubMed PMID: 11746576.
92. Price WS, Barzykin AV, Hayamizu K, Tachiya M. A model for diffusive transport through a spherical interface probed by pulsed-field gradient NMR. Biophysical journal. 1998;74(5):2259-71. Epub

1998/05/20. doi: 10.1016/S0006-3495(98)77935-4. PubMed PMID: 9591653; PubMed Central PMCID: PMC1299569.
93. Storas TH, Gjesdal KI, Gadmar OB, Geitung JT, Klow NE. Prostate magnetic resonance imaging: multiexponential T2 decay in prostate tissue. Journal of magnetic resonance imaging : JMRI. 2008;28(5):1166-72. Epub 2008/10/31. doi: 10.1002/jmri.21534. PubMed PMID: 18972358.
94. Karger J. Nmr Self-Diffusion Studies in Heterogeneous Systems. Adv Colloid Interfac. 1985;23(1-4):129-48. doi: Doi 10.1016/0001-8686(85)80018-X. PubMed PMID: WOS:A1985ARF1700006.
95. Novikov DS, Kiselev VG. Effective medium theory of a diffusion-weighted signal. NMR in biomedicine. 2010;23(7):682-97.
96. Armspach J-P, Gounot D, Rumbach L, Chambron J. In vivo determination of multiexponential T 2 relaxation in the brain of patients with multiple sclerosis. Magnetic resonance imaging. 1991;9(1):107-13.
97. Dumitresco BE, Armspach J-P, Gounot D, Grucker D, Mauss Y, Steibel J, et al. Multi-exponential analysis of T2 images. Magnetic resonance imaging. 1986;4(5):445-8.
98. Laule C, Vavasour IM, Mädler B, Kolind SH, Sirrs SM, Brief EE, et al. MR evidence of long T2 water in pathological white matter. Journal of Magnetic Resonance Imaging. 2007;26(4):1117-21.
99. Rumbach L, Armspach J-P, Gounot D, Namer IJ, Chambron J, Warter J-M, et al. Nuclear magnetic resonance T2 relaxation times in multiple sclerosis. Journal of the neurological sciences. 1991;104(2):176-81.
100. Acton FS. Numerical methods that work: Maa; 1970.
101. Lanczos C. Applied analysis. Prentice-Hall, Englewood Cliffs, NJ; 1956.
102. Golub GH, Pereyra V. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. SIAM Journal on numerical analysis. 1973;10(2):413-32.
103. Nash JC. Nonlinear Parameter Optimization Using R Tools: Wiley; 2014.
104. O'leary DP, Rust BW. Variable projection for nonlinear least squares problems. Computational Optimization and Applications. 2013;54(3):579-93.
105. Mulkern RV, Haker SJ, Maier SE. On high b diffusion imaging in the human brain: ruminations and experimental insights. Magnetic resonance imaging. 2009;27(8):1151-62.
106. Cohen AD, Schieke MC, Hohenwalter MD, Schmainda KM. The effect of low b-values on the intravoxel incoherent motion derived pseudodiffusion parameter in liver. Magnet Reson Med. 2015;73(1):306-11.
107. Mahmood F, Johannesen HH, Geertsen P, Opheim GF, Hansen RH. The effect of region of interest strategies on apparent diffusion coefficient assessment in patients treated with palliative radiation therapy to brain metastases. Acta Oncologica. 2015;54(9):1529-34.
108. Kristoffersen A. Estimating non-Gaussian diffusion model parameters in the presence of physiological noise and Rician signal bias. Journal of magnetic resonance imaging : JMRI. 2012;35(1):181-9. Epub 2011/10/06. doi: 10.1002/jmri.22826. PubMed PMID: 21972173.
109. Kiselev VG, Il'yasov KA. Is the "biexponential diffusion" biexponential? Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2007;57(3):464-9. doi: 10.1002/mrm.21164. PubMed PMID: 17326171.
110. Zhang JL, Sigmund EE, Rusinek H, Chandarana H, Storey P, Chen Q, et al. Optimization of b-value sampling for diffusion-weighted imaging of the kidney. Magnet Reson Med. 2012;67(1):89-97.
111. Istratov AA, Vyvenko OF. Exponential analysis in physical phenomena. Rev Sci Instrum. 1999;70(2):1233-57.
112. Ababneh Z, Beloeil H, Berde CB, Gambarota G, Maier SE, Mulkern RV. Biexponential parameterization of diffusion and T2 relaxation decay curves in a rat muscle edema model: decay curve components and water compartments. Magnetic resonance in medicine : official journal of the Society

of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2005;54(3):524-31. Epub 2005/08/09. doi: 10.1002/mrm.20610. PubMed PMID: 16086363.

113. Anderson SW, Barry B, Soto JA, Ozonoff A, O'Brien M, Jara H. Quantifying hepatic fibrosis using a biexponential model of diffusion weighted imaging in ex vivo liver specimens. Journal of magnetic resonance imaging : JMRI. 2012;30(10):1475-82. Epub 2012/08/28. doi: 10.1016/j.mri.2012.05.010. PubMed PMID: 22921938.

114. Andreou A, Koh D, Collins D, Blackledge M, Wallace T, Leach M, et al. Measurement reproducibility of perfusion fraction and pseudodiffusion coefficient derived by intravoxel incoherent motion diffusion-weighted MR imaging in normal liver and metastases. European radiology. 2013;23(2):428-34.

115. Bailey C, Vinnicombe S, Panagiotaki E, Waugh SA, Hipwell JH, Alexander DC, et al. Modelling Vascularity in Breast Cancer and Surrounding Stroma Using Diffusion MRI and Intravoxel Incoherent Motion. Breast Imaging: Springer; 2014. p. 380-6.

116. Bokacheva L, Kaplan JB, Giri DD, Patil S, Gnanasigamani M, Nyman CG, et al. Intravoxel incoherent motion diffusion-weighted MRI at 3.0 T differentiates malignant breast lesions from benign lesions and breast parenchyma. Journal of Magnetic Resonance Imaging. 2014;40(4):813-23.

117. Bourne RM, Kurniawan N, Cowin G, Stait-Gardner T, Sved P, Watson G, et al. Biexponential diffusion decay in formalin-fixed prostate tissue: Preliminary findings. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2012;68(3):954-9. doi: 10.1002/mrm.23291.

118. Chandarana H, Lee VS, Hecht E, Taouli B, Sigmund EE. Comparison of biexponential and monoexponential model of diffusion weighted imaging in evaluation of renal lesions: preliminary experience. Investigative radiology. 2011;46(5):285-91. Epub 2010/11/26. doi: 10.1097/RLI.0b013e3181ffc485. PubMed PMID: 21102345.

119. Cho GY, Moy L, Zhang JL, Baete S, Lattanzi R, Moccaldi M, et al. Comparison of fitting methods and b-value sampling strategies for intravoxel incoherent motion in breast cancer. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2014. doi: 10.1002/mrm.25484. PubMed PMID: 25302780.

120. Dopfert J, Lemke A, Weidner A, Schad LR. Investigation of prostate cancer using diffusion-weighted intravoxel incoherent motion imaging. Journal of magnetic resonance imaging : JMRI. 2011;29(8):1053-8. Epub 2011/08/23. doi: 10.1016/j.mri.2011.06.001. PubMed PMID: 21855241.

121. Dyvorne H, Jajamovich G, Kakite S, Kuehn B, Taouli B. Intravoxel incoherent motion diffusion imaging of the liver: Optimal b-value subsampling and impact on parameter precision and reproducibility. European journal of radiology. 2014;83(12):2109-13.

122. Forder JR, Bui JD, Buckley DL, Blackband SJ. MR imaging measurement of compartmental water diffusion in perfused heart slices. American journal of physiology Heart and circulatory physiology. 2001;281(3):H1280-5. Epub 2001/08/22. PubMed PMID: 11514298.

123. Grinberg F, Farrher E, Kaffanke J, Oros-Peusquens AM, Shah NJ. Non-Gaussian diffusion in human brain tissue at high b-factors as examined by a combined diffusion kurtosis and biexponential diffusion tensor analysis. NeuroImage. 2011;57(3):1087-102. Epub 2011/05/21. doi: 10.1016/j.neuroimage.2011.04.050. PubMed PMID: 21596141.

124. Guiu B, Cercueil J-P. Liver diffusion-weighted MR imaging: the tower of Babel? European radiology. 2011;21(3):463-7.

125. Heusch P, Wittsack H-J, Heusner T, Buchbender C, Quang MN, Martirosian P, et al. Correlation of biexponential diffusion parameters with arterial spin-labeling perfusion MRI: results in transplanted kidneys. Investigative radiology. 2013;48(3):140-4.

126. Hu G, Chan Q, Quan X, Zhang X, Li Y, Zhong X, et al. Intravoxel incoherent motion MRI evaluation for the staging of liver fibrosis in a rat model. Journal of Magnetic Resonance Imaging. 2014.

127. Ichikawa S, Motosugi U, Morisaka H, Sano K, Ichikawa T, Enomoto N, et al. MRI-based staging of hepatic fibrosis: Comparison of intravoxel incoherent motion diffusion-weighted imaging with magnetic resonance elastography. Journal of Magnetic Resonance Imaging. 2014.
128. Kakite S, Dyvorne H, Besa C, Cooper N, Facciuto M, Donnerhack C, et al. Hepatocellular carcinoma: Short-term reproducibility of apparent diffusion coefficient and intravoxel incoherent motion parameters at 3.0 T. Journal of Magnetic Resonance Imaging. 2015;41(1):149-56.
129. Klau M, Lemke A, Grünberg K, Simon D, Re TJ, Wente MN, et al. Intravoxel incoherent motion MRI for the differentiation between mass forming chronic pancreatitis and pancreatic carcinoma. Investigative radiology. 2011;46(1):57-63.
130. Lemke A, Laun FB, Simon D, Stieltjes B, Schad LR. An in vivo verification of the intravoxel incoherent motion effect in diffusion-weighted imaging of the abdomen. Magnet Reson Med. 2010;64(6):1580-5.
131. Ma C, Liu L, Li Yj, Chen Lg, Pan Cs, Zhang Y, et al. Intravoxel incoherent motion MRI of the healthy pancreas: Monoexponential and biexponential apparent diffusion parameters of the normal head, body and tail. Journal of Magnetic Resonance Imaging. 2015;41(5):1236-41.
132. Maier SE, Mulkern RV. Biexponential analysis of diffusion-related signal decay in normal human cortical and deep gray matter. Journal of magnetic resonance imaging : JMRI. 2008;26(7):897-904. Epub 2008/05/10. doi: 10.1016/j.mri.2008.01.042. PubMed PMID: 18467062; PubMed Central PMCID: PMC2782403.
133. Minati L, Zucca I, Carcassola G, Occhipinti M, Spreafico R, Bruzzone MG. Effect of diffusion-sensitizing gradient timings on the exponential, biexponential and diffusional kurtosis model parameters: in-vivo measurements in the rat thalamus. Magma. 2010;23(2):115-21. Epub 2010/04/09. doi: 10.1007/s10334-010-0208-9. PubMed PMID: 20376530.
134. Mulkern RV, Barnes AS, Haker SJ, Hung YP, Rybicki FJ, Maier SE, et al. Biexponential characterization of prostate tissue water diffusion decay curves over an extended b-factor range. Journal of magnetic resonance imaging : JMRI. 2006;24(5):563-8. Epub 2006/06/01. doi: 10.1016/j.mri.2005.12.008. PubMed PMID: 16735177; PubMed Central PMCID: PMC1880900.
135. Orsi G, Aradi M, Nagy SA, Perlaki G, Trauninger A, Bogner P, et al. Differentiating white matter lesions in multiple sclerosis and migraine using monoexponential and biexponential diffusion measurements. Journal of Magnetic Resonance Imaging. 2015;41(3):676-83.
136. Patel J, Sigmund EE, Rusinek H, Oei M, Babb JS, Taouli B. Diagnosis of cirrhosis with intravoxel incoherent motion diffusion MRI and dynamic contrast-enhanced MRI alone and in combination: Preliminary experience. Journal of Magnetic Resonance Imaging. 2010;31(3):589-600.
137. Rheinheimer S, Stieltjes B, Schneider F, Simon D, Pahernik S, Kauczor H, et al. Investigation of renal lesions by diffusion-weighted magnetic resonance imaging applying intravoxel incoherent motion-derived parameters—Initial experience. European journal of radiology. 2012;81(3):e310-e6.
138. Riches S, Hawtin K, Charles-Edwards E, De Souza N. Diffusion-weighted imaging of the prostate and rectal wall: comparison of biexponential and monoexponential modelled diffusion and associated perfusion coefficients. NMR in biomedicine. 2009;22(3):318-25.
139. Shinmoto H, Oshio K, Tanimoto A, Higuchi N, Okuda S, Kuribayashi S, et al. Biexponential apparent diffusion coefficients in prostate cancer. Journal of magnetic resonance imaging : JMRI. 2009;27(3):355-9. Epub 2008/09/05. doi: 10.1016/j.mri.2008.07.008. PubMed PMID: 18768281.
140. Suo S, Lin N, Wang H, Zhang L, Wang R, Zhang S, et al. Intravoxel incoherent motion diffusion-weighted MR imaging of breast cancer at 3.0 tesla: Comparison of different curve-fitting methods. Journal of Magnetic Resonance Imaging. 2014.
141. Tamura T, Usui S, Murakami S, Arihiro K, AKIYAMA Y, Naito K, et al. Biexponential signal attenuation analysis of diffusion-weighted imaging of breast. Magnetic Resonance in Medical Sciences. 2010;9(4):195-207.

142. Thoeny HC, De Keyzer F. Diffusion-weighted MR imaging of native and transplanted kidneys. Radiology. 2011;259(1):25-38. Epub 2011/03/26. doi: 10.1148/radiol.10092419. PubMed PMID: 21436095.
143. Vogel DWT, Zbaeren P, Geretschlaeger A, Vermathen P, De Keyzer F, Thoeny HC. Diffusion-weighted MR imaging including bi-exponential fitting for the detection of recurrent or residual tumour after (chemo) radiotherapy for laryngeal and hypopharyngeal cancers. European radiology. 2013;23(2):562-9.
144. Yoon JH, Lee JM, Baek JH, Shin C-i, Kiefer B, Han JK, et al. Evaluation of hepatic fibrosis using intravoxel incoherent motion in diffusion-weighted liver MRI. Journal of computer assisted tomography. 2014;38(1):110-6.
145. Yuan M, Zhang YD, Zhu C, Yu TF, Shi HB, Shi ZF, et al. Comparison of intravoxel incoherent motion diffusion-weighted MR imaging with dynamic contrast-enhanced MRI for differentiating lung cancer from benign solitary pulmonary lesions. Journal of Magnetic Resonance Imaging. 2015.
146. Zhang JL, Sigmund EE, Chandarana H, Rusinek H, Chen Q, Vivier PH, et al. Variability of renal apparent diffusion coefficients: limitations of the monoexponential model for diffusion quantification. Radiology. 2010;254(3):783-92. Epub 2010/01/22. doi: 10.1148/radiol.09090891. PubMed PMID: 20089719; PubMed Central PMCID: PMC2851010.
147. Isaacson E, Keller HB. Analysis of numerical methods: Courier Corporation; 2012.
148. Izenman A. Modern multivariate statistical techniques: Springer; 2008.
149. Farrar DE, Glauber RR. Multicollinearity in regression analysis: the problem revisited. The Review of Economic and Statistics. 1967:92-107.
150. Seber G, Wild C. Nonlinear regression. 1989. Wiley, New York; 1989.
151. Nash JC. Compact numerical methods for computers: linear algebra and function minimisation: CRC Press; 1990.
152. Landaw E, DiStefano JJ. Multiexponential, multicompartmental, and noncompartmental modeling. II. Data analysis and statistical considerations. American Journal of Physiology-Regulatory, Integrative and Comparative Physiology. 1984;246(5):R665-R77.
153. Mela CF, Kopalle PK. The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. Applied Economics. 2002;34(6):667-77.
154. Box M. Bias in nonlinear estimation. Journal of the Royal Statistical Society Series B (Methodological). 1971:171-201.
155. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. Measurement error in nonlinear models: a modern perspective: CRC press; 2006.
156. Bonate PL. The effect of collinearity on parameter estimates in nonlinear mixed effect models. Pharmaceutical research. 1999;16(5):709-17.
157. Mukherjee P, Berman J, Chung S, Hess C, Henry R. Diffusion tensor MR imaging and fiber tractography: theoretic underpinnings. American journal of neuroradiology. 2008;29(4):632-41.
158. Alin A. Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;2(3):370-4.
159. Chatterjee S, Hadi AS. Regression analysis by example: John Wiley & Sons; 2015.
160. Meloun M, Militký J, Hill M, Brereton RG. Crucial problems in regression modelling and their solutions. Analyst. 2002;127(4):433-50.
161. Reich JG. On parameter redundancy in curve fitting of kinetic data. Kinetic Data Analysis: Springer; 1981. p. 39-50.
162. Walter E, Pronzato L. On the identifiability and distinguishability of nonlinear parametric models. Mathematics and Computers in Simulation. 1996;42(2):125-34.
163. Marquaridt DW. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. Technometrics. 1970;12(3):591-612.

164. Belsley DA. Conditioning diagnostics: Wiley Online Library; 1991.
165. Belsley DA, Oldford R. The general problem of ill conditioning and its role in statistical analysis. Computational Statistics & Data Analysis. 1986;4(2):103-20.
166. Davison AC, Hinkley DV. Bootstrap methods and their application: Cambridge university press; 1997.
167. Cook RD, Weisberg S. Applied regression including computing and graphics: John Wiley & Sons; 1999.
168. Jones DK, Knösche TR, Turner R. White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI. NeuroImage. 2013;73:239-54.
169. Latt J, Nilsson M, Wirestam R, Johansson E, Larsson EM, Stahlberg F, et al. In vivo visualization of displacement-distribution-derived parameters in q-space imaging. Magnetic resonance imaging. 2008;26(1):77-87. doi: 10.1016/j.mri.2007.04.001. PubMed PMID: 17582719.
170. Ciris PA, Balasubramanian M, Seethamraju RT, Tokuda J, Scalera J, Penzkofer T, et al. Characterization of gradient echo signal decays in healthy and cancerous prostate at 3T improves with a Gaussian augmentation of the mono-exponential (GAME) model. NMR in biomedicine. 2016.
171. Rosenkrantz AB, Padhani AR, Chenevert TL, Koh DM, De Keyzer F, Taouli B, et al. Body diffusion kurtosis imaging: basic principles, applications, and considerations for clinical practice. Journal of Magnetic Resonance Imaging. 2015;42(5):1190-202.
172. Toivonen J, Merisaari H, Pesola M, Taimen P, Bostrom PJ, Pahikkala T, et al. Mathematical models for diffusion-weighted imaging of prostate cancer using b values up to 2000 s/mm : Correlation with Gleason score and repeatability of region of interest analysis. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2014. doi: 10.1002/mrm.25482. PubMed PMID: 25329932.
173. Grinberg F, Ciobanu L, Farrher E, Shah NJ. Diffusion kurtosis imaging and log-normal distribution function imaging enhance the visualisation of lesions in animal stroke models. NMR in biomedicine. 2012;25(11):1295-304.
174. Lu H, Jensen JH, Ramani A, Helpern JA. Three-dimensional characterization of non-gaussian water diffusion in humans using diffusion kurtosis imaging. NMR in biomedicine. 2006;19(2):236-47.
175. Jambor I, Merisaari H, Taimen P, Boström P, Minn H, Pesola M, et al. Evaluation of different mathematical models for diffusion-weighted imaging of normal prostate and prostate cancer using high b-values: A repeatability study. Magnet Reson Med. 2015;73(5):1988-98.
176. Glatting G, Kletting P, Reske SN, Hohl K, Ring C. Choosing the optimal fit function: comparison of the Akaike information criterion and the F-test. Medical physics. 2007;34(11):4285-92.
177. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
178. Spanos A. Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification. Journal of Econometrics. 2010;158(2):204-20.
179. Mayo D, Spanos A. Error statistics. Philosophy of statistics. 2011;7:152-98.
180. Murtaugh PA. In defense of P values. Ecology. 2014;95(3):611-7.
181. Burnham K, Anderson D. P values are only an index to evidence: 20th-vs. 21st-century statistical science. Ecology. 2014;95(3):627-30.
182. Yap B, Sim C. Comparisons of various types of normality tests. Journal of Statistical Computation and Simulation. 2011;81(12):2141-55.
183. Yazici B, Yolacan S. A comparison of various tests of normality. Journal of Statistical Computation and Simulation. 2007;77(2):175-83.
184. Charles N, Cowin G, Kurniawan N, Bourne R, editors. Biexponential modeling of diffusion in stroma and epithelium of prostate tissue. Joint Annual Meeting ISMRM-ESMRMB 2014; 2014.
185. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA: a cancer journal for clinicians. 2015;65(1):5-29.

186. Yu XQ, Luo Q, Smith DP, Clements MS, O'Connell DL. Prostate cancer prevalence in New South Wales Australia: A population-based study. Cancer epidemiology. 2015;39(1):29-36.
187. Bourne RM, Kurniawan N, Cowin G, Stait-Gardner T, Sved P, Watson G, et al. Microscopic diffusivity compartmentation in formalin-fixed prostate tissue. Magnetic resonance in medicine : official journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine. 2012;68(2):614-20. Epub 2012/07/19. doi: 10.1002/mrm.23244. PubMed PMID: 22807067.
188. Merisaari H, Movahedi P, Perez IM, Toivonen J, Pesola M, Taimen P, et al. Fitting methods for intravoxel incoherent motion imaging of prostate cancer on region of interest level: Repeatability and gleason score prediction. Magn Reson Med. 2016. doi: 10.1002/mrm.26169.
189. Taouli B, Beer AJ, Chenevert T, Collins D, Lehman C, Matos C, et al. Diffusion-weighted imaging outside the brain: Consensus statement from an ISMRM-sponsored workshop. Journal of Magnetic Resonance Imaging. 2016. doi: 10.1002/jmri.25196.
190. Cook JD. 2013. Available from: http://www.johndcook.com/blog/2013/03/05/data-calls-the-models-bluff/.
191. Hafidi B, Mkhadri A. The Kullback information criterion for mixture regression models. Statistics & probability letters. 2010;80(9):807-15.
192. Naik PA, Shi P, Tsai C-L. Extending the Akaike information criterion to mixture regression models. Journal of the American Statistical Association. 2007;102(477):244-54.
193. Huang HM, Shih YY, Lin C. Formation of parametric images using mixed-effects models: a feasibility study. NMR in biomedicine. 2015.
194. Estimating the reproducibility of psychological science. Science. 2015;349(6251). doi: 10.1126/science.aac4716.
195. Trafimow D, Marks M. Editorial. Basic and Applied Social Psychology. 2015;37(1):1-2. doi: 10.1080/01973533.2015.1012991.
196. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. The American Statistician. 2016:00-. doi: 10.1080/00031305.2016.1154108.
197. Gelman A, Loken E. The Statistical Crisis in Science Data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up.
198. Jones DK, Cercignani M. Twenty-five pitfalls in the analysis of diffusion MRI data. NMR in biomedicine. 2010;23(7):803-20. doi: 10.1002/nbm.1543. PubMed PMID: 20886566.
199. Sarewitz D. The pressure to publish pushes down quality. Nature. 2016;533(7602):147-.
200. Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biol. 2015;13(6):e1002165.
201. Health NIo. Rigor and Reproducibility.
202. Liu C, Liang C, Liu Z, Zhang S, Huang B. Intravoxel incoherent motion (IVIM) in evaluation of breast lesions: comparison with conventional DWI. European journal of radiology. 2013;82(12):e782-e9.
203. Malyarenko DI, Newitt D, J Wilmes L, Tudorica A, Helmer KG, Arlinghaus LR, et al. Demonstration of nonlinearity bias in the measurement of the apparent diffusion coefficient in multicenter trials. Magnet Reson Med. 2015;75(3):1312–23. doi: 10.1002/mrm.25754.
204. Jerome NP, Orton MR, d'Arcy JA, Collins DJ, Koh DM, Leach MO. Comparison of free-breathing with navigator-controlled acquisition regimes in abdominal diffusion-weighted magnetic resonance images: Effect on ADC and IVIM statistics. Journal of Magnetic Resonance Imaging. 2014;39(1):235-40.
205. Klauss M, Mayer P, Maier-Hein K, Laun FB, Mehrabi A, Kauczor H-U, et al. IVIM-diffusion-MRI for the differentiation of solid benign and malign hypervascular liver lesions—Evaluation with two different MR scanners. European journal of radiology. 2016;85(7):1289-94.
206. Lee Y, Lee SS, Kim N, Kim E, Kim YJ, Yun S-C, et al. Intravoxel incoherent motion diffusion-weighted MR imaging of the liver: effect of triggering methods on regional variability and measurement repeatability of quantitative parameters. Radiology. 2014;274(2):405-15.

207. Merisaari H, Jambor I. Optimization of b-value distribution for four mathematical models of prostate cancer diffusion-weighted imaging using b values up to 2000 s/mm2: Simulation and repeatability study. Magnet Reson Med. 2015;73(5):1954-69.