BOLD Parameter Estimation Using Particle Filters

Micah Chambers

Virginia Polytechnic Institute and State University

Contents

Chapter 1

Introduction

For the past twenty years Functional Magnetic Resonance Imaging (FMRI) has been at the forefront of cognitive research. Despite it's limited temporal resolution, FMRI is the standard tool for localizing neural activation. Whereas other methods of analyzing neural signals can be invasive or difficult to acquire, FMRI is relatively quick and cheap, and its analysis relatively straight forward.

By modeling the governing equations behind the neural response that drives FMRI, it is possible to make the analyses of FMRI even more powerful. The underlying state equations hold important information about how individual brain regions react to stimuli. The model parameters on the other hand, hold important information about the patients individual physiology including existing and future pathologies. In short, the long chain of events driving observable FMRI signal contains more information than which regions respond to which stimuli.

In the past fifteen years, a steady stream of studies have built on the original Blood Oxygen Level Dependent (BOLD) signal derivation first described by [1]. The seminal work by [2] attempted to explain the time evolution of the BOLD signal using a windkessel model to describe the time local changes in Deoxygenated Hemoglobin content. Various papers made substantial improvements to this model until [3] brought all the changes together into a single complete set of equations. And while there have been numerous adaptations in the model, many of them summarized in [4], even the most basic version has been shown to have less bias error than the convolution based "Canonical Hemodynamic Model" [4], [5]. On the other hand many of the BOLD signal models have far more parameters than a simple scaling parameter. In fact, the number of parameters range from seven [6] to 50 [7] per signal time course; a signal which may be as short as 100 samples long. Thus, even in a small FMRI image of the brain (20x20x20), the number of parameters could easily exceed 10,000. Clearly this number of parameters presents a significant risk of being under-determined and could suffer catastrophic variance error; to make no mention of the computation cost. In this work a method of countering these problem is presented with the use of a particle filter.

1.1 Overview

Detecting neural activity using the changes in FMRI images is based on the so called Blood Oxygen Level Dependent (BOLD) signal. The BOLD signal is caused by minute changes in the ratio of Deoxygenated Hemoglobin to Oxygenated Hemoglobin in blood vessels throughout the brain. Because Deoxygenated Hemoglobin (DHb) is paramagnetic, higher concentrations attenuate the signal detected during T2-weighted Magnetic Resonance Imaging (MRI) techniques. The most common FMRI imaging technique, due to its rapid repetition time (TR), is Echo Planar Imaging (EPI). When axons becomes active, a large amount of ions quickly flow out of the cell. In order for this action potential to occur again (and thus for the neuron to fire again), an active pumping process must move ions back into the axon. This process of recharging the axon requires extra energy, which temporarily increases the metabolic rate of oxygen. On a massive scale (cubic millimeter) this activation/recharge process happens continuously. However, when a particular region of the brain is very active, the action potentials occur significantly more often, resulting in significant local increase of the Cerebral Metabolic Rate of Oxygen (CMRO2). Thus, blood vessels in a very active area will tend to have less oxygenated hemoglobin (due to the increased rate at which oxygen is being consumed), and more deoxygenated hemoglobin, resulting in an attenuated FMRI signal. In compensation for activation, muscles that control blood vessels relax in that region to allow more blood flow, which actually overcompensates. This ultimately results in lower than average concentration of deoxyhemoglobin. Thus, the BOLD signal consists of a short initial dip in the MR signal, followed by a prolonged increase in signal that slowly settles out. It is this overcompensation that is the primary signal detected with FMRI imaging. This cascade of events is believed to consist of increased the local metabolism, blood flow, blood volume, and oxygenated hemoglobin. The differences in onsets of these various effects is what causes the overcompensation that is observable in FMRI. Unfortunately, FMRI has no inherent unit of measurement, and thus signal levels are all relative: within a particular person, scanner and run.

1.2 FMRI

Magnetic Resonance Imaging, MRI, is a method of building 3D images non-invasively, based on the difference between nuclear spin relaxation times in various molecules. First, the subject is brought into a large magnetic field which causes nuclear spins to align. Radio Frequency (RF) signals may then be used to excite nuclear spin away from the base alignment. As the nuclei precess back to the alignment of the magnetic field, they emit detectable RF signals. Conveniently, the excitation of nuclear spins return their original state at different rates, called the T1 relaxation time, depending on the atoms excited. Additionally, the coherence of the spins also decay differently (and quite a bit faster than T1 relaxation) based on the properties of the region. This gives two

primary methods of contrasting substances, which form the basis of T1 and T2 weighted images. Additionally, dephasing occurs at two different rates, the T2 relaxation time, which is unrecoverable, and T2* relaxation, which is much faster, but possible to recover from via special RF signals. T1 relaxation times are typically on the order of seconds if a sufficiently strong excitation was applied. In order to rapidly acquire entire brain images, as done in Functional MRI, a single large excitation pulse is applied to the entire brain, and the entire volume is acquired in a single T1 relaxation period. Because the entire k-space (spatial-frequency) volume is acquired from a single excitation, the signal-to-noise-ratio is very low in this type of imaging (Echo Planar Imaging).

Increasing the spatial resolution of EPI imaging necessarily requires more time or faster magnetic field switching. Increasing magnet switching rates though is difficult, because it can result in more artifacts, or even lower signal to noise ratios. The result is that at *best* FMRI is capable of 1 second temporal resolution. The signal is further diluted because each voxel contains the signal from a large number of neurons, capillaries and veins. Thus, the FMRI signal, which is sensitive to the chemical composition of materials, is the average signal from various types of tissue in addition to the blood. As mentioned in section 1.1, and explored in depth in section 1.3, the usefulness of FMRI comes from the discerning of changes in Deoxyhemoglobin/Oxyhemoglobin. Therefore, it is necessary to assume that in the short term the only chemical changes will be in capillary beds feeding neurons. In practice this may not be the case, for instance near significant veins, and it may explain some of the noise seen in FMRI imaging (see section 1.5. Because MRI lacks units and certain areas will have a higher base MR signal, all FMRI studies deal with percent change from the base signal; rather than raw values. This also removes most of the structural data which is not helpful in determining neural activity.

1.3   BOLD Physiology

It is well known that the two types of hemoglobin act as a contrast agents in EPI imaging [2], [8], [1], however the connection between Deoxyhemoglobin/Oxygenated Hemoglobin and neural activity is non-trivial. Intuitively, increased metabolism will increase Deoxyhemoglobin, however blood vessels are quick to compensate by increasing local blood flow. Increased inflow, accomplished by loosening capillary beds, precedes increased outflow, driving increased blood storage capacity. Since the local MR signal depends on the ratio of Deoxyhemoglobin to Oxygenated Hemoglobin, increased volume of blood can effect this ratio if metabolism doesn't exactly match the increased inflow of oxygenated blood. This was the impetus for the ground breaking balloon model ([2]) and windkessel model ([9]). These works derive from first principals the changes in deoxyhemoglobin ratio and volume of capillaries based on a given flow. These were the first two attempts to quantitatively account for the shape of the BOLD signal as a consequence of the lag between the cerebral blood volume

(CBV) and the inward cerebral blood flow (CBF). In fact [2] went so far as to show that a simple, well chosen blood flow waveform coupled with a square wave cerebral metabolic rate of oxygen (CMRO2) curve, in the context of a balloon model, could fully account for the BOLD signal.

Although [2] demonstrated that a well chosen flow waveform could explain most features of the BOLD signal, there was still a matter of proposing a realistic waveform for the CBF and for the CMRO2. [10] gave a reasonable and simple expression for CBF input, f, based on a flow inducing signal, s, in combination with the original balloon model where v is normalized cerebral blood volume (CBV), q is the normalized local deoxyhemoglobin/oxygenated hemoglobin ratio.

$$\dot{s} = \epsilon u(t) - \frac{s}{\tau_s} - \frac{f-1}{\tau_f} \tag{1.1}$$

$$\dot{f} = s \tag{1.2}$$

$$\dot{v} = \frac{1}{\tau_0}(f - v^\alpha) \tag{1.3}$$

$$\dot{q} = \frac{1}{\tau_0}\left(\frac{f(1 - (1 - E_0)^f)}{E_0} - \frac{q}{v^{1-1/\alpha}}\right) \tag{1.4}$$

where ϵ is a neuronal efficiency term, $u(t)$ is a stimulus, and τ_f, τ_s are both time constants, E_0 is the resting metabolic rate and α is Grubb's parameter controlling the balloon model.

This completed the basic balloon model, and was well summarized again in [11]. [12] refined the readout equation of the BOLD signal based on the deoxyhemoglobin content (q) and local blood volume (v), resulting in the final BOLD equation:

$$y = V_0((k_1 + k_2)(1 - q) - (k_2 + k_3)(1 - v)) \tag{1.5}$$

$$k_1 = 4.3 \times \nu_0 \times E_0 \times TE = 2.8 \tag{1.6}$$

$$K_2 = \epsilon_0 \times r_0 \times E_0 \times TE = .57 \tag{1.7}$$

$$k_3 = \epsilon_0 - 1 = .43 \tag{1.8}$$

Where $\nu_0 = 40.3s^{-1}$ is the frequency offset in Hz for fully de-oxygenated blood (at 1.5T), $r_0 = 25s^{-1}$ is the slope relating change in relaxation rate with change in blood oxygenation, and $\epsilon_0 = 1.43$ is the ratio of signal MR from intravascular to extravascular at rest. Although, these constants change with experiment (TE, ν_0, r_0), patient, and brain region (E_0, r_0), often the estimated values taken from [12] are taken as the constants $a_1 = k_1 + k_2 = 3.4$, and $a_2 = k_2 + k_3 = 1$ in studies using 1.5 Tesla scanners. While this model is very accurate, it is not perfect.

1.4 Post Stimulus Undershoot

Although the most widely used, the BOLD model described in Equation 1.4 and Equation 1.8 have been extensively added on to. The most significant feature missing from the original model is the "post-stimulus undershoot". The "post-stimulus" undershoot is the name for a prolonged subnormal BOLD response for a period of 10 to 60 seconds after stimulus has ceased ([13], [14]).

Because Equation 1.4 is not capable of producing such a prolonged undershoot, additional factors must exist. Two theories exist for the post stimulus undershoot. Recall that a lower than base signal means that there is an increased deoxyhemoglobin content in the voxel. The first and simplest explanation is that the post-stimulus undershoot is caused by a prolonged increase in CMRO2 after CBV and CBF have returned to their base levels. This theory is justified by quite a few studies that show CBV and CBF returning to the baseline before the BOLD signal ([15], [16], [3], [17], [18]). Unfortunately, because of limitations on FMRI and in vivo CBV/CBF measurement techniques it is difficult to isolate whether CBF and CBV truly have returned to their baseline. Other studies seems to indicate that there can be a prolonged supernormal CBV ([14], [7], [19]), although none of these papers completely rule out the possibility of increased CMRO2. The discrepancies may in part be explained by a spatial dependence in the post-stimulus undershoot; described by [20]. [13] makes a compelling case that most of the post stimulus undershoot can be explained by combination of a prolonged CBV increase, and a prolonged CBF undershoot, and that many of the previous measurements showing a quick recovery of CBV were in fact showing a return to baseline by arterial CBV.

Regardless of the probability that CMRO2 and CBF are detached, research into the post-stimulus undershoot has led to the creation of much more in depth models. In [?] additional state variables model oxygen transport, whereas [3] models CMR02 from a higher level, and somewhat more simply; though it still adds several new parameters. [7] introduces nonlinearities into the CBF equations as a method to explain the post-stimulus undershoot, which falls in line with a prolonged increase in CBF observed in [13]. Similarly [21] adds additional compartments to model the BOLD signal that result from venous and arterial blood. [4] compared these various models and though it did not deal extensively with the post-stimulus undershoot, it did show incremental improvements in quality from additional parameters (over the basic Balloon model); though at the cost of greatly increased complexity. Importantly,[4] did show that by simply adding viscoelastic terms from [3], a slowed return to baseline is possible to model, without greatly increasing complexity. Regardless, because of many of these models are extremely complex, and the benefits have yet to be proved, in this work the simple Balloon model will be used. Simplicity is even more important because it is the first time a particle filter has been used to calculate the BOLD parameters. Of course, future works could certainly benefit from the more advanced models, especially with the addition of viscoelastic effects.

In summary, there have been extensive attempts to refine the Balloon, many of them extremely precise. However, the trade off between increased complexity and increased flexibility is extremely important. Given that the basic BOLD model has 7 parameters, for this work, where computation time is very important, that model is the best fit.

1.5   Noise

As demonstrated in **??** the BOLD response has been extensively studied and despite minor discrepancies, the cause of the BOLD signal is well known. However, as FMRI detects an aggregate signal over the space of cubic centimeters, there are plenty of sources of noise. Though local neurons act "together" (i.e. around the same time), the density of neurons, the density of capillaries, and slight differences in activation across a particular voxel can all lead to signal attenuation and noise.

A particularly difficult form of noise present in FMRI is a low frequency drift, often characterized as a Wiener process ([6]). Though not present in all regions, as much as ten to fifteen percent of voxels can be affected ([22]), thus it is prevalent enough to cause significant inference problems [23]. It is still not clear what exactly causes this noise noise comes from, although it is possible it is the result of magnets heating up, or some distortion in magnetic fields [23]. It is clear that this drift signal is not solely due to a physiological effects, given its presence in cadavers and phantoms [24]. Interestingly, it is usually spatially correlated, and more prevalent at interfaces between regions, though by no means limited to such areas. Though one potential source could be slight movement, given that co-registration of volumes to a single time point is standard, this seems unlikely. Regardless, the problem mandates the use of a high pass filter to make inference reasonably powered [23].

In order to characterize the noise, I analyzed resting state data. During resting state, the patient is shown no images, and he is asked to avoid movement and complex thought, to the best of his abilities. Some limitations exist, for instance EPI sequences are extremely loud which could cause some stimulus in the auditory cortex, and of course very long sequences could give the patient time for his mind to wander. Overall though there should be no activation, and thus the signal will consist entirely of noise. Therefore resting state data is perfect for analyzing the noise distribution. The locations were chosen from various points around the brain, all in grey matter voxels (see **??** for a discussion on how grey matter voxels were found). The time series were also chosen because they were representative of different types of noise I found in the resting state data.

Because most methods (including the one used in this paper) assume the noise realizations are independent of each other, the auto- correlation is of particular interest (which is a necessary but not sufficient condition for independence). Gaussianity is also a common assumption made in studies of FMRI data, though that assumption is not made in this work. Regardless, comparing the distribution to a Gaussian is informative, so I used Q-Q

plots to compare example data with the Normal distribution. Additionally, in FMRI data the noise is often considered to be Wiener [11]. Recall that a Wiener random process is characterized by steps that are Gaussian; in other words the difference between any two samples is Normally distributed. The simulations discussed in **??** make use of this, by adding a Wiener random process to the overall signal. To determine whether the noise is in fact Wiener, the distribution of the difference between adjacent measurements was plotted against a Gaussian. Wiener processes further assume that the steps are independent; so the steps should have near flat autocorrelation functions.

Finally, removal of the so called "drift" is often performed with some variation of a high pass filter, so I checked the noise distribution after applying such a filter (in this case the subtraction of a spline, see section 4.2). Here I wanted to know how effective my high pass filter at the removal of Wiener noise.

Figure 1.1 shows the results with a regression line fit to the points on the Q-Q plot. Recall that in a Quartile-Quartile (Q-Q) plot, if the points plotted on the x-axis and the points on the y-axis come from the same *type* of distribution then all the points will fall on a line. Differences in the variance will cause the line to have a slope other than 1, while differences in the expected value will cause the fitted line to be shifted. In these Q-Q plots, the points are being compared to the standard Gaussian distribution, so the quality of the line fit determines how closely the points fit the Gaussian distribution. Note that in Figure 1.1 the points have all been normalized (changed to percent difference).

A Wiener process should still conform to the normal distribution, with a variance proportional to the run-time. Note that 1.1(a) and 1.1(b) are relatively well described by a Gaussian process with a small autocorrelation, 1.1(c) and 1.1(d) are not. In particular the tails of 1.1(c) do not seem to fit the Gaussian well. Also note the significant autocorrelation in 1.1(c) and 1.1(d). As expected, the noise is not strictly Gaussian white noise. On the other hand, the steps do conform rather closely to the normal distribution. As expected, most of the autocorrelation disappears for the step data. Given that the steps seem to fit the Normal distribution, the low autocorrelation indicates that the steps could be Independently Distributed. Therefore, the assumption that the noise follows a Wiener process seems to be correct.

De-trending the time-series by subtracting a spline fit to the distribution removed much of the autocorrelation present in 1.1(c) and 1.1(d), though clearly not perfectly. Though the distributions still do not exactly fit the Normal, 1.3(d) clearly is much improved compared to 1.1(d). In all, the de-trending is effectively removing Wiener noise.
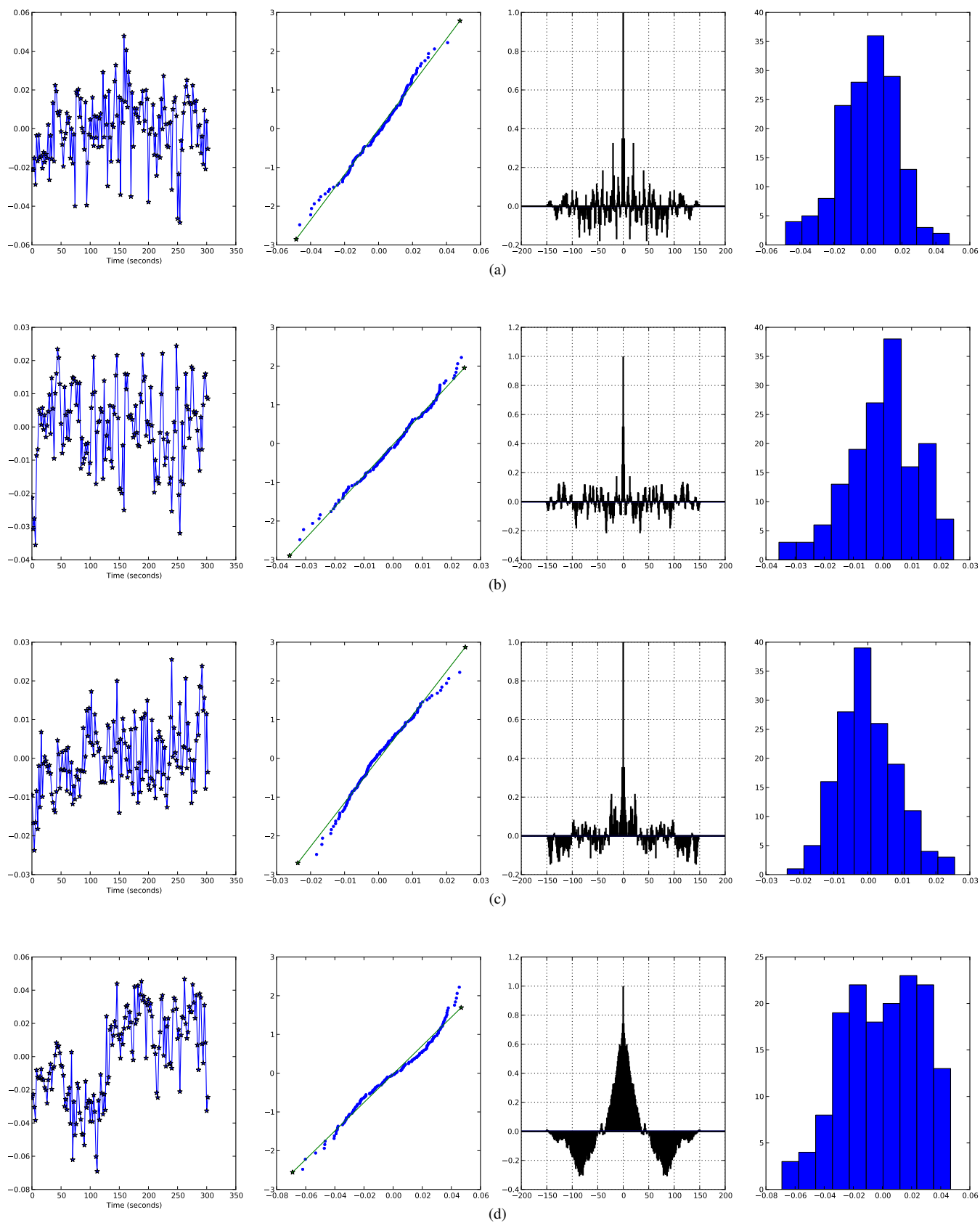
9

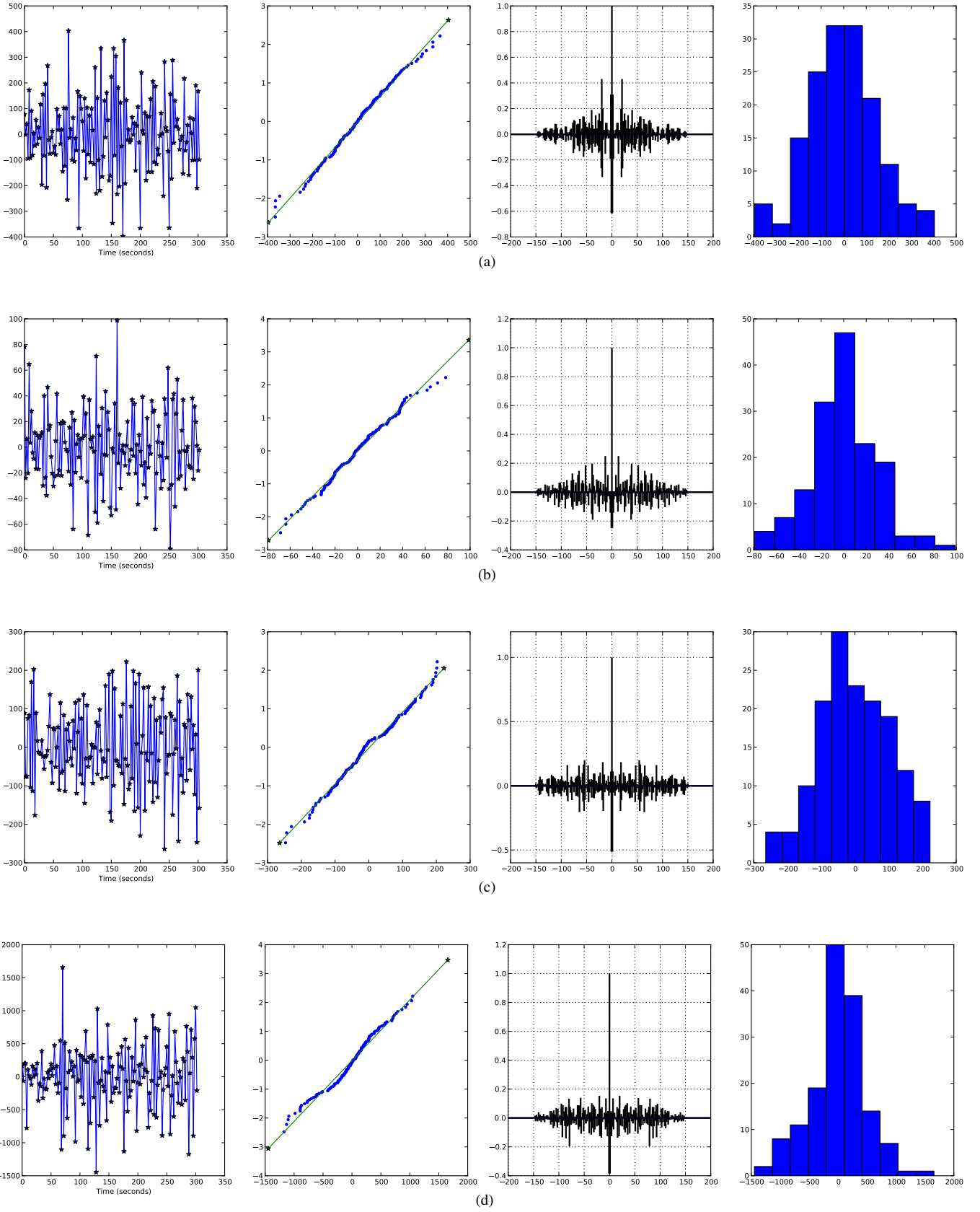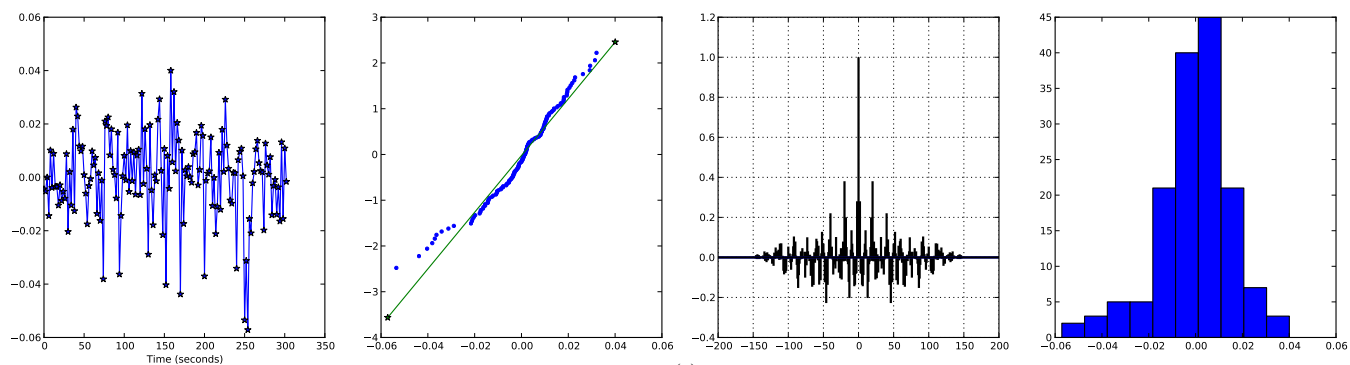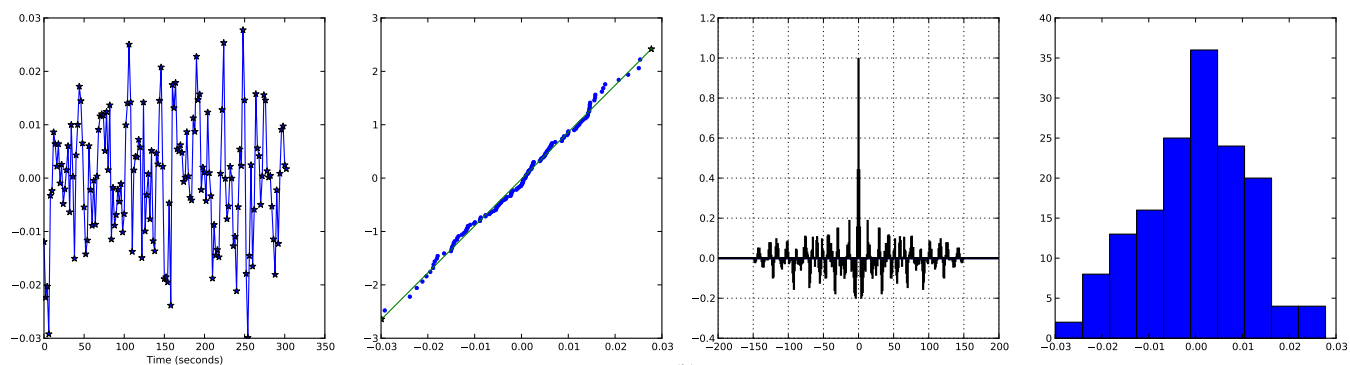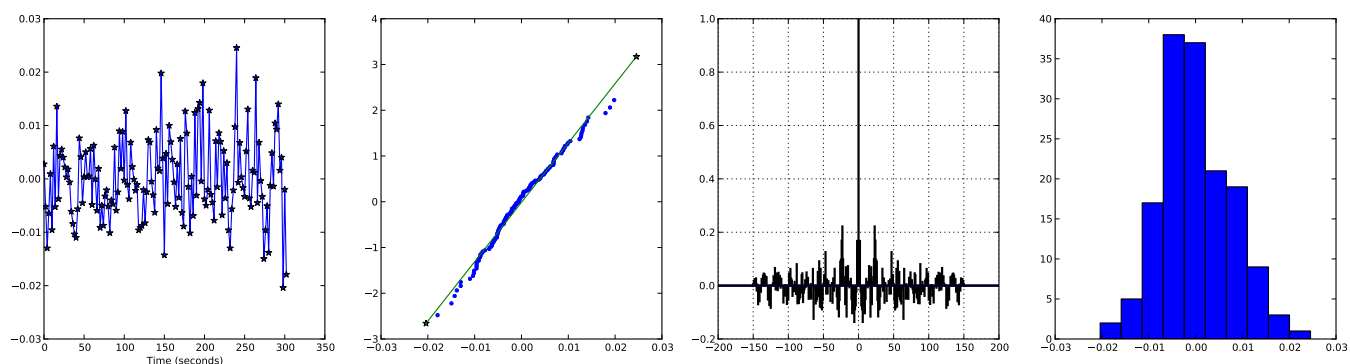Figure 1.1: Q-Q Plots of normalized resting state data

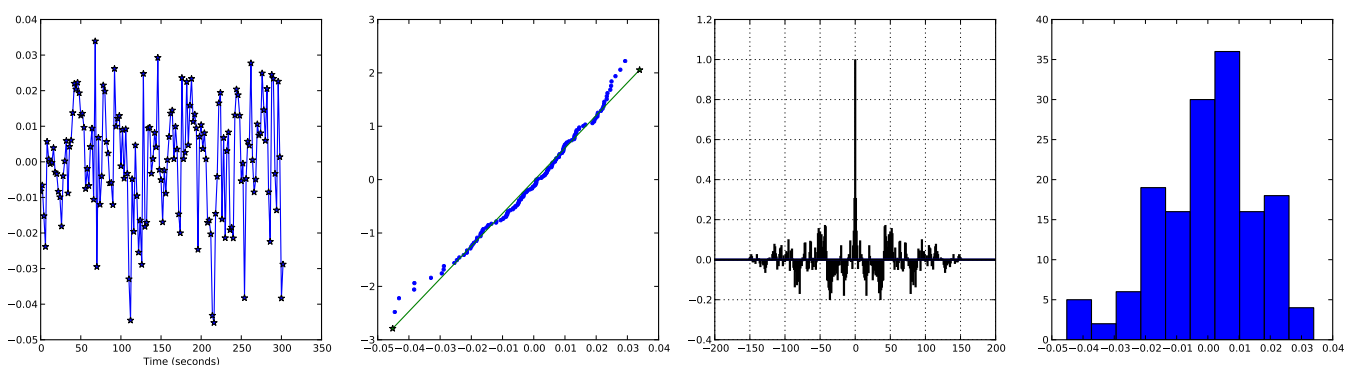Figure 1.2: Q-Q Plots of resting state data, using the BOLD signal changes

Figure 1.3: Q-Q Plots of resting state data, after the de-trending

12

Parameter	[25]	[29]	[30]
ϵ	$N(.54, .1^2)$	$.069 \pm .014$	(NC)
τ_s	$N(1.54, .25^2)$	4.98 ± 1.07	2.2
τ_f	$N(2.46, .25^2)$	8.31 ± 1.51	.45
τ_0	$N(.98, .25^2)$	8.38 ± 1.5	.94
α	$N(.33, .45^2)$	$.189 \pm .004$	.4 (NC)
E_0	$N(.34, .1^2)$	$.635 \pm .072$	.6 (NC)
V_0	.03 (NC)	$.0149 \pm .006$	(NC)

Table 1.1: Parameters found by various studies. (NC) indicates that the value wasn't calculated. [30] made use of the values from [25] where not explicitly stated

1.6   Properties of the BOLD Model

Since the first complete BOLD model was proposed by [25], several studies have endeavored to analyze properties of that model. The most important property is that the system is dissipative, and given enough time will converge to a constant value. This is found simply by analyzing the eigenvalues of the Jacobian of the state equations, ([4], [26]). The steady state of the Balloon model equations gives:

$$
\begin{aligned}
s_{ss} &= 0 \\
f_{ss} &= \tau_f \epsilon u + 1 \\
v_{ss} &= (\tau_f \epsilon u + 1)^\alpha \\
q_{ss} &= \frac{(\tau_f \epsilon u + 1)^\alpha}{E_0}(1 - (1 - E_0)^{1/(\tau_f \epsilon u + 1)}) \\
y_{ss} &= V_0((k_1 + k_2)(1 - q_{ss}) - (k_2 + k_3)(1 - v_{ss}))
\end{aligned}
\tag{1.9}
$$

In real FMRI data, there is a significant nonlinearity in response; with short sequences responding disproportionately strong ([27], [28], [4]). This nonlinearity is accounted for in the Balloon model, although [4] shows that if there will be large variance in the length of signals, modeling Neural Habituation may be necessary to fully capture the range of responses. Stimuli that last longer than 4 seconds tend to be more linear, which is why block designs are so well accounted for by the General Linear Model ([27], [4]). For this paper we will use only a simple version of the Balloon model, described by Equation 1.4, to keep the solution tractable.

Another interesting result of [4] was the sensitivity analysis. There it was found that many of the parameters compensate for each other, implying that exact inference of parameters may not be possible without constraint.

Several studies have been able to calculate distributions for the Balloon parameters, those results are summarized in Table 1.1

Necessary? [Image with two different α's] [image comparing the results of 10% changes in various signals]

Chapter 2

Prior Works

The ultimate purpose of this work is to provide a new tool for analyzing FMRI data.

Currently, FMRI is used to determine the location of responses to stimuli. The method of finding activation is discussed in **??**. The goal of this work is to move away from the question "Is this region active" and instead ask "What does the activation look like". Answering this question necessitates more complex models and certainly will result in longer run times. Already there have been several other attempts to model the BOLD response; these works will be discussed in [all other sections basically].

2.1 Statistical Parametric Mapping

Although not strictly the same as parameter calculation from FMRI, activation detection is similar and worth discussing. Estimation of parameters is a generalization of the idea of activation detection. Given the popularity of Statistical Parametric Mapping (SPM) it is important to draw a distinction between the methods proposed in this work.

2.1.1 Classical Activation Detection

The most basic method of analyzing FMRI data is through a standard T-test between "resting state" and "active state" samples. Simply put, the mean is calculated separately for non-stimulus and stimulus time intervals. A classic t-test may then be applied, giving the probability that the distributions actually have the same mean. Because of the correlated noise present in FMRI (section 1.5), it is necessary to apply some sort of high-pass filter to the data. Without applying such a filter, P values must be set extraordinarily high to prevent false positives [23]. If there truly is signal due to stimuli, the distributions will not actually be independent Gaussians

because activation does not fit a square wave (section 1.3). For this reason other methods are often more used, as discussed in subsection 2.1.3.

2.1.2 Random Field Theory

SPM methods make significant use of T-Tests across large regions; however, such T-tests work slightly differently than a single individual test. A t-test with a p-value of .05 over a modestly sized FMRI image, say 10,000 voxels, will on average generate 500 false positives. This is called the multiple comparison problem. Traditional Bonferroni Correction deals with this by requiring each test to pass with P value of $\frac{.05}{10000}$. The probability of a single false positive would then .05. Unfortunately this leads to unrealistically low p-values; so low that it would be impossible for any biological system to satisfy. To compensate, a Gaussian kernel is applied to smooth the image. This has the benefit of reducing the noise variance and decreasing the effective number of independent measurements. Because the number of independent measurements is smaller, Bonferroni correction can theoretically be applied with a lower scaling factor than the original voxel count [31]. A side effect of this, a single voxel activation is virtually impossible to detect.

2.1.3 General Linear Model

The most common FMRI analysis technique is SPM, though there are more advanced versions that the simple square wave method discussed in subsection 2.1.1. Hierarchal Models are one important improvement that allows researchers to combine data across multiple runs, patients and stimuli (see [32] for more on Hierarchical Modeling). Hierarchical Models concatenate all the data into a single dataset, then perform a linear fit between a design matrix and the data. The design matrix encapsulates all known experimental factors such as stimuli, young/old, etc. The equation for a general linear model is:

$$Y(t) = X(t)\beta + \epsilon(t) \tag{2.1}$$

where $Y(t)$ is the smoothed or de-trended time course of measurements, $X(t)$ is the design matrix, β is a column vector of weights, and ϵ is the error. Thus for every time, the measurement is assumed to be a weighted sum of the columns of X plus some error. The calculation of β is then performed using a maximum likelihood or gradient descent search to minimize the error.

As mentioned previously, a square wave stimulus does not result in a square wave in the activation of brain regions. The BOLD signal is in fact a significantly smoothed version of the stimuli. As such, when fitting an FMRI time course to the input, the input (X's columns) is usually smoothed to reduce bias error. The best method, that maintains a linear fit, is convolving the input with a Hemodynamic Response Function (HRF). The

Figure 2.1: Hemodynamic Response Function todo

so called "Hemodynamic Response Function" mimics the basic shape of BOLD activation, including a delay due to rise time and fall time. The fitting process is then a least squares fit over the space of the vector β. Therefore $Y(t)$ is estimated as a linear combination of the columns of X.

Smoothing the input with a single HRF poses certain problems. It is well known that different Hemodynamic Response Functions are necessary for different regions of the brain. The "Canonical" HRF that is most used, has been optimized for the visual cortex. As section 1.4 discussed, there are certainly variations in the shape of the BOLD response, both between brain regions and patients. [5] discusses the implications of choosing an incorrect HRF, the most important of which is a definite increase in false negatives. While an atlas of Hemodynamic Response Functions for each region could definitely mitigate this risk, it does not deal with variation between patients. Thus, the inability to fit parameters other than scale definitely hinders analysis of activation. As a result, there is significant bias error when using a linear fit to localize activation. Notably, studies will also have a bias toward the visual cortex, precisely because it is so well studied.

2.1.4 Hierarchical Linear Models

As mentioned previously and discussed extensively in [25] and [?], hierarchical models may be applied to account for systematic differences between subjects. For instance, if the study happens to be a mix between young and old, incorporating that into the model is wise, regardless if that is the purpose of the test. The reason to do this is to account for additional variance that may not be present in similar studies. The Hierarchical form used by [25] is shown in **??**.

$$Y(t) = X_1(t)\theta_1 + \epsilon_1(t)$$

$$\theta_1(t) = X_2(t)\theta_2 + \epsilon_2(t)$$

$$...$$

$$\theta_{n-1}(t) = X_n(t)\theta_n + \epsilon_n(t) \tag{2.2}$$

The Empirical Bayes algorithm is used in both in both [25] and [?]. As a consequence, point estimators are used for each θ, rather than the full distributions.

2.1.5 Conclusion

In all, the GLM is extremely useful for determining linear dependence of a set of regressors on the output. Unfortunately, as discussed in section 1.6 there are significant nonlinearities that almost certainly cause false negatives in the Statistical Parametric Maps. Unfortunately nonlinear analyses have only recently become feasible, so the scale of the problem is still unknown. The problem is highlighted by the relatively common scenario where no significant activation can be found in a single run [6] [29].

The static nature of the linear model also limits its inference power. Besides not permitting HRF differences between patients, there is no reasonable way to incorporate other forms of physiological data. Combined FMRI CBF or CBV imaging methods are rapidly getting better, as seen in [?]. These techniques could shed light on neural activation by providing extra measurements, yet a physiologically reasonable model is necessary for this. In reverse, activation detection methods also don't have the ability to identify pathologies based on state variables or parameters. For example decreased compliance of blood vessels could indicate, or even cause, a neurological condition that is not easily seen in other imaging modalities. Thus, the benefits of physiologically meaningful models are manifold.

2.2 Solving the BOLD Model

Unlike Statistical Parametric Mapping, the techniques described in this section are all attempts to learn some version of the BOLD model. There have been quite a few efforts to quantify the parameters of the various BOLD models. Although [2] and [10] both proposed physiologically reasonable values for the model parameters, [25] was the first paper to calculate the parameters based on actual FMRI data. However, in that case, the voxels were chosen from regions that were detected as active by the GLM. It is therefore possible that the parameters are biased toward parameters that fit the linear model.

2.2.1 Linear Approximation

In [33], a novel combination of linear and nonlinear modeling was used to generate parameter estimates. Because it is impossible to calculate the partial derivative of the output with respect to parameters, $\frac{\partial \theta}{\theta}$, [33] approximates the partial. At each step of the Expectation- Maximization algorithm, the differential equation is integrated, calculating the residuals. Then, for each θ_i surrounding the current estimate of θ, a Volterra-Kernel expansion of the output y is generated. Generation of the Volterra Kernel is relatively quick, and, since it is linear there is an analytical solution. Thus, once the Volterra Kernel exists, the value of y at that measurement point, for that θ_i can easily be found. Thus, one part of the partial is calculated at one point using $\frac{y(t)-y_i(t)}{\theta - \theta_i}$.

This is repeated for every dimension of θ to give the full $\frac{\partial y}{\theta}(t)$. This is only at time t though, and it must be repeated for every measurement. Finally the full derivate matrix is filled out, and the next step in the E-M algorithm can proceed. The full E-M algorithm for estimating the states is notation-heavy and can be found in [33].

Although this is certainly an interesting method of performing non-linear regression, there are a few caveats. First, the partials are numerical approximations, based on an approximate values (using Volterra-Kernels) of y. Importantly, the Volterra-Expansion of y is not able to model interactions between state variables; interactions that were found to cause interesting behavior in this work. In subsection 2.2.3, for the purpose of demonstrating nonlinearities in the signal, it was necessary to propagate state variables through 1 second of simulation. For the distribution of state variables mentioned in Table 2.1 it was necessary to use step sizes smaller than .001 to prevent unpredictable behavior. This is using the average parameters, thus the only complication came from the interplay between states, which can certainly lead to interesting behavior. Additionally, all the tests performed in [33] were on regions found to be active by the General Linear Model. For this reason, the reliability of the approximation is unknown for regions that are active but sufficiently nonlinear to avoid detection by conventional tests.

2.2.2 Nonlinear Least Squares

Rather than localizing activation, by using the physiologically plausible BOLD model, it is possible to determine the values of governing parameters.

Although there are certainly benefits to using a derived model, rather than a purely empirical model, there are serious implications. The first problem is that all the powerful classical techniques of gradient descent are off limits; since the model is a true nonlinear dynamical system with no closed form solution. The implication of this is that the calculation of a Jacobian for residuals won't work; and thus powerful techniques such as the Gauss-Newton method, which are helpful in many nonlinear problems, are off limits. Additionally, a gradient descent is difficult to perform without the ability to form partials of the output with respect to all the parameters.

Although anything requiring a Jacobian is out, there are other heuristic techniques that could potentially illuminate the BOLD response. Simulated Annealing (SA) is a common method of optimizing high dimensional regression problems. The idea is to pick a random start, and then at each iteration pick a random nearby point, and if that point is below some energy constraint (energy is a function of the residual), called the temperature, the algorithm moves to that point and continues with the next iteration. The temperature is slowly lowered until no nearby points below the temperature can be found (or the temperature drops below the current point). There are variations of this, for instance it is common to require every movement to be in the downward direction

(in terms of energy). Like most nonlinear optimization problems, there is no guarantee of an optimal solution, although the longer the algorithm is allowed to run, the better the solution will get. Since every step requires an entirely new run of the BOLD model, it can be extremely time consuming, which is why we are not using it here.

Algorithm 2.1 Simulated Annealing Algorithm

Initialize Θ, or if there exists a decent estimate start there
Initialize temperature, T to value above initial energy
while $E(\Theta) < T$ **do**
  repeat
    Pick θ near Θ
    Calculate energy, E, of θ
  until $E > T$
  Move to new estimate: set $\Theta = \theta$
end while

[simulated annealing image?]

Another potential method of interest is the use of Genetic Algorithms (GA). Genetic algorithms are similar to Simulated Annealing, in that they randomly move to better solutions based on a cost function. However; in genetic algorithms a single point estimate isn't used. Instead a population of estimates is generated, each with distinct parameters, and then each set of parameters is rated with a fitness function. Parameter sets that are good get a higher "fitness"; then new parameter sets are generated by randomly combining pieces of the old parameter sets. The pieces are typically chosen at a rate proportional to the fitness of the donor; thus "fit" parameter sets tend to pass on their properties. In addition to this, random mutations may be introduced that come from no existing parent. The new generation is then rated with the fitness function again, and the entire process starts over. The stop condition for a genetic algorithm is typically based on some threshold for fitness or a maximum number of generations. This is problematic, since it doesn't really provide any guarantee of even reaching a local minima, although in practice it can be quite effective.

[genetic algorithm picture]

Algorithm 2.2 Genetic Algorithm

Initialize N estimates, $E = \{\Theta_0, \Theta_1, ...\Theta_N\}$
for G generations **do**
  Calculate fitness for each Θ, Ex. for residual R, $1/R$ or, e^{-R}
  for i in N **do**
    Randomly select two parents (with higher probability for more fit Θ's)
    Randomly merge parts of the two parents to form a new Θ_i
    At some low probability change one or two parameters in Θ_i
  end for
end for

Although both these methods can be highly effective, they have the downside of requiring very high com-

putation time. In this case of the BOLD model, each time the energy or fitness needs to be calculated, a large number of cycles must be spent re-simulating the BOLD model for the set of parameters. As we'll discuss in **??**, the Particle Filter method is able to circumvent this re-calculation to some degree.

2.2.3 Unscented Kalman Filter

The Unscented Kalman Filter (UKF) is a powerful Gaussian/Bayes filter that attempts to model the posterior distribution of dynamical systems as a multivariate Gaussian. The Unscented Kalman Filter (UKF) generalizes the Extended Kalman Filter by allowing the state update to be a function, g,

$$X(t) = g(u(t), X(t-1)) \tag{2.3}$$

$$Y(t) = h(X(t) \tag{2.4}$$

In order to estimate the posterior at t, a deterministic set of "sigma" points (often 2 per dimension, plus 1 at the mode of the multivariate distribution) weighted according to a Gaussian estimate of $X(t-1)$ are passed through the update equation. This set of points are then used to estimate the mean and covariance of $X(t)$. The benefit of this, is that it requires no Jacobian and only a few extra calculations to get a decent esimate of a posterior Gaussian. In the BOLD case, the set of equations we are modeling have no closed form solution, and finding the Jacobian is impossible without approximations. Although [6], [26] mention a Jacobian of the BOLD response, this is not strictly the case and is rather $\frac{\partial J}{\partial t}$ rather than a true Jacobian. This is important because the Extended Kalman filter depends on the Jacobian to map a Gaussian through the advancement of time. Thus the popular Extended Kalman Filter won't work in this case, whereas the Unscented Kalman Filter still does. In fact [26] uses the UKF to perform a similar type of analysis to the one performed in this work.

The difficulty of using a Kalman Filter, however, is that it assumes a multivariate Gaussian for the state variables, $X(t-1)$. The problem with this is that when a system is significantly nonlinear, the state at $X(t)$ will almost certainly be non-Gaussian, and thus estimating $X(t+1)$ with $X(t)$ as a multivariate Gaussian will perpetually introduce error in the distribution. Furthermore, it is not really known what sort of underlying distributions may exist in such a mixed biological, mechanical, chemical system such as the brain. Assuming the parameters all to be Gaussian may in fact be a gross error. On the other hand, for small variances and short time steps the gaussiann distribution is a good fit, and so in some limited cases the Unscented Kalman Filter could work well. These are non-trivial issues given that the assumption of Gaussianity is what allows the UKF

Parameter	Run 1
τ_0	.98
α	.33
E_0	.34
V_0	.03
τ_s	1.54
τ_f	2.46
ϵ	.54
V_t	N(1, .09)
Q_t	N(1, .09)
S_t	N(1, .09)
F_t	N(1, .09)

Table 2.1: Parameters used to test Gaussianity of variables after being transitioned through the BOLD model

to estimate the posterior using only the first and second moments; two parameters that don't uniquely describe most distributions.

To determine the amount of error incurred in a Gaussian estimate during a typical sample period, the states of BOLD equations were assigned according to four dimensional Gaussian. The states were then propagated through two seconds of simulation (a typical TR in FMRI) and then the resulting marginal distributions were compared with a Gaussian distribution. The purpose is to determine the degree to which the results of simulation will result in non-Gaussian output, given a Gaussian input. This also demonstrates the degree of nonlinearity present in the system. The parameters used are shown in Table 2.1

Notably s_t has intentionally been set to a non-equilibrium, but physiologically plausible value. The value of u is left at zero the entire time, so the system will decay naturally (see section 1.3), though initializing s at a non-zero level will drive the system for several seconds. Figure 2.3 shows the results when the system is essentially left on for 100 milliseconds after setting the variables according to Table 2.1. The Q-Q plots fit very will with a Gaussian, demonstrating that at this short time interval nonlinearities have not yet begun to effect the distribution. However, Figure 2.4 us the result after 1 second, which is faster than most FMRI scanners are capable of sampling at. At that range the tails of the distributions for v and q are clearly starting to deviate from the Gaussian distribution. As a result the uncertainty in y is deviating from the Gaussian distribution as well. This is important, because although approximating the distribution with a Gaussian based on the first two moments will work in the short run, there will be residual error in the distribution.

On the other hand, this effect is more limited if the initial variance is somewhat smaller. In tests with those cases, it took much longer for the nonlinearities to skew the distribution. That result could be encouraging to those looking to use the UKF, if the distributions are kept relatively thin.

Another potential problem with the UKF is that typically the sigma points, used to estimate the posterior probability, $P(X(t)|X(t-1), u(t))$, are located on the main axes. As a result, the covariances are not al-
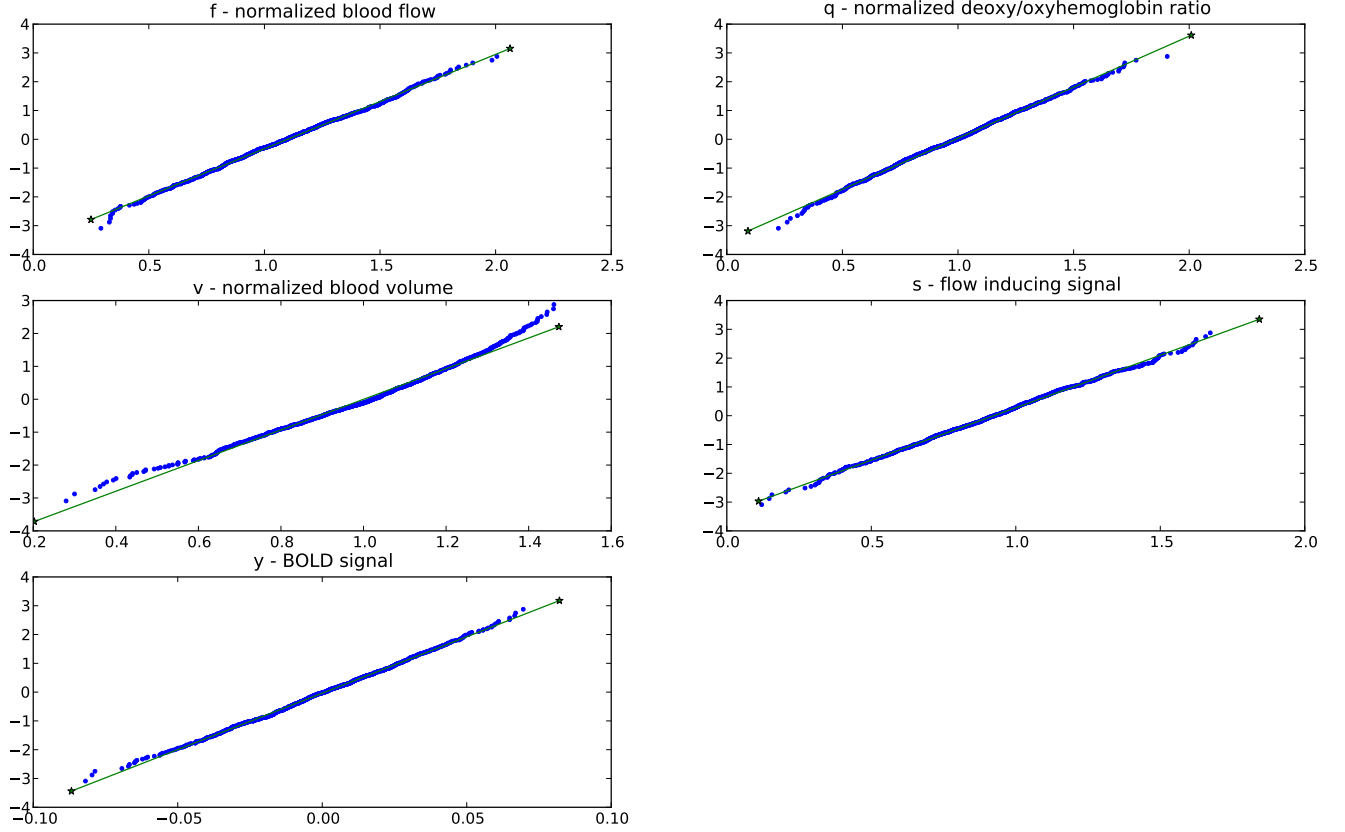
Figure 2.3: Distributions of state variables after simulating for .1s

lowed to be effected by the state transition the same way the variances are. While this may be reasonable in low-dimensional systems, high dimensional systems have a much greater potential for interplay between the variables. While this problem is relatively easy to fix (by using more sigma points off the main axes), there is a definite cost in complexity.

Ultimately, there is a distinct possibility that the nonlinearities of the BOLD model make Gaussian estimates unrealistic and thus less effective. More advanced tests where static variables such as α are varied as well could shed even more light on the issue. The trouble with using the UKF then to estimate parameters is that all eleven members of X would be treated like a single joint Gaussian distribution which certainly compound the issues of nonlinearity seen in Figure 2.3 and Figure 2.4
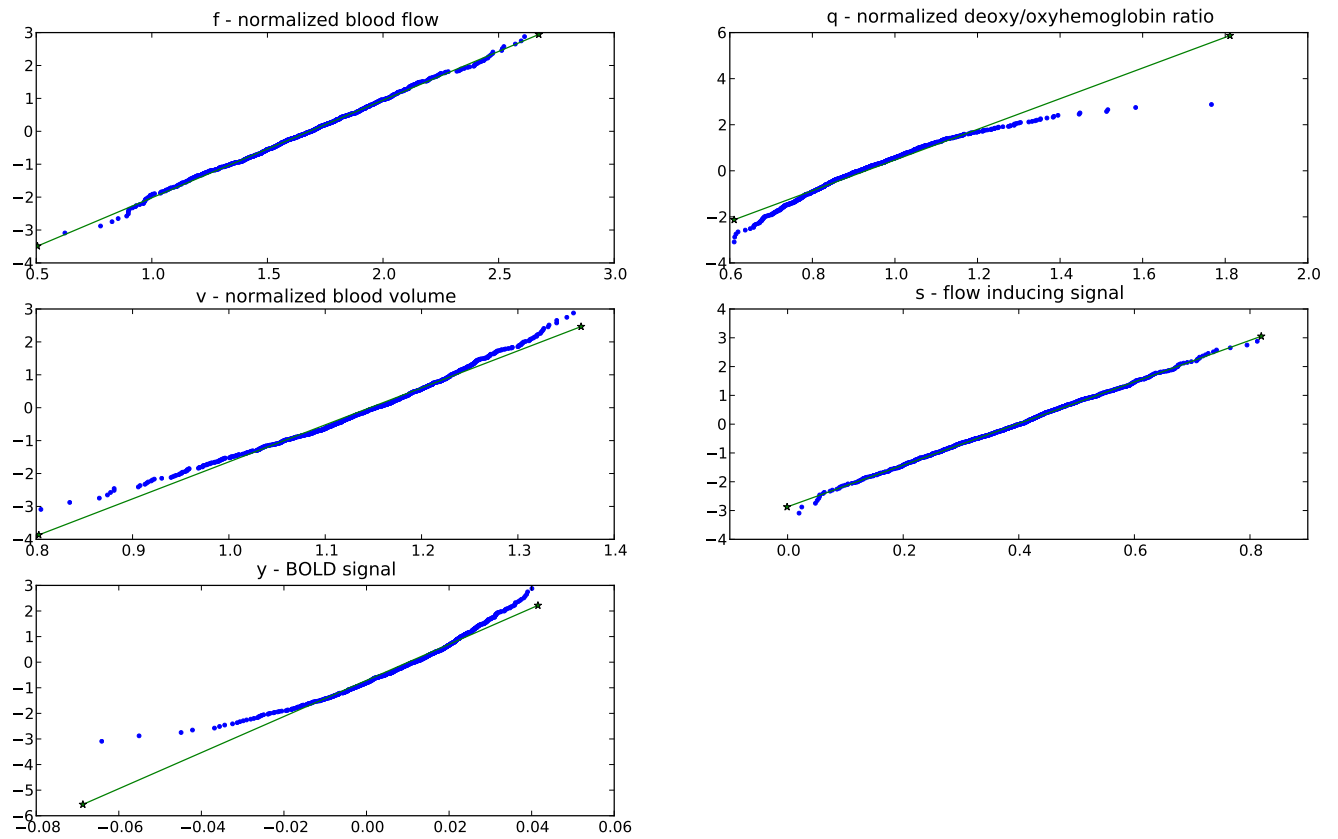
Figure 2.4: Distributions of state variables after simulating for 1s

2.3 Hybrid Methods

In [6], a maximum likelihood method for innovation processes was used, as described by [35]. [35] uses a similar construction to a Kalman filter, to break the time series into a series of innovations, for which Maximum Likelihood was performed. While this is probably the most likely solution to give the correct output, there are a few problems. First, every step in parameter space requires a recalculation of all the state variables. With two or three parameters this is fine, more than that, and calculations can become intractable. Additionally, there is no way to ensure a global minimum is reached. Although this is also true of the methods in subsection 2.2.2, those methods contained random elements designed to overcome this issue. [results?]

In [29], a hybrid particle filter/gradient descent algorithm was used to simultaneously derive the static and dynamic parameters, (classically known as parameters and state variables, respectively). Essentially a particle filter is used to calculate the state variables at each time; then the estimated distribution of the particles is used to find the most likely set of parameters that would give that distribution of state variables. This process is repeated until the parameters converge. Interestingly [29] comes to a very different set of parameter estimates as compared to the original [10] estimates (Table 1.1). In fact the results are significantly different from virtually every other study. The most obvious discrepancy is the significantly longer time constants, τ_f, τ_s and τ_0. While of course this could be poor convergence of the algorithm, there is another other possibility. Unlike all the other methods mentioned, excepting the methods in **??**, the algorithm described in [29] does not depend on prior distributions. It is possible then that the bias toward the prior in other methods screwed the results. While [29] is certainly in the minority; further exhaustive studies of the parameters, using unbiased techniques may be called for.

[graph of friston parameters vs. johnston]

[might want to mention dynamic causal modeling]

2.4 Conclusion

Clearly there is no ideal solution to solving this system of nonlinear equations. Exhaustive search type methods such as those employed by [29] and [30] can have extremely long run times even for a single voxel. While Volterra models are an interesting solution,there have not yet been exhaustive tests to determine whether such approximations work well throughout state space. The most promising method of those reviewed here is the Kalman filter based method. It is able to maintain a very fast runtime while still approaching the solution. While the reliance on a Gaussian estimate to the true posterior distribution could cause problems, some modifications could make it extremely powerful. The particle filter method proposed in the next section bears a strong

resemblance to the unscented Kalman filter; albeit with more point estimating the posterior.

Chapter 3

Particle Filters

3.1 Introduction

Particle filters, a type of Sequential Monte Carlo (SMC) methods, are a powerful way of estimating the posterior probability distribution a set of parameters given a timeseries of measurements. Unlike Markov Chain Monte Carlo (MCMC) estimation, particle filters are designed for time-varying random variables. The idea behind particle filters is extremely similar to Kalman Filters; however, unlike Kalman Filters, distributions are stored as an empircal distribution rather than the as the first two moments of a Gaussian. Thus particle filters are preferred when the model is nonlinear, and thus non-gaussian.

3.1.1 Model

The idea of the particle filter is to build an empirical distribution out of a large number of parameter *sets*, called particles. Each particle contains all the parameters and states needed to propagate the model forward. The particle filter begins with a wide distribution (called the Prior Distribution) of possible particles and then, as measurements come in, weights particles based on the quality of their output estimates. Thus parameter sets that tend to give good estimations of the measurements get weighted higher than parameter sets that give poor estimates. Although the reliance on a prior distribution can be troublesome, when the system being modeled has physical meaning, establishing reasonable ranges for parameters can actually be quite easy. Optimizing the prior can be more difficult though, unless the system has been extensively studied.

Suppose a discrete set or stream of measurements are given, $\{y_k, k = 1, 2, 3, ...K\}$, where K is infinite for a stream. Suppose also that there is a hidden set of state variables, denoted $\{x_k, k = 1, 2, 3, ...K\}$ the dictates the movement of y_k. The goal of the particle filter is to estimate the *distribution* of the true parameters

Θ that dictates the movement of x_k. Additionally the particle filter also estimates the distribution of x_k. The only difference between the members of parameter vector Θ and the members of x_k is that the memebers of Θ have no known update equation. Members of both vectors are permitted to have some time varying noise, although this is often not explicitly modeled. The generic, continuous, nonlinear system definition is shown in Equation 3.1.

$$\dot{x}(t) = f(t, x(t), u(t), \theta, \nu_x)$$

$$y(t) = g(t, x(t), u(t), \theta, \nu_y) \tag{3.1}$$

$x(t)$ is vector of state variables, θ is a vector of system constants, $u(t)$ is an input, $y(t)$ the observation, and ν_x and ν_y are random variates. Although any of these variables could be a vector, for the sake of simplicity only θ and $x(t)$ will be considered as such.

Although not strictly for particle filters in general, a few assumptions are made for this work. First, the systems are assumed to be time invariant. This assumption is based on the idea that if you paused the system for Δt seconds, when unfrozen the system would continue as if nothing happend. Few biological systems are predictable enough for them to be summarized by a time varying function, least of all the brain. While heart beats are certainly periodic and have an effect on the BOLD signal, the period varies too much for the system to be considered varying with time. Next, its assumes that input cannot directly influence the output, which in the case of the BOLD signal is a good assumption. Also, time is treated as discrete and noise is considered to be additive. Finally, because the only difference between the members of $x(t)$ and θ is an update function, from now on x will contain θ. The assumptions now allow for a simplified version of the state space equations:

$$\dot{x}_k = f(x_{k-1}, u_k) + \nu_x \tag{3.2}$$

$$y_k = g(x_k) + \nu_y \tag{3.3}$$

3.1.2 Prior

The goal of the particle filter is to evolve a probably distribution $Pr(\hat{x}(T)|u[0:T], y[0:T])$, that asymptotically approaches the probability distribution $Pr(x(T)|u[0:T])$. Considering that y contains measurement noise as well as noise from x, it is clear that $Pr(x(t)|u[0:T])$ is not a single true value but a true posterior. To begin with, the particle filter starts with a prior distribution, and N_p particles need to be drawn from that distribution,

$\alpha(x)$:

$$\{\hat{Pr}x_i(0), w_i] : x_i(0) \sim \alpha(x), w_i = \frac{1}{N_p}, i \in \{1, 2, ..., N_p\}\} \tag{3.4}$$

Where N_p is the number of particles or points used to describe the prior using a Mixture PDF.

$$\hat{Pr}(x(0) = \hat{x}) = \sum_{i=1}^{N_p} w_i \delta(\hat{x} - x_i(0)) dx \tag{3.5}$$

Where $\delta(x - x_0)$ is 1 if and only if $x = x_0$ (the Kronecker delta function).

If a true prior is preferred, then the weights should all be $1/N_p$, and since x_i was drawn from the prior, this will be an approximation of the prior distribution. If a relatively flat prior is preferred, then each particle's weight could be divided by the density, $\alpha(x_i)$, which creates a flat prior with support points in the region of $\alpha(x)$. Either way, $\alpha(x)$ should be much broader than the true posterior, $Pr(x(0))$, since the choice of support points is crucial to the convergence of any sampling importance algorithm. For the BOLD signal all the parameters have been studied and have relatively well known mean and variance, so a prior could be very helpful. We ran simulations for both normalized and un-normalized priors, although we believe in cases such as this, where a good prior exists, it should be used. For strictly positive parameters (members of x) we used a gamma distribution, whereas for parameters that could be negative, we used a Gaussian distribution. In both cases standard deviations twice that found in previous studies were used.

Note that all the probabilities implicitly depend on $u[0 : T]$, so those terms will be left off for simplicity. Once the probability, $\hat{Pr}(x(T)|x[0 : T - 1], y[0 : T - 1])$ has been found (initially this is just Mixture approximating the prior since no measurements are available and no previous probabilities are available), its possible to approximate the probability for short times between times when measurement is available, by shifting the probability according the progression of the state equations. This is only an approximate, since integrating ν_d should increase uncertainty as time without a measurement passes.

$$\hat{Pr}(x(T + \Delta t)) \approx \sum_{i=1}^{N_p} w_i \delta \left(x - (x_i(T) + \int_T^{T+\Delta} \dot{x}_i(t) dt) \right) \tag{3.6}$$

3.1.3 Weighting

When a measurement becomes available it is incorporated into the probability. This process of incorporating new data is called sequential importance sampling, and eventually causes the probability to converge. The weight is defined as

$$w_i(T) \propto \frac{\hat{Pr}(x_i[0 : T]|y[0 : T])}{q(x_i[0 : T]|y[0 : T])} \tag{3.7}$$

where q is called an *importance density*, meaning it decides where the support points for $x(T)$ are located. To remove the bias due to the location of the support points, we divide by $q(x_i[0:T]|y[0:T])$. By dividing by the posterior density of the support points (particles), the effect of the particle distribution may be removed from the posterior density. As a result the weight is dependent solely based on $\hat{P}r(x_i[0:T]|y[0:T])$, the probability of the ith particle's measurements being different from $y[0:T]$ due to noise alone. An example of an importance density would be drawing a large number of points from the standard normal, $N(0,1)$ and then weighting each point, l by $1/\beta(l), \beta \sim N(0,1)$. Of course if there is a far off peak in the posterior that q does not allocate support points in, there will be a quantization error, and that part of the density can't be modeled. This is why it is absolutely necessary that q covers $\hat{P}r(x_i[0:T]|y[0:T])$.

$q(x_i[0:T]|y[0:T])$ may be simplified by assuming that $y(T)$ doesn't contain any information about $x(T-1)$, which is more practical since knowledge of future measurements is impractical.

$$
\begin{aligned}
q(x[0:T]|y[0:T]) &= q(x(T)|x[0:T-1],y[0:T])q(x[0:T-1]|y[0:T]) \\
&= q(x(T)|x[0:T-1],y[0:T])q(x[0:T-1]|y[0:T-1]) \\
&= q(x(T)|x(T-1),y[0:T])q(x[0:T-1]|y[0:T-1])
\end{aligned}
\tag{3.8}
$$

In this paper we will use $q(x_i(T)|x_i(T-1),y[0:T]) = \hat{P}r(x_i(T)|x_i(T-1))$, based on the Markov assumption, and the belief that the state space model is able to approximate the true state. This means that prior to re-weighting particles, the particles will be distributed the same as the previous time but moved forward according to the integration of $f(x(t),u(t))$.

In addition to $q(x_i(T)|x_i[0:T-1],y[0:T])$, the weight is also based on $Pr(x_i[0:K]|y[0:K])$, which may be broken up as follows.

$$
\begin{aligned}
\hat{P}r(x[0:T]|y[0:T]) &= \frac{\hat{P}r(y[0:T],x[0:T])}{\hat{P}r(y[0:T])} \\
&= \frac{\hat{P}r(y(T),x[0:T]|y[0:T-1])\cancel{\hat{P}r(y[0:T-1])}}{\hat{P}r(y(T)|y[0:T-1])\cancel{\hat{P}r(y[0:T-1])}} \\
&= \frac{\hat{P}r(y(T)|x[0:T],y[0:T-1])\hat{P}r(x[0:T]|y[0:T-1])}{\hat{P}r(y(T)|y[0:T-1])} \\
&= \frac{\hat{P}r(y(T)|x[0:T],y[0:T-1])\hat{P}r(x(T)|x[0:T-1],y[0:T-1])\hat{P}r(x[0:T-1]|y[0:T-1])}{\hat{P}r(y(T)|y[0:T-1])}
\end{aligned}
\tag{3.9}
$$

Using the assumption that $y(t)$ is fully constrained by $x(t)$ Equation 3.3, and that $x(t)$ is fully constrained

by $x(t-1)$ Equation 3.2, we are able to make the reasonably good assumptions that:

$$\hat{Pr}(y(T)|x[0:T], y[0:T-1]) = \hat{Pr}(y(T)|x(T)) \tag{3.10}$$

$$\hat{Pr}(x(T)|x[0:T], y[0:T-1]) = \hat{Pr}(x(T)|x(T-1)) \tag{3.11}$$

Additionally, for the particle filter $y(T)$ and $y[0:T-1]$ are given, and therefore constant across all particles. Thus $\hat{Pr}(x[0:T]|y[0:T])$ may be simplified to:

$$
\begin{aligned}
\hat{Pr}(x[0:T]|y[0:T]) &= \frac{\hat{Pr}(y(T)|x[0:T], y[0:T-1])\hat{Pr}(x(T)|x[0:T-1], y[0:T-1])\hat{Pr}(x[0:T-1]|y[0:T-1])}{\hat{Pr}(y(T)|y[0:T-1])} \\
&= \frac{\hat{Pr}(y(T)|x(T))\hat{Pr}(x(T)|x(T-1))\hat{Pr}(x[0:T-1]|y[0:T-1])}{\hat{Pr}(y(T)|y[0:T-1])} \\
&\propto \hat{Pr}(y(T)|x(T))\hat{Pr}(x(T)|x(T-1))\hat{Pr}(x[0:T-1]|y[0:T-1]) \tag{3.12}
\end{aligned}
$$

Plugging these simplifications into Equation 3.7 leads to:

$$
\begin{aligned}
w_i(T) &\propto \frac{\hat{Pr}(y(T)|x(T))\cancel{\hat{Pr}(x(T)|x(T-1))}\hat{Pr}(x[0:T-1]|y[0:T-1])}{\cancel{\hat{Pr}(x_i(T)|x_i(T-1))}q(x[0:T-1]|y[0:T-1])} \\
&\propto w_i(T-1)\hat{Pr}(y(T)|x(T)) \tag{3.13}
\end{aligned}
$$

Thus, by making the following relatively weak assumptions, evolving a posterior density is easy and requires almost no knowledge of noise distribution.

1. $f(t, x(t), u(t)) = f(x(t), u(t))$ and $g(t, x(t), u(t)) = g(x(t))$ provide a sufficiently flexible model to encapsulate the true time series.

2. $E[\nu_d] = 0$ and $E[\nu_y] = 0$, and $\nu_x = d\nu_d$, ν_y are stationary

3. The PDF $q(x_i(0))$ (the prior) fully covers $Pr(x_i(0))$

4. Markov Assumption: $Pr(x(T)|x[0:T]) = Pr(x(T)|x(T-1))$

5. $q(x[0:T-1]|y[0:T]) = q(x[0:T-1]|y[0:T-1])$

3.1.4 Basic Particle Filter Algorithm

From the definition of w_i, the algorithm sequential importance sampling (SIS) is relatively simple.

Initialize N_p Particles: $\{x_i(0), w_i(0) : x_i(0) \sim \alpha(x), w_i(0) = \frac{1}{N_p}, i \in \{1, 2, ..., N_p\}\}$

$T = \{\text{Set of Measurement Times}\}$

for t in T **do**

    for i in N_p **do**

        $x_i(t) = x_i(t-1) + \int_{t-1}^{t} f(x(\tau), u(\tau))d\tau$

        $w_i(t) = w_i(t-1)\hat{P}r(y(t)|x(t))$

    end for

end for

At $t + \Delta t, t \in T, \hat{P}r(x(t+\Delta t)) \approx \sum_{i=1}^{N_p} w_i(t)\delta\left(x - (x_i(t) + \int_t^{t+\Delta t} f(x(\tau), u(\tau))d\tau)\right)$

The result is then a discrete approximation of the posterior distribution.

3.1.5 Resampling

As a consequence of the wide prior distribution (required for a proper discretization of a continuous distribution), there will be many particles with insignificant weights. While this does help describe the tails of the distribution very well, it means that only a small portion of the computation will be spent describing the most probable region. Ideally every particle would equally decrease the entropy of the distribution, thus the lower the variance of the weights, the more efficiently the discrete distribution is in describing the continuous distribution. A common measure of "Particle Degeneracy" is the effective number of particles, described in (Bergman "Navigation and Tracking Applications", 1999, J S Liu and R Chen "Sequential Monte Carlo Methods for Dynamical Systems", 1998), which is based on the "true weight" of each particle. Of course the true weight is unknown, so a heuristic approximating N_{eff} is used:

$$\hat{N}_{eff} \approx \frac{N_p}{\sum_{i=1}^{N_p} w_i^2} \tag{3.14}$$

Any quick run of a particle filter will reveal that unless the prior is particularly accurate, N_{eff} drops precipitously. To alleviate this problem a common technique known as resampling must be applied. The idea of re-sampling is to draw from the approximate posterior, thus generating a replica of the posterior with a support more suited to the distribution. Thus, if weights are all set to $1/N_p$, and N_p points are drawn from the posterior,

$$\hat{\chi}_j \sim \left(\sum_{i=1}^{N_p} w_i(t)\delta(x - x_i(t))\right), j \in \{1, ..., N_p\} \tag{3.15}$$

then $\hat{\chi} \sim \hat{x}$ should hold. Unfortunately, this isn't necessarily the truth: since the support is still limited to the original particles, the number of unique particles can only go down. This effect, often dubbed "particle impoverishment" can result in excessive quantization errors in the final distribution. However, there is a solution.

Instead of sampling from the discrete distribution, a smoothing kernel is applied, and $\hat{\chi}_j$ are drawn from that distribution. Because the distribution is continuous, there is no way for a collapse of the particles to occur. The question then, is how to decide on the smoothing kernel. Often times the easiest way to sample from the continous distribution is to break the re-sampling down into two steps. First a member of the discrete distribution is randomly selected based on the weights, and then based on the smoothing a nearby state variable is selected. The process of the selection will be defined as:

$$\chi_i = x_i + h\sigma\epsilon \tag{3.16}$$

Where h is the bandwidth, σ is the standard deviation such that $\sigma\sigma^T = cov(x)$ and ϵ is drawn from the chosen kernel. It has been proven that when all the elements of the mixture have the same weight, as is the case after basic resampling, the kernel that minimizes the MSE between the estimated and true posterior is the Epanechnikov Kernel (cite Improving Regularised Particle Filters, C Musso, N Oudjane and F LeGrand).

$$K = \begin{cases} \frac{n_x+2}{2c_{n_x}}(1 - \|x\|^2) & if \ \|x\| < 1 \\ 0 & otherwise \end{cases} \tag{3.17}$$

3.1.6 Weighting Function

Because $\hat{P}r(y(T)|x(T))$, what I will call the weighting function, is based on an unknown distribution, it is necessary to decide on a function that will approximate $\hat{P}r(y(t)|x(T))$. Obviously the function, $\omega(y(t), f(x(t))$ needs to be centered at zero and have a scale comparable to the signal levels. Obviously if the actual noise present in $y(t)$ were to be known, then that would be the best distribution for Ω. In that case, particles that fell far out on that distribution would be statistically impossible representations of the system, and it would be completely reasonable to throw such particles away. While a Gaussian function is the natural choice, because this distribution and weight are unknown we wanted to try distributions with wider tails, so that outliers don't completely destroy particle's weights (and thus convergence proceeds more slowly).

Another natural choice might be one of the robust estimator weight functions, for example the Huber or bi-square. For the purpose of this work we will stick with long tailed distributions, however it is worth noting that long tails may not be the optimal choice in all, or even this situation. The justification for long tailed distributions is that we believe the noise to be long tailed, and the variance of the noise is not well known.

Therefore, we tried three weighting functions based on three distributions: Gaussian, Laplace and the Cauchy. The standard deviation of the distribution is extremely important to the convergence of particle filter. A standard deviation that is too large will not allow the distribution to converge in any reasonable number
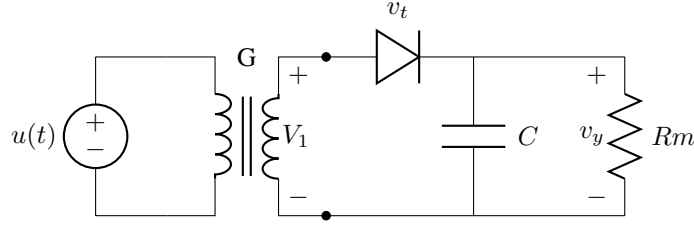
Figure 3.1: An Example Half Wave Rectifier Circuit, where G is the transformer gain, v_t is the activation voltage of the diode, $u(t)$ is the input at time t, C is the capacitance, R is the load resistance and v_y is the output voltage

Figure 3.2: Example Input/Output of the Half Wave Rectifier

of measurements. A standard deviation below the standard deviation of the noise will cause the algorithm to throw out perfectly acceptable particles. The weighting function ultimately will shape the output distribution, $P[y]$, into that distribution, however the distribution of x will still approach a reasonable estimate of its true distribution. Even an overly wide weighting function, will allow the Gaussian Mixture estimate of X to converge to the correct location parameters of the "real" posterior distribution, though the scale parameters may be overly large.

A reasonable method of setting the standard deviation of Ω may be by taking a small sample from "resting" data and using the sample standard deviation. Since this is the first attempt at using particle filters for modeling the BOLD model, in this work we set the standard deviation manually at ¡weight standard dev¿, because it gives a more consistency and control. Of course this could be taken further, by testing the sample data against a set of stock distributions and choosing the best fit. This of course depends on having enough samples to make a reasonable inference, which may not always exist.

3.2   Simple, Nonlinear Example

A typical half wave rectifier takes a AC voltage circuit and removes one half (say the negative half) of the signal. The resulting waveform is still not DC, however it is then possible to use a capacitor to smooth the signal into something similar to DC, as shown in Figure 3.2. There are other, more complex circuits that convert the negative portion into positive and waste less energy but but here we will keep the system simple. Thus, let us consider a simple half wave rectifier circuit, shown in Figure 3.1.

The half wave rectifier circuit smoothes the gaps between high voltage with a capacitor. Thus, when $u(t)G$ is less than v_t, the circuit will discharge the capacitor and maintain a non-zero voltage, but when $u(t)G$ is greater than v_t, the output voltage will be set by $u(t)G$ and the capacitor will charge up. We will assume a very simple model for all the components, ignoring the complex nonlinear behavior that can occur in a diode.

Although rectifiers are typically thought of as receiving an AC circuit 60 Hz, we will ignore such specifics and assume the voltage across the output of the transformer is simply a scalar multiple of the input voltage. As discussed in the subsection 3.1.1 any variable with uncertainty must be part of the state variable. Therefore the state variable will be: $X(t) = \{G, v_t, C, R_m, v_y\}$. Of course, $u(t)$ cannot be allowed to be a square wave in such a system, since that such a signal would never get across the transformer and regardless it would necessitate an unrealistic infinite current across the capacitor. The state equations would then be

$$v_y(t) = f(v_y(t-1), u(t)) = \begin{cases} u(t)G & \text{if } u(t)G - v_y \geq v_t \\ v_y(t-1)\left(1 - \frac{\delta t}{R_m C}\right) & \text{if } u(t)G - v_y < v_t \end{cases} \tag{3.18}$$

To run the particle filter is relatively easy then, since there exists a recursive definition of the dynamic state variable, v_y. To start with, an initial distribution must be assumed and while at first a Gaussian seems like a good idea, all the static state variables are strictly positive and thus not well suited to the Gaussian. In this case then, it would be wise to start with a Gamma distribution, and just be wary of any standard deviation that gets larger than the prior mean. We will define the gamma distribution as follows:

$$X \sim Gamma(k, \theta) \rightarrow f(x) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k \Gamma(k)} \tag{3.19}$$

where Γ is the gamma function. The in some ways the margin for error is decided by the weighting function, which here will define as $W(V_y, v_{yi})$, where V_y is the actual measurement, v_y is the estimate based on all the particles, and v_{yi} is the estimate by a particular (i^{th}) particle. The choice of this function is difficult, and although the Gaussian is typically used, we found the exponential helpful in dealing with particle deprivation. The algorithm will then look like the following,

Initialize N_p Particles:

for i in N_p **do**

$\quad G \sim Gamma(\frac{\mu_G^2}{\sigma_G^2}, \frac{\sigma_G^2}{\mu_G})$

$\quad v_t \sim Gamma(\frac{\mu_{v_t}^2}{\sigma_{v_t}^2}, \frac{\sigma_{v_t}^2}{\mu_{v_t}})$

$\quad C \sim Gamma(\frac{\mu_C^2}{\sigma_C^2}, \frac{\sigma_C^2}{\mu_C})$

$\quad R_m \sim Gamma(\frac{\mu_R^2}{\sigma_R^2}, \frac{\sigma_R^2}{\mu_R})$

$\quad v_y = 0$, (Assume the system has been off for a long time)

$\quad$ let $X_i(0) = \{G, v_t, C, R_m, v_y\}$

$\quad$ let $w_i(0) = 1$ or to make a flat prior, $w_i(0) = \frac{1}{Pr(X_i(0))}$

end for

Run the Filter:

for t in Set of Measurement Times **do**

    for i in N_p **do**

        $v_{yi}(t) = f(v_{yi}(t-1), u(t))$

        (All other members of $X_i(t)$ remain the same)

        $w_i(t) = w_i(t-1)W(V_y(t), v_y(t))$

    end for

end for

Initially the particles will have the same output, 0, however, as $u(t)$ changes, the response of each particle to that input will result in different outputs. Particles that have a v_{yi} near V_y will be weighted higher, and others farther away will be weighted lower. As the particle filter runs, weights will compound resulting in a distribution that asymptotically approaches the true joint distribution of the $X(t)$. Of course, as we mentioned in subsection 3.1.5, particles weighted zero do not significantly contribute to the empirical distribution, so re-sampling may be necessary. If the noise is assumed to be Gaussian then it is possible to further optimize. Thus we let h be defines as:

$$h = [N_s 8 c_{n_x}^{-1}(n_x + 4)(2\sqrt{\pi})^{n_x}]^{\frac{1}{n_x+4}} \tag{3.20}$$

and although it is very possible the underlying noise is non-gaussian, the Gaussian may work, but sub-optimally. It has been proposed that (Monte Carlo Approximations for General State-Space Models, markus Hurzeler and Hans R. Kunsch) if the underlying distribution is non-Gaussian, then using this bandwidth will oversmooth. In reality over smoothing should not be too great an issue because the smoothing is only being applied to find new particles. If the distribution is over smoothed then the algorithm may not converge as rapidly; however, because the bandwidth is still based on particle variance, which will decay as particles are ruled out, it is still able to converge. In fact over smoothing is preferrable to under smoothing, since the latter would result in false negatives, but the previous only results in a slower decay of the variance. At the same time, as n_x, the number of dimensions in x, goes to infinity, the standard deviation based approximation becomes less effective (cite a Tutorial on Particle Filters for on-line non-linear non-gaussian bayesian tracking, sanjeev arulampalam, simon maskell, neil gordon...). Because of the high dimensionality of our system, and limited measurements, it is helpful to have a broader bandwidth to explore the distribution. Nevertheless, because of the potentially wide smoothing factor applied by regularized resampling, performing this step at every measurement would allow particles a great deal of mobility. This mobility is the enemy of convergence, which is why regularized resampling should only be done when $\hat{N}_{eff}$ drops very low (say less than 50). Other than the periodic regularized resampling then, the regularized particle filter is nearly identical to the basic sampling importance sampling filter (SIS).

Initialize N_p Particles: $\{x_i(0), w_i(0) : x_i(0) \sim \alpha(x), w_i(0) = \frac{1}{N_p}, i \in \{1, 2, ..., N_p\}\}$

$T = \{\text{Set of Measurement Times}\}$

for t in T **do**

    for i in N_p **do**

        $x_i(t) = x_i(t-1) + \int_{t-1}^{t} f(x(\tau), u(\tau))d\tau$

        $w_i(t) = w_i(t-1)\hat{Pr}(y(t)|x(t))$

    end for

    Calculate N_{eff} with Equation 3.14

    if $N_{eff} < N_R$ (recommend $N_R = min(50, .1N_p)$) **then**

        Calculate empirical σ

        $h = [N_s 8 c_{n_x}^{-1}(n_x + 4)(2\sqrt{\pi})^{n_x}]^{\frac{1}{n_x+4}}$

        Redraw particles using (stratified) basic resampling

        for i in N_p **do**

            Draw $\epsilon \sim K$

            $x_i = x_i + h\sigma\epsilon$

        end for

    end if

end for

At $t + \Delta t, t \in T$, $\hat{Pr}(x(t+\Delta t)) \approx \sum_{i=1}^{N_p} w_i(t)\delta\left(x - (x_i(t) + \int_{t}^{t+\Delta t} f(x(\tau), u(\tau))d\tau)\right)$

The ultimate effect of this regularized resampling is a convergence similar to simulated annealing or a genetic algorithm. Versions of x that are "fit" (give good measurements) spawn more children nearby which allow for more accurate estimation near points of high likelihood. As the variance of the estimated x's decrease, the radius in which children are spawned also decreases. Eventually the radius will approach the width of the underlying uncertainty, ν_x and ν_y.

Chapter 4

Methods

4.1 Choosing $\hat{Pr}(y(T)|x(T))$

Choosing a representation of an unknown distribution is certainly tricky, and so the fact that $\hat{Pr}(y(T)|x(T)) = \nu_d + \nu_y$ means that there is a significant piece of the algorithm that is based primarily conjecture. Studies of the noise in FMRI typically attribute noise to a Gaussian random variable or an additive noise process with Gaussian steps.

4.1.1 Classical De-trending

The non-stationary aspect of a Weiner process as with ν_d is difficult to compensate for, and so various methods have been developed to compensate for it. [22] and [24] have demonstrated that this component is prevalent, and may in fact be a characteristic of FMRI. In some studies, as many as half the voxels benefit from detrending, meaning that this is certainly a serious barrier to inference. All the existing methods are performed during the preprocessing stage, rather than as an integral part of analyzing the BOLD signal. There is no shortage of theories on the "best" method of detrending; however a head to head comparison, [22], showed that in most cases subtracting off a spline works the best. The benefit of the spline versus wavelets, high pass filtering or other DC removal techniques is that the frequency response is not set. Rather, the spline is adaptive to the input, having a low cut off if the signal's median stays constant but a high cut off frequency if the signal's median tends to shift heavily over the course of time. In spite of this, the spline will still remove some amount of signal, just like all of these methods, which is why the the method proposed in subsection 4.1.2 was considered.

When using the median-based spline techniques, the resulting signal is will almost certainly end with a median of zero; in other words half the signal will be above zero, half below. The problem with this is this is

Figure 4.1: image of de-spline'd lines with "true" lines

not the "natural" BOLD signal. More specifically, when the signal is inactive, the BOLD response *should* be at 0% change from the base level; activation may then increase, or for short periods decrease from this base. After removing the spline, the BOLD resting state will be below 0%. This is troublesome because it reduces the power the BOLD model. One obvious solution is to add an arbitrary constant to each BOLD response. Of course this won't scale well to whole brain analysis. A more effective technique is adding a DC gain to the BOLD model. Essentially every measurement would then have a constant level added to it. Like all the other model parameters, enough measurements should allow the correct value to be found. On the downside adding another dimension increases the complexity of the model, for a variable that is relatively obvious to a human.

Thus, a more conventional approach was used. A robust estimator of scale, was used to determine roughly where the base level was. The Median Absolute Deviation (MAD) proved to be extremely accurate in determining how much to shift the signal up by. We tested both methods during the course of analysis, and found that the increase in model complexity far outweighted the slight increase in flexibility. Its possible that a more accurate method may exist; however, for this case the MAD seemed to work perfectly, as Figure 4.1 shows.

4.1.2 Delta Based Inference

I also propose and test a different method of dealing with the so called "drift". Instead of comparing the direct output of the particle filter with the direct measurement, the algorith compares the change in signal over a single TR. In most signal processing cases this would foolish, but that is because the general assumption that all noise is high frequency is not the case here. In fact, every pipeline for the analysis of BOLD signal uses a high pass filter, but low poss filters are rarely applied, because it is a well known fact that most of the signal is in the high frequency range and most of the noise is actually in the low frequency range. The particle filter is an extremely robust method of inference, and so I would assert that the particle filter ought to be given as *raw* data as possible. While taking direct measurements without de-trending would give awful results, using the difference removes the DC component and turns a Weiner process into a Gaussian random variable.

$$\Delta y = y(t) - y(t-1) = g(x(t)) - g(x(t-1)) + \nu_y(t) - \nu_y(t-1) + \nu_d(t) - \nu(t-1) \qquad (4.1)$$

Because ν_d is a Weiner process, then $\nu_d(t) - \nu_d(t-1)$ is simply a Gaussian step. If ν_d is some other additive process, the difference will still be one of a few stable distributions. If ν_y is i.i.d. then the resulting distribution will still be zero mean with a maximum variance of twice the original variance. All the assumptions made originally for the particle filter hold, and all of the parameters may be distinguished based on the step sizes, thus

Figure 4.2: frequency response graphs, highlighting noise frequency range and signal frequency range

it is not unreasonable to attempt to match the string of step sizes rather than string of direct readings.

4.2 Preprocessing

As discussed in the section on de-trending, the normal pipeline for analyzing FMRI involves a great deal of pre-processing. In this paper we make an effort to minimize any type of preprocessing that will degrade the signal. After FMRI data has been acquired it is always necessary to modify the data in some way to make different runs comparable. Because FMRI signal levels are unit-less, at the very least it is necessary to convert the data into % difference from the baseline. This process removes no data from signal since it merely subtracting then dividing by a constant. This is the signal that was input into the delta based particle filter. Of course there are much more advanced ways of performing this task. The generally accepted standard is actually to use a high pass filter, although the cutoff frequency is application dependent and often applied haphazardly. The high pass filter thus removes the DC component of the signal, and some amount of the so called "drift". The problem with this method is that it is not adaptive to the input. Huge variations in drift frequencies can exist in a single time-series. Thus, a single cutoff frequency could miss a significant drift component, or it could remove *actual* signal, if the cutoff frequency is set too high. This is why, as I mentioned in the De-trending section, a spline based detrending method will generally give better results.

For simulated and real images (tests with multiple time-series), tests were also run with and without Gaussian filtering with sigma of were run, since it is standard practice to apply a Gaussian spatial filter to the images at each timestep. Obviously a spatial filter such as Gaussian filtering increased SNR but can also lead to less precision in the output maps.

4.3 Simulation

We performed two types of simulations. First, we simulated a single BOLD time-series based on a random chosen set of model parameters. This process was relatively straight forward given the state-space equations for the BOLD signal. After a "true" signal was generated, we then added a carrier level, since BOLD is typically measured as a % difference from the base level. Finally, we added Gaussian noise, and a Weiner Process to the clean signal. The variance of the Gaussian noise may be expressed in terms of the desired noise SNR, R as:

$$var(y_{noisy}) = var(y)/R \qquad (4.2)$$

Since SNR doesn't have quite the same meaning for a Wiener process based noise, the variance of the Gaussian steps was set to be:

$$var(y_{noisy}) = var(y)/(4R) \tag{4.3}$$

Once this noisy simulated time series was generated, the exact same particle filter algorithm that would later be run on full sized images, was run on this single voxel image. We ran a series of tests to determine the convergence rate of the particle filter, the number of particles that were required, how weighting functions compared, how different de-trending methods compared with each other and, finally the variance of the result. By running the exact time-series with different noise realizations, it was possible to determine the model variance. As the reader may know, the error of an estimator may be calculated as:

$$MSE(\Theta) = Var(\Theta) + Bias(\Theta)^2 \tag{4.4}$$

The variance is an expression of how much the result would change for different noise realizations, whereas the bias is an expression of how well the model matches the true underlying model. In this case, because the same model is being used in the particle filter and underlying simulation, the bias is actually zero. Obviously when this is calculated using *real* data with an unknown underlying state space equation, there will be some amount of bias error, but assuming that the noise is similar to the noise used in these tests, the model variance will actually be about the same. Thus calculating the model variance is extremely helpful in calculating how well determined our model is, and how consistent it will be for real data. A single timeseries, as opposed to the thousands present in a real image, makes it easier to compare the output with the ground truth, with various parameters.

Second we used a modified version of the FSL tool POSSUM to generate an entire FMRI image from a parameter map. The parameter map was generated by creating a random image, smoothing it with a large Gaussian kernel, then thresholding the results. Finally connected regions were each given a set of parameters from a finite list of randomly chosen parameter sets. The result was a four dimensional (length x width x height x parameter) image with spatially varying parameters. Time-series of activation level was generated for each set of parameters, then activation levels were fed into POSSUM's function for generating frequency domain data. The patche for POSSUM will be made available. For each time-series in the simulate FMRI image, the final *static* parameters are saved into a parameter map. This parameter map may then be compared to the map used to generate the simulated data; additionally a new simulation using the calculated parameters may also be generated to test the functional difference between the two maps. This would give an absolute quantitative difference between the two parameter sets irrespective to parameter slopiness. So for instance, if V_0 is halved, ϵ

doubling may very well give a similar result. In this case the % difference between the parameters will be large in each case, but the functional difference between the parameters will not be great. This is obviously a bad situation, which is why we wanted to test for it.

4.4   Real Data

Finally, we also performed inference based on real FMRI data. The scanner we used...

The final result from calculating parameters with the real data was similar to that from the results from the POSSUM simulated data. The difference being that there was no ground truth the check it with.

Chapter 5

Results

5.1 Single Time-Series Simulation

Graphs:

For simulated data, single timeseries:

For {delta, DC/Spline}, {exponential, gaussian, cauchy}, {biased, unbiased initial}, {100, 500, 1000} particles

1. Ground truth vs. Estimated signal during particle filter run

2. Ground truth vs. Estimated signal with final parameter set

3. True Parameters vs. Final Parameter Sets

4. Variance of final parameters when faced with same ground truth, different noise

5. MSE of (a new timeseries based on X(t) vs. ground truth) for all t

6. Estimator Variance based on different noise runs

7. Final Particle Distribution

For Simulated Data, Full Volume:

5.2 Simulated Volume

1. Parameter Map

2. Error map of parameters

3. Histogram of %errors between parameters

4. Activation Map based on a single region with high ϵ, compared with linear

5.3   FMRI Data

....

image comparing epsilon-map with GLM activation map

Chapter 6

Conclusion

6.1 Future Work

*A more extensive investigation of the parameters distribution *Using a compliance model *restrict parameters more (may not be possible)

Bibliography

[1] S. Ogawa, R. S. Menon, D. W. Tank, S. Kim, H. Merkle, J. M. Ellermann, and K. Ugurbilt, "Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging A comparison of signal characteristics with a biophysical model," *Biophysical Journal*, vol. 64, pp. 803–812, 1993.

[2] R. B. Buxton, E. C. Wong, and L. R. Frank, "Dynamics of blood flow and oxygenation changes during brain activation: the balloon model," *Magn. Reson. Med.*, vol. 39, pp. 855–864, 1998.

[3] R. B. Buxton, K. Uludag, D. J. Dubowitz, and T. Lui, "Modeling the hemodynamic response to brain activation.," *NeuroImage*, vol. 23 Suppl 1, pp. S220–33, 2004.

[4] T. Deneux and O. Faugeras, "Using nonlinear models in fMRI data analysis: model selection and activation detection," *NeuroImage*, vol. 32, no. 4, pp. 1669–1689, 2006.

[5] D. a. Handwerker, J. M. Ollinger, and M. D'Esposito, "Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses.," *NeuroImage*, vol. 21, pp. 1639–51, Apr. 2004.

[6] J. J. Riera, J. Bosch, O. Yamashita, R. Kawashima, N. Sadato, T. Okada, and T. Ozaki, "fMRI activation maps based on the NN-ARx model," *NeuroImage*, vol. 23, pp. 680–697, 2004.

[7] Y. Behzadi and T. T. Liu, "An arteriolar compliance model of the cerebral blood flow response to neural stimulus," *NeuroImage*, vol. 25, pp. 1100–1111, 2005.

[8] R. M. Weisskoff, C. S. Zuo, J. L. Boxerman, and B. R. Rosen, "Microscopic susceptibility variation and transverse relaxation : theory and experiment," *Magnetic resonance in medicine*, vol. 31, no. 6, pp. 601–610, 1994.

[9] J. Mandeville, J. Marota, C. Ayata, G. Zaharchuk, M. Moskowitz, B. Rosen, and R. Weisskoff, "Evidence of a cerebrovascular postarteriole Windkessel with delayed compliance," *Journal of cerebral blood flow*

and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism, vol. 19, no. 6, pp. 679–689, 1999.

[10] K. J. Friston, A. Mechelli, R. Turner, and C. J. Price, "Nonlinear responses in fMRI: the Balloon model, Volterra kernels, and other hemodynamics," *NeuroImage*, vol. 12, pp. 466–477, 2000.

[11] J. J. Riera, J. Watanabe, I. Kazuki, M. Naoki, E. Aubert, T. Ozaki, and R. Kawashima, "A state-space model of the hemodynamic approach: nonlinear filtering of BOLD signals," *NeuroImage*, vol. 21, pp. 547–567, 2003.

[12] T. Obata, "Discrepancies between BOLD and flow dynamics in primary and supplementary motor areas: application of the balloon model to the interpretation of BOLD transients," *NeuroImage*, vol. 21, no. 1, pp. 144–153, 2004.

[13] J. J. Chen and G. B. Pike, "Origins of the BOLD post-stimulus undershoot.," *NeuroImage*, vol. 46, no. 3, pp. 559–68, 2009.

[14] J. B. Mandeville, J. J. A. Marota, C. Ayata, M. A. Moskowitz, R. M. Weisskoff, and B. R. Rosen, "MRI Measurement of the Temporal Evolution of Relative CMRO 2 During Rat Forepaw Stimulation," *Magnetic Resonance in Medicine*, vol. 951, pp. 944–951, 1999.

[15] J. Frahm, J. Baudewig, K. Kallenberg, A. Kastrup, K. D. Merboldt, and P. Dechent, "The post-stimulation undershoot in BOLD fMRI of human brain is not caused by elevated cerebral blood volume.," *NeuroImage*, vol. 40, pp. 473–81, Apr. 2008.

[16] M. J. Donahue, R. D. Stevens, M. de Boorder, J. J. Pekar, J. Hendrikse, and P. C. M. van Zijl, "Hemodynamic changes after visual stimulation and breath holding provide evidence for an uncoupling of cerebral blood flow and volume from oxygen metabolism.," *Journal of cerebral blood flow and metabolism : official journal of the International Society of Cerebral Blood Flow and Metabolism*, vol. 29, no. 1, pp. 176–85, 2009.

[17] H. Lu, X. Golay, J. J. Pekar, V. Zijl, and P.c.m, "Sustained poststimulus elevation in cerebral oxygen utilization after vascular recovery," *J. Cereb. Blood Flow Metab.*, vol. 24, pp. 764–770, 2004.

[18] Q. Shen, H. Ren, and T. Q. Duong, "CBF, BOLD, CBV, and CMRO(2) fMRI signal temporal dynamics at 500-msec resolution.," *Journal of magnetic resonance imaging : JMRI*, vol. 27, no. 3, pp. 599–606, 2008.

[19] J. J. Chen and G. B. Pike, "Origins of the BOLD post-stimulus undershoot," *NeuroImage*, vol. 46, no. 3, pp. 559–68, 2009.

[20] E. Yacoub, K. Ugurbil, and N. Harel, "The spatial dependence of the poststimulus undershoot as revealed by high-resolution BOLD- and CBV-weighted fMRI," *J. Cereb. Blood Flow Metab.*, vol. 26, pp. 634–644, 2006.

[21] Y. Zheng, D. Johnston, J. Berwick, D. Chen, S. Billings, and J. Mayhew, "A three-compartment model of the hemodynamic response and oxygen delivery to brain.," *NeuroImage*, vol. 28, no. 4, pp. 925–39, 2005.

[22] J. Tanabe, D. Miller, J. Tregellas, R. Freedman, and F. G. Meyer, "Comparison of detrending methods for optimal fMRI preprocessing," *NeuroImage*, vol. vol, pp. 15no4pp902–907, 2002.

[23] A. T. Smith, K. D. Singh, and J. H. Balsters, "A comment on the severity of the effects of non-white noise in fMRI time-series.," *NeuroImage*, vol. 36, no. 2, pp. 282–8, 2007.

[24] A. M. Smith, B. K. Lewis, U. E. Ruttimann, F. Q. Ye, T. M. Sinnwell, Y. Yang, J. H. Duyn, and J. A. Frank, "Investigation of Low Frequency Drift in fMRI Signal," vol. 533, pp. 526–533, 1999.

[25] K. J. Friston, W. Penny, C. Phillips, S. Kiebel, G. Hinton, and J. Ashburner, "Classical and Bayesian inference in neuroimaging: theory.," June 2002.

[26] Z. Hu, X. Zhao, H. Liu, and P. Shi, "Nonlinear Analysis of the BOLD Signal," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–14, 2009.

[27] R. M. Birn, Z. S. Saad, and P. a. Bandettini, "Spatial heterogeneity of the nonlinear dynamics in the FMRI BOLD response.," *NeuroImage*, vol. 14, pp. 817–26, Oct. 2001.

[28] T. D. Wager, A. Vazquez, L. Hernandez, and D. C. Noll, "Accounting for nonlinear BOLD effects in fMRI: parameter estimates and a model for prediction in rapid event-related studies.," *NeuroImage*, vol. 25, pp. 206–18, Mar. 2005.

[29] L. a. Johnston, E. Duff, I. Mareels, and G. F. Egan, "Nonlinear estimation of the BOLD signal.," *NeuroImage*, vol. 40, no. 2, pp. 504–14, 2008.

[30] V. a. Vakorin, O. O. Krakovska, R. Borowsky, and G. E. Sarty, "Inferring neural activity from BOLD signals through nonlinear optimization.," *NeuroImage*, vol. 38, pp. 248–60, Nov. 2007.

[31] K. J. Worsley, J. E. Taylor, F. Tomaiuolo, and J. Lerch, "Unified univariate and multivariate random field theory.," *NeuroImage*, vol. 23 Suppl 1, pp. S189–95, 2004.

[32] D. A. Hofmann, "An Overview of the Logic and Rationale of Hierarchical Linear Models," *Journal of Management*, vol. 23, no. 6, 1997.

[33] K. J. Friston, "Bayesian estimation of dynamical systems: an application to fMRI," *NeuroImage*, vol. 16, pp. 513–530, 2002.

[34] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. Cambridge, MA: MIT Press, 2005.

[35] T. Ozaki, "The Local Linearization Filter with Application to Nonlinear System Identifications," in *Proceedings of the first US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, (Dordrecht), pp. 217–240, Kluwer Academic Publishers, 1994.