Signal processing for magnetic resonance force microscopy

by

Michael Y. J. Ting

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2006

Doctoral Committee:

Professor Alfred O. Hero III, Chair
Professor Jeffrey A. Fessler
Professor Victor Solo
Associate Professor Douglas C. Noll
Dr. Daniel Rugar, International Business Machines

ACKNOWLEDGEMENTS

> Alice laughed: “There’s no use trying,” she said; “one can’t believe impossible things.”
>
> “I daresay you haven’t had much practice,” said the Queen. “When I was younger, I always did it for half an hour a day. Why, sometimes I’ve believed as many as six impossible things before breakfast.”
>
> — Lewis Carroll, Alice in Wonderland

The magnetic resonance force microscopy research effort is a fine example of the words of Georges Danton, “il nous faut de l’audace, encore de l’audace, toujours de l’audace. . . ” Being a part of the team that was involved in the quest for the single spin has been an enriching experience. At the University of Waterloo, where I received my undergraduate education, we were taught to be in harmony with the truth (Concordia cum veritate). The last five years has felt like a journey in its pursuit (Quaere verum). I am indebted to the people who have guided me in the right direction and those who have been an inspiration to me over the last several years.

This work would not have been possible without the support and encouragement of the following people.

University of Michigan Alfred O. Hero III, Jeffrey A. Fessler, Victor Solo, Douglas C. Noll, Cyrille Hory, Raviv Raich, and Chun-Yu Yip.

IBM Almaden Research Center Daniel Rugar, H. John Mamin, and Raffi Budakian.

University of Washington John A. Sidles.

There are scores of others I would like to acknowledge. Although they did not directly contribute to this thesis, it takes a village to raise a child. My parents, Joseph and Sharon Ting. EE: Systems colleagues, Raghuram Rangarajan, Derek Justice, Neal Patwari, Meng Fu Shih, Doron Blatt, Choon Yik Tang, and Kar Peo Yar. Cheers to all the Renaissance coopers and others who I've met and made great friends with. They constitute a long list; however, I shall attempt to mention several here. Gilles Castres Saint-Martin, Deborah van der Maas, Yustianto Tjiptowidjojo, Florian Becher, Regine Stoffel, David Winn, Irena Gershkovitch, Sébastien Cerbourg, Mohammed Elayan, Lauren Mitchell, Richard Tursky, el Mahdi Ahmed, and Hanh Pham. My sincere apologies to those I have omitted, but cheers all the same!

This work was supported in part by the DARPA Molecular Observation, Spectroscopy, and Imaging using Cantilevers (MOSAIC) program under ARO contract DAAD19-02-C-0055 (PI: D. Rugar, IBM), and by the ARO MURI Advancement of Magnetic Resonance Force Microscopy to Single Nuclear Spin Detection, funded by grant W911NF-05-1-0403 (PI: J. A. Sidles, Univ. of Washington). The ongoing collaboration with IBM is gratefully acknowledged.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

CHAPTER I

Introduction

1.1 Overview

In 1991, magnetic resonance force microscopy (MRFM) was conceived as a non-destructive method through which three dimensional images could be obtained with atomic (angstrom scale) resolution [59, 61]. The capability to directly look at individual atoms of a molecule is alluded to in Richard Feynman's classic 1959 talk, "There's Plenty of Room at the Bottom".

> We have friends in other fields—in biology, for instance. We physicists often look at them and say, "You know the reason you fellows are making so little progress?" (Actually I don't know any field where they are making more rapid progress than they are in biology today.) "You should use more mathematics, like we do." They could answer us—but they're polite, so I'll answer for them: "What you should do in order for us to make more rapid progress is to make the electron microscope 100 times better."

The lecture is a common reference mentioned to nanotechnology newcomers; not only does it mention the ability to look at very small objects, but it also mentions the miniaturization of machines and computers. It should be noted that the latter goal, i.e., the miniaturization of computers, has been achieved: computers in the

1950s were considerably larger.

If it were to be built, a MRFM microscope with angstrom scale resolution would represent an improvement over the scanning tunnelling microscope (STM), which, with a resolution of approximately 2 Å (1Å = 10^{-10} m), is the most powerful microscope currently available. There is, however, a significant advantage that a MRFM microscope would have over the latter. The STM is limited to imaging only the surface atoms of the sample. In contrast, a MRFM microscope would have the ability to look beneath the surface of the sample.

MRFM can be applied in various areas. The exploration of "chemical space" [11] will be greatly facilitated with the ability to obtain three dimensional images of biological molecules. A way to visualize the complicated interactions of these molecules in cells would aid the drug discovery process. Single spin MRFM can be applied to the creation of a high density storage device. Storage in a nutshell is the ability to write, retain, and read state information. In [5], the statistical fluctuation of an ensemble of electron spins was guided so that the aggregate polarity was positive. Although the creation of a MRFM storage device is still unrealizable, one can look to the IBM "millipede" project for inspiration [71]. The millipede project is an attempt to create a high density storage system based on principles of atomic force microscopy (AFM). AFM is a technology that pre-dates MRFM and from which MRFM borrows the idea of force sensing. Lastly, single nuclear spin MRFM can be applied to quantum computing [3].

Signal processing plays an essential role in the roadmap towards the realization of MRFM as an enabling technology for all of the applications previously mentioned. The first part of this thesis makes contributions to the areas of estimation and detection. A heuristic 0th order Extended Kalman Filter (EKF) for soft nonlinear systems that has less computational complexity than the standard EKF was developed. It showed better performance than EKF when applied to simulations of the classical model of the cantilever-electron spin interaction. This 0th order EKF can be applied to the estimation of other soft nonlinear systems.

The experiment in which the single electron spin was successfully detected was conducted under a SNR of -6.7 dB [54]. The detection of the single electron spin can be formulated as a binary detection test for a Markov signal in additive white Gaussian noise (AWGN). The optimal test for the aforementioned binary detection problem according to the Neyman-Pearson criterion was studied. Under low signal to noise ratio (SNR), an insightful interpretation was derived: the optimal test is approximately the matched filter statistic but with the one-step minimum mean-squared error (MMSE) predictor substituted for the known signal values. Approximations to the optimal test under the conditions of low SNR and long observation time were derived when the Markov signal was a random telegraph process and when it belonged to a certain class of random walk processes. These theoretical results confirm the optimality of the detection test used in [54]. The detection results have wide applicability to other fields where the detection of weak signals occurs, such as in landmine detection [20], in the study of particle tunnelling [7] and of fluctuations of the sun's magnetic field [74].

Medical imaging has revolutionized the field of medical diagnostics. The feat would not have been possible without the use of sophisticated imaging algorithms that were adapted to the imaging acquisition method and to the images of interest. If MRFM is to achieve its potential of being a looking glass through which one can study nanoscale structures, a similar development of imaging algorithms has to occur. The latter part of this thesis is a step in that direction. The salient feature of an atomic-level image is its sparsity. Most of the image would be empty space, and only a few spatial locations would be occupied by atoms. The premise taken by this work is that an image reconstruction algorithm that modelled the sparsity of the image can perform better. The other feature of an atomic-level image is its size. As the image is sparse, many voxels would be needed in order to obtain a good quality image. A useful image reconstruction method has to scale well.

This thesis proposes several sparse reconstruction methods that select the tuning parameters in a data-driven fashion. To address sparsity, the use of sparse priors and sparsifying penalty functions was employed. An example of the former is the weighted average of a Laplacian density with an atom at zero (LAZE); an example of the latter is the l_1 norm. The principle of selecting the tuning parameters in an empirical fashion is adhered to by employing marginal maximum likelihood (MML), maximum a posteriori (MAP), or Stein's unbiased risk estimate (SURE) [64]. Two of the proposed sparse reconstruction methods are extensions of existing sparse estimators. The first, known as EBD-LAZE, is an extension of the empirical Bayes denoising (EBD) method of [33]. Another is the L1 estimator with its regularization parameter selected via SURE. The proposed methods are more scalable and have lower computational complexity than sparse Bayesian learning (SBL), an existing method for solving sparse inverse problems. In a simulation study, L1-SURE and the MAP solution when used with the LAZE prior showed performance benefits over SBL. The proposed sparse reconstruction methods can potentially be used in other domains, like radioastronomy or sparse estimation of mixture models.

1.2 A short history of magnetic resonance force microscopy

The first MRFM experiment demonstrating the MRFM principle was reported in [55], and involved the detection of electronic spin resonance (ESR) in diphenylpicrylhydrazil (DPPH). Shortly thereafter, two DPPH particles were successfully imaged in [85]. The next step involved transferring the expertise gained in the ESR experiments to the detection of nuclear magnetic spins. As the "magnetic moment of common nuclei are at least 650 times smaller than the moment of the electron" [56], the detection of nuclear spins is, in general, more difficult. Nonetheless, detection of nuclear magnetic resonance (NMR) via the MRFM principle was successfully performed [56]. Paralleling the development of the ESR experiments, imaging of nuclear spins was later carried out [84]. Obtained in a span of several years in the early 1990s, these results were all positive and reinforced the feasibility of single spin detection.

However, the experiments up to this point had been performed with many millions of electron or nuclear spins. The estimated minimum number of protons required for detection in the experiment of [56] was 10^{13}, a number that is many orders of magnitude away from 1. In order to achieve single spin detection, the focus returned to electron spins. An improvement in the detectability of several orders of magnitude required countless improvements, among which was the fabrication of better cantilevers [66], a better understanding of the inversion and nutation of the electron spins [73], and spin relaxation effects [65].

In 2003, with the help of the "interrupted OSCAR" protocol, the detection threshold for ESR was lowered to about six electron spins [43]. This represented a tremendous improvement over previous experiments, and was arguably the turning point in the quest for the single spin. Although further improvements had to be made in going from six spins to the single spin, the big hurdle was going from millions of spins to six, and that was successfully cleared. The culmination of these efforts was the successful experiment at IBM in which individual electron spins associated with sub-surface atomic defects in silicon dioxide were detected [54]. This single spin detection milestone represented a factor of 10^7 improvement over conventional ESR detection techniques and was achieved using energy detection methods similar to those described in this thesis in Chapter IV. Other recent MRFM experiments have demonstrated the ability to detect and manipulate naturally occurring statistical fluctuations in small electron spin ensembles [5].

After having reached the single spin milestone, the current interest is to resolve multiple electron spins in a sample. Imaging can be performed with ESR detection: "it is standard laboratory practice to spin label proteins and DNA by attaching tempol and other compounds with unpaired electrons" [60]. However, the imaging of individual electron spins has not occurred. One would also like the single electron spin success to be replicated with nuclear spins. In this thesis, however, we consider only electron spin experiments, and so any reference to the word "spin" would implicitly mean an electron spin.

1.3 Outline of thesis

Chapter II reviews the basic principles of MRFM. The four single electron spin-cantilever models are introduced, and the single spin detection problem is formulated for each of the four models. The cantilever tip point spread function for a vertical cantilever is given.

Chapter III concerns the detection of the single spin for the continuous-time models. A heuristic argument led to the development of a 0th order Extended

Kalman Filter for the estimation of soft nonlinear systems, called the piecewise linear Kalman Filter (PLKF). It has a lower per iteration runtime than EKF. The idea that enabled its formulation was to treat the nonlinear system as a linear system in each sampling time interval.

The same principle was applied in adapting the KF detector to the detection of a soft nonlinear system vs. a linear system. The system governing the cantilever motion is linear in the case of the no spin (H_0) hypothesis. In contrast, under the spin (H_1) hypothesis, the system governing the cantilever motion is softly nonlinear. The KF detector is an optimal test if the two systems are linear and if the observations are zero mean. This last condition is approximately true under both the H_0 and H_1 hypotheses. Another issue that had to be dealt with was the lack of knowledge regarding the initial spin state. The Generalized Likelihood principle was applied, and the net result was called the KF/GLR innovations detector. It consists of one KF that is matched to the H_0 hypothesis, and several PLKF filters matched to the H_1 hypothesis. Simulations indicate that the PLKF outperformed the EKF, and that the KF/GLR innovations detector has good performance.

Chapter IV looks at the detection of the single spin for the discrete-time models. Three important results are contained here. Firstly, when used to detect the DT random telegraph in AWGN, the filtered energy (FE) detector is approximately optimal under the following four conditions: symmetric transition probabilities, low SNR, long observation time, and a small probability of transition between two consecutive instances. The FE detector is no longer approximately optimal when the transition probabilities are asymmetric. We extend the FE detector to a hybrid second-order detector that combines the filtered energy, amplitude, and energy statistics. It is shown that the hybrid detector is approximately optimal for the DT random telegraph model under only the last three conditions.

The second result of this chapter is a new interpretation of the optimal likelihood ratio test (LRT) for a DT finite state Markov signal under low SNR conditions. It is shown that, under low SNR, the LRT reduces to the matched filter statistic with the one-step MMSE predictor used in place of the known signal values. Single spin experiments operate under conditions of very low SNR; consequently, we are interested in the performance of detectors in the regime of low SNR and long observation time. Thirdly, necessary conditions are derived for the LRT of a certain class of random walks to be approximated by a bank of FE statistic generators, and by a single FE statistic.

Chapter V contains results regarding the reconstruction of sparse images. The problem involves simultaneous deconvolution and denoising. We propose several sparse reconstruction methods that select the tuning parameters in a data-driven fashion, i.e., without manual tuning. This empirical philosophy manifests itself in two ways. The first is to select a sparsifying prior, viz., a weighted average of a Laplacian density and an atom at zero, and estimate its unknown parameters empirically, e.g., using marginal maximum likelihood. This gave rise to three different sparse estimators: EBD-LAZE, MAP1, and MAP2. Note that MAP2 is unlike the others: it has a parameter that requires manual tuning. The second way the empirical philosophy is adhered to is by selecting a sparsifying penalty, e.g., the l_1 norm, and estimating the regularization parameter via Stein's unbiased risk estimator (SURE). This produced the L1-SURE and HHS-SURE estimators.

The scalability and computational complexity of the proposed sparse reconstruction methods was examined and compared with SBL. The computational complexity of the estimators can grouped in decreasing order as: SBL; EBD-LAZE and HHS-

SURE; L1-SURE, MAP1, and MAP2. The scalability of the methods in each group is roughly comparable.

A simulation study was conducted to compare: the proposed reconstruction methods; SBL; the standard and projected Landweber iteration (where the projection is on to the positive orthant). SBL had a lower l_1 and l_2 reconstruction error under high SNR when tested with an image that had both positive and negative pixel values. In the other scenarios, SBL had worse performance. It was observed that SBL did not produce a sparse estimate. The Landweber iteration, which produces the least-squares solution, was not competitive. Among the proposed reconstruction methods, L1-SURE had consistently good performance under the various error criteria considered. HHS-SURE, which can be regarded as a generalization of L1-SURE, achieved approximately similar l_1 and l_2 reconstruction errors but with a sparser estimate. MAP1 and MAP2 had good performance under low SNR, but their performance worsened under higher SNR. EBD-LAZE had performance that was generally worse than L1-SURE's, except for the detection error criterion. Finally, the projected Landweber method had good results for high SNR when the image was non-negative.

Note that the meaning of variables and the notation used are unique to each chapter. An appendix adheres to the conventions of the chapter to which it is attached.

1.4 List of publications

The publications that resulted from the work contained in this thesis are as follows:

- M. Ting and A. O. Hero. Detection of an electron spin in a MRFM cantilever experiment. In *Proc. of the IEEE Workshop on Stat. Sig. Proc.*, 2003.

- A. O. Hero, M. Ting, and J. A. Fessler. Two state Markov modelling and detection of single electron spin signals. In *Proc. of the XII European Sig. Proc. Conf.*, 2004.

- M. Ting, A. O. Hero, D. Rugar, C.-Y. Yip, and J. A. Fessler. Near optimal signal detection for finite state Markov signals with application to magnetic resonance force microscopy. *IEEE Trans. Sig. Proc.*, to appear in June 2006 issue.

- M. Ting and A. O. Hero. Detection of a random walk signal in the regime of low signal to noise ratio and long observation time. To appear in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech, and Sig. Proc.*, 2006.

- M. Ting, R. Raich, and A. O. Hero. Sparse image reconstruction using sparse priors. To appear in *Proc. of the IEEE Intl. Conf. on Image Proc.*, 2006.

CHAPTER II

Magnetic resonance force microscopy

Magnetic resonance force microscopy (MRFM) was conceived as a non-destructive method through which 3-d images with atomic (angstrom scale) resolution could be obtained. In this chapter, we shall provide a description of the MRFM setup used in the IBM experiment [54]. This is followed by a discussion of the various signal models used to model the single electron spin-cantilever interaction. Finally, we provide the point spread function of a vertical cantilever tip.

A general overview of MRFM can be obtained in [60, 24].

2.1 Description of the MRFM experiment

MRFM experiments, in general, involve the measurement of magnetic force between a submicron-size magnetic tip and spins in a sample. The details of spin manipulation and signal detection depend on the exact MRFM protocol used. One particularly successful protocol is called OSCAR, which stands for OScillating Cantilever-driven Adiabatic Reversal [65, 43]. A variation of this protocol, known as "interrupted OSCAR" (iOSCAR), was an important idea that enabled the detection of small spin ensembles [43] and was used in the single spin experiments [54].

A schematic diagram of an OSCAR-type MRFM experiment is shown in Fig. 2.1. As shown in the figure, a submicron ferromagnet is placed on the tip of a cantilever and positioned close to an unpaired electron spin contained within the sample. An applied radio-frequency (rf) field serves to induce magnetic resonance of the spin when the rf field frequency matches the Larmor frequency. Because the magnetic field emanating from the tip is highly inhomogeneous, magnetic resonance is spatially confined to a thin bowl-shaped region called the "resonant slice".

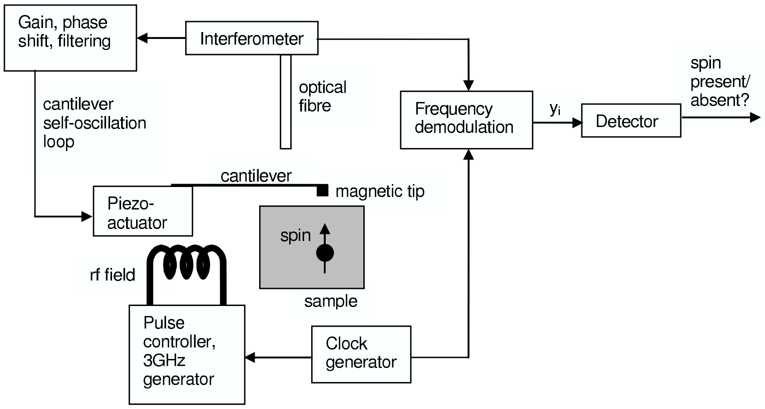


Figure 2.1: Schematic of an OSCAR-type MRFM experiment

If the cantilever is forced into mechanical oscillation by positive feedback, the tip motion will cause the position of the resonant slice to oscillate. As the slice passes back and forth through an electron spin in the sample, the spin direction will be cyclically inverted due to an effect called adiabatic rapid passage [73]. The cyclic inversion is synchronous with the cantilever motion and affects the cantilever dynamics by changing the effective stiffness of the cantilever. Therefore, the spin-cantilever interaction can be detected by measuring small shifts in the period of the cantilever oscillation. This methodology has been successfully used to detect small ensembles of electron spins [43, 65], and even a single spin [54].

We briefly review the single spin-cantilever interaction framework proposed by Rugar et al. [53] and Berman et al. [4]. Consider an electron spin in a coordinate frame that rotates at the frequency of the applied rf magnetic field $\underline{B}_1(t)$; see Fig. 2.2.

The effective magnetic field $\underline{B}_{\text{eff}}(t)$ in this rotating frame is given by

$$\underline{B}_{\text{eff}}(t) = B_1(t)\underline{i} + \Delta B_0(t)\underline{k}, \tag{2.1}$$

where $\underline{i}$ and $\underline{k}$ are the unit vectors in the x' and z directions of the rotating frame, $B_1(t)$ is the amplitude of the rf magnetic field, $B_0(t)$ is the amplitude of the tip magnetic field at the spin location, and $\Delta B_0(t) = B_0(t) - \omega_{\text{rf}}/\gamma$ is the off-resonance field amplitude. The constant $\gamma = 5.6\pi \times 10^{10}\,\text{s}^{-1}\text{T}^{-1}$ is the gyromagnetic ratio, and ω_{rf} is the frequency of the applied rf. In Fig. 2.2 below, $\underline{B}_0 = B_0\underline{k}$, i.e., it is aligned in the z direction, and $\Delta\underline{B}_0(t) = \Delta B_0(t)\underline{k}$.

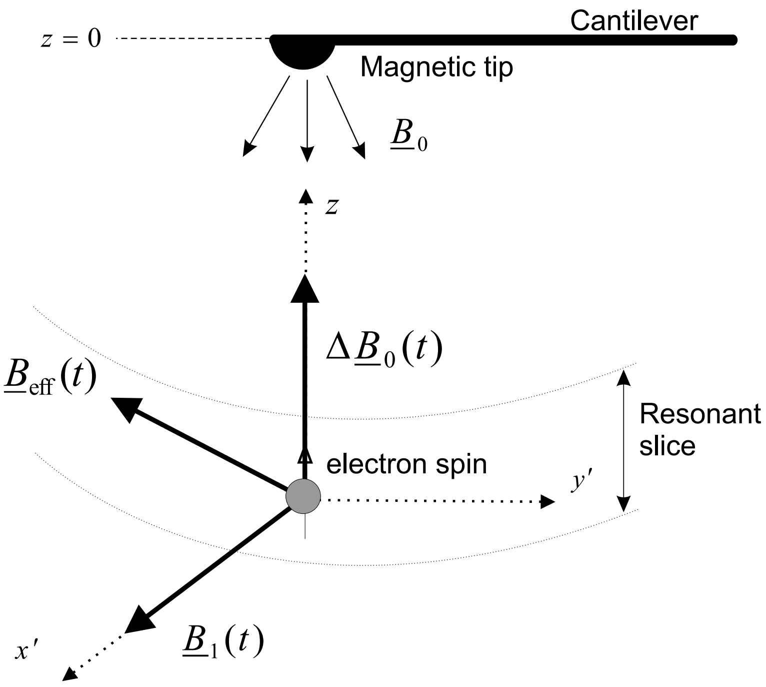


Figure 2.2: In the coordinate system rotating at ω_{rf}, the off-resonance field $\Delta\underline{B}_0(t)$, and therefore the effective field $\underline{B}_{\text{eff}}(t)$, oscillate synchronously with the cantilever. Under the spin-lock assumption, the electron spin aligns with $\underline{B}_{\text{eff}}(t)$; under the anti-spin-lock assumption, the electron spin aligns with $-\underline{B}_{\text{eff}}(t)$

The spins for which ω_{rf} approximately equals the Larmor frequency $\omega_L = \gamma B_0(t)$ are said to be in magnetic resonance. This condition defines a paraboloid-shaped slice under the cantilever in which the spins are in resonance. As the cantilever moves, so does this resonant slice. If $\Delta B_0(t)$ varies sufficiently slowly such that the adiabatic criterion

$$\frac{d\Delta B_0(t)}{dt} \ll \gamma B_1^2(t) \tag{2.2}$$

is satisfied, the spin can be assumed to remain aligned with either $\underline{B}_{\text{eff}}(t)$ or $-\underline{B}_{\text{eff}}(t)$. These are the *spin-lock* and *anti-spin-lock* conditions, respectively. Define $\underline{\mu}(t) \triangleq [\mu_x(t), \mu_y(t), \mu_z(t)]^T$ to be the electron spin moment, where the superscript $(\cdot)^T$ denotes the transpose operation. It is known that $\mu \triangleq \|\underline{\mu}\| = 9.28 \times 10^{-24}$J/T. Under the assumption that the spin is aligned with $\underline{B}_{\text{eff}}$, the z component of the field is given by [53]:

$$\mu_z(t) = \mu \frac{\Delta B_0(t)}{[(\Delta B_0(t))^2 + B_1^2]^{1/2}}. \tag{2.3}$$

Supposing that the motion of the cantilever is approximately sinusoidal, the off-resonance field amplitude can be written as $\Delta B_0(t) = B_{\text{mod}} \sin(\omega_{\text{mod}} t)$. If $\Delta B_0(t) \gg B_1$, (2.3) results in $\mu_z(t) \approx \mu \cdot \text{sgn}(B_0(t))$, where $\text{sgn}(\cdot)$ is the sign function, i.e., $\text{sgn}(x) = 1, x > 0$; $\text{sgn}(x) = -1, x < 0$; and zero otherwise.

Under the iOSCAR protocol, the rf signal $\underline{B}_1(t)$ is turned off after every N_{skip} cantilever cycles over a half-cycle duration to induce periodic transitions between the spin-lock and anti-spin-lock states. Let $\omega_{\text{skip}} \triangleq \omega_{\text{rf}}/N_{\text{skip}}$ be the frequency of the "off" pulses. Under the anti-spin-lock state and $\Delta B_0(t) \gg B_1$, $\mu_z(t) \approx -\mu \cdot \text{sgn}(B_0(t))$. In either spin-lock or anti-spin-lock state, the spin is said to be in cyclic adiabatic inversion (CAI). Under CAI, a rf signal that turns off with frequency ω_{skip} results in a $\mu_z(t)$ that is a square wave with frequency $\approx 2\omega_{\text{skip}}$.

2.2 MRFM single spin-cantilever signal models

Four single spin-cantilever signal models will be discussed in this section, and the detection problem formulated for each. Note that the notation used for the CT and

DT models are different. This difference in nomenclature will also manifest itself in the successive chapters to follow.

2.2.1 Continuous-time classical model (CTC)

The equations of the classical dynamics of a MRFM cantilever interacting with a single electron spin moment are described in [53]. Considering only the fundamental mode and ignoring the positive feedback term, the interaction is described by:

$$\begin{aligned}\Sigma_1 : \qquad \dot{\mu}_x &= \gamma\mu_y(Gz + \delta B_0)\\ \dot{\mu}_y &= \gamma\mu_z B_1(t) - \gamma\mu_x(Gz + \delta B_0)\\ \dot{\mu}_z &= -\gamma\mu_y B_1(t)\\ m\ddot{z} + \Gamma\dot{z} + kz &= G\mu_z + F_n(t)\end{aligned} \tag{2.4}$$

where $z(t)$ is the position of the cantilever ($z = 0$ is taken to be the equilibrium position); m is the cantilever's effective mass; k is the cantilever spring constant; Γ is a friction coefficient that is related to the cantilever quality factor; G is the magnetic field gradient; and δB_0 is whatever offset field that may be present. An overhead dot represents differentiation with respect to time. Recall from before that $B_1(t)$ is the rf signal, and $F_n(t)$ is AWGN which arises due to various noise sources in the experiment, e.g., background thermal noise.

The above equations omit the effect of the higher-order modes of the cantilever. This effect can be accommodated by adding more second order equations similar to the last equation in (2.4), and with $z_i, i = 2, 3, \ldots$ used in the i-th additional equation in place of z. Each additional 2nd order equation would have a different noise term $F_{ni}(t)$, and the z appearing in the first three equations of (2.4) will be replaced by $z + z_2 + \ldots + z_n$, where n is the number of cantilever modes considered. Note that $G \neq 0$, so that when a spin is present, $G\mu_z$ affects the dynamics of $z(t)$, and (2.4) is a system of nonlinear differential equations.

On the other hand, when a spin is not present, the $G\mu_z$ term vanishes, and we are left with the standard equation of motion for a cantilever, which is:

$$\Sigma_0 : \qquad m\ddot{z} + \Gamma\dot{z} + kz = F_n(t) \tag{2.5}$$

The observable output of the system are samples of the cantilever position $z(t)$ corrupted by observation noise, which is assumed to be AWGN. Define $t_i \triangleq iT_s$ to be the time instants at which $z(t)$ is sampled, where T_s is the sampling interval. Model the observation noise as w_i, where w_i is a sequence of independent and identically distributed (i.i.d.) Gaussian random variables (r.v.s) with zero mean and variance σ^2. Denote the observation sample at time t_i by y_i. Then $y_i = z(t_i) + w_i$. The detection problem for this signal model is as follows. Given the noisy observations $\underline{y} = [y_0, \ldots, y_{N-1}]^T$, classify the system that generated $\underline{y}$ as either:

$$\begin{aligned} H_0 \text{ (spin absent)} &: \quad \underline{y} \text{ generated by } \Sigma_0 \\ H_1 \text{ (spin present)} &: \quad \underline{y} \text{ generated by } \Sigma_1 \end{aligned} \tag{2.6}$$

2.2.2 Continuous-time random telegraph model (CTRT)

In [4], the classical CT model is used to obtain a simpler set of equations to describe the spin-cantilever interaction assuming that the CAI condition (2.2) holds. A perturbation analysis shows that the cantilever position can be described by:

$$m\ddot{z}(t) + \Gamma\dot{z}(t) + (k + \Delta k)z(t) = F_n(t) \tag{2.7}$$

Here, $\Delta k = -\mu G^2/|B_1|$. Note that the cantilever's natural mechanical resonance frequency is $\omega_0 = \sqrt{k/m}$. The shift in the spring constant results in a corresponding shift in ω_0 that is approximately given by

$$\omega_{0,s} = 2\mu \frac{\omega_0 G}{\pi k z_{\text{pk}}} \cos\theta \tag{2.8}$$

where we define $\omega_{0,s}$ to be the shift in ω_0, and z_{pk} is the peak amplitude of the cantilever vibration. The factor $\cos\theta$ represents the normalized projection of the spin in the direction of the effective field.

Under the iOSCAR protocol, ω_0 alternates between the two values $\omega_0 \pm \omega_{0,s}$ with frequency $2\omega_{\text{skip}}$ [43]. Defining $\Delta\omega_0(t) \triangleq \omega_0(t) - \omega_0$, we can equivalently say that $\Delta\omega_0(t)$ alternates between $\pm\omega_{0,s}$. By setting $F_n(t) = 0$ and ignoring the amplitude decay of z, the solution to (2.7) can be approximated as a frequency-modulated signal:

$$z(t) = Z_0 \cos\left[\omega_0 t + \int_0^t s(\xi) d\xi + \theta\right] \tag{2.9}$$

where Z_0 is the cantilever oscillation magnitude, θ is a random phase, and $s(t)$ is a square wave that is periodic with non-zero amplitude $\omega_{0,s}$ if a spin is present and zero amplitude otherwise. Thus, spin coupling, i.e., the presence of a spin, can be detected by frequency demodulating $z(t)$ to baseband and correlating the baseband signal with a known square wave signal derived from $B_1(t)$. Alternatively, we could frequency demodulate $z(t)$ to baseband and apply an energy test. Under the spin present (H_1) hypothesis, there exists a periodic square wave $s(t)$ whereas under the spin absent (H_0) hypothesis, one gets a zero signal. Consequently, the energy statistic in the presence of a spin should be higher. An additional improvement can be made: as the frequency of $s(t)$ is known to be $2\omega_{\text{skip}}$, one could filter the frequency demodulated $z(t)$ before applying the energy test. Since $s(t)$ is a lowpass signal, a natural filter to use would be a lowpass filter (LPF) with a -3 dB frequency of $2\omega_{\text{skip}}$.

Unfortunately, the effects of random thermal noise and spin relaxation decorrelate $s(t)$ and the square wave signal reference. The latter signal refers to the expected behaviour of $s(t)$ under the H_1 hypothesis. One model for this decoherence phenomenon is suggested by the Stern-Gerlach experiment [9]: the spins maintain either the spin-lock or anti-spin-lock states, but randomly change polarity during the course of the measurement. This leads to random transitions of $\Delta\omega_0(t)$ between $\pm\omega_{0,s}$; the transition times are assumed to be distributed according to a Poisson process with a rate of λ spin reversals/sec. Correlating the frequency demodulator output with the known square wave signal, as was described in the previous paragraph, has the effect of cancelling out the deterministic transitions in ω_0. What remains after correlation are the random transitions, and as the random transition times are generated by a Poisson process, the resultant signal takes the form of a so-called random telegraph process [63]. See Fig. 2.3.

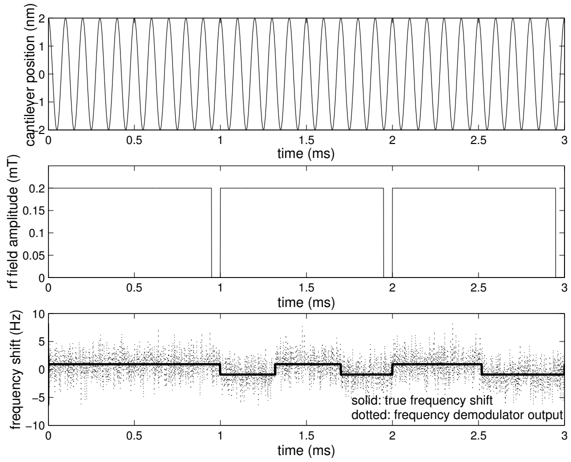


Figure 2.3: Top: Sample of an ideal cantilever position signal. Frequency shifts are not detectable by the eye. Middle: Amplitude of sample rf magnetic field, $B_1(t)$. It has synchronous half-cycle skips at 1 ms, 2 ms, and 3 ms for the creation of spin state transitions. Bottom: Ideal and noisy outputs of frequency demodulator under the spin present hypothesis. It has both deterministic transitions due to the rf skips at 1 ms, 2 ms and 3 ms, and random ones due to spin relaxation. The random transitions, τ_i, occur as a Poisson process. The initial polarity of the random telegraph is +1 for this example.

Let the baseband output of the frequency demodulator and correlator be denoted by $y(t)$. Let $[0, T]$ be the total measurement time period over which the correlator integrates the measurements, and let $\underline{\tau} = [\tau_1, \ldots, \tau_N]^T$, be the time instants within this period at which random spin reversals occur. As $\underline{\tau}$ are the arrival times of a Poisson process with intensity λ, N is a Poisson random variable with rate λT. Thus, the CT random telegraph model is: $y(t) = s(t) + w(t)$ where $w(t)$ is AWGN with variance σ^2, and $s(t)$ is a random telegraph signal containing only the random transitions. That is,

$$s(t) = \phi |w_{0,s}| \sum_{i=0}^{N} (-1)^i g\left(\frac{t - \tau_i}{\tau_{i+1} - \tau_i}\right) \tag{2.10}$$

where ϕ is a random variable that equals ±1 with equal probability (it represents the initial polarity of the spin), $\tau_0 = 0$, $\tau_{N+1} = T$, and $g(\cdot)$ is the unit rectangular function, i.e., $g(t) = 1$, $t \in [0, 1]$ and 0 otherwise [82].

The detection problem for the CT random telegraph model is to design a test between the two hypotheses:

$$\begin{aligned} H_0 \text{ (spin absent)} &: \quad y(t) = w(t) \\ H_1 \text{ (spin present)} &: \quad y(t) = s(t) + w(t) \end{aligned} \tag{2.11}$$

for $t \in [0, T]$.

2.2.3 Discrete-time random telegraph model (DTRT)

In the quantum measurement model, the frequency shift is characterized by random jumps between two discrete levels. The jumps are taken to be Poisson distributed. Suppose that the CTRT is sampled at times $t_i = iT_s$, where T_s is the sampling time interval. The result is a DT random telegraph signal, which we shall denote by X_i. In this paper, a Markovian process with a finite number of states will have a state space denoted by $\Psi = \{\psi_1, \ldots, \psi_r\}$, where r is the number of states. Let the state space of the DT random telegraph be Ψ_{rt}; it has $r = 2$ states and we shall take $\psi_1 = -A$, $\psi_2 = A$, where A is the amplitude of the random telegraph (A corresponds to $\omega_{0,s}$ for the case of a MRFM signal). As an initial condition, X_0 is equally likely to be either $\pm A$. A probability transition matrix $\mathbf{P}_{\text{rt}}$ can be associated with X_i such that the (j,k)-th value of $\mathbf{P}_{\text{rt}}$ equals $P(X_i = \psi_k | X_{i-1} = \psi_j)$ for $1 \leq j, k \leq 2$ and $i \geq 1$. Assume that $\mathbf{P}_{\text{rt}}$ has the form:

$$\mathbf{P}_{\text{rt}} = \begin{pmatrix} q & 1-q \\ 1-p & p \end{pmatrix}, \tag{2.12}$$

where $0 < p, q < 1$. If $p = q$, we say that the transition probabilities are *symmetric*, whereas if $p \neq q$, we shall say that they are *asymmetric*. Define the signal vector $\underline{x} = [x_0, \ldots, x_{N-1}]^T$, the noise vector $\underline{w} = [w_0, \ldots, w_{N-1}]^T$, and the observation vector $\underline{z} = [z_0, \ldots, z_{N-1}]^T$. The w_i values are modelled as i.i.d. Gaussian r.v.s with zero mean and variance σ^2. The detection problem is then to decide between:

$$\begin{aligned} H_0 \text{ (spin absent)} &: \quad \underline{z} = \underline{w} \\ H_1 \text{ (spin present)} &: \quad \underline{z} = \underline{x} + \underline{w} \end{aligned} \tag{2.13}$$

Examples of noiseless and noisy random telegraph signals are given in Fig. 2.4. For the random telegraph signal, the SNR is defined to be $\text{SNR} \triangleq A^2/\sigma^2$. The SNR in dB is defined in the usual way as $\text{SNR}_{\text{dB}} \triangleq 10 \log_{10} \text{SNR}$.

2.2.4 Discrete-time random walk model (DTRW)

In the classical spin detection model, the frequency shift signal is well approximated by a one dimensional random walk confined to the interval $I = [-A, A]$, where $A = \omega_{0,s}$ for the case of a MRFM signal. We discretize I into $(2M+1)$ states using a step size of s, where $M \in \mathbb{Z}$ and $M, s > 0$ and define X_i to be the random walk restricted to the discretized I. This model will be referred to as the DT random walk model. The state space Ψ_{rw} of the DT random walk will then have $r = 2M + 1$ states, where $\psi_j = (j - M - 1)s$ for $j = 1, \ldots, (2M+1)$. Associate with X_i the probability transition matrix $\mathbf{P}_{\text{rw}}$, so that, as before, the (j,k)-th element of $\mathbf{P}_{\text{rw}}$ is

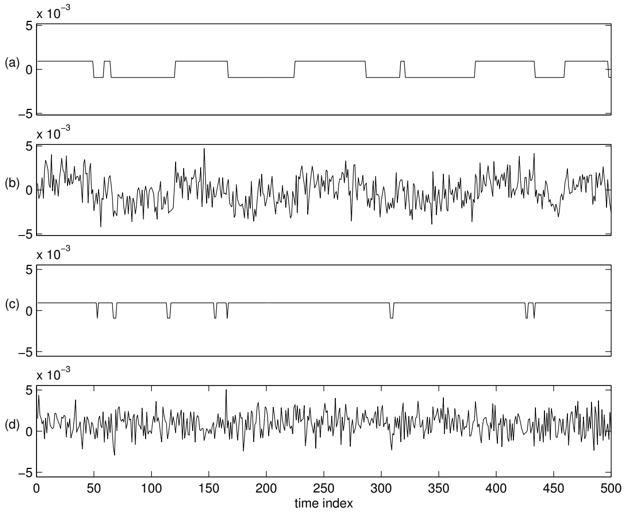


Figure 2.4: a: Noiseless random telegraph signal with symmetric transition probabilities $p = q = 0.98$; b: Noisy version of (a) at SNR $= -3$ dB; c: Noiseless random telegraph signal with asymmetric transition probabilities $p = 0.98, q = 0.6$; d: Noisy version of (c) at SNR $= -3$ dB.

$P(X_i = \psi_k | X_{i-1} = \psi_j)$ for $1 \leq j, k \leq (2M + 1)$ and $i \geq 1$. $\mathbf{P}_{\text{rw}}$ is defined such that, at each time step, X_i changes by either $\pm s$. We assume reflecting boundary conditions, and X_0 is equally likely to be either $\pm s$. These conditions imply that $\mathbf{P}_{\text{rw}}$ is a tridiagonal matrix.

The detection problem is now to test (2.13) when $\underline{x}$ is modelled by a random walk. The DT random walk process is not a multi-state generalization of the DT random telegraph process. In the former, X_i cannot stay in the same state for two consecutive time instants. As well, the random walk has reflecting boundary conditions, which the random telegraph does not have. In the limit as $s \to 0, M \to \infty$, the random walk converges to Brownian Motion over the interval I [63].

Examples of noiseless and noisy random walk signals are given in Figs. 2.5 and 2.6, where, at each state, a change of $\pm s$ is equally likely. Define the SNR for a random walk process X_i as SNR $\triangleq (\lim_{i\to\infty} E[X_i^2])/\sigma^2$. In other words, the SNR is the ratio of the steady-state expected energy of X_i to the noise variance. This definition is consistent with that provided for the random telegraph signal. If X_i represented the DT random telegraph signal, then $X_i^2 = A^2$ at all time instances, leading to $\mathrm{SNR} = A^2/\sigma^2$.

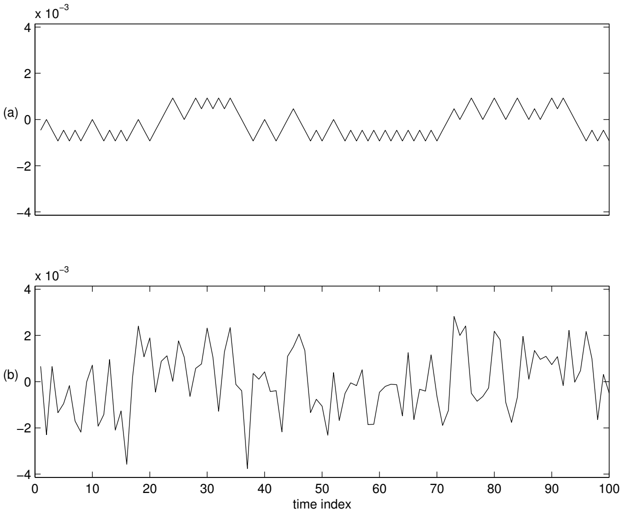


Figure 2.5: a: Noiseless random walk signal with 5 levels; b: Noisy version of (a) at SNR = −7.3 dB.

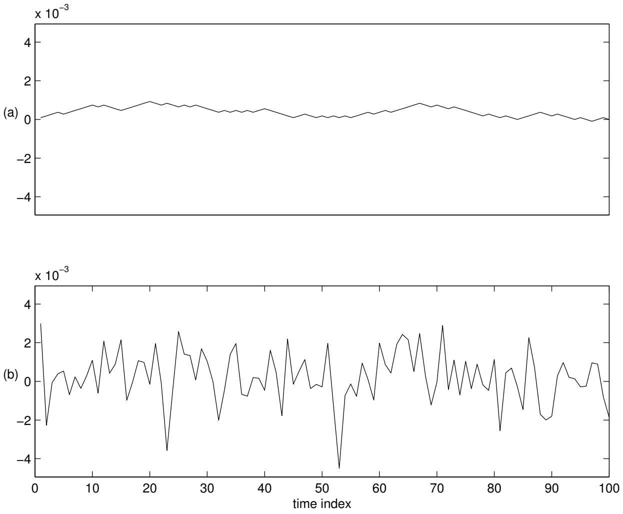


Figure 2.6: a: Noiseless random walk signal with 21 levels; b: Noisy version of (a) at SNR = −7.8 dB.

2.3 MRFM tip point spread function

In [44], the point spread function (psf) of a MRFM tip is derived under the following assumptions.

1. The tip can be modelled as a point dipole.
2. The spins are undergoing CAI.
3. There is no spin-spin coupling.
4. Energy-based measurements are taken.

Although Fig. 2.2 shows a horizontal cantilever which vibrates in the z direction, current experiments use a vertical cantilever vibrating in the x direction. A horizontal cantilever cannot vibrate too close to the sample surface; otherwise, van der Waals and electrostatic forces will draw the tip onto the surface and break the cantilever [66]. The psf for a vertical tip is given by

$$H(x,y,z) = \begin{cases} \left(\frac{G(x,y,z)}{G_0}\right)^2 \left(1 - \left[\frac{s(x,y,z)}{x_{\text{pk}}}\right]^2\right) & |s(x,y,z)| \le x_{\text{pk}} \\ 0 & |s(x,y,z)| > x_{\text{pk}} \end{cases} \tag{2.14}$$

where x_{pk} is the peak amplitude of the cantilever in the x direction and G_0 is a normalizing constant [44]. Let $r = \sqrt{x^2+y^2+z^2}$; then

$$s(x,y,z) = \frac{B_{\text{res}} - B_{\text{mag}}(x,y,z)}{G(x,y,z)}, \tag{2.15}$$

$$B_{\text{mag}}(x,y,z) = \sqrt{\left(\frac{3xzm}{r^5}\right)^2 + \left(\frac{3yzm}{r^5}\right)^2 + \left(\frac{m(2z^2-x^2-y^2)}{r^5} + B_{\text{ext}}\right)^2}, \tag{2.16}$$

and $G = \frac{\partial}{\partial x} B_{\text{mag}}$, which is

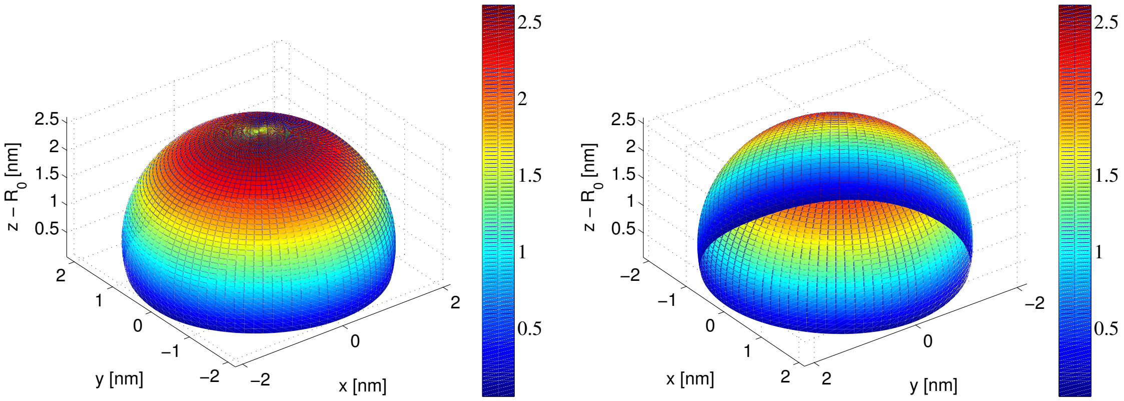
$$G(x,y,z) = \frac{1}{2B_{\text{mag}}(x,y,z)}\Bigg(-\frac{90m^2x^3z^2}{r^{12}} - \frac{90m^2xy^2z^2}{r^{12}} + \frac{18m^2xz^2}{r^{10}} \\ + 2\left[-\frac{2mx}{r^5} - \frac{5mx(-x^2-y^2+2z^2)}{r^7}\right]\left[B_{\text{ext}} + \frac{m(-x^2-y^2+2z^2)}{r^5}\right]\Bigg) \tag{2.17}$$

We shall use the parameter set listed in Table 2.1 to illustrate the MRFM psf. The plot of the resonant slice, defined by $B_{\text{mag}}(x, y, z) = B_{\text{res}}$ for $z \geq R_0$, is given in Fig. 2.7. The resonant slice is a bowl-shaped surface of non-zero thickness. Here, it is shown upside-down: a positive z value indicates a position below the cantilever tip.

Table 2.1: Parameters used to illustrate the MRFM psf.

Parameter		**Value**
Description	Name	
Amplitude of external magnetic field	B_{ext}	2×10^4 G
Value of B_{mag} in the resonant slice	B_{res}	2.25×10^4 G
Radius of tip when modelled as a sphere	R_0	2 nm
Distance from tip to sample	d	2 nm
Cantilever tip moment[†]	m	5.70×10^4 emu
Peak cantilever swing	x_{pk}	0.033 nm
Maximum magnetic field gradient[‡]	$G_{\max}$	610 G/nm

[†] Assuming a spherical tip.
[‡] Assuming optimal sample position.



(a) View from top

(b) View from bottom

Figure 2.7: Plot of surface $B_{\text{mag}}(x, y, z) = B_{\text{res}}$.

A 3-d contour plot of the normalized MRFM psf $H(x, y, z)$ is illustrated in Fig. 2.8 for the parameters listed in Table 2.1. The important thing to notice is that there is a slice of the yz plane where the response of the psf is zero. This is because the

x component of the gradient vector ∇B_{mag} is zero due to the symmetry of B_{mag}. A transverse xy slice of the MRFM psf is illustrated in Fig. 2.9. The gradient G is highest in a xz plane that slices through the middle of the bowl shaped B_{mag}. This is indeed the case in Fig. 2.9.

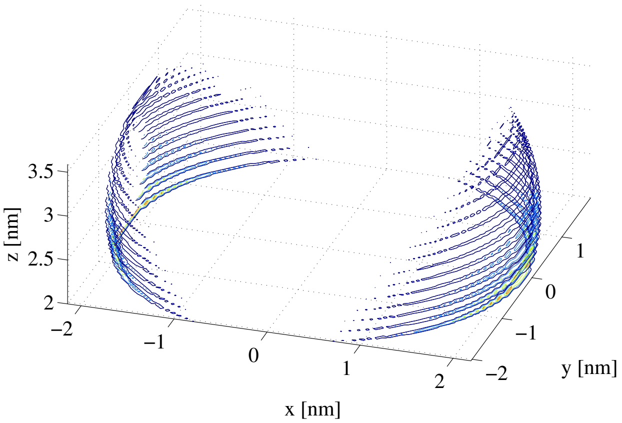


Figure 2.8: Three dimensional contour plot of MRFM psf.

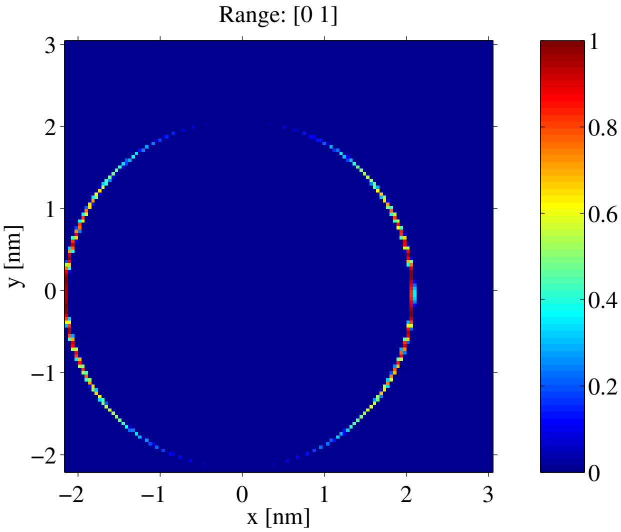


Figure 2.9: Transverse xy slice of the MRFM psf.

2.4 Conclusion

A description of the MRFM experiment was given, as well as an introduction to the single spin-cantilever interaction. Next, four models of the single spin-cantilever were described. The first two are continuous-time models, while the last two are discrete-time. The success of the single electron spin experiment lends strong support to the random telegraph model, as the detection algorithm used therein was based on the DT random telegraph process. We also give equations for the vertical cantilever tip psf.

CHAPTER III

Detection of the single spin in the continuous-time models

Detection of the single spin for the two continuous-time (CT) single spin-cantilever models are examined in this chapter. The main focus shall be on the CT classical (CTC) model. Results regarding the CT random telegraph (CTRT) model will be briefly mentioned.

3.1 Spin detection for the continuous-time random telegraph model

The development of the CTRT model in Chapter II suggests that almost all of the information pertaining to the presence or absence of a spin is contained in the frequency content of the cantilever position signal $z(t)$. Indeed, in the presence of a spin, we saw that $z(t)$, after being frequency demodulated and translated to baseband, consists of a deterministic, periodic square wave and a random signal component. In the absence of any randomness, optimal detection can be performed using a matched filter detector. When a random signal component is present, the deterministic part can be cancelled out and we are left with the detection of a random signal in AWGN. Detection methods based on the CTRT model assume that the approximation analysis used to show (2.7) holds.

3.1.1 Background

The form of the likelihood ratio (LR) for the detection of a CT random process in AWGN is established in [34, 35]. Its implementation requires the conditional mean estimate (CME) $\hat{s}(t) = E_1[s(t)|\{y(\xi) : \xi < t\}]$, where the subscript "1" denotes the expectation under the spin present (H_1) hypothesis. In particular, given the framework of (2.11), the LR is

$$\mathrm{LR} = \exp\left(\frac{1}{\sigma^2}\int_I \hat{s}(t)y(t)\,dt - \frac{1}{2\sigma^2}\int_I \hat{s}^2(t)\,dt\right) \tag{3.1}$$

where $\hat{s}(t)$ was previous defined, $I \triangleq [0,T]$, and the first integral is an Itô stochastic integral [34].

In [78], the filtering equations for obtaining $\hat{s}(t)$ for the CTRT are given. Therefore, it is possible to implement the optimal solution, as the likelihood ratio test (LRT) is optimal in the Neyman-Pearson sense. That is, the LR maximizes the probability of detection (P_D) subject to a constraint on the probability of false alarm (P_F) [70]. The solution is not finite-dimensional however; it has high computational complexity. The estimation of a CT random telegraph in AWGN was addressed in [80]: the performance of optimal filtering vs. smoothing was studied. An interesting result is that as the SNR approaches 0, linear estimates are asymptotically as efficient as the nonlinear estimates [80].

If only samples of the observation $y(t)$ are available, as in (2.6), then the LRT given in (3.1) is not directly applicable. However, it might be possible to approximate the integrals with sums. This would necessitate finding the CME $E_1[s(t)|\{y(t_i) : i \text{ s.t. } y_i < t\}]$. The exact solution is known [29, 16]; it too has high computational complexity. An approximation can be made if the stochastic differential equation is assumed to be time-invariant [42]. Unfortunately, system Σ_1 given in (2.4) is time-varying due to the presence of the rf field $B_1(t)$. One can approximately solve the partial differential equations governing the evolution of the posterior density using Galerkin's approximation when the state vector is one-dimensional [23]. The technique requires the usage of basis functions that can well approximate the posterior density. In the one-dimensional case, the class of complex exponentials was used, leading to an efficient implementation that used the Fast Fourier Transform (FFT). It is not clear what a suitable basis function set would be in higher dimensions. This would be needed to solve for Σ_1, as its state vector is in $\mathbb{R}^5$.

The drawback of the optimal solutions mentioned above is that they are not finite dimensional. This leads to a search for a suboptimal detector with a lower computational complexity.

3.1.2 Hybrid Bayes/Generalized likelihood Ratio detector

A hybrid Bayes/Generalized likelihood ratio (GLR) detector was developed in [81, 82]. The detector is essentially the LRT, but with the unknown phase of the random telegraph averaged out and the maximum likelihood (ML) estimates of N and $\underline{\tau}$ used. The detector is given by

$$\max_{N,\underline{\tau}} \left\{ \log\cosh\left[\frac{1}{\sigma^2}\int_0^T y(t)s^+(t;N,\underline{\tau})\,dt\right]\right\} \underset{H_0}{\overset{H_1}{\gtrless}} \xi \tag{3.2}$$

where $s^+(t;N,\underline{\tau})$ is the telegraph wave (2.10) generated by the parameters $N,\underline{\tau}$, and initial polarity $\phi = 1$.

One drawback to this method is that the parameter space of $\{N,\underline{\tau}\}$ is infinite dimensional: any maximization method will have to make some simplifying approximations. In [81, 82], a Gibbs sampler was used to efficiently search the parameter space.

In comparison to the optimal LRT discussed previously, the hybrid Bayes/GLR detector is suboptimal.

3.2 Spin detection for the continuous-time classical model

Detection schemes based on the CTRT model rely on the approximate analysis that single spin detection is equivalent to detection of a CT random telegraph in AWGN. In addition, implementation of a detector based on the CTRT model will require the frequency demodulation of the cantilever position signal $z(t)$. The performance of any frequency demodulation scheme (e.g., using a PLL) will degrade as the SNR decreases. In low SNR, the frequency demodulation of $z(t)$ will introduce inaccuracies that will degrade the performance of the hybrid Bayes/GLR detector.

In this section, we propose a more direct detection scheme that uses samples of the cantilever's position, and is based on the well-known Kalman Filter (KF) algorithm [45]. It is hoped that better detection performance can be achieved by using $z(t)$ directly without any additional assumptions.

We shall benchmark the KF based detection method against the energy based detection method mentioned in the development of the CTRT model. Let $s(t)$ be the signal $z(t)$ after it has been frequency demodulated and translated to baseband. We shall use a simple first-order, single-pole filter given by

$$H_{\mathrm{LP}}(z) = \frac{1-\alpha}{2}\frac{1+z^{-1}}{1-\alpha z^{-1}} \tag{3.3}$$

as the LPF to use in filtering $s[i] \triangleq s(t_i)$. The time constant α is chosen based on the bandwidth of the signal; if ω_c is the desired -3 dB bandwidth of the filter, one should set $\alpha = (1-\sin\omega_c)/\cos\omega_c$. Here, we shall use $\omega_c = 2\omega_{\mathrm{skip}}$, where ω_{skip} is the frequency of the rf "off" pulses mentioned in Chapter II. Let $h_{\mathrm{LP}}[i]$ be the impulse response of (3.3). The test statistic that will be used is

$$\sum_{i=1}^{n}(s * h_{\mathrm{LP}})_i^2 \underset{H_0}{\overset{H_1}{\gtrless}} \xi \tag{3.4}$$

where the "$*$" represents the convolution operator. The detector given by (3.4) will be called the post frequency demodulated filtered energy statistic (FDFE).

3.2.1 Analysis of the nonlinear system Σ_1

The system Σ_1 given by (2.4) is a nonlinear system. It is reproduced here for the convenience of the reader:

$$\begin{aligned} \Sigma_1 : \qquad \dot{\mu}_x &= \gamma\mu_y(Gz + \delta B_0) \\ \dot{\mu}_y &= \gamma\mu_z B_1(t) - \gamma\mu_x(Gz + \delta B_0) \\ \dot{\mu}_z &= -\gamma\mu_y B_1(t) \\ m\ddot{z} + \Gamma\dot{z} + kz &= G\mu_z + F_n(t) \end{aligned}$$

Recall from before that the electron spin moment $\underline{\mu}$ has a constant l_2 norm. The invariance of $\|\underline{\mu}\|$ can be verified by considering $V(\mu_x, \mu_y, \mu_z) = \mu_x^2 + \mu_y^2 + \mu_z^2 \Longrightarrow \dot{V} = 0$ by using the expressions for $\dot{\mu}_x$, $\dot{\mu}_y$, and $\dot{\mu}_z$ in (2.4). Setting $B_1(t) \equiv 0$ in the equations for Σ_1 enables us to solve for $\mu_x(t)$ and $\mu_y(t)$. We obtain the solution $\mu_x, \mu_y \approx C \cdot \cos\left[-\gamma G \int_0^t z(\tau)d\tau + \theta\right]$, C and θ being some constants which are different for μ_x and μ_y. Since $|\gamma G z_{ampl}|$ is a large quantity (let z_{ampl} be the amplitude of $z(t)$), the x and y components of $\underline{\mu}$ are oscillating very rapidly. With a non-zero $B_1(t)$, the same holds true for the simulated system. As a result, a small integration time step is required, on the order of 10^{-10} to 10^{-12} seconds.

Nonlinear systems analysis methods can be applied to analyze the reachability and observability properties of Σ_1 [72, 28]. Σ_1 can be written in state-space form by defining the state vector

$$\underline{v}(t) \triangleq [\mu_x(t), \mu_y(t), \mu_z(t), z(t), \dot{z}(t)]^T \tag{3.5}$$

The nonlinear system Σ_1 can be re-written as:

$$\dot{\underline{v}}(t) = f(\underline{v}, t) + Bw(t) \tag{3.6}$$

$$f(\underline{v}, t) \triangleq \begin{pmatrix} \gamma(Gv_4 + \delta B_0)v_2 \\ -\gamma(Gv_4 + \delta B_0)v_1 + \gamma B_1(t)v_3 \\ -\gamma B_1(t)v_2 \\ v_4 \\ \frac{G}{m}v_3 - \omega_0^2 v_4 - \frac{\omega_0}{Q}v_5 \end{pmatrix}, \ B \triangleq [0,0,0,0,1/m]^T \tag{3.7}$$

where $w(t) = F_n(t)$ is white noise. The time dependency in $f(\underline{v}, t)$ arises from $B_1(t)$, the rf signal. If we treat $\underline{u}(t) = [B_1(t), F_n(t)]$ as the control inputs, Σ_1 assumes the form of a linear-analytic system. That is, we can write $\dot{\underline{v}} = p(\underline{v}) + g(\underline{v})\underline{u}$ for suitable functions $p(\cdot)$ and $g(\cdot)$. In this formulation, Σ_1 is locally observable everywhere except for the set of points Ω_1, where $\Omega_1 = \{\underline{v} : Gv_4 + \delta B_0 = 0\}$. Σ_1 is not locally reachable. Instead, it is locally reachable on a sub-manifold of dimension 4 at almost all points except for those which are in either Ω_1 or $\Omega_2 = \{\underline{v} : v_3 = 0\}$. Note, however, that the solution manifold of Σ_1 is of dimension 4, as $[v_1, v_2, v_3]$ is constrained to lie on the unit sphere (where without loss of generality, we consider their appropriately scaled versions). Hence, Σ_1 is locally reachable in its solution manifold at almost all points contained within.

As Σ_1 is not locally reachable, it is not in minimal form. There is an obvious transformation that will bring it to minimal form. Namely, as $[v_1, v_2, v_3]$ lies on the unit sphere, one can take $v_3 = \pm\sqrt{1 - v_1^2 - v_2^2}$ and substitute it into the right-hand side of (2.4). The equation for $\dot{\mu}_z$ will no longer be needed, and the number of equations will decrease from 5 to 4. It is not clear, however, what advantage this reduction in dimensionality brings about.

Going back to the original formulation of Σ_1, although it is nonlinear, the nonlinearity is "soft". The $f(\underline{v}, t)$ term can be written as $f(\underline{v}, t) = \mathbf{F}(z(t), t)\underline{v}(t)$, where $\mathbf{F}(z, t)$ is given by

$$\mathbf{F}(z,t) = \begin{pmatrix} 0 & \gamma(Gz + \delta B_0) & 0 & 0 & 0 \\ -\gamma(Gz + \delta B_0) & 0 & \gamma B_1(t) & 0 & 0 \\ 0 & -\gamma B_1(t) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & G/m & -\omega_0^2 & -\omega_0/Q \end{pmatrix} \tag{3.8}$$

This is a reformulation that conveniently ignores the fact that z is in the state vector $\underline{v}(t)$. However, z is a quantity that is readily available to us, as it is the observed quantity. There are, however, non-ideal factors that are present. Firstly, we do not observe z continuously, but only a sampled version. Secondly, the samples are corrupted by noise. If the sampling frequency f_s is sufficiently high relative to the bandwidth of z, $z(t)$ will be a slowly-changing signal in each sampling interval $[t_i, t_{i+1})$, and can be approximated by a constant. Moreover, an estimator $\hat{z}$ can be employed instead of the noisy samples $x[i]$. This provides a "cleaner" version of $z(t_i)$. If the SNR is sufficiently high, we should have $\hat{z}(t_i) \approx z(t_i)$. This suggests that, in each sampling interval, the nonlinear system Σ_1 can be approximated by a linear system. We shall review some results on the estimation and detection of linear systems that will be used later on.

Henceforth in this chapter, we shall omit the vector notation in order to reduce clutter, and will only use it to emphasize that a particular variable is to be treated as a vector.

3.2.2 Estimation and detection of linear systems

Consider a CT linear system

$$dv(t) = \mathbf{F}(t)v(t)dt + \mathbf{G}(t)d\beta(t) \tag{3.9}$$

where $v(t)$ is a random state vector in $\mathbb{R}^m$, $\beta(t)$ is a s-dimensional Brownian motion process, $\mathbf{F}(t)$ is a time-varying $\mathbb{R}^{m\times m}$ matrix, and $\mathbf{G}(t)$ is a time-varying $\mathbb{R}^{m\times s}$ matrix. Assume that the statistics of $\beta(t)$ are given by

$$\begin{aligned} E[\beta(t)] &= 0 \\ E[(\beta(t) - E[\beta(t)])(\beta(t') - E[\beta(t')])^T] &= \int_{t'}^{t} \mathbf{Q}(\tau)d\tau. \end{aligned} \tag{3.10}$$

Observations of $v(t)$ are made at the time instances $t_1, t_2, \ldots, t_n$ according to

$$y[i] = \mathbf{H}(t_i)v(t_i) + w(t_i) \tag{3.11}$$

where $y[i] \triangleq y(t_i)$ is the p dimensional observation vector, $\mathbf{H}(t_i)$ is a $\mathbb{R}^{p\times m}$ matrix, and $w[i] \triangleq w(t_i)$ is a p-dimensional Gaussian random vector with the property that

$$\begin{aligned} E[w(t_i)] &= 0 \quad \text{for all } \ i = 1, \ldots, n \\ E[w(t_i)w(t_j)^T] &= \begin{cases} \mathbf{R}(t_i) & i = j \\ \mathbf{0} & i \neq j \end{cases}. \end{aligned} \tag{3.12}$$

The stochastic differential equation (SDE) describing the evolution of $v(t)$ requires an initial state $v(t_0)$. Assume that the following knowledge of $v(t_0)$ is available:

$$\begin{aligned} E[v(t_0)] &= \hat{v}(t_0) \\ E[(v(t_0) - E[v(t_0)])(v(t_0) - E[v(t_0)])^T] &= \mathbf{P}_0 \end{aligned} \tag{3.13}$$

If $v(t_0)$ is a Gaussian r.v. or a deterministic constant, $v(t)$ will be a Gaussian random process [45]. Consequently, each of the samples $v(t_i)$ is a Gaussian r.v., and is completely characterized by its first and second order moments.

Given the observations $O = (y[1], y[2], \ldots, y[n])$ and knowledge of the model as given by (3.9)–(3.13), what is the optimal estimate of $v(t_i)$ for $i = 1, \ldots, n$? In order to consider optimality, it is necessary to define the cost function that is applied to the estimates of $v(t_i)$. Let us use the mean squared error; in other words, i.e., we are interested in an estimator $\hat{v}(t_i)$ that minimizes $E\left[(v(t_i) - \hat{v}(t_i))^2\right]$ for each $i = 1, \ldots, n$. In general, $\hat{v}(t_i)$ will be a function of the observations O. Thus far, the values of the t_is have not been specified. Although they are likely to be regularly spaced time points, they do not have to be so. Henceforth, however, assume that the observation times are given by $t_i = iT_s$, where T_s is the sampling interval. Let $f_s = 1/T_s$ be the sampling frequency.

The other question that we would like to address is the following: given two known linear systems Π_0 and Π_1 of the form (3.9)–(3.13) and the observations O, deduce the system that produced the observation data in an optimal fashion. The dimensions of the state vectors need not be the same in both linear systems, and the same goes for the matrices $\mathbf{F}(t)$, $\mathbf{G}(t)$, $\mathbf{Q}(t)$. One can see that this question is a binary hypothesis testing problem, where hypothesis H_i corresponds to the observations originating from system Π_i, for $i = 0, 1$.

Kalman Filter

The expectation of the state vector $v(t_i)$ conditioned on the past observations $Y[i] \triangleq (y[1], \ldots, y[i])^T$ results in the minimum mean squared error (MMSE) estimate of $v(t_i)$. That is,

$$\hat{v}(t_i) = E[v(t_i)|Y[i]] \quad \text{for} \quad i = 1, \ldots, n \tag{3.14}$$

The Kalman Filter (KF) efficiently computes the conditional expectation $E[v(t_i)|Y[i]]$ in a recursive fashion. We will proceed to discuss the KF filtering equations. The notation used here is consistent with [45]: t_i^- denotes the time right before the i-th observation is available, and so $\hat{v}(t_i^-)$, the estimator of $v(t_i)$, does not incorporate information from $y(i)$. Similarly, t_i^+ is the time right after the i-th observation, and $\hat{v}(t_i^+)$ has information from $y(i)$ incorporated into it.

The KF has a predictor-corrector structure. In the interval $[t_{i-1}^+, t_i^-]$, the predicted value of $v(t_i)$ will be computed based on the state equation (3.9) and the estimate $\hat{v}(t_{i-1}^+)$. As the ith observation has not been used, this estimate of $v(t_i)$ will be denoted by $\hat{v}(t_i^-)$. At $t = t_i^+$, we know the value of $y(i)$; this will be used to correct the prediction $\hat{v}(t_i^-)$ via knowledge of the observation equation (3.11). The corrected estimate is denoted by $\hat{v}(t_i^+)$. The same type of behaviour occurs with the conditional covariance matrices

$$\mathbf{P}(t_i^-) \triangleq E[(v(t_i) - \hat{v}(t_i^-))(v(t_i) - \hat{v}(t_i^-))^T | Y[i-1] = Y_{i-1}]$$

$$\mathbf{P}(t_i^+) \triangleq E[(v(t_i) - \hat{v}(t_i^+))(v(t_i) - \hat{v}(t_i^+))^T | Y[i] = Y_i] \tag{3.15}$$

These are also called error covariance matrices, as they give a measure of the errors in $\hat{v}(t_i^-)$ and $\hat{v}(t_i^+)$.

The equations for the prediction or propagation step are

Algorithm 3.1 (Propagation equations for the KF).

(3.16)

$$\hat{v}(t_i^-) = \mathbf{\Phi}(t_i, t_{i-1})\hat{v}(t_{i-1}^+)$$

(3.17)

$$\mathbf{P}(t_i^-) = \mathbf{\Phi}(t_i, t_{i-1})\mathbf{P}(t_{i-1}^+)\mathbf{\Phi}^T(t_i, t_{i-1}) + \int_{t_{i-1}}^{t_i} \mathbf{\Phi}(t_i, \tau)\mathbf{G}(\tau)\mathbf{Q}(\tau)\mathbf{G}^T(\tau)\mathbf{\Phi}^T(t_i, \tau)d\tau$$

where $\mathbf{\Phi}(\cdot, \cdot)$ is the m-by-m state transition matrix for the system given by (3.9) [45]. The state transition matrix $\mathbf{\Phi}(\cdot, \cdot)$ satisfies the following differential equation

$$\frac{d}{dt}\mathbf{\Phi}(t,t_0) = \mathbf{F}(t)\mathbf{\Phi}(t,t_0)$$

$$\mathbf{\Phi}(t_0,t_0) = \mathbf{I} \tag{3.18}$$

If $\mathbf{F}(t) = \mathbf{F}$, a constant matrix, then $\mathbf{\Phi}(t,t_0) = \exp[\mathbf{F}(t-t_0)]$.

The propagation step can also be implemented as the solution to two differential equations [45]. Define $\hat{v}_i(t)$ for $t \in [t_i, t_{i+1})$ as the estimate of $v(t)$ conditioned on all the observations up until $y[i]$. Similarly, let $\mathbf{P}_i(t), t \in [t_i, t_{i+1})$ to be the conditional covariance conditioned on all observations up until $y[i]$.

Algorithm 3.2 (Differential equations for the Propagation step in the KF).

$$\dot{\hat{v}}_i(t) = \mathbf{F}(t)\hat{v}_i(t), \quad \hat{v}_i(t_i) = \hat{v}(t_i^+) \tag{3.19}$$

$$\dot{\mathbf{P}}_i(t) = \mathbf{F}(t)\mathbf{P}_i(t) + \mathbf{P}_i(t)\mathbf{F}^T(t) + \mathbf{G}(t)\mathbf{Q}(t)\mathbf{G}^T(t), \quad \mathbf{P}_i(t_i) = \mathbf{P}(t_i^+) \tag{3.20}$$

with $\hat{v}(t_{i+1}^-) = \hat{v}_i(t_{i+1})$ *and* $\mathbf{P}(t_{i+1}^-) = \mathbf{P}_i(t_{i+1})$.

The equations for the correction or measurement update step are

Algorithm 3.3 (Measurement update equations for the KF).

$$\mathbf{K}(t_i) = \mathbf{P}(t_i^-)\mathbf{H}^T(t_i)\left[\mathbf{H}(t_i)\mathbf{P}(t_i^-)\mathbf{H}^T(t_i) + \mathbf{R}(t_i)\right]^{-1} \tag{3.21}$$

$$\hat{v}(t_i^+) = \hat{v}(t_i^-) + \mathbf{K}(t_i)(y[i] - \mathbf{H}(t_i)\hat{v}(t_i^-)) \tag{3.22}$$

$$\mathbf{P}(t_i^+) = (\mathbf{I} - \mathbf{K}(t_i)\mathbf{H}(t_i))\mathbf{P}(t_i^-) \tag{3.23}$$

The $m \times p$ matrix $\mathbf{K}(t_i)$ is known as the *Kalman gain.* Its usage in (3.22) illustrates why such a nomenclature is appropriate. The quantity

$$\eta[i] = y[i] - \mathbf{H}(t_i)\hat{v}(t_i^-) = y[i] - \hat{z}(t_i^-) \tag{3.24}$$

is the prediction error, also known as the *innovations.* From (3.22), one sees that the estimate $\hat{v}(t_i^+)$ is formed by adding the predicted value to the innovations scaled by the Kalman gain. A large Kalman gain is an indication that we are not confident in the predicted value $\hat{v}(t_i^-)$, and so the observation $y[i]$ plays a major role in the estimation of $v(t_i)$. On the other hand, a small Kalman gain downplays the effect of the innovations. It signifies that we believe that the predicted value $\hat{v}(t_i^-)$ is a good estimate of $v(t_i)$.

From (3.17), (3.21), and (3.23), one realizes that computing $\mathbf{P}(t_i^-)$ and $\mathbf{P}(t_i^+)$ does not require the observations $y(t_i)$. It follows that their calculation can be done offline.

Kalman Filter detection

Let Π_0 and Π_1 denote two systems of the form (3.9)–(3.13), and suppose that $p = 1$, so that the observations are scalars. Assume that the observations produced by both systems are zero mean. Having observed the sequence $(y[1], y[2], \ldots, y[n])$, we would like to determine which system most likely produced these values. We are faced with the hypothesis testing problem

$$\begin{aligned} H_0 : &\quad y[i] = s_0(t_i) + w[i] \\ H_1 : &\quad y[i] = s_1(t_i) + w[i] \end{aligned} \tag{3.25}$$

where $s_0(t_i)$ and $s_1(t_i)$ are zero mean Gaussian signals with covariances $\mathbf{R}_0$ and $\mathbf{R}_1$ respectively.

The LRT is shown in [25, 70] to be

$$y^T(R_0^{-1} - R_1^{-1})y \underset{H_0}{\overset{H_1}{\gtrless}} \xi \tag{3.26}$$

where $y \triangleq (y[1], \ldots, y[n])^T$. It can be equivalently written in terms of the innovations produced by a KF matched to Π_0 and another matched to Π_1. Let $\eta_k[i]$ denote the innovations produced by the KF matched to Π_k for $k = 0, 1$. Note that $\eta_k[i] \triangleq \eta_k(t_i)$. Then, the LRT can also be expressed as

$$\sum_{i=1}^{n} \frac{\eta_0^2[i]}{\operatorname{var}(\eta_0[i])} - \sum_{i=1}^{n} \frac{\eta_1^2[i]}{\operatorname{var}(\eta_1[i])} \underset{H_0}{\overset{H_1}{\gtrless}} \xi \tag{3.27}$$

where the threshold ξ is time-varying, i.e. a function of n [25]. The detector can be implemented with a dual KF setup: see Figure 3.1 below.

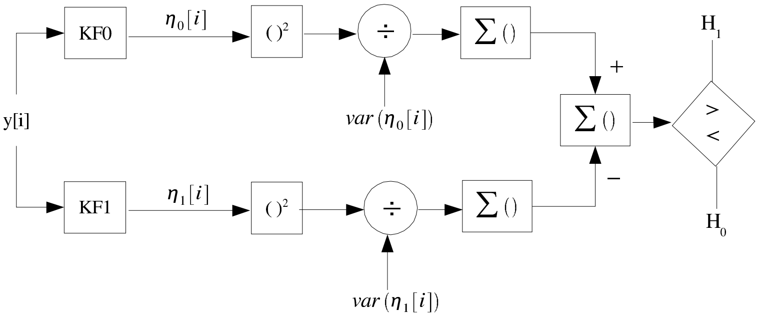


Figure 3.1: Dual Kalman Filter detector

The variances of the innovations $\eta_0[i]$ and $\eta_1[i]$ appear in the decision rule (3.27). They are the variance of $\eta_0[i]$ assuming the H_0 hypothesis and the variance of $\eta_1[i]$ assuming the H_1 hypothesis respectively. As these values do not depend on the observations, they can be computed offline.

3.2.3 Estimation of nonlinear systems

Linear systems are tractable to work with. Indeed, the previous section characterized the estimation of a linear system with discrete-time sampled observations. Furthermore, the detection of linear systems with scalar observations was addressed. Unfortunately, nonlinear systems are less tractable. There are two places where the nonlinearity could potentially appear: either in the state equation or in the observation equation. For the CTC model, the nonlinearity appears in the former. Suppose we extend the linear system described by (3.9) and (3.11) to the following nonlinear system:

$$dv(t) = f(v(t), t) + \mathbf{G}(t)d\beta(t)$$

$$y[i] = \mathbf{H}(t_i)v(t_i) + w[i] \tag{3.28}$$

and pose the question: what form does the MMSE estimator of $v(t_i)$ take? The answer to this question can be applied to finding the optimal MMSE state estimator of Σ_1, the CTC model under the spin present hypothesis.

The optimal mean squared estimate of nonlinear systems with CT state equations and DT observations is known; however, the computational complexity can be prohibitive [29, 16, 42]. In [42], an approximation to the optimal filtering equations are given for the following system

$$dv(t) = b(v(t))dt + \sigma(v(t))d\beta(t) \tag{3.29}$$

$$y[i] = h(v(t_i)) + w[i] \tag{3.30}$$

where $v(t), \beta(t), y[i], w[i]$ retain their meaning from the discussion on the KF. The state equation in (3.29) is time invariant; however, the state equation in the nonlinear system (3.28) is not. For this reason, the approximation in [42] cannot be directly used without some modification. Another attempt at approximating the optimal filtering equations was done in [23]. Galerkin's method was used here to approximately solve the partial differential equation (PDE) that the posterior density $p(v_t|\{y(t_i) : t_i < t\})$ satisfied. A choice of suitable basis functions with which to approximate the posterior density has to be made: in [23], only $v(t) \in \mathbb{R}$ was considered, and the complex exponentials were chosen as basis functions. This choice conveniently led to the usage of the Fast Fourier Transform (FFT) and inverse FFT in computing approximations to the posterior density. For a general multidimensional random process $v(t)$, however, it is not clear what would be a good choice of basis functions.

Extended Kalman Filter

The KF can be extended in a heuristic way to address the estimation of a nonlinear system. The extended version is aptly called the Extended Kalman Filter (EKF) [46]. The principle behind the EKF is to linearize any nonlinearities that appear in the state or observation equation around the previous best state estimate. The measurement update equations are the same as those of the KF; however, the propagation equations are different. The propagation equations for the EKF are

Algorithm 3.4 (Propagation equations for the EKF).

$$\dot{\hat{v}}_i(t) = f(\hat{v}_i(t), t), \quad \hat{v}_i(t_i) = \hat{v}(t_i^+) \tag{3.31}$$

$$\dot{\mathbf{P}}_i(t) = \mathbf{J}(\hat{v}_i(t), t)\mathbf{P}_i(t) + \mathbf{P}_i(t)\mathbf{J}^T(\hat{v}_i(t), t) + \mathbf{Q}(t), \quad \mathbf{P}_i(t_i) = \mathbf{P}(t_i^+) \tag{3.32}$$

with $\hat{v}(t_{i+1}^-) = \hat{v}_i(t_{i+1})$ and $\mathbf{P}(t_{i+1}^-) = \mathbf{P}_i(t_{i+1})$. The matrix $\mathbf{J}(v,t)$ is the Jacobian of $f(v,t)$.

Note that (3.31) and (3.32) are coupled differential equations, and so must be solved simultaneously. It is interesting to compare them with their KF counterparts in (3.19) and (3.20) respectively. As noted in [46], the coupling in the EKF propagation equations may produce numerical difficulties. Another undesirable side effect of the coupling is that the computation of the error covariance matrix cannot be done offline, as it depends on the observations through the state estimate. A way of decoupling the two differential equations is to use the approximation $\hat{v}(t) \approx \hat{v}(t_i^+)$ for $t \in [t_i, t_{i+1})$.

3.2.4 Piecewise linear Kalman Filter

We shall consider the case when the nonlinearity in the state equation is soft. Suppose that $f(\cdot,\cdot)$ in (3.28) could be written as $f(v(t),t) = \mathbf{F}(v_{k_1}(t),\ldots,v_{k_d}(t),t)v(t)$, where the k_i's are distinct, $1 \le k_i \le n$ for each i, and $1 \le d < n$. Moreover, assume that the collection of state components $V_s \triangleq \{v_{k_1}(t),\ldots,v_{k_d}(t)\}$ are all slowly varying relative to the sampling frequency f_s of the observations. Define $v_s(t) \triangleq [v_{k_1}(t),\ldots,v_{k_d}(t)]^T$. Then, in the sampling interval $[t_i, t_{i+1})$, one could make the approximation

$$f(v(t),t) = \mathbf{F}(v_s(t),t)v(t) \approx \mathbf{F}(\hat{v}_s(t_i^+),t)v(t). \tag{3.33}$$

This has the effect of linearizing the nonlinear system (3.28) in each sampling interval. It is then possible to apply the KF to the linearized system. A modification can be made to propagate the state estimate $\hat{v}(t)$ through the original nonlinear system, as in (3.32). The estimator for nonlinear systems with the specified softness will be called the piecewise linear Kalman Filter (PLKF).

One way to ensure that each $v_{k_i}(t)$ is sufficiently slowly varying is to stipulate that each $v_{k_i}(t) \in V_s$ have a bandlimited spectrum, say $[-B_i, B_i]$, and that $f_s \gg B_i$. Clearly, a larger f_s is desirable. Another necessary condition for the approximation (3.33) is that the state estimates $\hat{v}_{k_1}(t_i^+) \approx v_{k_1}(t_i),\ldots,\hat{v}_{k_d}(t_i^+) \approx v_{k_d}(t_i)$. Thus, a sufficiently high SNR is desirable.

The measurement update equations for the PLKF are the same as for the KF. The propagation equations are given in differential form by

Algorithm 3.5 (Propagation equations for the PLKF).

$$\dot{\hat{v}}_i(t) = f(\hat{v}_i(t),t), \quad \hat{v}_i(t_i) = \hat{v}(t_i^+) \tag{3.34}$$

$$\dot{\mathbf{P}}_i(t) = \mathbf{F}(\hat{v}_s(t_i^+),t)\mathbf{P}_i(t) + \mathbf{P}_i(t)\mathbf{F}^T(\hat{v}_s(t_i^+),t) + \mathbf{Q}(t), \quad \mathbf{P}_i(t_i) = \mathbf{P}(t_i^+) \tag{3.35}$$

with $\hat{v}(t_{i+1}^{-}) = \hat{v}_i(t_{i+1})$ *and* $\mathbf{P}(t_{i+1}^{-}) = \mathbf{P}_i(t_{i+1})$.

Comparing (3.34), (3.35) with the filtering equations for EKF, we see that the difference lies in the propagation of the error covariance matrix in time. Rather than using the Jacobian $\mathbf{J}(v,t)$, the PLKF uses $\mathbf{F}(v_s(t),t)$. Therefore, the PLKF can be viewed as a 0th order EKF method, whereas the EKF is of 1st order.

3.2.5 Detection with piecewise linear Kalman Filtering

Let Π_0 and Π_1 denote two continuous-discrete systems that are either linear or softly nonlinear (in the sense that the assumptions underlying the application of the PLKF apply) and with zero mean observations. Let F_i denote the KF or PLKF matched to model Π_i for $i = 0, 1$. If one (or both) of the systems is (are) softly nonlinear, the decision rule (3.27) is no longer applicable. Nevertheless, we shall invoke the piecewise linear assumption and apply the rule to the innovations generated by F_0 and F_1.

3.2.6 Application to spin detection for the continuous-time classical model

At this point, it becomes a matter of assembling the methods that have been discussed in order to construct a single spin detector. We shall take Π_0 to be Σ_0, which is described by (2.5). The system can be rewritten in state-space form as:

$$\begin{aligned} \Sigma_0 : dv(t) &= \mathbf{A}_0 v(t) + \mathbf{B}_0 d\beta(t) \\ y[i] &= \mathbf{H}_0 v(t_i) + w[i] \end{aligned} \tag{3.36}$$

where

$$\mathbf{A}_0 = \begin{pmatrix} 0 & 1 \\ -\omega_0^2 & -\omega_0/Q \end{pmatrix}, \ \mathbf{B}_0 = \begin{pmatrix} 0 \\ 1/m \end{pmatrix}, \mathbf{H}_0 = [1, 0],$$

and recall that $y[i] = y(t_i)$, $w[i] = w(t_i)$. Note that in (3.36) above, the state vector $v(t) = [z(t), \dot{z}(t)]^T \in \mathbb{R}^2$.

Π_1 will be taken to be Σ_1, the CTC model of the cantilever-spin dynamics. As noted in 3.2.1, Σ_1 is softly linear. If f_s is sufficiently high compared to the bandwidth of the observation $y(t)$, we can apply the piecewise linear (indeed, constant) approximation (3.33) and make use of the PLKF. The two systems Π_0 and Π_1 have observations that are approximately zero mean. From the last subsection, the dual KF structure can be used to obtain an approximate detector of the signal model.

It is instructive to compare $\mathbf{F}(z,t)$ with $\mathbf{J}(v,t)$, the Jacobian of $f(v,t)$. As was previously mentioned, the only difference between EKF and PLKF is in the propagation of the error covariance matrix. The propagation equations for both filters are of the same form: the dissimilarity arises from the usage of $\mathbf{F}(z,t)$ vs. $\mathbf{J}(v,t)$. Now, $\mathbf{F}(z,t)$ has already been given in (3.8); in contrast, the Jacobian $\mathbf{J}(v,t)$ is

$$(3.37) \qquad \mathbf{J}(v,t) = \begin{pmatrix} 0 & \gamma(Gz+\delta B_0) & 0 & \gamma G\mu_y & 0 \\ -\gamma(Gz+\delta B_0) & 0 & \gamma B_1(t) & -\gamma G\mu_x & 0 \\ 0 & -\gamma B_1(t) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & G/m & -\omega_0^2 & -\omega_0/Q \end{pmatrix}$$

Comparing (3.8) and (3.37), we see there are two additional terms in the Jacobian as compared to $\mathbf{F}(z,t)$.

Although implementation of the dual KF structure for detection is straightforward, there are some ways in which the implementation can be simplified. Since Σ_1 is not linear, the computation of $\mathrm{var}(\eta_1[i])$ under the H_1 hypothesis is difficult. As a simplification, it can be obtained either empirically or assuming that $\mathrm{var}(\eta_1[i]) \approx \mathrm{var}(\eta_0[i])$ for $i = 1, \ldots, n$.

A further simplification to (3.27) can be made by assuming that $\mathrm{var}(\eta_0[i])$ and $\mathrm{var}(\eta_1[i])$ are approximately constant and that the two constants are equal to each other. Then, one obtains an innovations energy test

$$\sum_{i=1}^{n} \eta_0^2[i] - \sum_{i=1}^{n} \eta_1^2[i] \mathop{\gtrless}_{H_0}^{H_1} \xi \tag{3.38}$$

An issue that must be examined more closely is the initial values used in the KF for Σ_0 and the PLKF for Σ_1. We assume that $z(0)$ and $\dot{z}(0)$ are approximately known. This would be the case if σ^2 is small and $z(t)$ is an approximately sinusoidal signal. For $x(t) = C\sin(\omega_0 t)$, $\dot{x} = \omega_0 C \cos(\omega_0 t) = \pm\omega_0\sqrt{1-x^2}$. The initial spin moment $\underline{\mu}(0)$ is not known, however. Since the PLKF is sensitive to $\underline{\mu}(0)$, an attempt is made to guess $\underline{\mu}(0)$. We shall apply the GLR principle, which entails replacing $\underline{\mu}(0)$ with its ML estimate. The lower KF1 branch in the electron spin detector of Fig. 3.1 is replaced by a filter bank of p PLKF1s, each initialized with a different $\underline{\mu}(0)$. Denote by $\underline{\mu}_k(0)$ the value of $\underline{\mu}(0)$ with which the kth PLKF1 is initialized. The minimum output value of all p PLKF1s will be selected and compared with the output of KF0. By doing this, (3.27) is modified to be

$$\sum_{i=1}^{n} \frac{\eta_0^2[i]}{\mathrm{var}(\eta_0[i])} - \min_{1\le k\le p}\left(\sum_{i=1}^{n} \frac{\eta_{1,k}^2[i]}{\mathrm{var}(\eta_1[i])}\right) \mathop{\gtrless}_{H_0}^{H_1} \xi \tag{3.39}$$

and the innovations energy test (3.38) becomes

$$\sum_{i=1}^{n} \eta_0^2[i] - \min_{1\le k\le p}\left(\sum_{i=1}^{n} \eta_{1,k}^2[i]\right) \mathop{\gtrless}_{H_0}^{H_1} \xi \tag{3.40}$$

where $\eta_{1,k}[i]$, $i = 1, \ldots, n$ is the innovations sequence produced by the kth PLKF1, for $k = 1, \ldots, p$. We shall use these hybrid KF/GLR detectors (3.39) or (3.40) in the simulations.

3.2.7 Simulations

In this section, the CTC model was simulated with only the fundamental mode, and the performance of the PLKF vs. EKF was examined. The simulation parameters used are given in Table 3.1.

Table 3.1: Simulation parameters used in the CTC model in the comparison of the PLKF vs. EKF filters

Parameter		**Value**
Description	Name	
Sampling frequency	f_s	250 kHz
Cantilever oscillation frequency	ω_0	10 kHz
Cantilever rms amplitude	z_{rms}	1 nm
Noise temperature[†]	T_f	1 K
Skip period of rf field	T_{skip}	5 cycles[‡]
Amplitude of rf field	$B_{1\mathrm{ampl}}$	4 G
Q of cantilever	Q	10^5
Variance of observation noise	σ^2	10^{-20}
Magnetic field gradient	G	10 G/Å
Cantilever spring constant	k	10^{-4} N/m
Length of each simulation	T_{sim}	5 ms
Integration time step	T_{int}	10^{-9} s

[†] Of fundamental mode of cantilever.
[‡] Each cycle refers to a period of the cantilever oscillation.

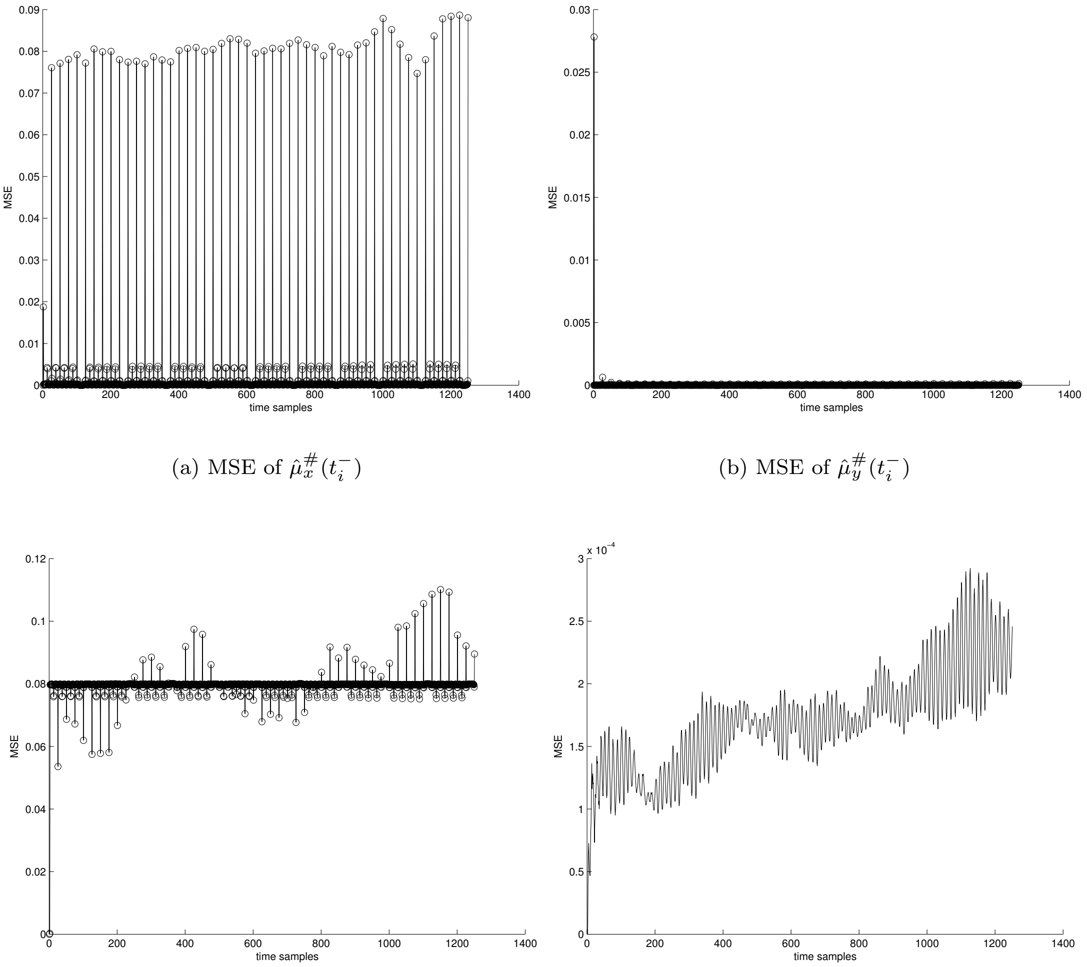
One hundred simulations of the CTC model under the H_1 spin hypothesis were generated with the parameters of Table 3.1. Before proceeding to discuss the results, a discussion of the simulation of the H_0 and H_1 systems is in order. The 4th order Runge-Kutta algorithm (RK4) is a common algorithm used to solve first-order differential equations. The systems H_0 and H_1 are stochastic. However, RK4 was developed to solve deterministic differential equations. It is not at all clear that using RK4 will produce a realization that is representative of the statistics of the random process.

The first issue that must be resolved is an error criterion for a stochastic simulation. One such criterion is given in [37], and is referred to as weak or wide-sense simulation, where the first and second properties of the processes are matched. This was used in deriving the coefficients of a 4th-order RK algorithm for linear stochastic systems [38]. The covariance matrix of the stochastic process is matched to that of the RK solution at multiples of the integration step size T_{int}. For nonlinear stochastic systems, however, only a 2nd-order RK algorithm has been solved [36].

The simulation of Π_0 makes use of the coefficients derived in [38] for a linear, time-varying stochastic differential equation. The simulation of Π_1 uses RK4 in the "classical" method as it is called in [38], [36]. For a small integration step size T_{int}, this zero-order reduction of the Brownian motion term $w(t)$ will result in a small error for Π_1.

The initial values of $z(t_0)$ and $\dot{z}(t_0)$ were the same for all one hundred simulations (NB. $t_0 = 0$). Specifically, $z(t_0) = 0$ and $\dot{z}(t_0) = \sqrt{2} z_{\text{rms}} \omega_0$. Let $\underline{\mu}^{\#} = \underline{\mu} / \|\underline{\mu}\|$, so that $\underline{\mu}^{\#}$ is a vector on the 3-d unit sphere (in the simulations, we work with $\underline{\mu}^{\#}$ as opposed to $\underline{\mu}$). The initial spin moment $\underline{\mu}^{\#}(t_0)$ was assumed to be uniformly distributed over the unit sphere. The EKF filter was applied to the first 50 realizations, while the PLKF filter was applied to all 100 realizations. Both filters were initialized with the true value of $v(t_0)$, and the empirical covariance matrix over the 100 realizations was used to initialize $\mathbf{P}_0$.

The MSE for the estimators of $\mu_x^{\#}(t_i^-)$, $\mu_y^{\#}(t_i^-)$, $\mu_z^{\#}(t_i^-)$, and a normalized version of $z(t_i^-)$ are plotted as functions of the discrete time index i. Let $z^{\#}(t) = z(t)/z_{\text{rms}}$ denote the normalized version of $z(t)$. MSEs were obtained by averaging the squared error over all of the different realizations: 50 for EKF and 100 for PLKF. Note that EKF takes significantly longer than PLKF to run. The difference is due to the propagation of the error covariance matrix. While EKF has to solve (3.32), there exists a closed form solution for the corresponding PLKF equation (3.35). The MSE for EKF is given in Fig. 3.2, while the MSE for PLKF is given in Fig. 3.3.



(a) MSE of $\hat{\mu}_x^{\#}(t_i^-)$

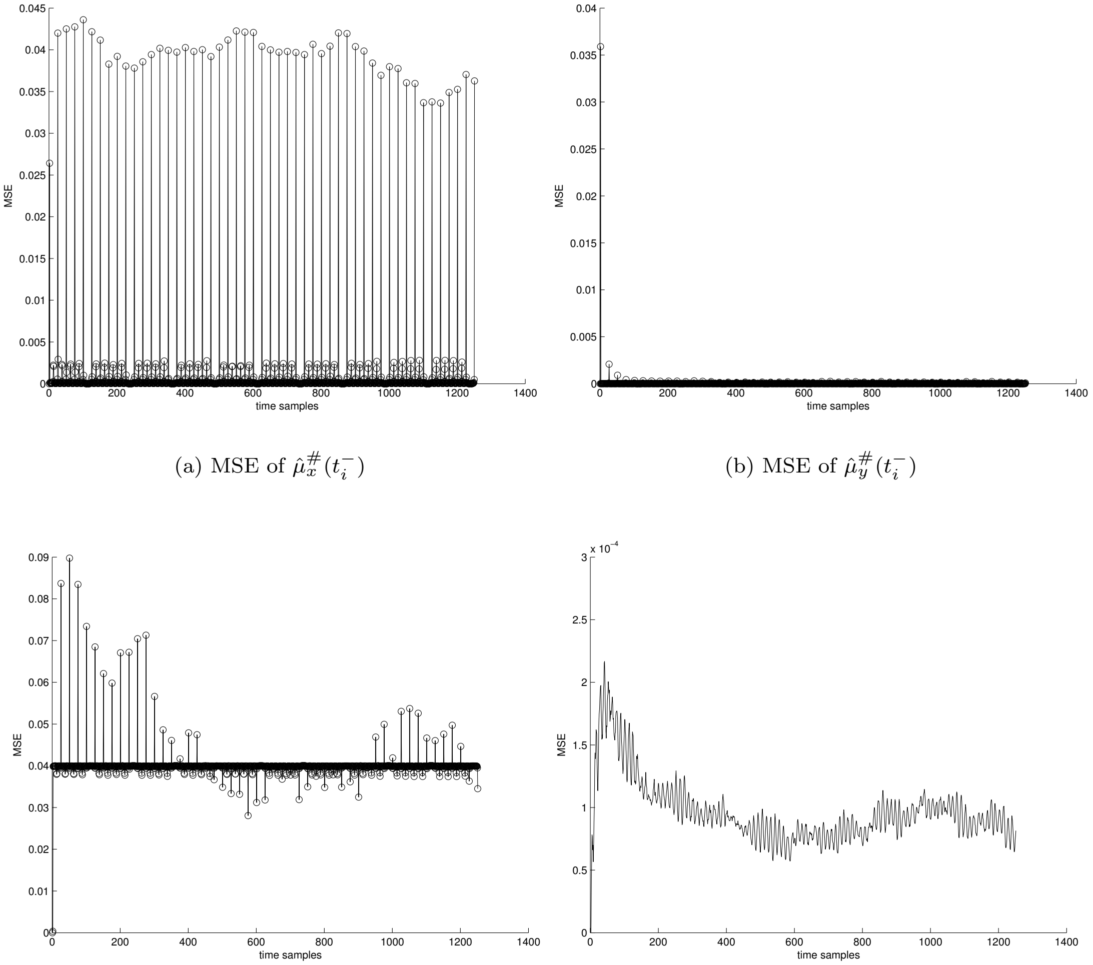
(b) MSE of $\hat{\mu}_y^{\#}(t_i^-)$

(c) MSE of $\hat{\mu}_z^{\#}(t_i^-)$

(d) MSE of $\hat{z}^{\#}(t_i^-)$

Figure 3.2: MSE of the EKF estimator when applied to the CTC model. The average of the squared error over 50 realizations is displayed.

In $\hat{\mu}_x^{\#}$, $\hat{\mu}_z^{\#}$, and $\hat{z}^{\#}$, the PLKF filter is doing better than the EKF filter. On average, the MSE of $\hat{\mu}_z^{\#}$ for the PLKF is half that of EKF. Both filters track $\mu_y^{\#}$ equally well. The EKF's MSE of $\hat{z}^{\#}$ appears to be increasing. In contrast, the MSE for PLKF appears to be oscillating around a steady state value. One can see the effect of a smaller number of realizations (50 for EKF vs. 100 for PLKF) has



(a) MSE of $\hat{\mu}_x^{\#}(t_i^-)$

(b) MSE of $\hat{\mu}_y^{\#}(t_i^-)$

(c) MSE of $\hat{\mu}_z^{\#}(t_i^-)$

(d) MSE of $\hat{z}^{\#}(t_i^-)$

Figure 3.3: MSE of the PLKF estimator when applied to the CTC model. The average of the squared error over 100 realizations is displayed.

on EKF's MSE of $\hat{z}^{\#}$, as it has a higher variance than the PLKF's MSE. It is not surprising that the MSE for $\hat{\mu}_z^{\#}$ has the same general trend as that of $\hat{z}^{\#}$. Referring back to (2.4), the z component of the spin is an input to the second-order equation for $z(t)$. Consequently, the estimation error in $\mu_z^{\#}$ will propagate to the estimation of $z^{\#}$.

Since $y(t_i) - \hat{z}(t_i^-) = \eta(t_i) = \eta[i]$ (see (3.24)), the innovations $\eta[i] = (z(t_i) - \hat{z}(t_i^-)) + w[i]$. Consequently, the MSE of $\hat{z}(t_i^-)$ is equal to $(E[\eta^2[i]] - \sigma^2)$. It follows from Figs. 3.2 and 3.3 that, for the parameters given in Table 3.1, the PLKF produces innovations that are lower in mean-squared value. Under the spin present hypothesis, one would like $\sum_i \eta_1^2[i]/\mathrm{var}(\eta_1[i])$ to be smaller than $\sum_i \eta_0^2[i]/\mathrm{var}(\eta_0[i])$; refer back to (3.27) for the dual KF test statistic. Therefore, innovations that are smaller in mean-squared value are desirable under the spin present hypothesis. The simulations presented suggest that PLKF is a better candidate than EKF to use in the hybrid KF/GLR detectors.

Secondly, the performance of the KF/GLR detector was compared with the FDFE detector, which is given by (3.4). In the latter, demodulation is performed on samples of the cantilever signal via a zero crossings method. The -3 dB bandwidth ω_c of the LPF filter (3.3) was set to $2\omega_{\text{skip}}$, where $\omega_{\text{skip}} = 2\pi/T_{\text{skip}}$.

The PLKF filter was used in KF1 branch(es) of the hybrid KF/GLR detectors. Two scenarios with different values of k (cantilever spring constant) were considered. The first case was $k = 10^{-3}$ N/m, while the second case was $k = 10^{-4}$ N/m. The other simulation parameters that are shared by both cases are given in Table 3.2.

For each of the two cases, 40 realizations were generated: 20 under the spin present (H_1) hypothesis and 20 under the no-spin (H_0) hypothesis. In the spin present realizations, $\underline{\mu}^{\#}(t_0)$ was assumed to be uniformly distributed over the unit sphere. The same initial conditions as mentioned above for $z(t_0)$ and $\dot{z}(t_0)$ were used for both H_0 and H_1 realizations.

We applied the hybrid KF/GLR detectors given by (3.39) and (3.40) to each of the 40 realizations; p set to 2. The initial spin moment vectors were $\underline{\mu}_1^{\#} = (\sqrt{3}/2, 0, 1/2)^T$ and $\underline{\mu}_2^{\#} = (\sqrt{3}/2, 0, -1/2)^T$. The true values of $z(t_0)$ and $\dot{z}(t_0)$ were used to initialized the PLKF and KF. The initial covariance matrix $\mathbf{P}_0$ used for the PLKF was $\mathbf{P}_0 = \mathrm{diag}(0.3, 0.3, 0.3, s_z, s_{\dot{z}})$, where $s_z, s_{\dot{z}}$ are the empirical variances of z and $\dot{z}$

Table 3.2: Simulation parameters used in the CTC model in the comparison of the dual KF detector vs. the post frequency demodulated filtered energy statistic.

Parameter		**Value**
Description	Name	
Sampling frequency	f_s	250 kHz
Cantilever oscillation frequency	ω_0	10 kHz
Cantilever rms amplitude	z_{rms}	1 nm
Noise temperature[†]	T_f	0.6 K
Skip period of rf field	T_{skip}	10 cycles[‡]
Amplitude of rf field	$B_{1\text{ampl}}$	4 G
Q of cantilever	Q	10^5
Variance of observation noise	σ^2	10^{-22}
Magnetic field gradient	G	10 G/Å
Length of each simulation	T_{sim}	15 ms
Integration time step	T_{int}	10^{-11} s

[†] Of fundamental mode of cantilever.
[‡] Each cycle refers to a period of the cantilever oscillation.

respectively under the H_1 hypothesis. The empirical covariance matrix under the H_0 hypothesis was used to initialize $\mathbf{P}_0$ for KF0. The empirical variance of the innovations under H_1 was used in the hybrid KF/GLR detectors.

In Fig. 3.4, the receiver operating characteristic (ROC) curve for the KF/GLR innovations energy detector vs. the FDFE detector are presented for $k = 10^{-3}$ N/m, and in Fig. 3.5, the plots of their test statistics under H_0 and H_1. The ROC curve is a plot of the probability of detection (P_D) vs. the probability of false alarm (P_F). For values of $P_F < 0.35$, the KF/GLR innovations energy detector has better performance. However, for $P_F > 0.35$, the FDFE detector has better performance.

The ROC curve for the KF/GLR innovations detector is presented in Fig. 3.6b. The test accurately distinguishes the spin and no-spin realizations. A glance at the test statistic values of the KF/GLR innovations detector under H_0 and H_1 in Fig. 3.6a reveals the reason: there is a wide separation in the values between these two hypotheses. Overall, the KF/GLR innovations detector has the best performance

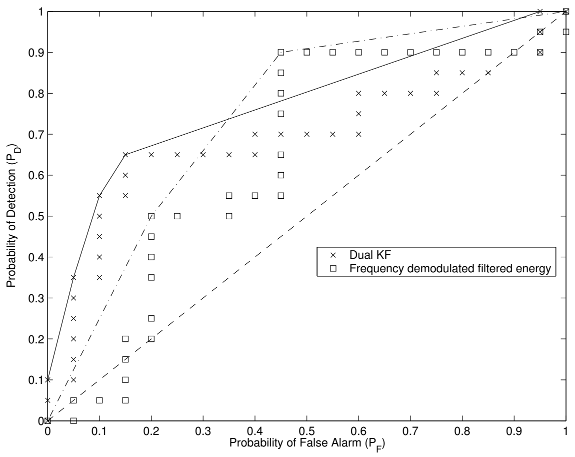
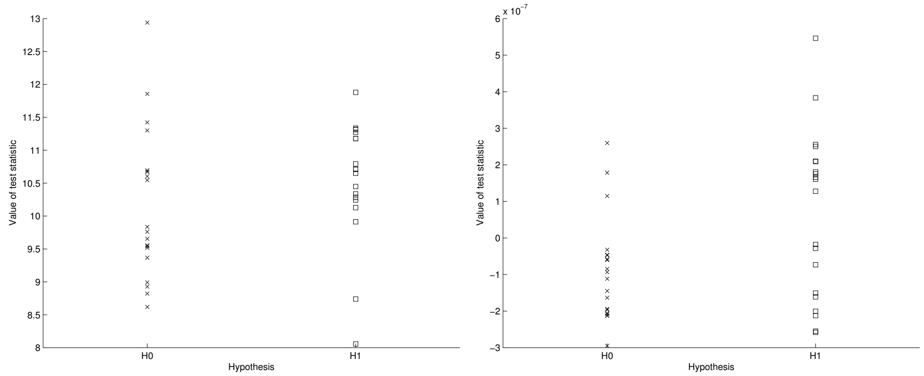


Figure 3.4: ROC curve of the KF/GLR innovations energy detector vs. the FDFE detector for $k = 10^{-3}$ N/m. The KF/GLR innovations energy detector has better performance for $P_F < 0.35$.



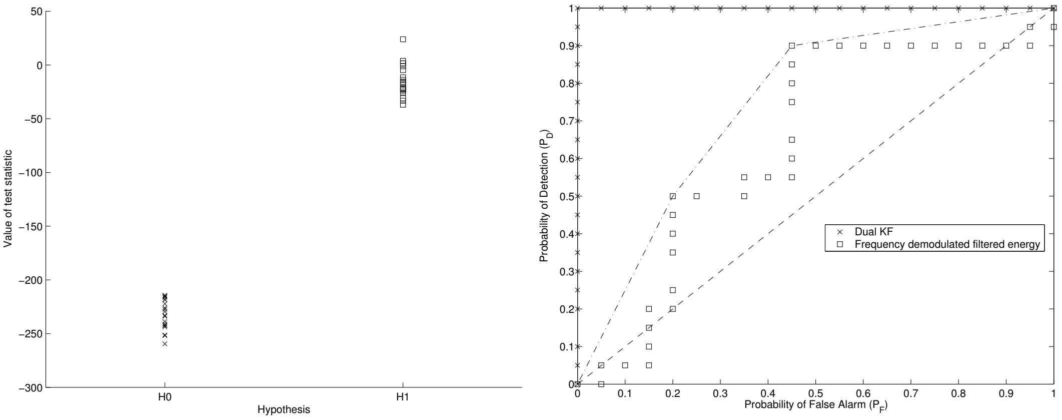
(a) Test statistic for the FDFE detector

(b) Test statistic for the KF/GLR innovations energy detector

Figure 3.5: Test statistics of the KF/GLR innovations energy detector vs. the FDFE detector for $k = 10^{-3}$ N/m.

for this case.

The second case considered is $k = 10^{-4}$ N/m. Paralleling the presentation for the previous case, Fig. 3.7 illustrates the ROC curve for the KF/GLR innovations energy detector vs. the FDFE detector. Again, there seems to be a point where the two ROC curves intersect. For $P_F < 0.25$, the KF/GLR innovations energy detector has better performance, while for $P_F > 0.25$, the FDFE detector is superior. The test statistics of the two detectors are given in Fig. 3.8. They look similar. The



(a) Test statistic for the KF/GLR innovations detector

(b) ROC curve

Figure 3.6: Comparison of the KF/GLR innovations detector vs. the FDFE detector for $k = 10^{-3}$ N/m. The KF/GLR innovations detector accurately distinguishes the spin present vs. the spin absent realizations.

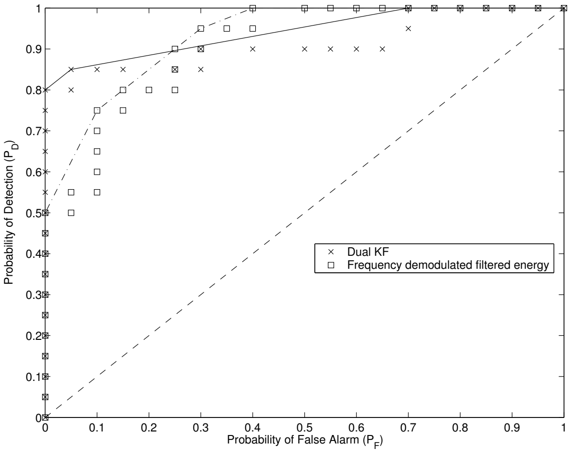
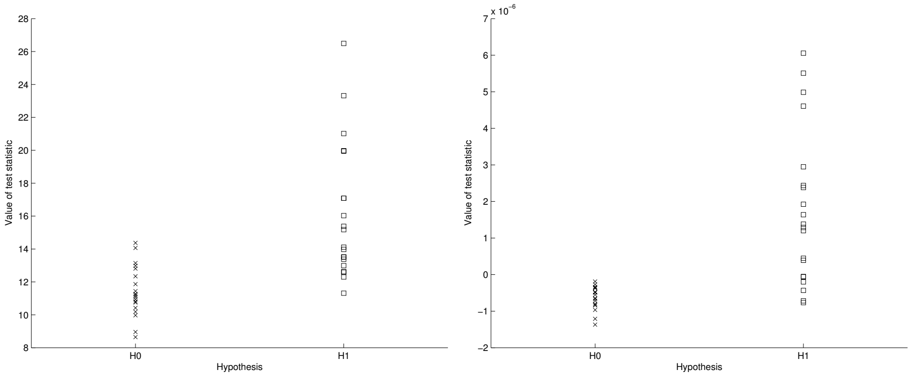


Figure 3.7: ROC curve of the KF/GLR innovations energy detector vs. the FDFE detector for $k = 10^{-4}$ N/m. The KF/GLR innovations energy detector has better performance for $P_F < 0.25$.

H_0 statistics for the KF/GLR innovations energy test have smaller variance, which explains why it is doing better for small P_F values.

These two detectors have better performance than for the case of $k = 10^{-3}$ N/m. In the development of the CTRT model in Chapter 2.2.2, the shift in the cantilever frequency is inversely proportional to k; refer to (2.8). Therefore, a smaller k results in a bigger frequency shift, which would make the observations under the H_1



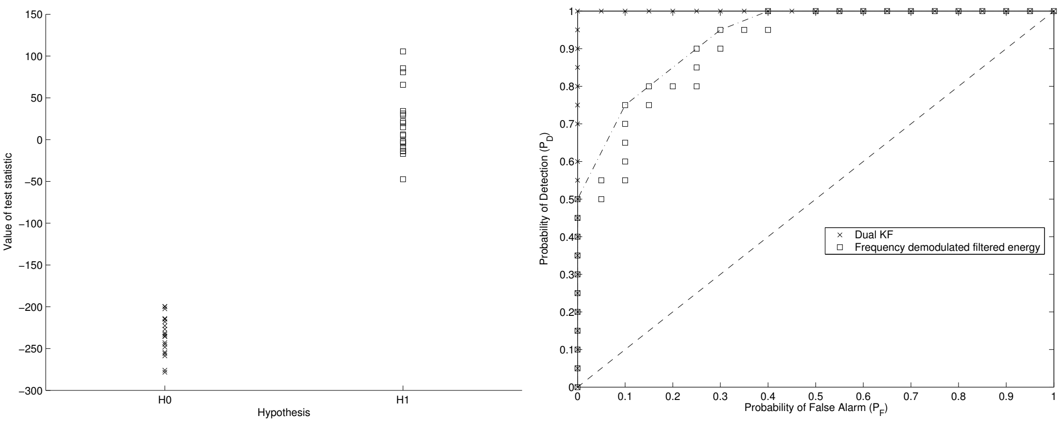
(a) Test statistic for the FDFE detector

(b) Test statistic for the KF/GLR innovations energy detector

Figure 3.8: Test statistics of the KF/GLR innovations energy detector vs. the FDFE detector for $k = 10^{-4}$ N/m.

hypothesis more dissimilar than under the H_0 hypothesis.

The ROC curve for the KF/GLR innovations detector is plotted in Fig. 3.9b. As in the case of $k = 10^{-3}$ N/m, the test accurately distinguishes between the spin and no-spin realizations. The test statistic values of the KF/GLR innovations detector is illustrated in Fig. 3.9a. There is a wide separation in the values between the two hypotheses. The KF/GLR innovations detector again has the best performance for this case. We see that knowledge of the variance of the innovations $\eta_1[i]$ under the H_1 hypothesis is critical in the performance of the KF/GLR detector. Once that information is discarded, we are left with the innovations energy detector. The simulations have demonstrated that it is comparable to the FDFE detector for the parameters of Table 3.2 and $k = 10^{-3}, 10^{-4}$ N/m. As the FDFE detector is more computationally inexpensive, it would be preferred over the KF/GLR innovations energy detector.



(a) Test statistic for the KF/GLR innovations detector

(b) ROC curve

Figure 3.9: Comparison of the KF/GLR innovations detector vs. FDFE detector for $k = 10^{-4}$ N/m. The KF/GLR innovations detector accurately distinguishes between the spin present and spin absent realizations.

3.3 Conclusion

The optimal detection test for the two CT models if their observations were continuously available is given by (3.1). The optimal solution, however, is not finite dimensional, and its exact implementation is computationally prohibitive. In the continuous-time classical model, which is the main focus of this chapter, the observations are sampled. Consequently, the LRT in (3.1) cannot be used.

A filter is heuristically derived for a class of "soft" nonlinear systems, of which the nonlinear CTC model is one of them. The result, called the piecewise linear Kalman filter, is a filter whose state propagation step is similar to EKF, but whose error covariance propagation step is similar to the KF. It can be regarded as a 0th order EKF. For the CTC model, the PLKF's error covariance propagation step was solvable in closed form. In contrast, EKF had to solve equations for the propagation of its error covariance matrix. Consequently, PLKF's per iteration runtime was shorter than EKF's.

With a filter for the nonlinear CTC model, we applied the piecewise linear idea and used the KF detector, which optimally distinguishes between two linear models with zero mean observations. The PLKF was substituted for KF1, while an ordinary KF was substituted for KF0; see Fig. 3.1. It should be noted that the model describing the no-spin hypothesis is linear. Unfortunately, the PLKF requires initial conditions for the spin, which are not available. One observes noisy versions of $z(t)$ and $\dot{z}(t)$, and so can form rough estimates. The Generalized Likelihood principle was applied to the KF detector; the result was a KF/GLR innovations detector with several PLKF branches matched to the spin present model. Each PLKF is initialized with a different initial spin value; in essence, we are attempting to "guess" the initial spin value.

The KF/GLR innovations detector requires knowledge of the variance of two innovations: the innovations produced by KF0 under H_0, and the innovations produced by the PLKF under H_1. The variance of the first innovations sequence is straightforward to compute. However, while the variance of the second innovations sequence is computable in theory, it is difficult in practice. As an approximation, one could obtain it empirically or assume that the variance is equal to the variance of the first innovations sequence. A simpler KF/GLR detector was formulated that assumed that the variance of both innovations were approximately constant and equal to each other. This resulted in a test that just involved the sum of squares of the innovations produced by KF0 and the other PLKFs. To differentiate both versions, the original detector was called the KF/GLR innovations detector, while the simpler detector was called the KF/GLR innovations energy detector.

Simulations were presented comparing PLKF vs. EKF. For the parameter set given in Table 3.1, the PLKF has lower MSE than the EKF in three out of the four state variables considered. In particular, it has innovations that are smaller in mean-squared value. This indicates that the filter is better at tracking the nonlinear CTC model, which is essential if the innovations KF/GLR detector is to work.

In the next set of simulations, the following three detectors were compared: the FDFE statistic, the KF/GLR innovations detector, and the KF/GLR innovations energy detector. The parameter set given in Table 3.2 was used, and two different values of k were considered. In both cases, the KF/GLR innovations detector had the best performance, and was able to accurately differentiate the H_0 and H_1 models. The KF/GLR innovations energy detector was comparable to the FDFE detector; as the latter is simpler to implement, it would be preferable. The simulations indicate that knowledge of the variance of the innovations is crucial to the performance of the KF/GLR detector.

CHAPTER IV

Detection of the single spin in the discrete-time models

The detectors considered here can be placed into three categories: versions of existing detectors that are currently in use for MRFM; LRTs; and approximations to the LRT for the DT random telegraph (DTRT) and DT random walk (DTRW) model.

The LRT is a most powerful test that satisfies the Neyman-Pearson criterion: it maximizes the probability of detection (P_D) subject to a constraint on the probability of false alarm (P_F) [70], which is set by the user. Consequently, it can be used as a benchmark with which to compare the other detectors.

Firstly, a recursive form of the LRT for a general DT Markov process whose observations are perturbed by AWGN is presented. The derivation comes from [57]. We then extend it to derive a closed form of the LRT and also show that, under the regime of low SNR, the LRT is approximately the matched filter statistic with the one-step MMSE estimator used in place of the known signal values.

Next, we discuss the detectors that are currently used in MRFM experiments. Approximations to the LRT for the DTRT model are then introduced: these will be of the form of a filtered energy statistic. Lastly, we ascertain sufficient conditions under which a certain class of DTRW models admits a filtered energy (FE) statistic approximation.

4.1 Likelihood ratio test for a Markov process in AWGN

4.1.1 Description

In this section, we shall consider a general Markov process and derive its LRT. The formulas that provide an initial starting point are given in [57]. We shall use the notation in [57]: while it is slightly different, the differences are superficial.

The hypothesis test is between $H_0 : z_k = w_k$ and $H_1 : z_k = x_k + w_k$, where the observations are $(z_k)_{k=0}^{N-1}$. $(X_k)_{k=0}^{N-1}$ is a Markov sequence; let its state space be denoted by $\Psi = \{\psi_1, \ldots, \psi_r\}$, where r is the number of possible values that X_k can assume. Let $\mathbf{P}^{(k)}$ be the probability transition matrix associated with the process X_k at the k-th time step, so that $\mathbf{P}_{ij}^{(k)} = P(X_k = \psi_j | X_{k-1} = \psi_i)$. Let $f_0(\cdot)$ and $f_1(\cdot)$ be pdfs induced under hypothesis H_0 and H_1 respectively. Similarly, let $E_i[\cdot]$ denote the expectation under hypothesis H_i for $i = 0, 1$. The noise is denoted by w_k: for the moment, we shall not specify its p.d.f.

4.1.2 Derivation of the LRT

Define $p_k, q_k, r_k \in \mathbb{R}^r$ and $\mathbf{\Omega}^{(k)} \in \mathbb{R}^{r \times r}$ in the following way:

$$p_k \triangleq [P(X_k = \psi_1), \ldots, P(X_k = \psi_r)]^T$$

$$q_k \triangleq [P(X_k = \psi_1 | z^{k-1}), \ldots, P(X_k = \psi_r | z^{k-1})]^T$$

$$r_k \triangleq [P(X_k = \psi_1 | z^{k}), \ldots, P(X_k = \psi_r | z^{k})]^T$$

$$\mathbf{\Omega}^{(k)} \triangleq \text{diag}\left[\frac{f_1(z_k | X_k = \psi_1, z^{k-1})}{f_1(z_k | z^{k-1})}, \ldots, \frac{f_1(z_k | X_k = \psi_r, z^{k-1})}{f_1(z_k | z^{k-1})}\right] \tag{4.1}$$

where the vector notation is dropped for the sake of brevity. The variable p_k is the probability of states at the kth time; q_k is the one-step prediction probabilities for time k; r_k consists of the filtered state probabilities at time k.

Proposition 4.1. $p_k^T = p_{k-1}^T \mathbf{P}^{(k)}$

Examine the j-th element of p_k:

$$\begin{aligned} P(X_k = \psi_j) &= \sum_{i=1}^{r} P(X_k = \psi_j | X_{k-1} = \psi_i) P(X_{k-1} = \psi_i) \\ &= p_{k-1}^T \left(\text{j-th column of } \mathbf{P}^{(k)}\right) \quad \blacksquare \end{aligned}$$

Proposition 4.2. $q_k^T = r_{k-1}^T \mathbf{P}^{(k)}$

Examine the j-th element of q_k. By the Markov assumption,

$$\begin{aligned} P(X_k = \psi_j | z^{k-1}) &= \sum_{i=1}^{r} P(X_k = \psi_j | X_{k-1} = \psi_i, z^{k-1}) P(X_{k-1} = \psi_i | z^{k-1}) \\ &= r_{k-1}^T \begin{pmatrix} P(X_k = \psi_j | X_{k-1} = \psi_1) \\ \vdots \\ P(X_k = \psi_j | X_{k-1} = \psi_r) \end{pmatrix} \\ &= r_{k-1}^T \left(\text{j-th column of } \mathbf{P}^{(k)}\right) \quad \blacksquare \end{aligned}$$

Proposition 4.3. $r_k^T = q_k^T \mathbf{\Omega}^{(k)}$

Under H_1 [57] (39):

$$P(X_k = \psi_j | z^k) = \frac{f_1(z_k | X_k = \psi_j, z^{k-1}) P(X_k = \psi_j | z^{k-1})}{f_1(z_k | z^{k-1})} \quad \blacksquare$$

Define the *likelihood ratio statistic* $L_{N-1}(z^{N-1})$ and the *transition likelihood ratio* $l(z_k|z^{k-1})$ as:

$$L_{N-1}(z^{N-1}) = \frac{f_1(z^{N-1})}{f_0(z^{N-1})} = \prod_{k=1}^{N-1} l(z_k|z^{k-1}) \cdot L_0(z^0) \text{where}$$

$$l(z_k|z^{k-1}) = \frac{f_1(z_k|z^{k-1})}{f_0(z_k|z^{k-1})} \tag{4.2}$$

We shall now *specialize* to the case where the noise w_k are independent Gaussian r.v.s with variance R_k. Then,

$$f_1(z_k|X_k = \psi_j, z^{k-1}) \sim \varphi(z_k; \psi_j, R_k)$$

$$f_0(z_k|z^{k-1}) \sim \varphi(z_k; 0, R_k)$$

where $\varphi(x; \mu, \sigma^2) \triangleq \exp[-(x-\mu)^2/2\sigma^2]/\sqrt{2\pi}\sigma$.

Let $\pi_j = P(X_0 = \psi_j)$ for $j = 1, \ldots, r$. Define $n_k, \pi \in \mathbb{R}^r$ and $\mathbf{H}^{(k)} \in \mathbb{R}^{r \times r}$ as:

$$\begin{aligned} n_k &\triangleq [\varphi(z_k; \psi_1, R_k), \ldots, \varphi(z_k; \psi_r, R_k)]^T \\ \pi &\triangleq [\pi_0, \ldots, \pi_r]^T \\ \mathbf{H}^{(k)} &\triangleq \operatorname{diag}(n_k)/\varphi(z_k; 0, R_k) \\ &= \operatorname{diag}\left[\frac{\varphi(z_k; \psi_1, R_k)}{\varphi(z_k; 0, R_k)}, \ldots, \frac{\varphi(z_k; \psi_r, R_k)}{\varphi(z_k; 0, R_k)}\right] \end{aligned} \tag{4.3}$$

From Props. 4.2 and 4.3,

$$q_k^T = r_{k-1}^T \mathbf{P}^{(k)} = q_{k-1}^T \mathbf{\Omega}^{(k-1)} \mathbf{P}^{(k)}. \tag{4.4}$$

But

$$\begin{aligned} \mathbf{\Omega}^{(k-1)} &= \frac{f_0(z_{k-1}|z^{k-2})}{f_1(z_{k-1}|z^{k-2})} \frac{\operatorname{diag}(n_{k-1})}{f_0(z_{k-1}|z^{k-2})} \\ &= \frac{1}{l(z_{k-1}|z^{k-2})} \mathbf{H}^{(k-1)} \\ \Longrightarrow l(z_k|z^{k-1}) &= \frac{q_k^T n_k}{\varphi(z_k; 0, R_k)} \\ &= \frac{q_{k-1}^T \mathbf{H}^{(k-1)} \mathbf{P}^{(k)} n_k}{\varphi(z_k; 0, R_k)} \cdot \frac{1}{l(z_{k-1}|z^{k-2})} \end{aligned} \tag{4.5}$$

We notice that the $(k-1)$-th transition likelihood ratio appears in the denominator of the RHS expression of (4.5). This suggests a cancellation effect when forming $L_{N-1}(z^{N-1})$; see (4.2). Indeed, that is the case. One obtains:

$$L_{N-1}(z^{N-1}) = \pi^T \mathbf{H}^{(0)} \mathbf{P}^{(1)} \mathbf{H}^{(1)} \ldots \mathbf{H}^{(N-1)} \underline{1} \tag{4.6}$$

This is a nice, compact expression for the LRT of a general DT Markov process.

A recursive form of the LRT can be written based on the previous results. Let $q_0 = p_0$, as the definition of q_k only makes sense for $k \geq 1$. Then, $L_0(z^0) = f_1(z_0)/f_0(z_0) = q_0^T n_0/\varphi(z_0; 0, R_0)$, which is the same form that $l(z_k|z^{k-1})$ assumes in (4.5). The log LRT is

$$\log L_{N-1}(z^{N-1}) = \sum_{k=0}^{N-1} \log \frac{q_k^T n_k}{\varphi(z_k; 0, R_k)} \tag{4.7}$$

The q_k's can be computed recursively by using (4.4) and noticing that $\mathbf{\Omega}^{(k)} = \text{diag}(n_k)/(q_k^T n_k)$. It follows that

$$q_k^T = q_{k-1}^T \frac{\text{diag}(n_{k-1})}{q_{k-1}^T n_{k-1}} \mathbf{P}^{(k)} \tag{4.8}$$

4.1.3 The LRT under low SNR

Let us consider the log LRT: from (4.2) and (4.5),

$$\log L_{N-1}(z^{N-1}) = \sum_{k=1}^{N-1} \log l(z_k|z^{k-1}) + \log \left(\frac{f_1(z_0)}{f_0(z_0)} \right)$$

Each transition likelihood ratio can be simplified as follows:

$$\begin{aligned} l(z_k|z^{k-1}) &= \sum_{i=1}^{r} P(X_k = \psi_i|z^{k-1}) \exp\left[-\frac{1}{2R_k}(-2z_k\psi_i + \psi_i^2) \right] \\ &\approx \sum_{i=1}^{r} P(X_k = \psi_i|z^{k-1}) \left(1 + \frac{1}{R_k} z_k \psi_i - \frac{1}{2R_k}\psi_i^2 \right) \\ &= 1 + \frac{1}{R_k} z_k E_1[X_k|z^{k-1}] - \frac{1}{2R_k} E_1[X_k^2|z^{k-1}] \end{aligned} \tag{4.9}$$

where the approximation $e^\delta \approx 1 + \delta$ for small δ. Next, we shall use the approximation $\log(1+\delta) \approx \delta$ for small δ. This is justified if the SNR is low so that $|\psi_i/\sqrt{R_k}| \ll 1$ for all $i = 1, \ldots, r$. So

$$\log l(z_k|z^{k-1}) \approx \frac{1}{R_k} z_k E_1[X_k|z^{k-1}] - \frac{1}{2R_k} E_1[X_k^2|z^{k-1}], \quad k \geq 1 \tag{4.10}$$

As well, the same approximation can be applied to $\log L_0(z^0)$, so that

$$\log L_0(z_0) \approx \frac{1}{R_0} z_0 E[X_0] - \frac{1}{2R_0} E[X_0^2]. \tag{4.11}$$

Define the MMSE estimator of X_k as follows: $\hat{x}_k = E_1[X_k|z^{k-1}]$ for $k \geq 1$ and $\hat{x}_0 = E[X_0]$. Use a similar notation for X_k^2, so that $\widehat{x_0^2} = E[X_0^2]$ and $\widehat{x_k^2} = E_1[X_k^2|z^{k-1}]$ for $k \geq 1$. Using (4.10) and (4.11), the log LRT can be approximately written under low SNR as

$$\log L_{N-1}(z^{N-1}) \approx \sum_{k=0}^{N-1} \frac{1}{R_k} z_k \hat{x}_k - \frac{1}{2} \sum_{k=0}^{N-1} \frac{1}{R_k} \widehat{x_k^2} \tag{4.12}$$

The right hand side of (4.12) is similar to the matched filter statistic, but with the MMSE estimates of X_k and X_k^2 used instead. The CT analog was discussed in the previous chapter—see (3.1). There is a noteworthy difference: in the second term, the expected value of X_k^2 conditioned on the past observations is used vs. the square of the expected value of X_k conditioned on the past observations. In general, for $k \geq 1$, $E_1[X_k^2|z^{k-1}] \neq (E_1[X_k|z^{k-1}])^2$. Indeed, for a r.v. X,

$$E[X^2] = (E[X])^2 \text{ iff } \operatorname{var}(X) = 0, \tag{4.13}$$

By the Chebyshev inequality, for $\delta > 0$, $P[|X - E[X]| \geq \delta] \leq \operatorname{var}(X)/\delta^2 = 0$. So (4.13) holds iff X is some value $c \in \mathbb{R}$ w.p. 1. As a result $E_1[X_k^2|z^{k-1}] = (E_1[X_k|z^{k-1}])^2$ iff X_k is a function of z^{k-1} w.p. 1.

The result (3.1) in CT is exact, while in the DT case, we have only shown that (4.12) approximately holds under low SNR.

In [58], it is shown that in detecting a finite state Markov process, the LR is in general not expressible as the known form LR with an estimator of x_k, i.e., the RHS of (4.12). For the detection of Gauss-Markov processes, [58] derives a "locally stable estimator" of x_k such that the LR can be expressed as a known form LR. A r-dimensional Gauss-Markov process X_k is one that is generated by the following equation

$$X_k = \mathbf{A}X_{k-1} + W_k,\ k \geq 1 \tag{4.14}$$

where $\mathbf{A} \in \mathbb{R}^{r \times r}$, $W_k \sim \varphi(w_k; 0, \mathbf{Q}_k)$ are independent Gaussian random vectors, and $X_0 \sim \varphi(x_0; 0, \mathbf{P}_0)$ for some positive definite matrix $\mathbf{P}_0$.

4.2 Detectors currently used in MRFM experiments: the amplitude, energy, and filtered energy detectors

The DT amplitude detector is

$$\left| \frac{1}{N} \sum_{i=0}^{N-1} z_i \right| \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.15}$$

where η is set to satisfy the constraint on P_F. This is the optimal test under the assumption that z_i is the sum of an unknown constant and AWGN. This assumption would be true if there were no random spin flips. In this case, the amplitude detector is simply a MF detector. However, as the number of random transitions in z_i increases, the performance of the amplitude detector degrades. An alternative test statistic is the DT signal energy, i.e., the sum of the squares of the z_i instead of the magnitude of the sum in (4.15). As the signal and noise are assumed to be independent, under hypothesis H_1, one would expect $\underline{z} = [z_0, \ldots, z_{N-1}]^T$ to have a higher energy on average than under hypothesis H_0. This can be reliably detected under a sufficiently high SNR. A natural improvement to the energy detector is to reject out-of-band noise by prefiltering $\underline{z}$ over the signal passband. As the signal X_i is baseband, a LPF is appropriate. In particular, one might use a simple first-order, single-pole filter given by

$$H_{\mathrm{LP}}(z) = \frac{1-\alpha}{2}\frac{1+z^{-1}}{1-\alpha z^{-1}} \tag{4.16}$$

where we require $|\alpha| < 1$ for stability [47]. The time constant α should be chosen based on the bandwidth of the signal; if ω_c is the desired -3 dB bandwidth of the filter, one should set $\alpha = (1 - \sin\omega_c)/\cos\omega_c$. The -3 dB bandwidth depends on the mean number of transitions, i.e., $(1-p)/T_s$ for the DT random telegraph model. Note that this LPF is the same that was used to benchmark the KF/GLR detector for the CTC model, cf. (3.3). When the transition probabilities are symmetric, $(1-p)/T_s$ corresponds to the mean number of transitions per second λ of the CT random telegraph model. Given a value $p = q = p_0$, one can equate the expected number of transitions in both DT and CT models to obtain $\lambda = (1-p_0)/T_s$. In practice, since the mean number of transitions is only approximately known to the experimenter, a bank of LPFs with different α's are used to perform detection [54]. The energy and filtered energy detector can then be expressed as

$$\sum_{i=0}^{N-1}(z*h)_i^2 \;\underset{H_0}{\overset{H_1}{\gtrless}}\; \eta \tag{4.17}$$

where "$*$" represents the convolution operator. For the energy detector, h_i is taken to be the unit impulse function $\delta[i]$, while for the filtered energy detector, $h_i = h_{\mathrm{LP}}[i]$, the impulse response of $H_{\mathrm{LP}}(z)$ in (4.16).

Note that the computational complexity for the amplitude, filtered energy, and energy detectors is $\mathcal{O}(N)$.

4.3 LRT for the Discrete-time random telegraph model

4.3.1 The LRT

One can use the formulas (4.7) and (4.8) and specialize to the DTRT model. For $\psi \in \Psi$, let $q_k(\psi) = P(X_k = \psi | z^{k-1})$ for $k \geq 1$ and $q_0(\psi) = P(X_0 = \psi)$. Applying the aforementioned equations results in the LRT being

$$\Lambda_{\text{rt}} = \prod_{i=0}^{N-1} \left[q_i(A) e^{\frac{A}{\sigma^2} z_i} + q_i(-A) e^{-\frac{A}{\sigma^2} z_i} \right] \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{4.18}$$

where $q_0(A) = q_0(-A) = 1/2$ and $q_i(A)$ and $q_i(-A)$ can be computed for $i \geq 1$ by using

$$\begin{pmatrix} q_i(-A) \\ q_i(A) \end{pmatrix} = \mathbf{P}_{\text{rt}}^T \begin{pmatrix} 1 - \# \\ \# \end{pmatrix} \quad \text{where}$$

$$\# = \frac{e^{\frac{A}{\sigma^2} z_{i-1}} q_{i-1}(A)}{e^{\frac{A}{\sigma^2} z_{i-1}} q_{i-1}(A) + e^{-\frac{A}{\sigma^2} z_{i-1}} q_{i-1}(-A)} \tag{4.19}$$

Note that $q_0(A) = q_0(-A) = 1/2$ arises from the fact that we assume the initial spin state is equally likely to be either $\pm A$. From (4.18) and (4.19), the running time of the LRT for the DTRT is $\mathcal{O}(N)$.

Under low SNR conditions ($|A/\sigma| \ll 1$), the log LRT becomes

$$\begin{aligned} \log \Lambda_{\text{rt}} &\approx \sum_{i=0}^{N-1} \log \left[q_i(A)(1 + \frac{A}{\sigma^2} z_i) + q_i(-A)(1 - \frac{A}{\sigma^2} z_i) \right] \\ &= \sum_{i=0}^{N-1} \log \left[1 + \frac{1}{\sigma^2} z_i (A q_i(A) - A q_i(-A)) \right] \end{aligned}$$

where we use the approximation $e^{\delta} \approx 1 + \delta$ for small δ. Now, $Aq_0(A) - Aq_0(-A) = E_1[X_0] = \hat{x}_0$ and $Aq_i(A) - Aq_i(-A) = E_1[X_i | z^{i-1}] = \hat{x}_i$ for $i \geq 1$, i.e., it is the MMSE predictor of X_i given the past observations.

$$\therefore \log \Lambda_{\text{rt}} \approx \sum_{i=0}^{N-1} \log \left(1 + \frac{1}{\sigma^2} z_i \hat{x}_i \right) \approx \frac{1}{\sigma^2} \sum_{i=0}^{N-1} z_i \hat{x}_i \tag{4.20}$$

where the approximation $\log(1+\delta) \approx \delta$ for small δ was used. Under low SNR, the LRT is effectively correlating the observations z_i with the MMSE predictor of X_i. This is the matched filter statistic with the MMSE predictor substituted in place of the "known" values of X_i, and is known as the estimator-correlator detector. In particular, the estimator-correlator structure is known to be optimal for Gaussian signals in AWGN [58].

This result is consistent with the approximation (4.12) that was earlier derived for a general Markov process in AWGN. Since $|\psi_1| = |\psi_2| = A$, $\widehat{x_i^2} = A^2 \ \forall \, i$. So the second term in (4.12) will become a constant independent of the observations and can be omitted.

4.3.2 Approximate second order expansion of the LRT

Under the regime of low SNR, long observation time ($N \gg 1$), and the probability of transition between consecutive samples is small ($p + q \approx 2$), the second order expansion of $\log \Lambda_{\mathrm{rt}}$ is approximately equal to the hybrid detector with test statistic

$$\sum_i (z_i * h_{\mathrm{LP}}[i])^2 + K_a \sum_i z_i + K_e \sum_i z_i^2 \tag{4.21}$$

where $K_a = K_a(p, q, A, \sigma)$ and $K_b = K_e(p, q)$ are constants independent of the data. Therefore, in the aforementioned regime, one expects the hybrid detector to have performance similar to the optimal LRT. When $p = q$, the second order expansion of the LRT is approximately equal to the filtered energy detector considered in (4.17). See Appendix A for more details. In light of the computation complexities for the filtered energy, energy, and amplitude statistics, the complexity of the hybrid detector is also $\mathcal{O}(N)$.

4.4 LRT for the Discrete-time random walk model

Consider a random walk process with r states $\psi_1, \ldots, \psi_r$. Let $\mathbf{P}$ be its probability transition matrix. Let the set of all such random walks (equivalently the set of all such matrices $\mathbf{P}$) be denoted by $\mathcal{R}^r$. In this section, we shall only consider random walks in a subset $\mathcal{R}_0^r \subseteq \mathcal{R}^r$ that have the following properties:

1. it has no self-loops, i.e., it is not possible to remain in the same state for two consecutive times $\iff$ $p_{ii} = 0$ for all $1 \leq i \leq r$
2. when in state ψ_k, $1 < k < r$, it is possible to reach either ψ_{k-1} or ψ_{k+1} with non-zero probability $\iff$ $p_{i,i-1} > 0$ and $p_{i,i+1} > 0$ for all $1 < i < r$
3. when in state ψ_1 (ψ_r), the random walk must proceed to ψ_2 (ψ_{r-1}), i.e., "reflecting" boundary conditions $\iff$ $p_{1,2} = 1$ and $p_{r,r-1} = 1$
4. there exists a stationary probability distribution $p_{\text{ss}} \in \mathbb{R}^r$ such that (s.t.) $\lim p_j = p_{\text{ss}}$

As with the DTRT model, the LRT for the DTRW model can be written by applying the formulas (4.7) and (4.8). Let Λ_{rw} denote the LRT for the DTRW model.

We will obtain an approximate form for the LRT of random walks in $\mathcal{R}_0 = \bigcup_{r=2}^{\infty} \mathcal{R}_0^r$ under the same two conditions considered for the random telegraph, i.e., the regime of low SNR and long observation time. The notation from the previous section will be retained. The matrix results used in this section are covered in Appendix B.

4.4.1 First and second-order moments

First moment

Let us compute the first moment of the observations Z_j under the H_1 hypothesis. Now, $E_1[Z_j] = E[X_j + W_j] = E[X_j]$ since the noise process is zero-mean. But $E[X_j] = \sum_{n=1}^{r} P(X_j = \psi_n)\psi_n = p_j^T \mathbf{M}_\psi \underline{1}$, where $\mathbf{M}_\psi \triangleq \text{diag}(\psi_1, \ldots, \psi_r)$. Also, $p_j^T = \pi^T \mathbf{P}^j$ for $j \geq 0$. Putting it together:

$$E_1[Z_j] = \pi^T \mathbf{P}^j \mathbf{M}_\psi \underline{1} \text{ for } j \geq 0 \tag{4.22}$$

Note that the first moment, in general, depends on the time index j. Consequently, it is not stationary. By assumption, however, $\lim p_j = p_{\text{ss}}$. So $\lim E_1[Z_j] = p_{\text{ss}}^T \mathbf{M}_\psi \underline{1}$. In other words, the first moment is approximately stationary if we wait for a while after the start of the process. In addition, $\lim E_1[Z_j] = 0 \iff p_{\text{ss}}^T \mathbf{M}_\psi \underline{1} = 0$.

Second moment

Here, evaluate $E[Z_j Z_k]$. Consider two cases.

Case 1: $j = k$. Then, $E_1[Z_j^2] = E[(X_j + W_j)^2] = E[X_j^2] + E[W_j^2]$ since the noise process is independent of the random walk and is zero-mean. Assume that $R_j = \sigma^2$ for all j. Then, $E_1[Z_j^2] = E[X_j^2] + \sigma^2$. Now, $E[X_j^2] = \sum_{n=1}^{r} P(X_j = \psi^n)\psi_n^2 = p_j^T \mathbf{M}_\psi^2 \underline{1}$, which results in:

$$E_1[Z_j^2] = p_j^T \mathbf{M}_\psi^2 \underline{1} + \sigma^2 = \pi^T \mathbf{P}^j \mathbf{M}_\psi^2 \underline{1} + \sigma^2$$

Case 2: $j \neq k$. Suppose for now that $j < k$. Then, $E_1[Z_j Z_k] = E[X_j X_k]$, since the noise process is independent and identically distributed (i.i.d.). Using conditional expectation, one can show that when $j < k$, $E[X_j X_k] = p_j^T \mathbf{M}_\psi \mathbf{P}^{k-j} \mathbf{M}_\psi \underline{1}$. So $E_1[Z_j Z_k] = \pi^T \mathbf{P}^j \mathbf{M}_\psi \mathbf{P}^{k-j} \mathbf{M}_\psi \underline{1}$. In general then,

$$E_1[Z_j Z_k] = \pi^T \mathbf{P}^{\min(j,k)} \mathbf{M}_\psi \mathbf{P}^{|j-k|} \mathbf{M}_\psi \underline{1}$$

Combining the results of both cases:

$$E_1[Z_j Z_k] = \begin{cases} \pi^T \mathbf{P}^j \mathbf{M}_\psi^2 \underline{1} + \sigma^2 & j = k \\ \pi^T \mathbf{P}^{\min(j,k)} \mathbf{M}_\psi \mathbf{P}^{|j-k|} \mathbf{M}_\psi \underline{1} & j \neq k \end{cases} \tag{4.23}$$

We see that the second moment of the observations is not stationary either. However, using $\lim p_j = p_{\text{ss}}$, we obtain:

$$\lim_{\min(j,k)\to\infty} E_1[Z_j Z_k] = \begin{cases} p_{\text{ss}}^T \mathbf{M}_\psi^2 \underline{1} + \sigma^2 & j = k \\ p_{\text{ss}}^T \mathbf{M}_\psi \mathbf{P}^{|j-k|} \mathbf{M}_\psi \underline{1} & j \neq k \end{cases} \tag{4.24}$$

As with the first moment, the second moment is approximately stationary if we wait for a while after the process starts.

4.4.2 Approximate second order expansion of the LRT

Start with (4.6) from the previous section, which is:

$$L_{N-1}(z^{N-1}) = \pi^T \mathbf{H}^{(0)} \mathbf{P}^{(1)} \mathbf{H}^{(1)} \ldots \mathbf{P}^{(N-1)} \mathbf{H}^{(N-1)} \underline{1}$$

Since we are considering the random walk whose transition matrix does not change over time: $\mathbf{P}^{(1)} = \ldots = \mathbf{P}^{(N-1)} = \mathbf{P}$. Assume that $R_k = \sigma^2$ for $0 \leq k \leq N-1$. Then,

$$\begin{aligned} \mathbf{H}^{(k)} &= \operatorname{diag}\left[\frac{\mathcal{N}(z_k; \psi_1, R_k)}{\mathcal{N}(z_k; 0, R_k)}, \ldots, \frac{\mathcal{N}(z_k; \psi_r, R_k)}{\mathcal{N}(z_k; 0, R_k)}\right] \\ &= \operatorname{diag}\left[\exp\left(\frac{2z_k\psi_1 - \psi_1^2}{2\sigma^2}\right), \ldots, \exp\left(\frac{2z_k\psi_r - \psi_r^2}{2\sigma^2}\right)\right] \\ &\approx \operatorname{diag}\left[e^{-\psi_1^2/2\sigma^2}\left(1 + \psi_1 \frac{z_k}{\sigma^2}\right), \ldots, e^{-\psi_r^2/2\sigma^2}\left(1 + \psi_r \frac{z_k}{\sigma^2}\right)\right] \end{aligned} \tag{4.25}$$

where the last statement is justified by using the low SNR assumption so that $|\frac{\psi_j}{\sigma}| \ll 1$ for $1 \leq j \leq r$ and applying the approximation $e^\delta \approx 1 + \delta$ for small δ. Define the matrices

$$\mathbf{M}_{\psi 1} \triangleq \operatorname{diag}(e^{-\psi_1^2/2\sigma^2}, \ldots, e^{-\psi_r^2/2\sigma^2})$$

$$\mathbf{M}_{\psi 2} \triangleq \frac{1}{\sigma}\mathbf{M}_\psi = \operatorname{diag}(\frac{\psi_1}{\sigma}, \ldots, \frac{\psi_r}{\sigma})$$

Using (4.25), $\mathbf{H}^{(k)} \approx \mathbf{M}_{\psi 1}(\mathbf{I} + \frac{z_k}{\sigma}\mathbf{M}_{\psi 2})$. If we define

$$\mathbf{Q} \triangleq \mathbf{P}\mathbf{M}_{\psi 1} \approx \mathbf{P}\left(\mathbf{I} - \frac{1}{2\sigma^2}\mathbf{M}_{\psi}^2\right) \tag{4.26}$$

$$\mathbf{R} \triangleq \mathbf{P}\mathbf{M}_{\psi 1}\mathbf{M}_{\psi 2} = \mathbf{Q}\mathbf{M}_{\psi 2} \tag{4.27}$$

then $\mathbf{P}\mathbf{H}^{(k)} \approx \mathbf{Q} + \frac{z_k}{\sigma}\mathbf{R}$.

Some comments are in order regarding the matrices $\mathbf{P}$ and $\mathbf{Q}$ before proceeding. We shall see in the following equations that both these matrices play a crucial role in the approximate form of the LRT. The reason why we have restricted the analysis to a certain subset of probability transition matrices $\mathbf{P}$ is that its eigendecomposition is well characterized. Because $\mathbf{Q}$ assumes the same structure as $\mathbf{P}$, it too is nicely characterized. Two of its properties are briefly mentioned: first, if the additional assumption that $\psi_2, \psi_{r-1} \neq 0$, then by Prop. B.13, the eigenvalues of $\mathbf{Q}$ are strictly less than 1 in absolute value. Second, by Prop. B.14, the eigenvalues of $\mathbf{Q}$ can be made arbitrarily close to $\mathbf{P}$ by decreasing the SNR. This is not a surprising result, as $\mathbf{M}_{\psi 1} \to \mathbf{I}$ as the SNR decreases to zero.

For $N \geq 3$,

(4.28)

$$\begin{aligned}
L_{N-1}(z^{N-1}) &\approx \pi^T \mathbf{H}^{(0)} \prod_{j=1}^{N-1} \left(\mathbf{Q} + \frac{z_j}{\sigma}\mathbf{R}\right) \underline{1} \\
&\approx \pi^T \mathbf{H}^{(0)} \Bigg\{ \mathbf{Q}^{N-1} + \frac{1}{\sigma} \sum_{j} z_j \mathbf{Q}^{j-1}\mathbf{R}\mathbf{Q}^{N-1-j} \\
&\qquad + \frac{1}{\sigma^2} \sum_{j<k} z_j z_k \mathbf{Q}^{j-1}\mathbf{R}\mathbf{Q}^{k-1-j}\mathbf{R}\mathbf{Q}^{N-1-k} \Bigg\} \underline{1} \\
&= \pi^T \mathbf{H}^{(0)} \mathbf{Q}^{N-1} 1 + \frac{1}{\sigma} \sum_{j=1}^{N-1} z_j \pi^T \mathbf{H}^{(0)} \mathbf{Q}^{j-1}\mathbf{R}\mathbf{Q}^{N-1-j} \underline{1} \\
&\qquad + \frac{1}{\sigma^2} \sum_{1 \leq j < k \leq N-1} z_j z_k \pi^T \mathbf{H}^{(0)} \mathbf{Q}^{j-1}\mathbf{R}\mathbf{Q}^{k-1-j}\mathbf{R}\mathbf{Q}^{N-1-k} \underline{1} + \text{ h.o.t.}
\end{aligned}$$

Using the approximation for $\mathbf{H}^{(0)}$ in (4.28) and re-grouping:

(4.29)

$$\begin{aligned} L_{N-1}(z^{N-1}) \approx \pi^T \mathbf{M}_{\psi 1} \mathbf{Q}^{N-1} \underline{1} + \frac{z_0}{\sigma} \pi^T \mathbf{M}_{\psi 2} \mathbf{Q}^{N-1} \underline{1} + \sum_{j=1}^{N-1} \frac{z_j}{\sigma} \pi^T \mathbf{M}_{\psi 1} \mathbf{Q}^{j-1} \mathbf{R} \mathbf{Q}^{N-1-j} \underline{1} \\ + \sum_{j=1}^{N-1} \frac{z_0 z_j}{\sigma^2} \pi^T \mathbf{M}_{\psi 2} \mathbf{Q}^{j-1} \mathbf{R} \mathbf{Q}^{N-1-j} \underline{1} + \\ + \sum_{1 \le j < k \le N-1} \frac{z_j z_k}{\sigma^2} \pi^T \mathbf{M}_{\psi 1} \mathbf{Q}^{j-1} \mathbf{R} \mathbf{Q}^{k-1-j} \mathbf{R} \mathbf{Q}^{N-1-k} \underline{1} + \text{ h.o.t.} \end{aligned}$$

The first term in (4.29) is a constant and plays no role in the test statistic. Consequently, we can ignore it. From here onwards, only the terms of second-order or less are retained.

One crucial tool that will be used here is the eigendecomposition of $\mathbf{Q}$. Under the small SNR assumption, $|\frac{\psi_j}{\sigma}| \ll 1 \Rightarrow \psi_j^2/2\sigma^2 < 1$ for all $1 \le j \le r$. Apply Prop. B.6 to $\mathbf{Q}$ with $\mathbf{D}_\delta = \frac{1}{2\sigma^2}\mathbf{M}_\psi^2$. Then, $\mathbf{Q}$ has eigenvalues $\kappa_1, \ldots, \kappa_r$ which satisfies the first two statements of Prop. B.5. As a result, $\mathbf{Q}$ is diagonalizable. Consequently, we can write $\mathbf{Q} = \mathbf{U}_Q \mathbf{\Lambda}_Q \mathbf{U}_Q^{-1}$, where $\mathbf{U}_Q$ is a matrix which contains the eigenvectors of $\mathbf{Q}$ and $\mathbf{\Lambda}_Q = \text{diag}(\kappa_1, \ldots, \kappa_r)$, where $\kappa_1 > \ldots > \kappa_r$. Let $\kappa'_n \triangleq \kappa_n/\kappa_1$. We can do this since $\kappa_1 \neq 0$ by the second statement of Prop. B.5.

The key result is as follows. Suppose $\lim E[X_j] = 0$, and $c_{jk} = \pi_\alpha^T \mathbf{\Upsilon}[j,k] d$ is approximately a function of $(k-j)$, where $\mathbf{\Upsilon}[j,k]$ is defined in (C.6). Then, the LRT is approximately

$$L_{N-1}(z^{N-1}) \approx \sum_{n=1}^{r} \sum_{j<k} B_n (\kappa'_n)^{k-j} z_j z_k \tag{4.30}$$

for some constants $B_n, n = 1, \ldots, r$. The filters for $n = 2, \ldots, (r-1)$ can be approximated by the FE statistic given by (4.17), while the filters for $n = 1, r$ can be generated as second order polynomials in z_i. For $n \in \{1, r\}$, $|\kappa'_n| = 1$. The reader is referred to Appendix C for more details.

4.4.3 Discussion and comparison to the filtered energy detector

The conditions under which the random walk LRT can be realized by FE statistics are discussed here. From (C.2) and (C.3), the first order terms of the random walk LRT are unaccounted for in the filtered energy statistic. For it to “disappear”, we require that $\lim E_1[Z_j] = 0 \iff \lim E[X_j] = 0$, i.e., in steady-state, the expected value of the random walk is zero. In Appendix C, we derived conditions for the second order terms of the LRT to be approximated by a bank of filtered energy statistics as well as by a single filtered energy statistic.

Necessary conditions for the FE statistic to approximate the LRT for a random walk in $\mathcal{R}_0$ in the regime of low SNR and large N are:

1. $\lim E_1[X_j] = 0$; if the states are symmetric about zero, then $\lim E_1[X_j] = 0$ iff the steady-state probability distribution is symmetric about zero
2. the coefficient c_{jk} must be well approximated by an exponential function of the form $C\alpha^{k-j}$ for some $C, \alpha \in \mathbb{R}$

Define

$$M(u) \triangleq \sum_{i=1}^{r} \frac{u\lambda_i}{1-u\lambda_i} w_i, \tag{4.31}$$

where λ_is are the eigenvalues of $\mathbf{P}$ defined according to Prop. B.5, and w_i is given in (C.12). Suppose the necessary two conditions above exist. If the α in the second condition satisfies

$$M(\alpha) \gg \frac{\alpha}{1+\alpha} \lim E[X_j^2] \tag{4.32}$$

then we have sufficient conditions for a single FE statistic to approximate the LRT.

It is perhaps surprising that the filtered energy statistic can, under certain conditions, approximate the LRT of a certain class of random walks. However, in [80], it has been observed that as the SNR approaches 0, linear estimators of the CT (symmetric) random telegraph process in noise have performances that are comparable to non-linear estimators. As estimation is intimately linked with detection—see [34, 35] for CT and under low SNR, (4.12) for DT—this result suggests that simpler detectors can be formulated that are approximately optimal. Indeed, the results that we have presented so far lend strong evidence to this notion.

4.5 Simulations

The objective in this section is to compare the detection methods discussed in the previous section. The class of LRT detectors is optimal for their respective signal models, and provides a good benchmark for comparison. Comparison of the various detectors is done using: (1) ROC curves, each of which is a plot of probability of detection (P_D) vs. probability of false alarm (P_F), and (2) power curves, each of which is a plot of P_D vs. SNR at a fixed P_F. Some of the parameters used in the simulation of the DT random telegraph and random walk models are as follows: $k = 10^{-3}$ Nm^{-1}, $\omega_0 = 2\pi \cdot 10^4$ rads^{-1}, $B_1 = 0.2$ mT, $G = 2 \cdot 10^6$ Tm^{-1}. The sampling period was $T_s = 1$ ms, and signal durations of $T = 60$ s and $T = 150$ s were used. The performance of the detectors varies as a function of T; in general, a larger T results in better performance. Values of T used in iOSCAR MRFM experiments are on the order of tens of hours [54]. Nevertheless, the comparative results obtained from using the two values of T above are representative of larger values. Indeed, our approximations to the optimal detectors improve with increased T.

4.5.1 Discrete-time random telegraph model

First, consider the DT random telegraph. Figure 4.1 depicts the simulated ROC curves at SNR $= -35$ dB, $\lambda = 0.5$ s^{-1}, and with symmetric transition probabilities

$(p = q)$. With $T_s = 1$ ms, this results in $p = q = 0.9995$. We examine the matched filter, DT random telegraph LRT (RT-LRT), filtered energy, hybrid, amplitude, and unfiltered energy detectors. The RT-LRT, filtered energy, and hybrid detector curves are virtually identical, which is consistent with our analysis. The unfiltered energy and amplitude detectors have performance that is poorer than the RT-LRT, as it should be since the RT-LRT is the optimal detector. The unfiltered energy detector has the worst performance out of the five detector methods considered. Lastly, the omniscient MF detector has the best performance.

A power curve was generated over a range of SNRs under the same conditions as before with a fixed $P_F = 0.1$; it is illustrated in Fig. 4.2. For spin detection, an acceptable range for P_F is on the order of 0.05 to 0.1. The RT-LRT, filtered energy, and hybrid detector have similar performance from -30 dB to -45 dB. With this particular value of P_F and λ, the RT-LRT, filtered energy, and hybrid detector perform from 5 dB to 10 dB worse than the MF detector. Although the amplitude detector has worse performance than the RT-LRT and filtered energy detector, all three have comparable performance at -45 dB.

Figure 4.3 shows the power curves generated using the bigger value of $T = 150$ s. Again, the RT-LRT, filtered energy, and hybrid detectors have the same performance from -30 dB to -45 dB. Note that the values of P_D have increased as compared to Fig. 4.2.

The ROC and power curve simulations were repeated with different values of λ, and the same relative performance was observed. In the interest of space, however, they will not be shown. Note that performance degrades as λ increases while T_s is held constant.

In the second set of simulations, we investigate the case in which the transition

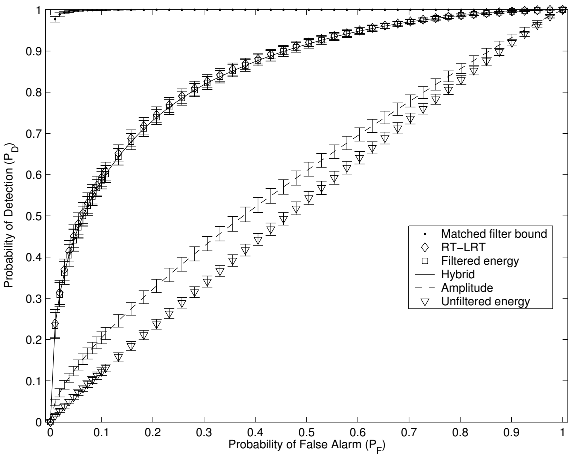


Figure 4.1: Simulated ROC curves for the DT random telegraph model with symmetric transition probabilities at SNR $= -35$ dB, $T = 60$ s, and $\lambda = 0.5$ s^{-1} for the omniscient matched filter, DT random telegraph LRT (RT-LRT), filtered energy, hybrid, amplitude, and unfiltered energy detectors. The RT-LRT is theoretically optimal.

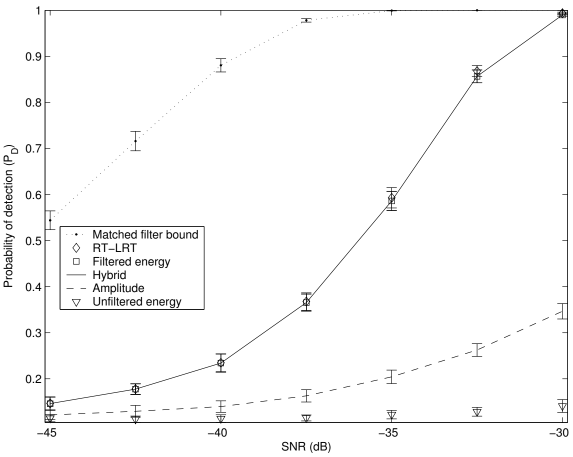


Figure 4.2: Simulated power curves (P_D vs. SNR) for the DT random telegraph model with P_F fixed at 0.1 and $\lambda = 0.5$ s^{-1}, $T = 60$ s. The RT-LRT is theoretically optimal.

probabilities are asymmetric, i.e., $p \neq q$. Consider the scenario where $p = 0.9998$, $q = 0.9992$, and all of the other parameters values are unchanged. The ROC curves for these parameter values are presented in Fig. 4.4. There is a noticeable difference between the curves of the RT-LRT and filtered energy detectors. The hybrid detector's curve is slightly below that of the LRT, and it is better than that of the filtered energy detector. In fact, the filtered energy detector has worse performance

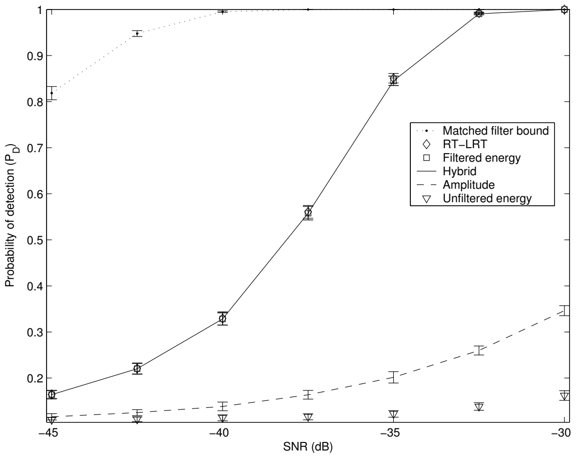


Figure 4.3: Simulated power curves (P_D vs. SNR) for the DT random telegraph model with P_F fixed at 0.1 and $\lambda = 0.5\ \mathrm{s}^{-1}$, $T = 150$ s. The RT-LRT is theoretically optimal.

than the amplitude detector. An asymmetry in p, q leads to a non-zero mean signal, which is why the amplitude detector's performance improves. Indeed, for the DT random telegraph model, $\lim_{i\to\infty} E[X_i] = A\frac{p-q}{2-p-q} = 0.6A$ for the values of p and q used here. Asymmetric transition probabilities can arise in some situations, e.g., the feedback-cooling-of-spins MRFM protocol proposed by Budakian [5].

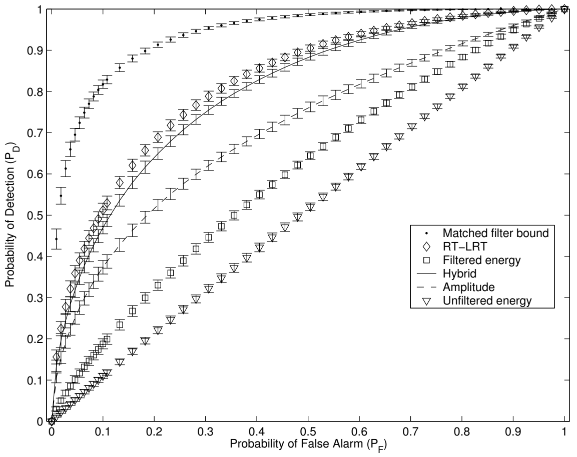


Figure 4.4: Simulated ROC curves for the DT random telegraph model with asymmetric transition probabilities ($p = 0.9998, q = 0.9992$) at SNR $= -45$ dB, $T = 150$ s. The RT-LRT is theoretically optimal.

Power curves from SNR $= -55$ dB to -35 dB were generated for the asymmetric case in Fig. 4.5. A larger value of T is required when $p \neq q$ for the hybrid detector to be a good approximation to the optimal LRT; hence, $T = 150$ s was used for simulations of the asymmetric random telegraph model. The hybrid detector has better performance than the amplitude and filtered energy detectors. It has performance that is comparable to the RT-LRT for lower SNR values.

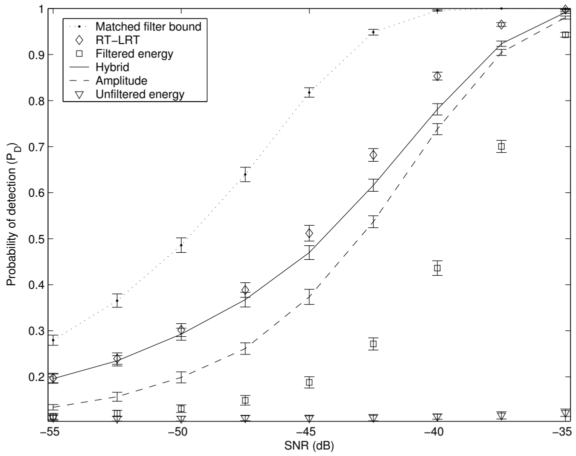


Figure 4.5: Simulated power curves (P_D vs. SNR) for the DT random telegraph model with P_F fixed at 0.1, $p = 0.9998, q = 0.9992$, and $T = 150$ s. The RT-LRT is theoretically optimal.

4.5.2 Discrete-time random walk model

Recall that for the DT random walk model, $\mathbf{P}_{\mathrm{rw}}$ is tridiagonal. For the simulations, a particular subset of tridiagonal matrices was studied. Suppose for the moment that M is even. Recall that the random walk X_i is confined to the interval $[-Ms, Ms]$. Define the lower-quartile transition probabilities as $p_{l,1}, p_{l,2}$ and the upper-quartile transition probabilities as $p_{u,1}, p_{u,2}$. Let $\mathbf{P}_{\mathrm{rw}}^{(j,k)}$ be the (j,k)-th element of $\mathbf{P}_{\mathrm{rw}}$. Here, we examine the performance of the detectors assuming the following reflecting boundary conditions: $\mathbf{P}_{\mathrm{rw}}^{(1,2)} = 1, \mathbf{P}_{\mathrm{rw}}^{(1,i)} = 0$ for $i \neq 2$ and

$\mathbf{P}_{\text{rw}}^{(2M+1,2M)} = 1, \mathbf{P}_{\text{rw}}^{(2M+1,i)} = 0$ for $i \neq 2M$. The rest of $\mathbf{P}_{\text{rw}}$ is

$$\mathbf{P}_{\text{rw}}^{(j,k)} = \begin{cases} p_{l,1} & 2 \leq j < M/2+1, k = j-1 \\ p_{l,2} & 2 \leq j < M/2+1, k = j+1 \\ 0.5 & M/2+1 \leq j \leq 3M/2+1, k = j \pm 1 \\ p_{u,1} & 3M/2+1 < j \leq 2M, k = j-1 \\ p_{u,2} & 3M/2+1 < j \leq 2M, k = j+1 \end{cases} \tag{4.33}$$

Let $\mathbf{A}_n(p_1, p_2)$ be a $n \times (n+2)$ matrix that looks like:

$$\mathbf{A}_n(p_1,p_2) = \begin{pmatrix} p_1 & 0 & p_2 & & & \\ & p_1 & 0 & p_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & p_1 & 0 & p_2 \end{pmatrix}$$

where the unspecified parts of the matrix are taken to be all zeros. In this section, the following subset of transition matrices for the DT random walk was studied:

$$\mathbf{P}_{\text{rw}} = \begin{pmatrix} 0 & 1 & & \\ \mathbf{A}_{\frac{M}{2}-1}(p_{l,1}, p_{l,2}) & & & \\ & \mathbf{F} & & \\ & & \mathbf{A}_{\frac{M}{2}-1}(p_{u,1}, p_{u,2}) & \\ & & 1 & 0 \end{pmatrix},$$

where $\mathbf{F} = \mathbf{A}_{M+1}(0.5, 0.5)$. Note that since each row of a probability transition matrix must sum to 1, one has $p_{l,1} + p_{l,2} = 1$ and $p_{u,1} + p_{u,2} = 1$.

In the case of M odd, the ranges for the indices j, k would change in an obvious way. When $p_{l,1} = p_{u,2} \iff p_{l,2} = p_{u,1}$, we say that the transition probabilities are *symmetric*, and if not, that they are *asymmetric*. The matched filter, DT random walk LRT (RW-LRT), DT random telegraph LRT (RT-LRT), filtered energy, amplitude, and unfiltered energy detectors are compared. In order to run the RT-LRT in the case of the symmetric DT random walk, an average autocorrelation function of the random walk was empirically generated; then p was selected (and one used $q = p$) so that the autocorrelation function of the symmetric DT random telegraph matched the empirical result. From this, the optimal α for the LPF of the filtered energy detector was also obtained.

The ROC curves for two symmetric cases are illustrated in Figs. 4.6 and 4.7. In the former, $p_{l,1} = p_{l,2} = p_{u,1} = p_{u,2} = 0.5$, while in the latter, $p_{l,1} = p_{u,2} = 0.52$ and $p_{l,2} = p_{u,1} = 0.48$. In both cases, the performance of the RW-LRT, RT-LRT, and filtered energy detector are all approximately the same, i.e., the latter two detectors are nearly optimal. When the transition probabilities of the DT random walk are asymmetric, however, as in the case of Fig. 4.8, the DT random walk LRT is noticeably better than the filtered energy detector.

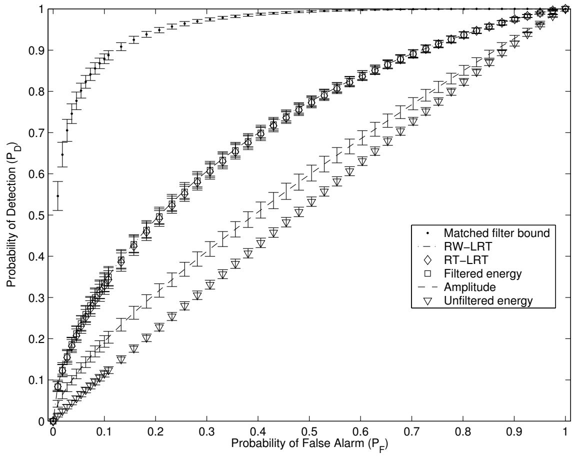


Figure 4.6: Simulated ROC curves for the DT symmetric random walk $p_{l,1} = p_{l,2} = p_{u,1} = p_{u,2} = 0.5$ at SNR $= -39.9$ dB, $T = 60$ s for the matched filter, RW-LRT, RT-LRT, filtered energy, amplitude, and unfiltered energy detector. The RW-LRT is theoretically optimal.

Let us consider why the RW-LRT in the two DT symmetric random walks are well approximated by a filtered energy detector, whereas the DT asymmetric random walk is not. The two symmetric walks satisfy the condition that $\lim E[X_j] = 0$,

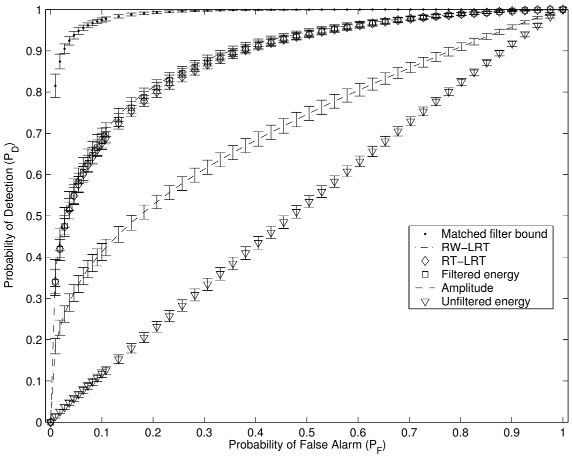


Figure 4.7: Simulated ROC curves for the DT symmetric random walk $p_{l,1} = p_{u,2} = 0.52, p_{l,2} = p_{u,1} = 0.48$ at SNR $= -37.4$ dB, $T = 60$ s. The RW-LRT is theoretically optimal.

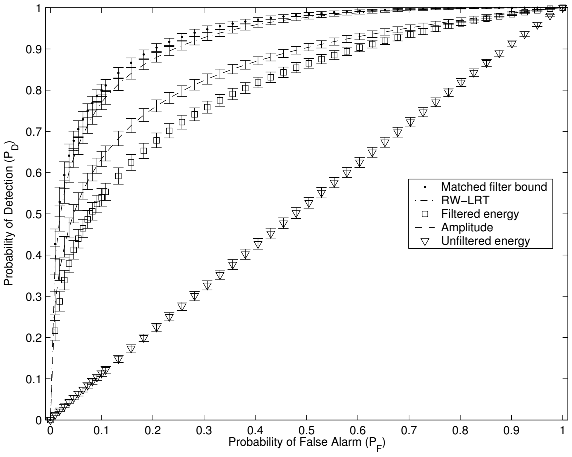


Figure 4.8: Simulated ROC curves for the DT asymmetric random walk $p_{l,1} = p_{u,1} = 0.45$, $p_{l,2} = p_{u,2} = 0.55$ at SNR $= -41.0$ dB, $T = 60$ s. The RW-LRT is theoretically optimal.

while the asymmetric walk does not. The steady state probability distribution of the asymmetric walk is plotted in Fig. 4.9 below. We would not expect the LRT for the asymmetric DTRW to be well approximated by a single FE statistic. Let us proceed to check the second condition for the three DTRWs.

Plots of $(\rho_{r*}^T \odot \rho_{*1})$, $(\rho_{1*}^T \odot \rho_{*r})$, $(\rho_{1*}^T \odot \rho_{*1})$, and $(\rho_{r*}^T \odot \rho_{*r})$ were generated for the three DT random walks. The case for the symmetric random walk with $p_{l,1} = p_{u,1} = 0.5$ is illustrated in Fig. 4.10 below. We see that the largest element of the

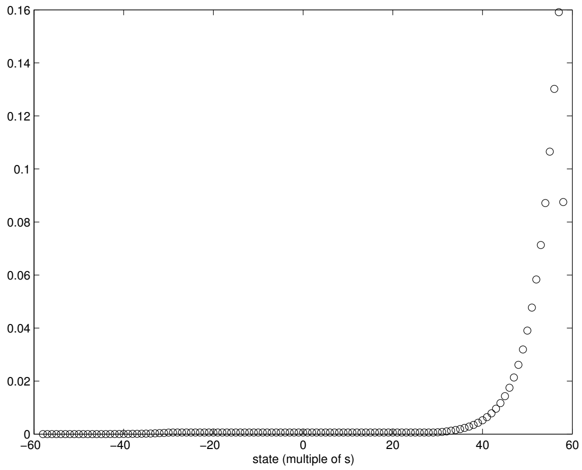


Figure 4.9: Steady state probability distribution of the DT asymmetric random walk $p_{l,1} = p_{u,1} = 0.45$ and $p_{l,2} = p_{u,2} = 0.55$.

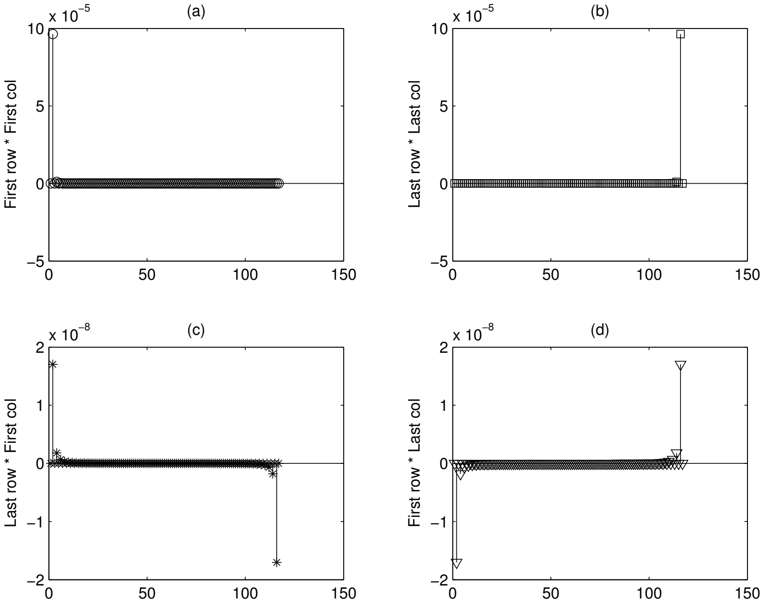


Figure 4.10: Plots of a: $(\rho_{1*}^T \odot \rho_{*1})$, b: $(\rho_{r*}^T \odot \rho_{*r})$, c: $(\rho_{r*}^T \odot \rho_{*1})$, d: $(\rho_{1*}^T \odot \rho_{*r})$ for the DT symmetric random walk with $p_{l,1} = p_{u,1} = 0.5$.

vectors $(\rho_{1*}^T \odot \rho_{*1})$ and $(\rho_{r*}^T \odot \rho_{*r})$ are approximately four orders of magnitude larger than the largest element of the vectors $(\rho_{r*}^T \odot \rho_{*1})$ and $(\rho_{1*}^T \odot \rho_{*r})$. Condition (C.19) is therefore satisfied; we obtained $i = 2$ and $\kappa_2' = 0.999601$. The plot of $M(\alpha)$ is depicted in Fig. 4.11, and at $\alpha = 0.999601$, $M(\alpha)$ is well above the steady state expected energy of the DTRW. So (C.15) is also satisfied. Since $\lim E[X_j] = 0$, sufficient conditions for the LRT of the DTRW in consideration to be approximated by a single FE statistic do indeed hold.

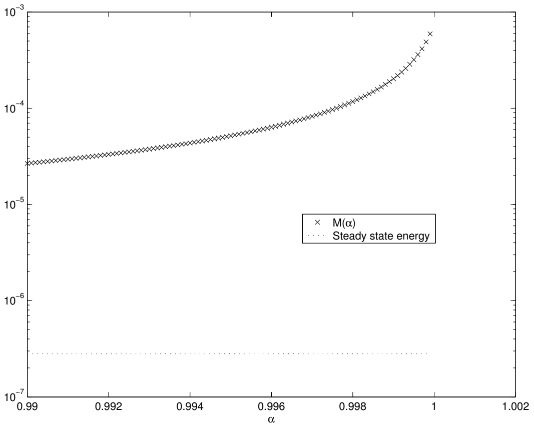


Figure 4.11: Plot of $M(\alpha)$ for the DT symmetric random walk with $p_{l,1} = p_{u,1} = 0.5$

Next, we examine the same four plots for the symmetric random walk with $p_{l,1} = 0.52, p_{u,1} = 0.48$. They are given in Fig. 4.12. Like the previous case, the plots show

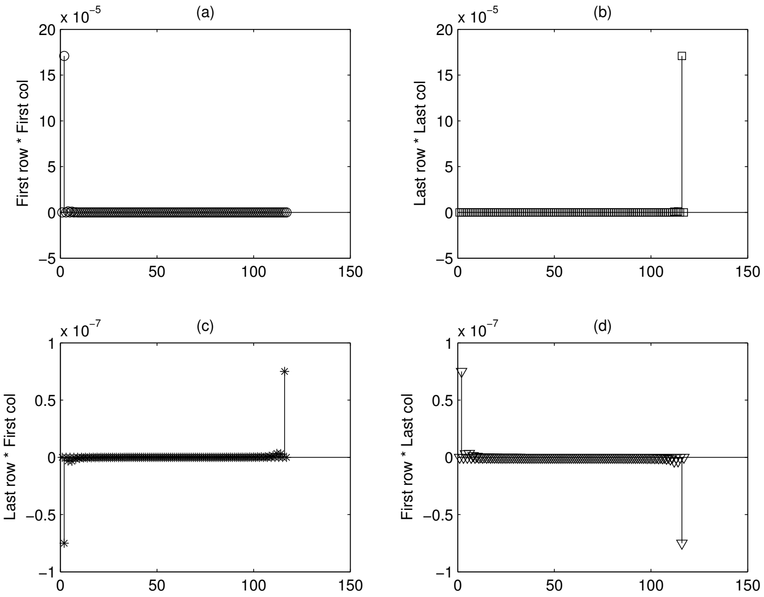


Figure 4.12: Plots of a: $(\rho_{1*}^T \odot \rho_{*1})$, b: $(\rho_{r*}^T \odot \rho_{*r})$, c: $(\rho_{r*}^T \odot \rho_{*1})$, d: $(\rho_{1*}^T \odot \rho_{*r})$ for the DT symmetric random walk with $p_{l,1} = 0.52, p_{u,1} = 0.48$.

that $(\rho_{1*}^T \odot \rho_{*1})$ and $(\rho_{r*}^T \odot \rho_{*r})$ are more "dominant" than $(\rho_{1*}^T \odot \rho_{*r})$ and $(\rho_{r*}^T \odot \rho_{*1})$. The value of $i = 2$ was obtained, which corresponded to $\kappa_2' = 0.999866$. The plot of $M(\alpha)$ is given in Fig. 4.13, and at $\alpha = 0.999866$, the plot is several orders of magnitude larger than the steady state expected energy. So both (C.19) and (C.15) are satisfied for this case, and we expect a single FE statistic to approximate the RW-LRT quite well. Figure 4.7 is in agreement with this expectation.

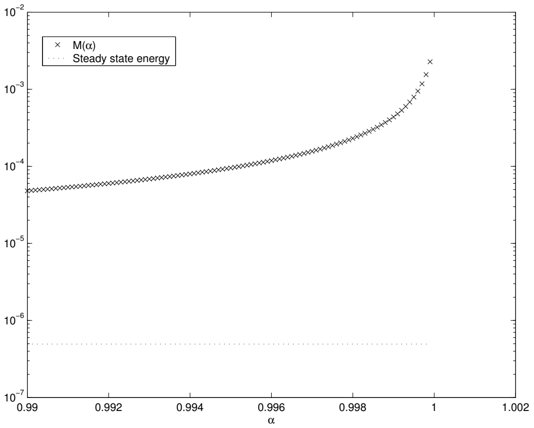


Figure 4.13: Plot of $M(\alpha)$ for the DT symmetric random walk with $p_{l,1} = 0.52, p_{u,1} = 0.48$

Lastly, the plots for the asymmetric random walk with $p_{l,1} = p_{u,1} = 0.45$ are given in Fig. 4.14. While the condition that the two "symmetric" vectors $(\rho_{1*}^T \odot \rho_{*1})$ and

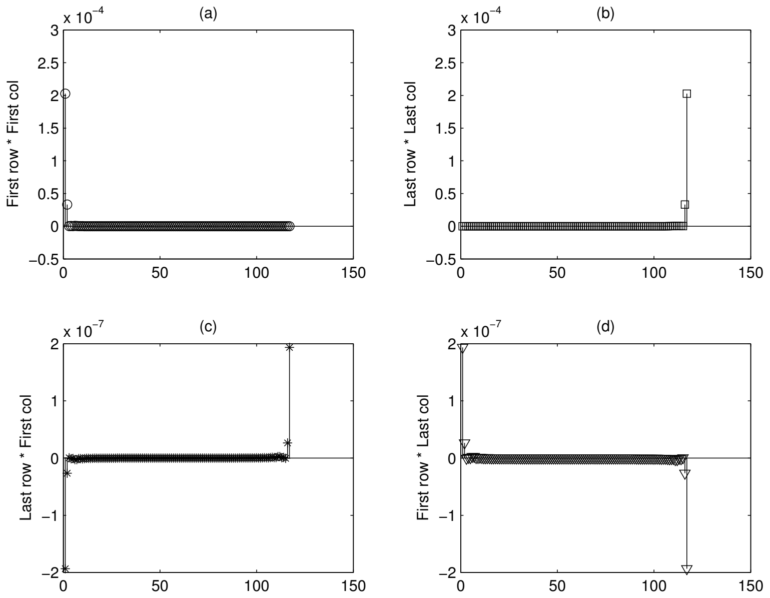


Figure 4.14: Plots of a: $(\rho_{1*}^T \odot \rho_{*1})$, b: $(\rho_{r*}^T \odot \rho_{*r})$, c: $(\rho_{r*}^T \odot \rho_{*1})$, d: $(\rho_{1*}^T \odot \rho_{*r})$ for the DT asymmetric random walk with $p_{l,1} = p_{u,1} = 0.45$.

$(\rho_{r*}^T \odot \rho_{*r})$ are dominant is satisfied, condition (C.19) is not. As was pointed out before, because $\lim E[X_j] \neq 0$, one would not expect a single FE statistic to well approximate the RW-LRT. In Fig. 4.8, we see that this is indeed the case. It is interesting that the amplitude statistic has good performance—indeed, better than the filtered energy statistic. This might indicate that the violation $\lim E[X_j] \neq 0$ is more pertinent than (C.19) in this case.

4.6 Conclusion

In this chapter, we studied the detection of a DT finite state Markov process in AWGN. From the work of [57], which contained recursive equations of the optimal LRT, we derived a closed form expression for the LRT. Under low SNR, we showed that the LR takes the form of the matched filter statistic, but with the one-step MMSE predictor used instead of the known signal value. This parallels the result in CT, which is given by (3.1). There are two differences: (1) $E_1[X_k^2|z^{k-1}]$ is used vs. $(E_1[X_k|z^{k-1}])^2$; (2) the DT result is an approximation, whereas the CT result is exact.

A second order approximation to the LRT for the DT random telegraph model was obtained under the conditions of low SNR, long observation time, and small probability of transition between two consecutive time instances. The approximation is a hybrid statistic that combines the FE, amplitude, and energy statistics. When the transition probabilities are symmetric, the LRT is approximately equal to the FE statistic. This is an intuitively pleasing result; in the presence of the DT random telegraph signal, one would expect a higher energy statistic. It is also natural to filter the observation with an appropriate filter before computing the energy statistic; as the DT random telegraph is a lowpass signal, a LPF is desirable. The approximation tells us that a first-order LPF is adequate, which is surprising, as one would expect "sharper" LPFs to work better.

Next, the detection of a certain class of DT random walks was considered. A second order approximation to the LRT was derived, and conditions obtained for the LRT to be approximated by a bank of filtered energy statistics as well as by a single filtered energy statistic.

Lastly, simulations were performed. They are in excellent agreement with the analysis.

CHAPTER V

Sparse image reconstruction

5.1 Introduction

In most image reconstruction problems, the images of interest are not directly observable. Instead, one observes a transformed version of the image, possibly corrupted by noise. In the general case, the estimation of the so-called original image from the noisy, blurred image can be regarded as a simultaneous deconvolution and denoising problem. Intuitively, a better reconstruction can be obtained by incorporating knowledge of the original image into the reconstruction algorithm. The MRFM image reconstruction problem is different from most other types of imaging problems. The difference lies in the sparsity of the image of interest. As we are interested in imaging molecules, most of the image values will be zero, indicating empty space. Only a few elements will be non-zero, indicating the presence of spin centres. Sparse images also appear naturally in radioastronomy.

In this chapter, the images of interest to be reconstructed are assumed to be sparse and positive valued. No other prior knowledge will be assumed. We consider the model where the observation is a linear transformation of the original image, and corrupted by additive white Gaussian noise (AWGN). Note that the reconstruction methods mentioned here can also be used to solve the sparse denoising problem with coloured Gaussian noise. We consider two issues. Firstly, what is a good sparse prior or penalty, and is there a way to design this? Secondly, given a prior or penalty parameterized by tuning parameters, which shall be called hyperparameters, how does one choose or learn them? The chapter proposes several sparse image reconstruction in attempting to answer these two questions.

There are several existing methods that address the sparse image reconstruction problem. The first is sparse Bayesian learning (SBL) [76]. The voxels to be estimated are modelled as independent, zero-mean Gaussian random variables (r.v.s), each with an unknown variance. The unknown variances in the image prior are learned empirically. The p.d.f. of the observation conditioned on the prior variances can be obtained in closed form. Then, marginal maximum likelihood (MML) is used to learn the prior variances. This empirical Bayes approach does not require any manual tuning. The second existing method is the estimator formed by maximizing the penalized likelihood criterion with a l_1 penalty on the original image values [1, 68]. The aforementioned error criterion is known to promote sparsity in the estimate [13, 22]. This estimator shall be called the L1 estimator; it is also known as the LASSO estimator. For the L1 estimator, one must choose a suitable regularization parameter.

We primarily consider the l_1 norm penalty and the sparse prior used in the empirical Bayes denoising (EBD) method of [33]. For a fixed thresholding rule that satisfies certain conditions, it can be shown that the iterative thresholding framework proposed in [19, 10] minimizes a convex cost function in a monotone fashion. This permits the design of a sparse prior/penalty via selection of a good thresholding function. Three methods of estimating the hyperparameters are investigated: marginal MML, maximum a posterior (MAP), and Stein's unbiased risk estimate (SURE). Several reconstruction methods are proposed based on these three methods.

In EBD, the sparse prior which consists of a weighted average of a Laplacian p.d.f. and an atom at zero (LAZE) was used. The hyperparameters of the LAZE prior are estimated via MML. SBL also estimates its hyperparameters using MML. In this chapter, EBD, which only performs denoising, is extended to perform simultaneous deconvolution and denoising by using a plug-in estimator. This method will be referred to as EBD-LAZE. The LAZE prior is also used to produce a MAP estimator.

The next two estimators are maximum penalized likelihood (MPL) estimators with penalty functions that encourage sparsity, e.g., the l_1 norm penalty. The tuning parameters in the MPL estimators are estimated by minimizing SURE of the l_2 risk between the transformed image (i.e., the linear transformation applied to the original image) and the estimated transformed image. The method when applied to the L1 estimator is called L1-SURE. The hybrid hard-soft (HHS) thresholding function that appeared in the MAP solution obtained with the LAZE prior is used to derive a penalty term. The tuning parameters of the penalty term can also be estimated via SURE, which results in an estimator that will be called HHS-SURE. A simulation study was conducted comparing: SBL; the standard and projected Landweber iteration; and the proposed reconstruction methods.

5.2 Problem formulation

In image reconstruction problems, the image is typically a 2-dimensional or 3-dimensional array. By enumerating the elements of the array in a fashion, one can equivalently represent the image by a vector. Without loss of generality, then, we shall take the observation $\underline{y} \in \mathbb{R}^N$. Let $\underline{\theta}$ be the image that we would like to reconstruct. Similarly, without loss of generality, let $\underline{\theta} \in \mathbb{R}^M$.

Consider the conditional probability density of $\underline{y}$ given $\underline{\theta}$, i.e., $p(\underline{y}|\underline{\theta})$, or equivalently the density of $\underline{y}$ parameterized by $\underline{\theta}$, i.e., $p(\underline{y};\underline{\theta})$. Suppose that we would like to estimate $\underline{\theta}$ under the condition that it is sparse, i.e., most of the values of θ_i are zero. This chapter examines the case when $p(y_i;\underline{\theta}), 1 \leq i \leq N$ represents a sequence of independent Gaussian r.v.'s. Let

$$y_i \sim \mathcal{N}(\underline{h}_i^T \theta, \sigma^2) \tag{5.1}$$

where $\mathcal{N}(\underline{\mu}, \mathbf{\Sigma})$ is the Gaussian distribution with mean $\underline{\mu}$ and covariance matrix $\mathbf{\Sigma}$. The model can be more familiarly written as

$$\underline{y} = \mathbf{H}\underline{\theta} + \underline{w}, \text{ where } \underline{w} \sim \mathcal{N}(\underline{0}, \sigma^2\mathbf{I}), \tag{5.2}$$

with $\mathbf{H} \triangleq (\underline{h}_1|\underline{h}_2|\ldots|\underline{h}_N)^T$. Note that $\underline{w}$ represents AWGN, and $\mathbf{H} \in \mathbb{R}^{N\times M}$.

If $\mathbf{H}$ had full column rank, $(\mathbf{H}^T\mathbf{H})$ would be invertible, and (5.2) could be rewritten as

$$\underline{\tilde{y}} = \underline{\theta} + \underline{\tilde{w}}, \text{ where } \underline{\tilde{w}} \sim \mathcal{N}(\underline{0}, \sigma^2\mathbf{H}^\dagger(\mathbf{H}^\dagger)^T) \tag{5.3}$$

where $\underline{\tilde{y}} \triangleq \mathbf{H}^\dagger\underline{y}$; $\mathbf{H}^\dagger \triangleq (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T$ is the pseudo-inverse of $\mathbf{H}$; and $\underline{\tilde{w}} \triangleq \mathbf{H}^\dagger\underline{w}$ is coloured Gaussian noise. In this case, (5.2) would be equivalent to denoising $\underline{\theta}$ in coloured Gaussian noise. If $\mathbf{H}$ were orthonormal, the estimation of $\underline{\theta}$ in (5.2) would be a denoising problem in i.i.d. Gaussian noise.

It should be noted that, while the sparsity considered here is in the natural (amplitude) domain of the image $\underline{\theta}$, the results here are applicable where sparsity is present in some other domain, e.g., in an appropriate wavelet basis [12].

5.3 Discussion of model

We shall discuss the validity of the linear model (5.2) in this section. This chapter assumes that the four conditions under which the MRFM vertical tip psf was derived in section 2.3 hold. Then, in the noiseless scenario, the observation is given by $\underline{y} = \mathbf{H}\underline{\theta}$. The matrix $\mathbf{H}$ implements convolution with the cantilever tip psf $H(x, y, z)$, e.g., for the vertical tip, $H(x, y, z)$ is given by (2.14).

Later on, in the derivation of the SURE expressions for the L1 estimator and the MPLE based on the cost function derived from the HHS thresholding function, the assumption of linear independence of the columns of $\mathbf{H}$ is made. This is an *observability* assumption, in that knowing $\underline{y}$ is sufficient information to uniquely solve for $\underline{\theta}$. If the noiseless system were not observable, then multiple $\underline{\theta}$ have the ability to produce the observation $\underline{y}$. Ideally, one would like to have many more rows of $\mathbf{H}$ than columns, i.e., $N \gg M$. That is, a desirable situation is for the dimension of the projection of the observations to be greater than the number of observations. When $N \gg M$, the linear independence condition is not that restrictive. The other proposed methods in this chapter do not use the linear independence assumption.

Consider the i.i.d. Gaussian noise model used in (5.2). One of the assumptions used in deriving the MRFM vertical tip psf in section 2.3 was that energy-based measurements are used. In the previous chapter, we showed the optimality, under certain conditions, of the filtered energy statistic in detecting a DT random telegraph signal in AWGN. There is therefore strong justification for using energy-based measurements. In the experiment that uses the iOSCAR detection protocol, the measurements y_i are taken according to the schematic illustrated in Fig. 5.1 below. The clock signal used in the generation of the in-phase and quadrature-phase signals $s_I(t)$ and $s_Q(t)$ respectively come from the pulses of the rf field $B_1(t)$.

Among the sources of noise are: noise due to the phase-lock loop (PLL), noise from the interferometer measurements, and system thermal noise. The major noise contributor is the PLL, and its phase noise can be characterized as approximately

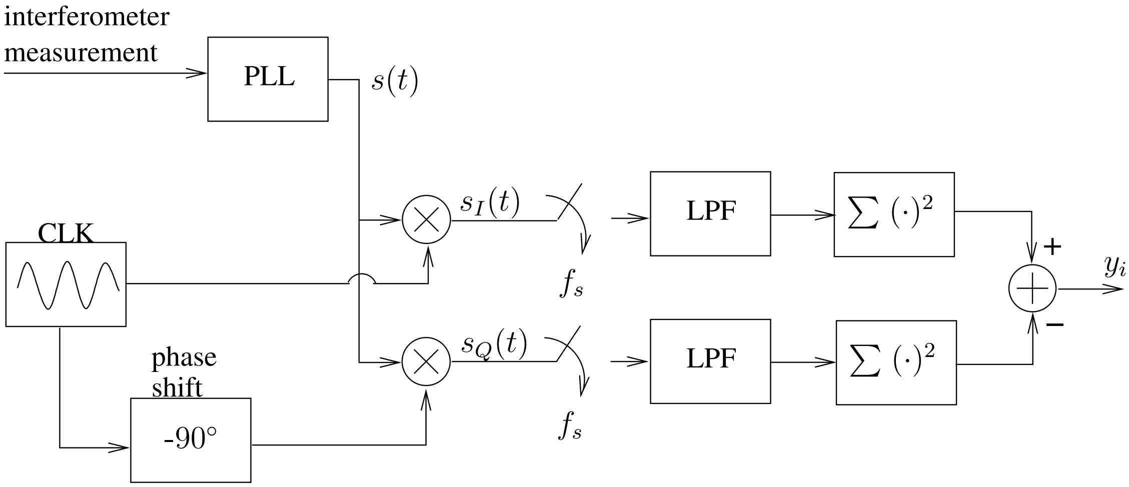


Figure 5.1: Schematic of energy-based measurements for image reconstruction under iOSCAR.

narrowband Gaussian around the bandwidth of interest. Note that there are additional filters not shown in the schematic.

Let us compute the output y_i in the absence of the MRFM psf. For simplicity, make the following assumptions:

1. The noise variance in the samples of the in-phase and quadrature-phase signals $s_I(t)$ and $s_Q(t)$ is unity.
2. There are no random spin flips.
3. Ignore the effect of the LPF, which will introduce correlation across the samples.

Let G_i, $1 \le i \le M$ denote i.i.d. Gaussian r.v.s $G_i \sim \mathcal{N}(0,1)$. Under the above assumptions, the output of the lower branch of Fig. 5.1 after M samples is $\sum_{i=1}^M G_i^2$. We recognize this quantity as a central chi-squared r.v., i.e.,

$$\sum_{i=1}^{M} G_i^2 = \chi_M^2. \tag{5.4}$$

Under the same assumption, the upper branch of Fig. 5.1 is $\sum_{i=1}^M (G_i+\delta_i)^2$, where the δ_is denote the telegraph signal induced by the iOSCAR protocol. This is a noncentral chi-squared r.v. with noncentrality parameter $\lambda = \sum_{i=1}^{M} \delta_i^2$.

$$\sum_{i=1}^{M}(G_i + \delta_i)^2 = \chi_M^2(\lambda) \tag{5.5}$$

The difference between the upper and lower branch statistics is the observed quantity y_i, so

$$y_i \approx \sum_{i=1}^{M}(G_i + \delta_i)^2 - \sum_{i=1}^{M} G_i^2 = \lambda + 2\sum_{i=1}^{M}\delta_i G_i \sim \mathcal{N}(\lambda, 4\lambda). \tag{5.6}$$

One sees that y_i has a Gaussian density where the mean and variance are related. The latter fact is not used in the formulation of the linear model (5.2). Assuming that the PLL phase noise is stationary in time leads to the w_is in the linear model being i.i.d. Gaussian.

For large M, an asymptotic approximation exists for the central chi-squared r.v.: $\chi_M^2 \to \mathcal{N}(M, 2M)$ as $M \to \infty$ [30]. However, no such comparable result exists for the noncentral chi-squared r.v. $\chi_M^2(\lambda)$ [31].

5.4 A general separation principle

The separation of the deconvolution and denoising sub-problems can be achieved through the use of the Expectation-Maximization (EM) algorithm. This was noted in [19] in the case of Gaussian statistics, i.e., under the model given by (5.2). A general form is presented here that includes the case of Poisson statistics. From the system

$$\underline{\theta} \longrightarrow \underline{y}, \tag{5.7}$$

suppose that there exists an intermediate random variable $\underline{z}$ such that, conditioned on $\underline{z}$, $\underline{y}$ is independent of $\underline{\theta}$, i.e., $p(\underline{y}|\underline{z};\underline{\theta}) = p(\underline{y}|\underline{z})$. We now have

$$\underline{\theta} \longrightarrow \underline{z} \longrightarrow \underline{y}. \tag{5.8}$$

In EM terminology, $\underline{z}$ plays the role of the *complete data* [17]. As $\underline{z}$ is an admissible complete data for $p(\underline{y};\underline{\theta})$, $p(\underline{y},\underline{z};\underline{\theta}) = p(\underline{y}|\underline{z})p(\underline{z};\underline{\theta})$, and so $\log p(\underline{y},\underline{z};\underline{\theta}) = \log p(\underline{y}|\underline{z}) + \log p(\underline{z};\underline{\theta})$. Consider applying the EM algorithm to obtain the MAP/MPLE estimate of θ, which is

$$\hat{\underline{\theta}} = \text{argmax}_{\underline{\theta}}(\log p(\underline{y}|\underline{\theta}) - \text{pen}(\underline{\theta})) \tag{5.9}$$

where $\text{pen}(\underline{\theta})$ is a penalty function imposed on $\underline{\theta}$. Let $\hat{\underline{\theta}}^{(n)}$ denote the estimate of $\underline{\theta}$ at the nth iteration. At this point, the Q function of the EM algorithm is

$$Q(\underline{\theta},\hat{\underline{\theta}}^{(n)}) = E_z[\log p(\underline{z};\underline{\theta})|\underline{y},\hat{\underline{\theta}}^{(n)}] - \text{pen}(\underline{\theta}) + K \tag{5.10}$$

where K is a constant independent of $\underline{\theta}$.

Suppose a judicious choice of $\underline{z}$ can be made such that the RHS of (5.10) assumes the form $f(\underline{\theta},\hat{\underline{z}}^{(n)})$, for some suitable function $f(\cdot,\cdot)$ and $\hat{\underline{z}}^{(n)}$ is an estimator of $\underline{z}$ at the nth iteration. Then, a two-step estimation procedure occurs in each EM iteration. In the nth iteration, the estimate $\hat{\underline{z}}^{(n)}$ is first formed; then, the estimate $\hat{\underline{\theta}}^{(n)}$ is formed from

$$\hat{\underline{\theta}}^{(n)} = \text{argmax}_{\underline{\theta}} f(\underline{\theta},\hat{\underline{z}}^{(n)}) \tag{5.11}$$

Note that the separation principle does not enforce any sparsity in $\hat{\underline{\theta}}^{(n)}$. Sparsity is encouraged by appropriate selection of the penalty function, $\text{pen}(\underline{\theta})$. As well, while the EM algorithm ensures a monotonic increase in the objective function, it does not guarantee convergence to a maximizer in general [17, 19]. However, if the likelihood function is unimodal and certain differentiability criteria are met, then the EM algorithm will converge to the unique maximum [79].

5.4.1 Gaussian statistics

Consider the model given by (5.2). In [19], the separation principle was obtained by selecting $\underline{z} = \underline{\theta} + \alpha \underline{w}_1$, where $\underline{w}_1 \sim \mathcal{N}(\underline{w}_1; \underline{0}, \alpha^2 \mathbf{I})$. Then, $\underline{y}$ and $\underline{z}$ are related as follows

$$\underline{y} = \mathbf{H}\underline{z} + \underline{w}_2 \tag{5.12}$$

The noise $\underline{w}_2 \sim \mathcal{N}(\underline{w}_2; \underline{0}, \sigma^2\mathbf{I} - \alpha^2\mathbf{H}\mathbf{H}^T)$. Note that $\underline{w}_1$ is AWGN, but $\underline{w}_2$ is coloured Gaussian noise. The decomposition only works if $(\alpha/\sigma)^2 \leq \rho(\mathbf{H}\mathbf{H}^T)^{-1}$, where $\rho(\mathbf{A})$ is the spectral radius of the square matrix $\mathbf{A}$. The spectral radius $\rho(\mathbf{A}) \triangleq \max_i |\lambda_i|$, where λ_i are the eigenvalues of $\mathbf{A}$. The decomposition is visualized in Fig. 5.2.

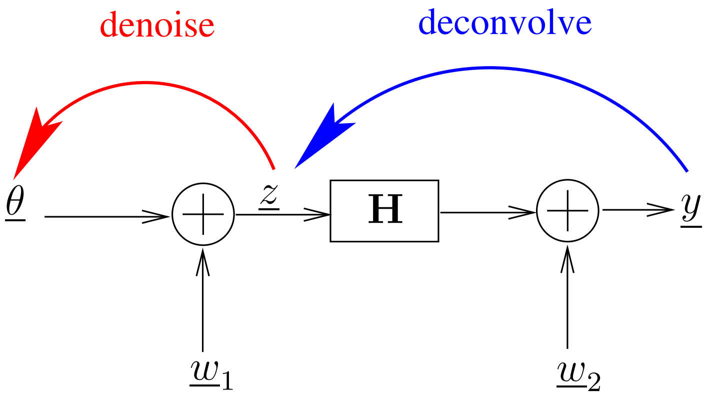


Figure 5.2: Decomposition of the deconvolution and denoising steps for Gaussian statistics.

With $\underline{z} = \theta + \alpha \underline{w}_1$, the Q function assumes the form of

$$Q(\underline{\theta}, \underline{\hat{\theta}}^{(n)}) = f(\underline{\theta}, \underline{\hat{z}}^{(n)}) \tag{5.13}$$

with

$$\underline{\hat{z}}^{(n)} = E[\underline{z} | \underline{y}, \underline{\hat{\theta}}^{(n)}] = \underline{\hat{\theta}}^{(n)} + \left(\frac{\alpha}{\sigma}\right)^2 \mathbf{H}^T(\underline{y} - \mathbf{H}\underline{\hat{\theta}}^{(n)}), \tag{5.14}$$

and $f(\underline{\theta}, \cdot)$ is a quadratic function. One realizes that (5.14) is a Landweber iteration: consequently, it can be viewed as the *deconvolution* step. The maximization of

$Q(\underline{\theta}, \hat{\underline{\theta}}^{(n)})$ can be regarded as a *denoising* step. The two-step estimation procedure that occurs in each EM iteration can be interpreted as a separation of the denoising and deconvolution subproblems.

Solving for the maximum of $Q(\underline{\theta}, \hat{\underline{\theta}}^{(n)})$,

$$\hat{\underline{\theta}}^{(n+1)} = \text{argmax}_{\underline{\theta}} \left[-\frac{1}{2\alpha^2} \|\underline{\theta} - \hat{\underline{z}}^{(n)}\|^2 - \text{pen}(\underline{\theta}) \right] \tag{5.15}$$

In this chapter, the norm $\|\cdot\|$ without a subscript indicates the l_2 norm. Equations (5.14) and (5.15) can be written more succinctly as

$$\hat{\underline{\theta}}^{(n+1)} = \mathcal{D}\left(\hat{\underline{\theta}}^{(n)} + c\mathbf{H}^T(\underline{y} - \mathbf{H}\hat{\underline{\theta}}^{(n)}) \right) \tag{5.16}$$

where $\mathcal{D}(\cdot)$ is a denoising operation that depends on the form of $\text{pen}(\cdot)$, and $c = (\alpha/\sigma)^2$.

Remark 1 If $\mathbf{H}$ implements convolution with a psf H and $M = N$, the computation of $\rho(\mathbf{H}\mathbf{H}^T)$ can be done using the Discrete Fourier Transform (DFT). In such a scenario, there exists a unitary matrix $\mathbf{Q} \in \mathbb{C}^{N\times N}$ such that $\mathbf{H} = \mathbf{Q\Gamma Q}^H$, where $(\cdot)^H$ denotes the complex conjugate transpose [17, p. 10]. The diagonal matrix $\mathbf{\Gamma} = \text{diag}(\gamma_1, \ldots, \gamma_N) \in \mathbb{C}^{N\times N}$ contains the DFT coefficients of the psf H arranged in a lexicographic order. Then,

$$\mathbf{H}\mathbf{H}^T = \mathbf{Q\Gamma\Gamma}^H\mathbf{Q}^H \tag{5.17}$$

which results in

$$\rho(\mathbf{H}\mathbf{H}^T) = \max_{1\le i\le N} |\gamma_i|^2. \tag{5.18}$$

Remark 2 Under the assumption that $\mathbf{H}$ implements the convolution operation with a psf H and $M = N$, the induced l_2 norm of $\mathbf{H}$ can also be efficiently computed using the DFT.

$$\|\mathbf{H}\|_2 = \sqrt{\max_{1\le i\le N} \rho(\mathbf{H}^T\mathbf{H})} = \sqrt{\max_i |\gamma_i|^2} = \max_i |\gamma_i| \tag{5.19}$$

5.4.2 Poisson statistics

Consider the case when $p(y_i;\underline{\theta}), 1 \le i \le N$ represents a sequence of independent Poisson random variables (r.v.). In particular, suppose that each y_i is an observation from a Poisson r.v., so that

$$y_i \sim P_0(y_i; \underline{a}_i^T \underline{\theta}) \tag{5.20}$$

where $P_0(n;\lambda) = e^{-\lambda}\lambda^n/n!, n \in \mathbb{N}$ is the Poisson probability mass function. This model appears in emission tomography reconstruction problems [39], where y_i represents the number of photons/positrons counted at the ith detector, θ_i represents the emission density of the ith voxel, and a_{ij} is the conditional probability that a photon/positron emitted from the jth voxel is detected by the ith detector. As a result, $p(\underline{y};\underline{\theta}) = \prod_i P_0(y_i;\underline{a}_i^T\underline{\theta})$, and the log likelihood is

$$\log p(\underline{y};\underline{\theta}) = -\sum_i \underline{a}_i^T\underline{\theta} + \sum_i y_i \log(\underline{a}_i^T\underline{\theta}) \tag{5.21}$$

An admissible complete data is $\{z_{ij}\}$, $1 \le i \le N, 1 \le j \le M$, where z_{ij} is the number of photons/positrons emitted from the jth voxel and recorded at the ith detector [39]. Note that $y_i = \sum_j z_{ij}$, and $z_{ij} \sim P_0(z_{ij};\theta_j a_{ij})$. Then, in computing the Q function of the EM algorithm at the nth iteration, $n \ge 1$, one gets

$$Q(\underline{\theta},\underline{\hat{\theta}}^{(n)}) = \sum_j -\left(\sum_i a_{ij}\right)\theta_j + \log\theta_j \sum_i \hat{z}_{ij}^{(n)} - \mathrm{pen}(\underline{\theta}) + K \tag{5.22}$$

$$\text{where:}\quad \hat{z}_{ij}^{(n)} = \frac{y_i a_{ij}\hat{\theta}_j^{(n-1)}}{\sum_k a_{ik}\hat{\theta}_k^{(n-1)}} \tag{5.23}$$

and K is independent of θ. If $\text{pen}(\underline{\theta}) \equiv 0$, a closed form solution for the maximization of $Q(\theta, \hat{\theta}^{(n)})$ is available:

$$\hat{\theta}_j^{(n)} = \hat{\theta}_j^{(n-1)} \frac{\sum_i y_i \cdot (a_{ij}/\sum_k a_{ik}\hat{\theta}_k^{(n-1)})}{\sum_i a_{ij}} \tag{5.24}$$

The above application of the EM algorithm results in the same separation principle as previously discussed. Each iteration of the EM algorithm results in the estimation of the intermediate variable $\underline{z}$, followed by the estimation of $\underline{\theta}$. As $y_i = \sum_j z_{ij}$, the estimation of $\underline{z}$ can be regarded as a deconvolution step.

For the rest of the chapter, we shall focus on the sparse image reconstruction problem in the setting of Gaussian statistics, i.e., (5.2).

5.5 Literature review

5.5.1 Sparse denoising

When $\mathbf{H}$ is orthonormal, (5.2) reduces to denoising a sparse $\underline{\theta}$ from the observation $\underline{y}$. The latter would be the sum of $\underline{\theta}$ and i.i.d. Gaussian noise. This problem appears in the context of wavelet regression, where one would like to estimate an unknown function in noise [32, 8]. Here, it is assumed that the unknown function has only several non-zero wavelet coefficients in a suitable wavelet transform. One way of encouraging sparsity in $\underline{\theta}$ is to model $\underline{\theta}$ with a sparse distribution. As we would like the distribution to adapt to the sparsity of $\underline{\theta}$, the use of tuning parameters is necessary. The question that naturally follows is how to best select the tuning parameters of the sparse distribution so that they are a good fit to the true $\underline{\theta}$. As $\underline{\theta}$ is the unknown quantity that we would like to estimate, however, a method that does not rely on knowing $\underline{\theta}$ must be used.

Bayesian methods have been successfully applied in model selection [6, 2]. The ability of Bayesian methods to learn the statistical properties of data naturally led to its use in the estimation of the tuning parameters, which we shall call the *hyperparameter*. An empirical Bayes (EB) approach was adopted in [32, 8], where the hyperparameter was estimated either through maximum likelihood (ML) or method of moments (MOM). The estimated hyperparameters are then used as if they were known a priori, and a suitable thresholding rule applied. The following sparse prior was used in [32]:

$$\theta_i|\nu^2, w \overset{\text{i.i.d.}}{\sim} (1-w)\delta(\theta_i) + w\mathcal{N}(\theta_i; 0, \nu^2) \tag{5.25}$$

with the hyperparameter $\underline{\phi} = (\nu^2, w)$. In general, the thresholding rule will be a function of the hyperparameter $\underline{\phi}$.

This naturally leads to the question of how accurate the hyperparameter estimate is. In [18], this question was avoided by using a Jeffreys' prior for θ_i. The resulting MAP estimate of θ_i was independent of any tuning parameter, and consequently can be regarded as being data-independent. However, there is still the outstanding issue of which prior is best suited for the θ_is. What types of images or coefficients are well modelled by a Jeffreys' prior?

Recent work with the prior

$$\theta_i|a, w \overset{\text{i.i.d.}}{\sim} (1-w)\delta(\theta_i) + w\gamma(\theta_i; a), \tag{5.26}$$

where $\gamma(x; a) = (1/2)ae^{-a|x|}$ is the Laplacian p.d.f. with shape parameter a resulted in an estimator with performance that is within a constant of the asymptotic minimax error under certain conditions [33]. Moreover, the error is bounded. We shall call this denoising method empirical Bayes denoising (EBD). Recall that (5.26) is referred to as the LAZE density.

The primary result of [33] will be mentioned here. The performance of EBD is measured by the following risk function

$$R_q(\underline{\theta}, \underline{\hat{\theta}}) \triangleq M^{-1} \sum_{i=1}^{M} E_{\underline{Y}} |\hat{\theta}_i(\underline{y}) - \theta_i|^q \tag{5.27}$$

for $0 < q \leq 2$, where recall that $\mathbf{H}$ is orthonormal in the context of the problem statement (5.2). The notation $E_{(\cdot)}$ is used to denote the expectation with respect to the subscripted random variable. EBD is compared to the asymptotic minimax risk for $\underline{\theta}$ belonging to the l_p norm ball of radius η for $0 \leq p \leq 2$, defined as

$$l_p[\eta] \triangleq \left\{ \underline{\theta} : M^{-1} \sum_i |\theta_i|^p \leq \eta^p \right\}, \; p > 0 \tag{5.28}$$

$$\text{and} \;\; l_0[\eta] \triangleq \left\{ \underline{\theta} : M^{-1} \sum_i I(\theta_i \neq 0) \leq \eta \right\} \tag{5.29}$$

Let $r_{p,q}(\eta)$ be the asymptotic minimax risk. Then, under the conditions of [33, Thm. 1], there exists $C_i(p, q, \gamma)$ for $i = 1, 2$ such that for $\eta \leq \eta_0(p, q, \gamma)$ and $n \geq n_0(p, q, \gamma)$, EBD's risk function satisfies

$$\sup_{\underline{\theta} \in l_p[\eta]} R_q(\underline{\theta}, \underline{\hat{\theta}}) \leq C_1 r_{p,q}(\eta) + C_2 M^{-1} (\log M)^{2+(q-p)/2} \tag{5.30}$$

This result holds only when $\underline{y} \sim \mathcal{N}(\underline{\theta}, \sigma^2 \mathbf{I})$, i.e., the noise $\underline{w}$ is i.i.d.

The hyperparameter $\underline{\phi} = (a, w)$ has an intuitive interpretation: w represents the fraction of non-zero values in $\underline{\theta}$, and a represents the variability of non-zero θ_i's. The hyperparameter is first estimated using marginal ML. Namely, the ML estimate of $\underline{\phi}$ using the marginalized density $p(\underline{y}|\underline{\phi})$ is taken, i.e., $\hat{\underline{\phi}} = \text{argmax}_{\underline{\phi}} \log p(\underline{y}|\underline{\phi})$. In the next step, $\underline{\hat{\theta}}$ is formed by applying a thresholding rule $T(\cdot; \phi)$ (e.g., posterior median) to each y_i, i.e., $\hat{\theta}_i = T(y_i; \hat{\phi})$. This is illustrated in Fig. 5.3 below. An example of the posterior median thresholding rule when the LAZE prior is used to model $\underline{\theta}$ is given in Fig. 5.4.

The asymptotic results are not limited to using the Laplacian p.d.f. for γ. A permissible γ has the following properties: (1) it is heavy-tailed, (2) its support is

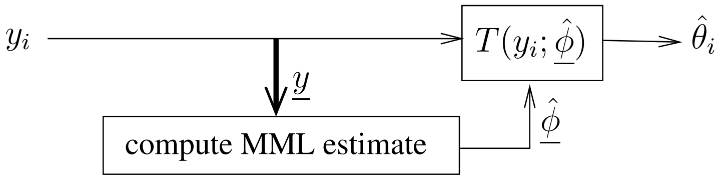


Figure 5.3: Block diagram of EBD.

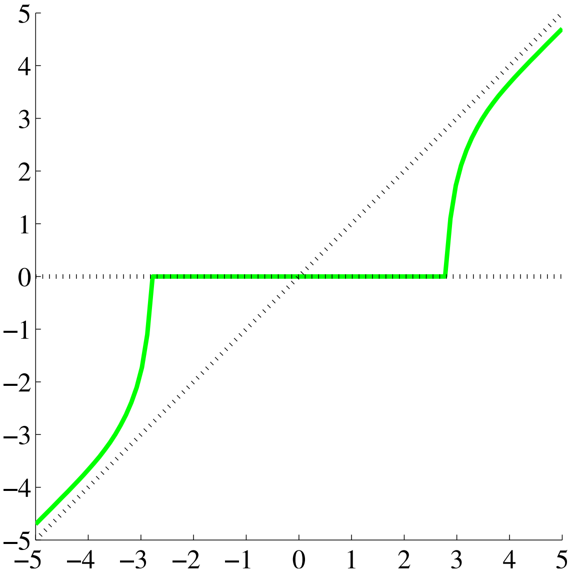


Figure 5.4: Posterior median with LAZE prior with $a = 0.3$, $w = 0.1$, $\sigma = 1$.

unimodal and symmetric, and (3) it satisfies some regularity conditions. As well, the posterior median is but one thresholding rule that can be used. Others are also permissible. For further details, refer to [33]. For the rest of this chapter, we shall assume that γ is the Laplacian p.d.f. unless otherwise mentioned.

5.5.2 Sparse basis representation

When no noise is present, (5.2) reduces to finding the sparsest representation of $\underline{y}$ in terms of the column vectors of $\mathbf{H}$. In other words, the objective is to solve

$$\text{P0: minimize } \|\underline{\theta}\|_0 \text{ such that } \underline{y} = \mathbf{H}\underline{\theta} \tag{5.31}$$

where the l_0 counting measure is defined as $\|\underline{x}\|_0 \triangleq \#\{i : x_i \neq 0\}$. Typically in the sparse basis representation problem, $M > N$, i.e., the columns of $\mathbf{H}$ are an overcomplete basis.

Under certain conditions [13, 22], (5.31) is equivalent to the problem

$$\text{P1: minimize } \|\underline{\theta}\|_1 \text{ such that } \underline{y} = \mathbf{H}\underline{\theta} \tag{5.32}$$

More generally, solving P1 is a suboptimal way of finding the solution of P0. The l_1 norm is defined as $\|\underline{x}\|_1 \triangleq \sum_i |x_i|$. The advantage of (5.32) is that it is a convex optimization problem [13], and can be solved via linear programming techniques. On the other hand, the general solution of (5.31) requires an enumerative approach, resulting in a search that is exponential in M [13].

Sparse Bayesian learning (SBL) is an alternative method to find a sparse basis representation [75, 76, 77]. A MAP framework was used to find $\underline{\theta}$, with $\underline{\theta}$ modelled as independent but *not* identical r.v.s

$$\theta_i | \zeta_i \sim \mathcal{N}(0, \zeta_i) \tag{5.33}$$

An empirical Bayes approach was employed to find an estimate of the hyperparameter $\underline{\phi} = (\zeta_1, \ldots, \zeta_M)$. The details are as follows: a closed-form expression of the marginalized likelihood was derived, and the hyperparameter $\underline{\phi}$ estimated via MML through the use of the EM algorithm. Then, the posterior mean of $\underline{\theta}$ was used as $\hat{\underline{\theta}}$. One would not normally think of the Gaussian distribution as a sparse density. However, the following sparsifying effect occurs: if a particular $\hat{\zeta}_i = 0$, the corresponding $\hat{\theta}_i$ will also be driven to 0.

5.5.3 Sparse deconvolution and denoising

The relationship between P0 and P1 in (5.31) and (5.32) respectively motivates the following sparse estimator: the MAP/MPLE estimator with a l_1 norm penalty. Specifically, the estimate of θ is obtained by using $\text{pen}(\theta) = \beta' \sum_i |\theta_i|$ in (5.9), for some regularization parameter β'. Define the MAP/MPLE estimate of θ with a l_1 penalty to be

$$\hat{\underline{\theta}}_{l1}(y;\beta) \triangleq \operatorname{argmin}_{\underline{\theta}} \left\{ \|\mathbf{H}\underline{\theta} - \underline{y}\|^2 + \beta\|\underline{\theta}\|_1 \right\}, \tag{5.34}$$

where the noise variance σ^2 has been absorbed in the regularization parameter β. The parameter β is restricted to non-negative values. We shall call $\hat{\underline{\theta}}_{l1}$ the L1 estimator, and omit $\underline{y}$ from its argument list for the sake of brevity at times. Note that the objective function is convex. The estimator $\hat{\underline{\theta}}_{l1}$ can also be viewed as the MAP estimator of $\underline{\theta}$ when a Laplacian prior is imposed on $\underline{\theta}$ [27, 21]. The l_1 penalty seems to have been first suggested in [1], and was further developed in [68].

Define a denoising operator that operates on each individual element of $\underline{x} \in \mathbb{R}^M$ as

$$\mathcal{D}_T(\underline{x}) \triangleq \sum_i T(x_i)\underline{e}_i \tag{5.35}$$

where the $\underline{e}_i$s are the standard unit norm basis vectors in $\mathbb{R}^M$.

If $\|\mathbf{H}\| < 1$, the optimization of the objective function in (5.34) can be achieved in the framework of (5.16) by setting $c = 1$ and $\mathcal{D}(\cdot) = \mathcal{D}_{T_\mathrm{s}}(\cdot)$, where

$$\mathcal{D}_{T_\mathrm{s}}(\underline{x}) = \sum_i T_\mathrm{s}(x_i;\beta/2)\underline{e}_i, \tag{5.36}$$

where $T_\mathrm{s}(x;t) \triangleq (x - \operatorname{sgn}(x)t)I(|x| > t)$ is the *soft-thresholding rule* [10]. In the special case that $\mathbf{H}$ implements convolution with a psf and $M = N$, the condition $\|\mathbf{H}\| < 1$ can be efficiently checked via the DFT: see (5.19). If $\mathbf{H}$ has a trivial nullspace, the iterations will converge to the unique minimizer; otherwise, the minimizer is not necessarily unique [10].

More generally, the iterative thresholding algorithm in [10] seeks to minimize the cost function

$$\Psi(\underline{\theta}) = \|\mathbf{H}\underline{\theta} - \underline{y}\|^2 + \|\underline{\theta}\|_{w,p}^p \tag{5.37}$$

$$\text{where: } \|\underline{\theta}\|_{w,p} \triangleq \left(\sum_i w_i |\theta_i|^p\right)^{1/p}, \ 1 \leq p \leq 2, \tag{5.38}$$

and the w_is are uniformly bounded away from 0, i.e., $\exists\, c > 0$ such that all $w_i \geq c$. The algorithm is of the form (5.16), where $\mathcal{D}$ depends on w and p.

In order to solve for the L1 estimator for arbitrary $\|\mathbf{H}\| < C$, one can reparameterize the L1 cost function with the variables

$$\widetilde{\mathbf{H}} \triangleq C^{-1}\mathbf{H} \text{ and } \underline{\tilde{\theta}} \triangleq C\underline{\theta} \tag{5.39}$$

to get the L1 estimator cost function

$$\Psi_{l1}(\underline{\theta}) = \|\mathbf{H}\underline{\theta} - \underline{y}\|^2 + \beta\|\underline{\theta}\|_1 = \|\widetilde{\mathbf{H}}\underline{\tilde{\theta}} - \underline{y}\|^2 + \beta C^{-1}\|\underline{\tilde{\theta}}\|_1.$$

The iterative minimization can then be applied to $\widetilde{\mathbf{H}}$ and $\underline{\tilde{\theta}}$. The denoising operator $\mathcal{D}_{T_\mathrm{s}}$ will use the soft-thresholding function with a new threshold of $t = \beta C^{-1}/2$.

Least angle regression (LARS) is an efficient method that solves for the L1 estimator [15]. NB. the more accurate term is LARS-LASSO; however, we shall henceforth omit the LASSO suffix for the sake of brevity. Although iterative in nature, the number of iterations required is approximately equal to the desired number of non-zero values of the L1 solution, i.e., $\|\underline{\hat{\theta}}_{l1}\|_0$. If the desired $\underline{\hat{\theta}}_{l1}$ is highly sparse, only a small number of iterations is required. In contrast, the number of iterations needed in the iterative thresholding framework of (5.16) depends on its rate of convergence. In practice, we have observed that LARS is a faster implementation than the framework of (5.16). Another benefit of LARS is that it solves for the exact L1 estimator. On the other hand, if one stopped the iterations (5.16) prematurely, the output would not be close to the L1 solution. The disadvantage of LARS, however, is that it requires the columns of $\mathbf{H}$ to be linearly independent. The iterative thresholding framework does not have this requirement.

There are various ways of choosing the regularization parameter β [67]. In the next section, we propose using Stein's unbiased risk estimator (SURE) to select β [64, 62]. In [51], it was reported that using SURE in an optical flow estimation problem produced better results than generalized cross-validation (GCV). The SURE criterion has also been successfully used in SureShrink, which is a sparse denoising method for wavelet coefficients [14].

5.5.4 Commentary

In the context of MAP/MPLE estimators, each estimate satisfies an optimality criteria. However, there are many such criteria to choose from, and it is not certain which is the best for an sparse, arbitrary $\underline{\theta}$. Out of the estimators previously mentioned, EBD is the only method that has minimax properties in terms of an error bound between the true value of $\underline{\theta}$ and its estimate.

In empirical Bayes, a hierarchical effect occurs in that variables that were initially deterministic are substituted by r.v.s with the a desirable p.d.f., e.g., a sparse distribution. In the EB methods previously discussed, there are two levels in the hierarchy: refer to Fig. 5.5. The hyperparameter $\underline{\phi}$ are the parameters of a model $\mathcal{M}$ that describe $\underline{\theta}$. In turn, $\underline{y}$ is the result of a linear mapping $\mathcal{H}$ applied to $\underline{\theta}$ in the absence of noise. It is possible to go one step further and impose a hyperprior on the hyperparameter [48, 49].

$$\begin{tikzcd}
\underline{\phi} \arrow[r, "\mathcal{M}"] & \underline{\theta} \arrow[r, "\mathcal{H}"] & \underline{y}
\end{tikzcd}$$

Figure 5.5: Hierarchy of parameters in empirical Bayes.

5.6 Proposed sparse reconstruction methods

We propose four image reconstruction methods. The first method uses EBD as $\mathcal{D}(\cdot)$ in (5.16); in general, this will be labelled EBD-X, where “X” refers to the prior distribution assumed on the image $\underline{\theta}$. In this chapter, we shall use the LAZE prior (5.26), and so the method will be called EBD-LAZE. The second method is the penalized L1 estimator with the regularization parameter selected via SURE. The third method involves using the prior (5.26) in a MAP framework. Lastly, the fourth method involves using the hybrid hard-soft thresholding function that appears in the MAP solution and selecting the tuning parameters via SURE.

5.6.1 Shrinkage and thresholding rule

We shall review the definition of a shrinkage and thresholding rule, as given in [33]. The function $T(\cdot)$ is a *shrinkage rule* iff $T(\cdot)$ is anti-symmetric and increasing on $(-\infty, \infty)$. A shrinkage rule that satisfies the property that $T(x) = 0$ iff $|x| < t$ will be called a *thresholding rule* with threshold t.

Henceforth, we only consider rules with the property that $T(\cdot)$ is strictly increasing outside of $(-t, t)$. Consequently, $T(\cdot)$ is a bijection on $\mathbb{R} \setminus (-t, t)$. With abuse of notation, let $T^{-1}(\cdot)$ denote the “inverse” of $T(\cdot)$, which will be discontinuous at 0.

5.6.2 EBD based methods

There are several interesting facts about EBD based methods.

Proposition 5.1. *An EBD-based reconstruction method can be regarded as an EM-like iteration, but where the prior on $\underline{\theta}$ is designed in each iteration.*

The results of [27] can be used to create a prior for $\underline{\theta}$ such that the estimate $\hat{\underline{\theta}}$ formed via EBD can be regarded as a MAP estimate. Let $\tilde{p}(\theta_i; \underline{\phi})$ be the prior on θ_i

induced by the thresholding rule $T(\cdot;\underline{\phi})$. Then,

$$\tilde{p}(\theta_i;\underline{\phi}) \propto \exp\left(-\frac{1}{\alpha^2}\int [T^{-1}(\theta_i;\underline{\phi}) - \theta_i]\, d\theta_i\right). \tag{5.40}$$

Let $M(x) \triangleq \int [T^{-1}(x;\underline{\phi}) - x]dx$. For the prior $\tilde{p}(\cdot;\underline{\phi})$ to exist, $\exp(-M(x)/\alpha^2)$ must be integrable over $\mathbb{R}$, as $\tilde{p}(x;\underline{\phi}) = \exp(-M(x)/\alpha^2)/\int_{-\infty}^{\infty}\exp(-M(x)/\alpha^2)dx$. Since the θ_i's are i.i.d. in EBD,

$$\tilde{p}(\underline{\theta};\underline{\phi}) = \prod_{i=1}^{M}\tilde{p}(\theta_i;\underline{\phi}) \tag{5.41}$$

Note that $\tilde{p}(\theta_i;\underline{\phi})$ might not bear any resemblance to the prior on $\underline{\theta}$ used in EBD, e.g., (5.26).

The second step is to fit EBD in the EM framework. Previously, the EBD estimator was interpreted as a MAP estimator. This makes it compatible with the MAP based optimality criterion of (5.9). Let $\underline{z}$ be the complete data, as before. The Q function is

$$Q(\underline{\theta},\underline{\hat{\theta}}^{(n)}) = -\frac{1}{2\alpha^2}\|\underline{\theta} - \underline{\hat{z}}^{(n)}\|^2 - \mathrm{pen}(\underline{\theta}) \tag{5.42}$$

By setting $\mathrm{pen}(\underline{\theta}) = -\log\tilde{p}(\underline{\theta};\underline{\hat{\phi}})$, where $\underline{\hat{\phi}}$ is the nth estimate of the hyperparameter obtained via EBD,

$$Q(\underline{\theta},\underline{\hat{\theta}}^{(n)}) = -\frac{1}{2\alpha^2}\|\underline{\theta} - \underline{\hat{z}}^{(n)}\|^2 + \log\tilde{p}(\underline{\theta};\underline{\hat{\phi}}) \tag{5.43}$$

Therefore, EBD iterations can be considered to be an EM-like iteration, but where the prior on $\underline{\theta}$ is *re-designed* at each iteration. ■

Remark Convergence of the EBD iterations has yet to be established.

Proposition 5.2. *If the hyperparameter $\underline{\phi}$ is fixed, an EBD-based method can be regarded as an EM iteration that maximizes $[\log p(\underline{y}|\underline{\theta}) - pen(\underline{\theta})]$ for a suitably defined penalty function $pen(\underline{\theta})$.*

This follows from (5.43). If the hyperparameter $\underline{\phi}$ is fixed, then EBD iterations are EM iterations with regard to the objective function $[\log p(\underline{y}|\underline{\theta}) + \log \tilde{p}(\underline{\theta}; \underline{\phi})]$. ■

Proposition 5.3. *Suppose that $\|\mathbf{H}\| < 1$, and that the hyperparameter $\underline{\phi}$ is fixed to $\underline{\phi}^0$, i.e., a fixed thresholding rule $T(\cdot) = T(\cdot; \underline{\phi}^0)$ is used. Assume that the thresholding rule $T(\cdot)$ satisfies the properties mentioned in section 5.6.1. The iterations (5.16) with $c = 1$ decrease the convex cost function $\Psi(\underline{\theta})$ in a monotonic fashion, where $\Psi(\underline{\theta})$ is given by*

$$\Psi(\underline{\theta}) = \|\mathbf{H}\underline{\theta} - \underline{y}\|^2 + J(\underline{\theta})$$

(5.44)

$$\textit{where: } J(\underline{\theta}) \triangleq \sum_{i=1}^{M} J_1(\theta_i), \;\; \textit{and } J_1(x) \triangleq 2T^{-1}(x)x - x^2 - 2\int T(\xi)d\xi \Big|_{\xi = T^{-1}(x)}$$

Notice that $J_1'(x) = 2(T^{-1}(x) - x)$. A thresholding rule $T(x)$ is anti-symmetric and satisfies $0 \leq T(x) \leq x$ for all $x \geq 0$. Therefore, $J_1'(x) \geq 0$ for $x > 0$ and $J_1'(x) \leq 0$ for $x < 0$. So $J(\underline{\theta})$ is convex, and since $\|\mathbf{H}\underline{\theta} - \underline{y}\|^2$ is also convex, that makes $\Psi(\underline{\theta})$ convex as well.

The concept of surrogate functions was used to reverse engineer the function $\Psi(\underline{\theta})$. Its usage here is inspired by [10]. For details of the derivation, refer to Appendix D.1. An intuitive understanding can be gained by computing the gradient of Ψ:

$$\nabla\Psi(\underline{\theta}) = 2\mathbf{H}^T\mathbf{H}\underline{\theta} - 2\mathbf{H}^T\underline{y} + 2(T^{-1}(\theta_1) - \theta_1, \ldots, T^{-1}(\theta_M) - \theta_M)^T \tag{5.45}$$

and so

$$\begin{aligned} \nabla\Psi(\underline{\theta}) = \underline{0} \iff & \mathbf{H}^T\mathbf{H}\underline{\theta} - \mathbf{H}^T\underline{y} + (T^{-1}(\theta_1), \ldots, T^{-1}(\theta_M))^T - \underline{\theta} = \underline{0} \\ \iff & T^{-1}(\theta_i) = \theta_i + (\mathbf{H}^T(y - \mathbf{H}\theta))_i, \; 1 \leq i \leq M \\ \iff & \theta_i = T\left(\theta_i + (\mathbf{H}^T(y - \mathbf{H}\theta))_i\right), \; 1 \leq i \leq M \qquad (5.46) \end{aligned}$$

We see that a stationary point of $\Psi(\underline{\theta})$ is a fixed point of the function

$$m(\underline{\theta}) \triangleq \sum_{i=1}^{M} T\left[(\underline{\theta} + \mathbf{H}^T(\underline{y} - \mathbf{H}\underline{\theta}))_i\right] \underline{e}_i \tag{5.47}$$

The function m above corresponds to one step of the iteration (5.16), i.e., the iterations are $\hat{\underline{\theta}}^{(n+1)} = m(\hat{\underline{\theta}}^{(n)})$. Consequently, the stationary points of $\Psi(\underline{\theta})$ are the fixed points of the iteration (5.16). ■

Remark 1 The following are true: (i) the iteration map m is continuous with $\Psi(\underline{\theta}) \geq \Psi(m(\underline{\theta}))$, where equality holds iff $\underline{\theta}$ is a fixed point of m and (ii) the stationary points of $\Psi(\underline{\theta})$ are fixed points of the iteration $\underline{\theta}^{(n+1)} = m(\underline{\theta}^{(n)})$. Therefore, any limit point of the sequence generated by $\underline{\theta}^{(n+1)} = m(\underline{\theta}^{(n)})$ is a stationary point of $\Psi(\underline{\theta})$ [41].

Remark 2 If the columns of $\mathbf{H}$ are linearly independent, then $\|\mathbf{H}\underline{\theta} - \underline{y}\|^2$ is strictly convex, which makes the cost function $\Psi(\underline{\theta})$ strictly convex as well. In this case, the minimizer of $\Psi(\underline{\theta})$ is unique. It was noted in the previous remark that the sequence $\underline{\theta}^{(n)}$ generated by m converged to a minimizer of $\Psi(\underline{\theta})$. There, however, we did not have uniqueness. Under the condition of linear independence of the columns of $\mathbf{H}$, the minimizer is unique.

Remark 3 Proposition 5.3 enables us to *design* a sparse prior or penalty by first specifying a good thresholding rule. Then, the iterative optimization of alternating Landweber and thresholding steps minimizes the cost function $\Psi(\underline{\theta})$ given by (5.44). This is illustrated in Fig. 5.6. Once $\Psi(\underline{\theta})$ is available, any other optimization method can be used to find a minimizer. In particular, if there is a more efficient optimization method, or if there is a closed-form solution, that should be used.

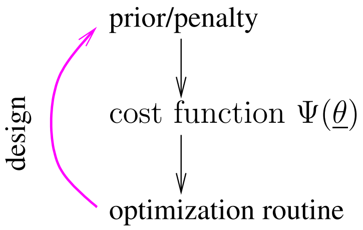


Figure 5.6: Design of a sparse prior/penalty

In addition to the interpretation of the EBD-based methods given by Prop. 5.1-5.3, there is another interpretation that is specific to EBD methods that use a γ which satisfies [33, Thm. 1]. If the complete data $\underline{z}$ were observable, EBD could be applied with the accompanying optimality result (5.30). However, since $\underline{z}$ is not directly observable, the EBD-based methods are using the estimate of $\underline{z}$ given by (5.14). This is the best estimate of $\underline{z}$ given $\hat{\underline{\theta}}^{(n)}$ and $\underline{y}$ in terms of mean-squared error (MSE). The proposed EBD-based method can therefore be regarded as applying EBD on a sequence of minimum MSE estimates of $\underline{z}$.

5.6.3 L1-SURE

The L1 estimator, i.e., (5.34), depends on the regularization parameter β. Different values of β will result in different sparsity levels of the reconstruction. We propose using the SURE criterion to select β. The quality of β will be quantified by the following risk function:

$$R(\underline{\theta}, \beta) \triangleq E_{\underline{Y}} \|\mathbf{H}\hat{\underline{\theta}}_{l1}(\underline{y}; \beta) - \mathbf{H}\underline{\theta}\|^2 \tag{5.48}$$

One cannot compute $R(\underline{\theta}, \beta)$, as that would require knowledge of $\underline{\theta}$. Instead, SURE permits the construction of an unbiased estimator of $R(\underline{\theta}, \beta)$, which is denoted by $\hat{R}(\beta)$. Select $\beta \geq 0$ that minimizes $\hat{R}(\beta)$. Under the assumption that the columns of $\mathbf{H}$ are linearly independent, the expression for $\hat{R}(\beta)$ is given by

$$\hat{R}(\beta) = N\sigma^2 + \|\underline{y} - \mathbf{H}\hat{\underline{\theta}}_{l1}(\beta)\|^2 + 2\sigma^2 \|\hat{\underline{\theta}}_{l1}(\beta)\|_0. \tag{5.49}$$

See Appendix D.2 for the derivation. To find the optimum β, it is necessary to compute $\underline{\hat{\theta}}_{l1}(\beta)$ for different $\beta \geq 0$, evaluate the $\hat{R}(\beta)$s, and select the β that minimizes $\hat{R}(\beta)$.

The expression for SURE of the soft-thresholding estimator used in SureShrink, i.e., [14, (11)], can be obtained from (5.49) by setting $\mathbf{H} = \mathbf{I}$. A SURE expression similar to (5.49) was derived in the case of a diagonal $\mathbf{H}$ and where the l_1 penalty was imposed on the coefficients of a 2-d wavelet transform of $\underline{\theta}$ [50, (10)–(11)].

Remark 1 The risk function of $R(\underline{\theta}, \beta) = E_{\underline{Y}} \|\underline{\hat{\theta}}_{l1}(\underline{y}; \beta) - \underline{\theta}\|^2$ would be more ideal, as it measures the mismatch between $\underline{\hat{\theta}}_{l1}$ and $\underline{\theta}$. However, one could not apply Stein's results to form an unbiased risk estimate. Let $\underline{\mu} \triangleq \mathbf{H}\underline{\theta}$ and $\underline{\hat{\mu}} \triangleq \mathbf{H}\underline{\hat{\theta}}$. If the columns of $\mathbf{H}$ were invertible, the pseudoinverse $\mathbf{H}^\dagger$ would exist, and we would get that

$$\|\underline{\hat{\theta}} - \underline{\theta}\| \leq \|\mathbf{H}^\dagger\| \cdot \|\underline{\hat{\mu}} - \underline{\mu}\|. \tag{5.50}$$

Thus, minimizing the risk $E\|\underline{\hat{\mu}} - \underline{\mu}\|^2$, e.g., (5.48), minimizes the upper bound on $E\|\underline{\hat{\theta}} - \underline{\theta}\|^2$.

Remark 2 Least angle regression (LARS), an efficient method for computing the L1 estimator, can be employed to determine the optimal $\beta \geq 0$ [15]. The columns of $\mathbf{H}$ must be linearly independent in order for LARS to be applied. This is the same condition used in deriving the expression (5.49).

5.6.4 MAP based methods

Instead of computing $\underline{\hat{\theta}} = \text{argmax}_{\underline{\theta}}[\log p(\underline{y}|\underline{\theta}) - \text{pen}(\underline{\theta})]$, simultaneously estimate $\underline{\theta}$ and $\underline{\phi}$:

$$\underline{\hat{\theta}}, \underline{\hat{\phi}} = \underset{\underline{\theta}, \underline{\phi}}{\text{argmax}} \, [\log p(\underline{y}, \underline{\theta}|\underline{\phi}) - \text{pen}(\underline{\theta}, \underline{\phi})]] \tag{5.51}$$

For the rest of the section, we shall not use $\text{pen}(\underline{\theta}, \underline{\phi})$, i.e., set it to zero. Define $\underline{\rho} \triangleq (\underline{\theta}, \underline{\phi})$. Using the decomposition (5.12) and applying the EM algorithm, the resulting Q function is

$$Q(\underline{\rho}, \underline{\hat{\rho}}^{(n)}) = -\frac{1}{2\alpha^2}\|\underline{\theta} - \underline{\hat{z}}^{(n)}\|^2 + p(\underline{\theta}|\underline{\phi}). \tag{5.52}$$

The optimization $\text{argmax}_{\underline{\rho}}\, Q(\underline{\rho}, \underline{\hat{\rho}}^{(n)})$ is equivalent to the denoising of $\underline{\hat{z}}^{(n)}$ under a MAP criterion. We would like to use the LAZE p.d.f. for $p(\underline{\theta}|\underline{\phi})$. However, the delta function in (5.26) is difficult to work with, so define the random variables $\tilde{\theta}_i$ and I_i such that $\theta_i = I_i\tilde{\theta}_i$, $1 \leq i \leq M$. The r.v.s $\tilde{\theta}_i, I_i$ have the following density:

$$I_i = \begin{cases} 0 & \text{with probability } (1-w) \\ 1 & \text{with probability } w \end{cases} \tag{5.53}$$

$$p(\tilde{\theta}_i|I_i) = \begin{cases} g(\tilde{\theta}_i) & I_i = 0 \\ \gamma(\tilde{\theta}_i; a) & I_i = 1 \end{cases}, \tag{5.54}$$

where $g(\cdot)$ is some p.d.f. that will be specified later on. It is assumed that $(\tilde{\theta}_i, I_i)$ are i.i.d. This is the discrete-continuous version of the prior (5.26) with one exception: the introduction of g.

Redefine the optimality criterion as

$$\underline{\hat{\tilde{\theta}}}, \underline{\hat{I}}, \underline{\hat{\phi}} = \underset{\underline{\tilde{\theta}}, \underline{I}, \underline{\phi}}{\text{argmax}}\ \log p(\underline{\tilde{\theta}}, \underline{I}|\underline{y}, \underline{\phi}) = \underset{\underline{\tilde{\theta}}, \underline{I}, \underline{\phi}}{\text{argmax}}\ \log p(\underline{\tilde{\theta}}, \underline{I}, \underline{y}|\underline{\phi}). \tag{5.55}$$

Let $\underline{\rho} = (\tilde{\theta}, I, \phi)$. Define $\mathcal{I}_1 \triangleq \{i : I_i = 1\}$ and $\mathcal{I}_0 \triangleq \overline{\mathcal{I}}_1$. The maximization of (5.55) is equivalent to the maximization of

$$-\frac{1}{2\sigma^2}\|\mathbf{H}\underline{\theta} - \underline{y}\|^2 + \|\underline{I}\|_0 \log w + (M - \|\underline{I}\|_0)\log(1-w) + \sum_{i\in\mathcal{I}_1}\log\left(\frac{1}{2}ae^{-a|\tilde{\theta}_i|}\right) + \sum_{i\in\mathcal{I}_0}\log g(\tilde{\theta}_i) \tag{5.56}$$

We propose to perform the maximization of (5.56) in a block coordinate-wise fashion [17]. The maximizing $\underline{\rho}$ is obtained by alternately (i) maximizing the hyperparameter $\underline{\phi}$ while holding $(\underline{\tilde{\theta}}, \underline{I})$ fixed, and (ii) maximizing $(\underline{\tilde{\theta}}, \underline{I})$ while holding

$\underline{\phi}$ fixed. Consider two cases: firstly, let $g(x) = \gamma(x; a)$. This will give rise to the algorithm MAP1. Secondly, let $g(x)$ be an arbitrary p.d.f. such that: (1) $|g(x)| < \infty$ for all $x \in \mathbb{R}$; (2) $\sup g(x)$ is attained for some $x \in \mathbb{R}$; and (3) $g(x)$ is *independent* of a, w. This will give rise to the algorithm MAP2.

MAP1

Since $g(x) = \gamma(x; a)$, the criterion that is maximized is

(5.57)

$$\Psi_1(\tilde{\underline{\theta}}, \underline{I}, \underline{\phi}) \triangleq -\frac{1}{2\sigma^2}\|\mathbf{H}\underline{\theta} - \underline{y}\|^2 + \|\underline{I}\|_0 \log w + (M - \|\underline{I}\|_0)\log(1-w) + M \log\frac{a}{2} - a\|\tilde{\underline{\theta}}\|_1$$

The Hessian of Ψ_1 with respect to (w.r.t.) $\underline{\phi}$ is

$$\nabla_{\underline{\phi}}\nabla_{\underline{\phi}}^T \Psi_1 = \begin{pmatrix} -\frac{M}{a^2} & 0 \\ 0 & -\frac{\|\underline{I}\|_0}{w^2} - \frac{M - \|\underline{I}\|_0}{(1-w)^2} \end{pmatrix} \tag{5.58}$$

which is clearly negative definite for all $a > 0$ and $0 < w < 1$. The solution to $\nabla_{\underline{\phi}}\Psi_1 = 0$ maximizes Ψ_1, which results in step (i) being

$$\hat{a} = \frac{M}{\|\hat{\tilde{\underline{\theta}}}\|_1} \text{ and } \hat{w} = \frac{\|\hat{\underline{I}}\|_0}{M}. \tag{5.59}$$

Now, given n samples $x_1, \ldots, x_n$ drawn from a Laplacian p.d.f. $\gamma(\cdot; a)$, the ML estimate of a is $\hat{a}_{\mathrm{ML}} = n(\sum_{i=1}^n |x_i|)^{-1}$. The estimate $\hat{a}$ in (5.59) is therefore the ML estimate of a given that all of the $\hat{\tilde{\theta}}_i$s are used.

Next, the maximization in step (ii) can be obtained by applying the EM algorithm with the complete data $\underline{z} = \underline{\theta} + \alpha\underline{w}_1$. In this step, $\underline{\phi}$ is held fixed. The E-step is given in (5.14). For the M-step, the Q function at the $(n+1)$th iteration is

(5.60)

$$Q(\tilde{\underline{\theta}}, \underline{I}; \hat{\tilde{\underline{\theta}}}^{(n)}, \hat{\underline{I}}^{(n)}) = -\frac{1}{2\alpha^2}\|\underline{\theta} - \hat{\underline{z}}^{(n)}\|^2 + \|\underline{I}\|_0 \log w + (M - \|\underline{I}\|_0)\log(1-w) + M\log\frac{a}{2} - a\|\tilde{\underline{\theta}}\|_1.$$

Since $(-Q)$ is convex in $\tilde{\underline{\theta}}$, the maximizing $\tilde{\underline{\theta}}$ can be obtained by solving $\nabla_{\tilde{\underline{\theta}}} Q = 0$. Because Q is the sum of identical expressions of $\tilde{\theta}_i$, each $\tilde{\theta}_i$ can be solved separately.

Let $I(\cdot)$ be the indicator function. One obtains

$$\tilde{\theta}_i = \begin{cases} T_{\mathrm{s}}(\hat{z}_i^{(n)}; a\alpha^2) & I_i = 1 \\ 0 & I_i = 0 \end{cases} . \tag{5.61}$$

By using (5.61) and comparing $\Delta Q = Q|_{I_i=1} - Q|_{I_i=0}$ to zero, the maximizing I_i can be found as

$$I_i = \begin{cases} I(|\hat{z}_i^{(n)}| > a\alpha^2 + \sqrt{2\alpha^2 \log(\frac{1-w}{w})}) & 0 < w \leq \frac{1}{2} \\ 1 & \frac{1}{2} < w \leq 1 \end{cases} . \tag{5.62}$$

Now, the so-called *hard-thresholding rule* is given by $T_{\mathrm{h}}(x;t) \triangleq xI(|x| > t)$. Define a *hybrid hard-soft thresholding rule* as $T_{\mathrm{hs}}(x; t_1, t_2) \triangleq (x - \mathrm{sgn}(x)t_2)I(|x| > t_1)$. See Fig. 5.7. We restrict $t_1 \geq 0$ and $0 \leq t_2 \leq t_1$. The soft and hard-thresholding rule can be expressed as $T_{\mathrm{hs}}(x; t, t)$ and $T_{\mathrm{hs}}(x; t, 0)$ respectively.

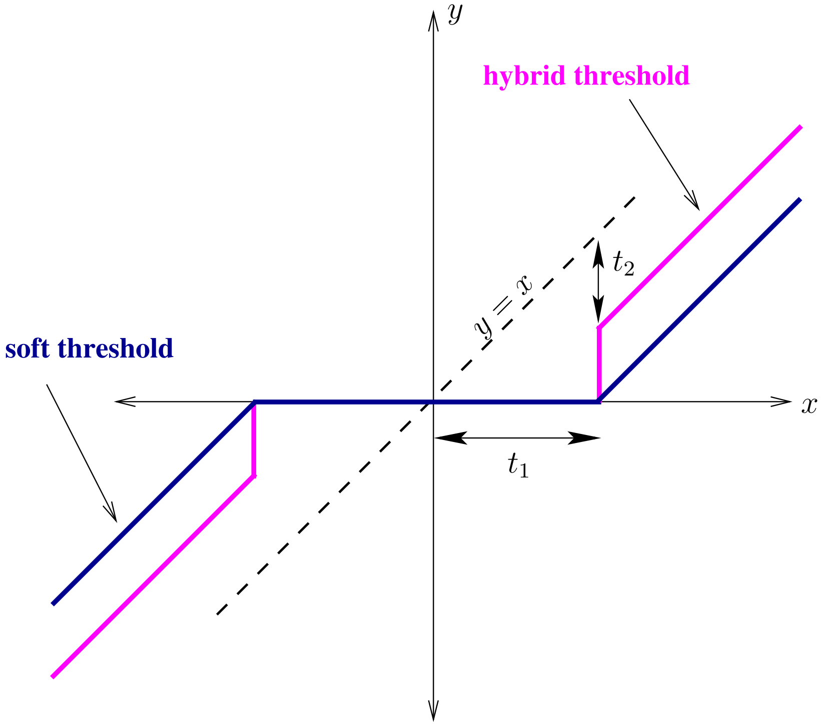


Figure 5.7: Hybrid hard-soft thresholding rule.

Equations (5.61) and (5.62) can be combined to yield

$$\theta_i = \tilde{\theta}_i I_i = \begin{cases} T_{\mathrm{hs}}(\hat{z}_i^{(n)}; a\alpha^2 + \sqrt{2\alpha^2 \log(\frac{1-w}{w})}, a\alpha^2) & 0 < w \leq \frac{1}{2} \\ T_{\mathrm{s}}(\hat{z}_i^{(n)}; a\alpha^2) & \frac{1}{2} < w \leq 1 \end{cases} \tag{5.63}$$

If $w > 1/2$, the soft-thresholding rule is applied in the Q-step of the EM iterations of MAP1. From earlier discussion, these iterations produce the L1 estimate (5.34) with regularization parameter $\beta = 2a\alpha^2$. However, if $0 < w \leq 1/2$, a larger thresholding value is used that increases the smaller w becomes. This is intuitively pleasing, as it is what we would expect.

MAP2

Since $\tilde{\theta}_i \neq 0$ w.p. 1, the set

$$\mathcal{I}_1 = \{i : I_i = 1\} = \{i : \theta_i \neq 0\} \text{ w.p. } 1. \tag{5.64}$$

This implies $\|\underline{I}\|_0 = \|\underline{\theta}\|_0$ w.p. 1. Applying (5.64) to the criterion to maximize, i.e., (5.56),

$$\begin{aligned} \Psi_2(\underline{\tilde{\theta}}, \underline{I}, \underline{\phi}) \triangleq -\frac{1}{2\sigma^2}\|\mathbf{H}\underline{\theta} - \underline{y}\|^2 + \|\underline{I}\|_0 \log w + (M - \|\underline{I}\|_0)\log(1-w) \\ + \|\underline{\theta}\|_0 \log\frac{a}{2} - a\|\underline{\theta}\|_1 + \sum_{\{i:I_i=0\}} \log g(\tilde{\theta}_i) \end{aligned} \tag{5.65}$$

The Hessian $\nabla_{\underline{\phi}}\nabla_{\underline{\phi}}^T\Psi_2$ is the same as (5.58) except that the $(1,1)$ entry is $-\|\underline{\theta}\|_0/a^2$; clearly, it is also negative definite for all $a > 0$ and $0 < w < 1$. The maximization in step (i) is obtained by solving for $\nabla_{\underline{\phi}}\Psi_2 = 0$, which produces

$$\hat{a} = \frac{\|\hat{\underline{\theta}}\|_0}{\|\hat{\underline{\theta}}\|_1} \text{ and } \hat{w} = \frac{\|\hat{\underline{\theta}}\|_0}{M}. \tag{5.66}$$

It is instructive to compare the equations for the hyperparameter estimates of MAP1 vs. MAP2. The main difference lies in the estimation of a. Assuming that the estimates $\hat{\underline{I}}$ and $\hat{\underline{\theta}}$ obey (5.64), one can re-write $\hat{a} = |\mathcal{I}_1|/\sum_{i\in\mathcal{I}_1}|\hat{\tilde{\theta}}_i|$. The MAP2 estimate of a can then be interpreted as the ML estimate using only the $\hat{\tilde{\theta}}_i$ for $i \in \mathcal{I}_1$. This is a consequence of g being independent of a, w.

On the other hand, the MAP1 estimate of a can be written as

$$\hat{a} = \frac{|\mathcal{I}_1| + |\mathcal{I}_0|}{\sum_{i\in\mathcal{I}_1} |\hat{\tilde{\theta}}_i| + \sum_{i\in\mathcal{I}_0} |\hat{\tilde{\theta}}_i|}. \tag{5.67}$$

As was previously noted, all of the $\hat{\tilde{\theta}}_i$ are used, even those $i \in \mathcal{I}_0$. This is contrary to the intent of a, which is to model the variance of the non-zero values of $\underline{\theta}$. The estimate $\hat{\tilde{\theta}}_i$ for $i \in \mathcal{I}_0$ should not affect $\hat{a}$. We will see later on that there is a price to be paid for the condition on g assumed by MAP2.

As with MAP1, the maximization in step (ii) can be obtained by applying the EM algorithm with the complete data $\underline{z} = \underline{\theta} + \alpha \underline{w}_1$. The E-step is given in (5.14), and the Q function at the $(n+1)$th step is

$$\begin{aligned} Q(\underline{\tilde{\theta}}, \underline{I}; \underline{\hat{\tilde{\theta}}}^{(n)}, \underline{\hat{I}}^{(n)}) = -\frac{1}{2\alpha^2}\|\underline{\theta} - \underline{\hat{z}}^{(n)}\|^2 + \|\underline{I}\|_0 \log w + (M - \|\underline{I}\|_0)\log(1-w) \\ + \sum_{i\in\mathcal{I}_1} \log(\frac{1}{2} a e^{-a|\tilde{\theta}_i|}) + \sum_{i\in\mathcal{I}_0} \log g(\tilde{\theta}_i). \end{aligned} \tag{5.68}$$

Define

$$g^* \triangleq \sup_x g(x), \;\; r \triangleq \frac{g^*}{a/2}\frac{1-w}{w}, \;\text{ and } G(y) \triangleq \{x : g(x) = y\}. \tag{5.69}$$

The maximization of Q in the nth iteration results in

$$\tilde{\theta}_i = \begin{cases} T_{\mathrm{s}}(\hat{z}_i^{(n)}; a\alpha^2) & I_i = 1 \\ \text{any element of } G(g^*) & I_i = 0 \end{cases}, \tag{5.70}$$

$$\text{and:} \quad I_i = \begin{cases} I(|\hat{z}_i^{(n)}| > a\alpha^2 + \sqrt{2\alpha^2 \log r}) & r \geq 1 \\ 1 & 0 \leq r < 1 \end{cases}. \tag{5.71}$$

The resulting θ is given by the following thresholding rule

$$(5.72) \qquad \theta_i = \begin{cases} T_{\mathrm{hs}}(\hat{z}_i^{(n)}; a\alpha^2 + \sqrt{2\alpha^2 \log r}, a\alpha^2) & r \geq 1 \\ T_{\mathrm{s}}(\hat{z}_i^{(n)}; a\alpha^2) & 0 \leq r < 1 \end{cases}$$

which is similar to the MAP1 solution in (5.63). Indeed, the MAP1 solution can be obtained by setting $g^* = a/2$. In this case, $r = (1-w)/w$, and $r \geq 1 \iff w \leq 1/2$.

The tuning parameter g^* is an extra degree of freedom that arises due to g being independent of a, w. This makes the MAP2 solution a function of g^*, and it is incumbent on the practitioner to select a suitable g^*. In contrast, MAP1 has no free tuning parameters: they are all automatically estimated.

Just like in MAP1, the EM iterations of MAP2 produce a larger threshold the sparser the hyperparameter w is. As well, if a is smaller, r increases. Since the variance of the Laplacian $\gamma(\cdot; a)$ is $2/a^2$, a smaller a implies a larger variance of the Laplacian. It makes sense that a larger threshold is used. It is not clear how to select g^*. One could suppose that g were a uniform distribution with finite support. As the prior $p(\tilde{\theta}_i | I_i = 0) = g(\tilde{\theta}_i)$ becomes more uninformative, $g^* \to 0$. The thresholding rule for $r \geq 1$ then becomes the same as for $0 < r < 1$, and the EM iterations will produce a L1 estimate. The MAP2 estimator when $g^* = 0$ will *not* be the L1-SURE estimator: in the former, the regularization parameter is chosen via MAP, while in the latter, it is chosen via minimizing the SURE criterion. It is interesting to note that the point(s) at which g attains g^* do not play a role in the estimate of $\underline{\theta}$ in step (ii).

5.6.5 Hybrid hard-soft thresholding function and SURE

The appearance of the hybrid hard-soft thresholding function in the MAP-based solution poses the question of whether or not it is a “better” thresholding function than the soft-threshold. Certainly, the collection of soft-thresholding functions is but a subset of all the possible HHS thresholding functions. Recall that the HHS thresholding function came from the MAP solution when a weighted average of a Dirac delta at zero and the Laplacian p.d.f. was used as the prior on $\underline{\theta}$. In this light, one would expect the HHS thresholding function to be a generalization of the soft-thresholding function. The collection of hard-thresholding functions is also a subset of all possible HHS thresholding functions. It has been noted in [26] that the hard-thresholding rule used in the framework of (5.16) can exactly recover a sparse representation under certain conditions.

Just like the SURE criterion was used to select the regularization parameter β for the L1 estimator, we propose to use SURE to select the parameters t_1 and t_2 of the hybrid thresholding function.

Proposition 5.4. *When $\|\mathbf{H}\| < 1$, the iterations (5.16) with $c = 1$ and $\mathcal{D}(\cdot)$ taken to be the hybrid thresholding function minimizes the cost function*

$$\Psi_{hs}(\underline{\theta}) = \|\mathbf{H}\underline{\theta} - \underline{y}\|^2 + \sum_i J_1(\theta_i)$$

(5.73)

where: $J_1(x) = I(|x| < t_1 - t_2)[-(x - sgn(x)t_1)^2 + 2t_1t_2] + I(|x| \geq t_1 - t_2)(2t_2|x| + t_2^2)$

See Appendix D.3 for the derivation. This result was obtained by applying Prop. 5.3. We shall call this estimator the hybrid hard-soft estimator. This is an example of where Prop. 5.3 is used to design a sparse penalty from a thresholding rule. ■

As a check, it was previously noted that the soft-thresholding function $T_s(\cdot; t) = T_{\text{hs}}(\cdot; t, t)$. When $t_1 = t_2 = t$, (5.73) produces

$$J_1^{\text{s}}(x) = 2t|x| + t^2 \tag{5.74}$$

which gives rise to the L1 estimator, as expected. The difference between the J_1^{s} of the L1 estimator and the J_1 of the hybrid thresholding function is that, with the latter, there is a region $\{x : |x| < t_1 - t_2\}$ where the cost is quadratic as opposed to linear.

The concave-shaped quadratic cost can be used to encourage a higher level of sparsity. Since J_1 is symmetric, suppose without any loss of generality that $x > 0$. The quadratic part has derivative $-2(x - t_1)$. For $x \in [0, t_1 - t_2) \implies 2t_1 \geq -2(x - t_1) > 2t_2$. Consider the case when the L1 estimator is computed with $t_1 = t_2 = t$. If we were to hold t_2 fixed but increase t_1, the slope of J_1 in $x \in [0, t_1 - t_2)$ would become larger than $2t = 2t_2$. In this manner, the hybrid thresholding function has the ability to produce a sparser solution. The penalty function $J_1(x)$ is plotted in Fig. 5.8 for $t_1 = 1$ and various $t_2 \in [0, 1]$. When $t_1 = t_2$, $J_1(x) = |x| + \text{const}$, which is consistent with our expectations. When $t_2 = 0$, the HHS thresholding rule is the hard threshold. The corresponding $J_1(x)$ is like a 0-1 penalty term.

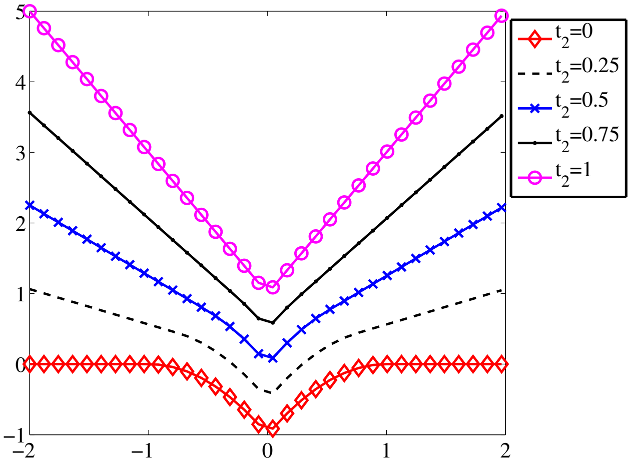


Figure 5.8: Penalty function $J_1(x)$ derived from the HHS thresholding rule for $t_1 = 1$ and various $t_2 \in [0, 1]$.

For an arbitrary $\|\mathbf{H}\| < C$, the normalization (5.39) is first carried out before the iterative framework of (5.16) is applied. For the remainder of section 5.6.5, we shall assume that $\|\mathbf{H}\| < 1$ without loss of generality.

Let $\underline{t} = (t_1, t_2)$. SURE can be used to construct an unbiased estimate of the risk

$$R(\underline{\theta}, \underline{t}) \triangleq E_{\underline{Y}} \|\mathbf{H}\hat{\underline{\theta}}_{hs}(\underline{y}; \underline{t}) - \mathbf{H}\underline{\theta}\|^2 \tag{5.75}$$

where $\hat{\underline{\theta}}_{hs}(\underline{y}; \underline{t})$ minimizes the cost $\Psi_{hs}(\underline{\theta})$. The expression for SURE of the HHS estimator is not as tractable as SURE of the L1 estimator. In the derivation of SURE of the HHS estimator, assume that the columns of $\mathbf{H}$ are linearly independent and that the Gram matrix $\mathbf{G}(\mathbf{H}) \triangleq \mathbf{H}^T\mathbf{H}$ does not have an eigenvalue of $1/2$. The latter condition is equivalent to $\mathbf{H}$ not having any singular values of $1/\sqrt{2}$.

Several definitions are in order before giving the SURE result. Suppose that $\hat{\underline{\theta}}$ has r zero values and $(M - r)$ non-zero values. Define the permutation matrices $\mathbf{P}, \mathbf{Q} \in \mathbf{R}^{M \times M}$ such that $\mathbf{P}\text{diag}(\hat{\underline{\theta}})\mathbf{Q} = \text{diag}(0, \ldots, 0, x_1, \ldots, x_{M-r})$, where $x_i \neq 0$. In other words, $\mathbf{P}$ and $\mathbf{Q}$ re-arrange $\text{diag}(\hat{\underline{\theta}})$ so that all of the zero-valued $\hat{\theta}_i$s are in the front. Let the subscript $(\cdot)_{22}$ denote the lower $(M - r) \times (M - r)$ submatrix of the argument. Define

$$\mathbf{U}(\underline{\theta}) \triangleq \text{diag}\left(\text{rect}\left(\frac{\theta_1}{2(t_1 - t_2)}\right), \ldots, \text{rect}\left(\frac{\theta_M}{2(t_1 - t_2)}\right)\right). \tag{5.76}$$

when $t_2 < t_1$ and $\mathbf{0}$ when $t_2 = t_1$. Let $\mathbf{K} = (\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{Q})_{22}$ and $\mathbf{J} = -(1/2)(\mathbf{P}\mathbf{U}(\hat{\underline{\theta}}_{hs})\mathbf{Q})_{22}$. SURE of the HHS estimator is

$$\hat{R}(\underline{t}) = N\sigma^2 + \|\underline{y} - \mathbf{H}\hat{\underline{\theta}}_{hs}(\underline{t})\|^2 + 2\sigma^2 \text{tr}(\mathbf{K}[\mathbf{K} + \mathbf{J}]^{-1}) \tag{5.77}$$

To evaluate (5.77), one would have to build the matrices $\mathbf{P}, \mathbf{Q}$, and invert a $(M - r) \times (M - r)$ matrix. If $\hat{\underline{\theta}}_{hs}$ is sparse, $(M - r)$ will be small, and the inversion would not be that computationally demanding. It was previously noted that when $t_1 = t_2$, the HHS estimator reduces to the L1 estimator. We expect SURE of the HHS estimator, i.e., (5.77), to reduce to SURE of the L1 estimator when $t_1 = t_2$.

That is indeed the case: when $t_1 = t_2$, $\mathbf{J} = \mathbf{0}$ and the last term of (5.77) equals $2\sigma^2(M-r) = 2\sigma^2\|\hat{\underline{\theta}}\|_0$. $\hat{R}(\underline{t})$ can be approximated as

$$\hat{R}(\underline{t}) \approx N\sigma^2 + \|\underline{y} - \mathbf{H}\hat{\underline{\theta}}_{hs}(\underline{t})\|^2 + 2\sigma^2\|\mathcal{D}_{T_s}(\hat{\underline{\theta}}_{hs}(\underline{t}); t_1 - t_2)\|_0, \tag{5.78}$$

We expect the approximation to be relatively good if $t_1 \approx t_2$. See App. D.4 for details.

The optimum $\underline{t}$ corresponds to the $\underline{t} \in \mathcal{T} \triangleq \{(t_1, t_2) : t_1 \geq 0, 0 \leq t_2 \leq t_1\}$ that minimizes $\hat{R}(\underline{t})$. The corresponding $\hat{\underline{\theta}}_{hs}(\underline{t})$ would be the output. We shall call this method HHS-SURE.

Rather than do a two dimensional search for the minimum of SURE of the HHS estimator, we suggest a suboptimal search for the minimum that consists of a sequence of line searches. Notice that L1-SURE finds the optimal parameters $\underline{t}$ when $t_1 = t_2$ according to the HHS SURE criterion. This is equivalent to a line optimization of HHS-SURE in the direction of $(1, 1)$. Another line optimization can be performed from the L1-SURE solution, say in the direction of $(1, 0)$. Here, t_2 is kept fixed while the minimum of the HHS SURE criterion is sought for larger t_1. From previous discussion, this is expected to produce a sparser solution. We need not restrict the second line optimization to $(1, 0)$, but can also consider $(0, -1)$, etc. The advantage of leveraging the L1-SURE solution is that it can be efficiently computed via the LARS algorithm (under the assumption that the columns of $\mathbf{H}$ are linearly independent). So the first line optimization will be fairly inexpensive, computationally speaking. Subsequent line optimizations have to be done in the iterative framework of (5.16), and will be more computationally expensive.

5.6.6 Computational aspects of proposed methods and SBL

The memory requirements of the proposed algorithms is dominated in the general case by the storage of $\mathbf{H}$, which is of $\mathcal{O}(MN)$. Manipulations of vectors with dimensions of $\underline{y}$ and $\underline{\theta}$ involve memory of $\mathcal{O}(\max(M,N))$. If $\mathbf{H}$ is sparse or implements the convolution operator with a point spread function (psf), and the support of the psf is smaller than $\mathcal{O}(\max(M,N))$, then the overall memory requirement is $\mathcal{O}(\max(M,N))$.

We shall discuss the computational complexity of the proposed algorithms in the remainder of the section.

EBD and SBL

An optimization is performed in each iteration of the EBD-based methods. Specifically, the optimization is to find the MML estimate of the hyperparameter given an estimate of the complete data $\underline{z}$. Fortunately, in the case of EBD-LAZE, the hyperparameter $\underline{\phi}$ is two-dimensional, and an increase in the image size M will not affect the dimension of the search space. It is possible to decrease the computational cost of finding the MML estimate of the hyperparameter in each iteration by updating it only every nth iteration, for $n > 1$. If the hyperparameter estimate changes slowly, such a modification should not overly affect the performance.

It would be desirable to obtain a single estimate of the hyperparameter, and to use it for all of the iterations. In the EBD-based framework, this would ensure convergence, according to Prop. 5.2. As well, it obviates the need to perform a search for the MML estimate of the hyperparameter in each iteration. SBL is an example where this was successfully done. The approach adopted there would work only if one had a computationally efficient way of computing the marginalized likelihood

$p(\underline{y}|\underline{\phi})$. In the case of SBL, a closed-form expression of $p(\underline{y}|\underline{\phi})$ was obtained. We were not able to obtain a closed-form expression of $p(\underline{y}|\underline{\phi})$ for the LAZE sparse prior. Obtaining an exact MML estimate of the hyperparameter $\underline{\phi}$ from the joint density $p(\underline{y}, \underline{z}|\underline{\phi})$ is more often intractable than tractable [83].

The disadvantage of SBL is that the dimension of the hyperparameter search space is M; contrast that with a fixed number of two for EBD-LAZE. Moreover, each iteration of SBL requires the inversion of a $P \times P$ matrix, where $P = \min(M, N)$. As M and N increase, SBL becomes more computationally expensive. Additionally, the inversion of a large matrix is, in general, numerically unstable. This would affect the optimization of the hyperparameter.

MAP1 and MAP2

MAP1 and MAP2 are computationally thrifty algorithms. The EM iterations are in closed form, as are the optimal hyperparameter estimates. For MAP2, however, there is a regularization parameter that needs to be tuned. An automatic tuning method will naturally increase the computational complexity of MAP2.

L1-SURE and HHS-SURE

The LARS algorithm can be used to implement L1-SURE efficiently. As was previously mentioned in the literature review, LARS solves for the L1 estimator in an iterative fashion. The number of iterations is approximately on the order of M. In the version that is considered, the iterations essentially sweeps the regularization parameter β from a large number (thus producing a solution of all zeros) to zero (thus producing the least-squares solution). The version of LARS-LASSO that we use starts with the zero vector, and adds a coefficient to $\underline{\hat{\theta}}$ at each step in most cases, although it might zero out a location in $\underline{\hat{\theta}}$ once in a while. See [15] for the details. Let a be the number of non-zero elements of $\underline{\hat{\theta}}$ in the current LARS step. Then, the inversion of a matrix of size $a \times a$ is required in order to proceed to the next step. In practice, one would not need to run LARS all the way to the least-squares solution, as perhaps an upper bound on $\|\underline{\theta}\|_0$ is available. One would stop the LARS iterations once $\|\underline{\hat{\theta}}\|_0$ achieved this upper bound. If the expected $\underline{\theta}$ is highly sparse, then the inversion will only be applied to matrices of small sizes. L1-SURE can also be implemented as a parallel operation of several sub-tasks. Each sub-task would compute the estimator $\underline{\hat{\theta}}_{l1}(t)$ for different values of t, and evaluate the SURE criterion. At the end, the different criterion values are compared; the sub-task with the lowest value would have its estimator selected as the L1-SURE estimator.

The HHS-SURE algorithm in its general form requires the minimization of the HHS SURE criterion. One implementation would be to lay out a grid of points covering the area $\{(t_1, t_2) : 0 < t_1 < t_{\max}, 0 \leq t_2 \leq t_1\}$ for some $t_{\max} > 0$, and compute $\underline{\hat{\theta}}_{hs}(\underline{t})$ for each $\underline{t}$ using the iterative thresholding framework. This will be computationally expensive. The run time of the implementation can be decreased via parallel processing. In a cluster of computers, each node can be assigned to compute $\underline{\hat{\theta}}_{hs}(\underline{t})$ for one or two $\underline{t}$ values. The HHS-SURE algorithm that suboptimally finds the minimum of the HHS SURE criterion via a sequence of line searches decreases the computational complexity at the expense of accuracy. If the L1-SURE solution is used as the first line optimization step, the computational complexity can be reduced by an additional amount. The overall complexity reduction might be minor if many line optimizations are performed. If only several are performed, the reduction will be more significant.

5.7 Simulations

The performance of the following methods will be compared over a range of SNRs: (i) the proposed sparse reconstruction methods; (ii) SBL; (iii) the standard and projected Landweber iteration. The projected Landweber iteration will use a projection on to the positive orthant so that the estimate $\hat{\underline{\theta}}$ is positive-valued. It implements a constrained optimization, as the iterations converge to a minimizer of $\|\mathbf{H}\underline{\theta} - \underline{y}\|^2$ subject to $\underline{\theta} \succeq \underline{0}$, i.e., each $\theta_i \geq 0$. The pseudocode for the various methods is given in App. E.

Two sparse images are investigated: a sparse binary-valued image, and an image that is based on the realization of the prior (5.26). The latter will be called the LAZE image. The image $\underline{\theta}$ studied was of size 32×32, and the noisy observation $\underline{y}$ was also of the same size. So $M = N = 1024$.

The binary-valued image has 12 pixels set to one, for a sparsity level of 1.2%. For the LAZE image, only the inner 26×26 pixels were drawn from the LAZE prior with parameters $a = 1$ and $w = 0.04$. This left a space of 3 pixels on each side of the image, so that convolution with the psf would not cause any wrap-around effects. This second image can be regarded as being approximately a realization of the LAZE prior with $a = 1$ and $w = 0.026$. The matrix $\mathbf{H}$ represented convolution with a Gaussian blur psf; its columns were linearly independent and its Gram matrix did not have an eigenvalue of 1/2. The Gaussian blur is illustrated in Fig. 5.9 below; its exact specification is given in App. F.

Define the SNR as $\text{SNR} \triangleq (N^{-1}\|\mathbf{H}\underline{\theta}\|^2)/\sigma^2$, and the SNR in dB as $\text{SNR}_{\text{dB}} \triangleq 10\log_{10}\text{SNR}$. The following error criteria will be used to assess the performance of the reconstruction methods:

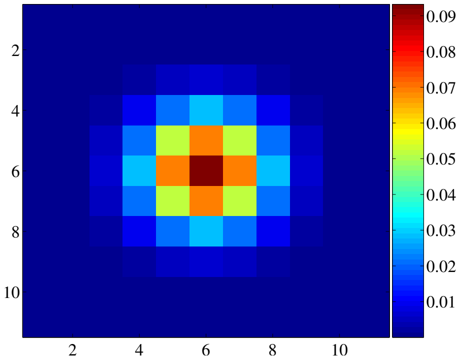


Figure 5.9: Gaussian blur used in the sparse image reconstruction simulations.

- The l_0, l_1, and l_2 measures of the *reconstruction error*, which is defined by $\underline{e} = \underline{\theta} - \hat{\underline{\theta}}$.

- A *detection error* criterion defined by

$$E_d(\underline{\theta}, \hat{\underline{\theta}}; \delta) \triangleq \sum_{i=1}^{M} |I(\theta_i = 0) - I(|\hat{\theta}_i| < \delta)| \tag{5.79}$$

  The threshold δ is used to select the value at which $\hat{\theta}_i$ is assumed to be zero. This is used to handle the round-off effects etc. in computers. As well, it addresses the fact that, to the human observer, small non-zero values are not discernible from zero values. We took $\delta = 10^{-2}\|\underline{\theta}\|_\infty$. The error criterion (5.79) measures the ability of the reconstruction technique to discriminate between the zero and non-zero values of $\underline{\theta}$. Accurately determining the support of a sparse $\underline{\theta}$ is more critical than the actual values [26, 69]. If the support is accurately determined, the non-zero coefficients can be optimized with respect to a fit with the observations, e.g., $\|\underline{y} - \mathbf{H}\hat{\underline{\theta}}\|^2$.

- Although not strictly an error criterion, the number of non-zero values of $\hat{\underline{\theta}}$, i.e., $\|\hat{\underline{\theta}}\|_0$, will be considered. We are interested in sparse solutions, and so would like a small number here.

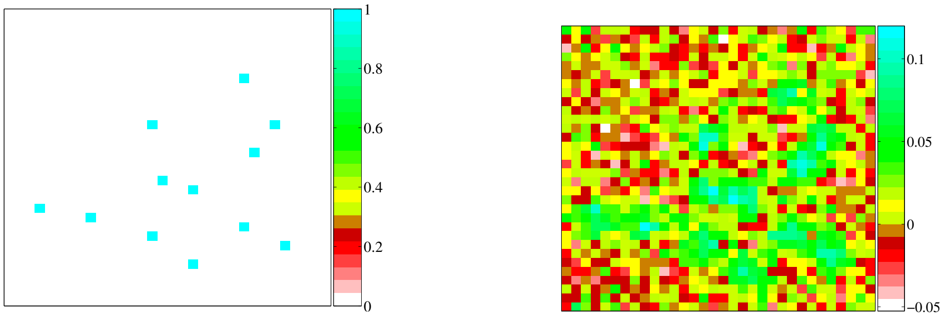
The proposed algorithms were implemented as outlined in the previous section.

In the case of MAP2, where the tuning parameter g^* needs to be specified, several different values were used. The HHS-SURE implementation that was used consisted of two line searches. First, the L1-SURE solution was computed. Then, a line search in the direction $(1, 0)$ was performed, i.e., t_2 was kept constant and t_1 was increased.

This section is divided into four parts. In the first, SBL will be compared to the following methods: the standard and projected Landweber iterations; and the proposed methods: EBD-LAZE, MAP1, MAP2, L1-SURE, and HHS-SURE. Because SBL is computationally intensive and takes a long time to run, only two different SNRs were investigated, viz. , SNR = 1.76 dB and 20 dB. The exact expression for SURE of the HHS estimator was used, i.e., (5.77). In the second part, the five proposed methods and the two Landweber iterations will be compared to each other in terms of performance vs. SNR for SNR values ranging from 1.76 dB to 20 dB. The noise variances corresponding to these SNR values are given in App. F. The exact SURE expression for the HHS estimator was used. It was observed that the SURE approximation (5.78) matched the exact expression for the range of SNRs considered. In the third part, the quality of the hyperparameter estimates of EBD-LAZE, MAP1, and MAP2 will be investigated. Finally, reconstruction examples involving the MRFM psf are provided.

5.7.1 Performance of the reconstruction methods under low and high SNR

The performance of the estimators is given in the Table 5.1 for the binary-valued $\underline{\theta}$ with the SNR equal to 1.76 dB (low SNR) and 20 dB (high SNR). The number reported in Table 5.1 is the mean over the simulation runs. The best mean number for each criterion is underlined. In terms of the error criteria, this is the lowest number. However, in terms of $\|\hat{\underline{\theta}}\|_0$, it would be the number closest to $\|\underline{\theta}_{\text{true}}\|_0$. The binary-valued θ is displayed in Fig. 5.10a, and a realization of the noisy observation under an SNR of 1.76 dB is displayed in Fig. 5.10b. The number of simulation runs executed for each row of the table is given in Table 5.2.



(a) Image $\underline{\theta}$

(b) Noisy observation $\underline{y}$

Figure 5.10: Binary-valued $\underline{\theta}$ and a realization of $\underline{y}$ under an SNR of 1.76 dB.

Table 5.1: Performance of the reconstruction methods for the binary-valued $\underline{\theta}$.

Method	$\|\|\underline{e}\|\|_0$	$\|\|\underline{e}\|\|_1$	$\|\|\underline{e}\|\|_2$	$E_d(\underline{\theta}, \hat{\underline{\theta}})$	$\|\|\hat{\underline{\theta}}\|\|_0$
SNR = 1.76 dB					
Landweber	1024	578.5	22.6	1000	1024
Proj. Landweber	88.9	10.3	1.66	68.5	88.8
SBL	1024	13.8	2.35	58.7	1024
EBD-LAZE	27.3	7.55	1.69	15.6	26.9
MAP1	12	12	3.46	12	0
MAP2, $g^* = (\sqrt{2})^{-1}$	15.49	2.72	0.912	3.68	15.3
L1-SURE	60	7.83	1.51	44.2	60.6
HHS-SURE	39.3	7.25	1.51	27.0	39.3
SNR = 20 dB					
Landweber	1024	86.1	3.67	929.3	1024
Proj. Landweber	84.9	1.20	0.20	27.0	84.9
SBL	1024	1.19	0.184	32.2	1024
EBD-LAZE	41.8	1.93	0.337	29.7	41.8
MAP1	43.9	1.07	0.209	22.9	43.9
MAP2, $g^* = (\sqrt{2})^{-1}$	229.5	3.82	0.380	114.4	229.5
L1-SURE	61.2	0.923	0.176	15.7	61.8
HHS-SURE	22.0	0.584	0.152	7.5	22.0

Some observations can be made regarding SBL. Firstly, SBL does not produce a strictly sparse solution for the two SNR conditions considered. In both, $||\hat{\underline{\theta}}||_0 = 1024$,

Table 5.2: Number of simulation runs for the comparison of the binary-valued $\underline{\theta}$ and the LAZE distributed $\underline{\theta}$.

Method	# runs for binary $\underline{\theta}$		# runs for LAZE $\underline{\theta}$	
	SNR = 1.76dB	SNR = 20dB	SNR = 1.76dB	SNR = 20dB
SBL	30	30	22	22
EBD-LAZE	30	50	30	30
MAP1	30	30	30	30
MAP2, $g^* = (\sqrt{2})^{-1}$	100	50	30	30
L1-SURE	30	30	30	30
HHS-SURE	30	30	30	30
Landweber	30	30	30	30
Proj. Landweber	30	30	30	30

which is the length of $\underline{\theta}$. Secondly, it does not have better performance than the proposed reconstruction methods. The same observations apply to the Landweber iterations. Under the low SNR value, Landweber has the worst performance, followed by SBL. SBL is more competitive under the high SNR value: its $||\underline{e}||_2 = 0.184$ is close to the lowest value of 0.163. Nonetheless, even here, its $||\underline{e}||_0 = 1024$ and its $||\hat{\underline{\theta}}||_0 = 1024$. SBL is not producing a sparse estimator even under the higher SNR number. On the other hand, the proposed methods produce sparse estimates under both the low and high SNR values. The Landweber iteration has lower numbers for $||\underline{e}||_1$ and $||\underline{e}||_2$ under the high SNR value, but is still not competitive with the proposed reconstruction methods.

The projected Landweber method, interestingly enough, sparsifies the estimate $\hat{\underline{\theta}}$ even though it does not enforce sparsity. Rather, it enforces non-negativity of $\hat{\underline{\theta}}$. In the low SNR case, there is a significant improvement over the standard Landweber method. It has a low $||\underline{e}||_2$ error; however, the other error metrics are far from the best values. Despite this, the error numbers are better than SBL's. In the high SNR case, the projected Landweber estimate is not competitive in so far as the $||\underline{e}||_0$, E_d, and $||\hat{\underline{\theta}}||_0$ criteria are concerned. However, it exhibits good $||\underline{e}||_2$ error, and its $||\underline{e}||_1$

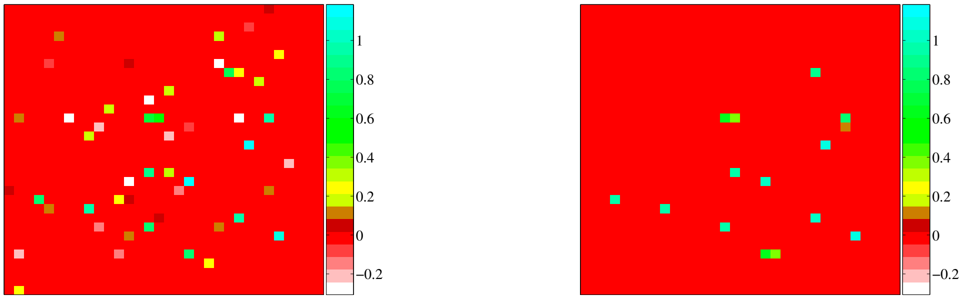
error puts it in the middle of the pack. In the high SNR case, SBL and the projected Landweber method have comparable performance in terms of $\|\underline{e}\|_1$. However, the latter produces a sparser result than SBL, as is evidenced by the $\|\hat{\underline{\theta}}\|_0$ values.

We shall now comment on the performance of the proposed reconstruction methods. In the low SNR case, MAP1 is consistently producing an output of all zeros. Effectively, MAP1 takes the conservative approach that $\underline{\theta}$ is $\underline{0}$. This results in MAP1's $\|\underline{e}\|_1 = \|\underline{\theta}\|_1$ and $\|\underline{e}\|_2 = \|\underline{\theta}\|_2$. MAP1 has the lowest $\|\underline{e}\|_0$ of all the estimators considered. However, the $\|e\|_0$ criterion is not always very meaningful. It measures the $0-1$ risk between $\underline{\theta}$ and $\hat{\underline{\theta}}$, which is a stringent error criterion. MAP2 with $g^* = (\sqrt{2})^{-1}$ has the lowest numbers for $\|\underline{e}\|_1$, $\|\underline{e}\|_2$, and $E_d(\underline{\theta}, \hat{\underline{\theta}})$, even though there is a mismatch between $\underline{\theta}$ and the prior that MAP2 uses (the discrete-continuous form of the LAZE prior). In the low SNR case, MAP2 has the best performance.

In the high SNR case, the HHS-SURE estimator has the best performance. The mean values of all the error criteria decrease as compared to L1-SURE. The greatest decreases are in $\|\underline{e}\|_0$, E_d, and $\|\hat{\underline{\theta}}\|_0$. They indicate that the HHS-SURE estimator is properly zeroing out spurious non-zero values and producing a sparser estimate than L1-SURE. Recall that the HHS thresholding function originated from the MAP solution when using the LAZE prior for $\underline{\theta}$. The incorporation of the Dirac delta at zero in the prior of $\underline{\theta}$ does indeed permit a sparser estimate.

The number $\|\hat{\underline{\theta}}\|_0$ does not necessarily give an accurate assessment of the *perceived* sparsity of the reconstruction. Consider the SBL reconstruction and the MAP2 reconstruction under the SNR of 1.76 dB in Figs. 5.11a and 5.11b respectively. The SBL reconstruction looks sparse despite having $\|\hat{\underline{\theta}}\|_0 = 1024$. This is because many of the non-zero pixel values have a small magnitude, and are visually indistinguishable from zero. We note that the MAP2 reconstruction closely resembles the original

$\underline{\theta}$, although there is some blurring around several non-zero pixel locations. On the other hand, the SBL reconstruction has many spurious non-zero pixels, in addition to blurring around several non-zero pixel locations. The SBL estimate contains negative values. This is not surprising, as positivity of $\underline{\theta}$ is not taken into account. Recall that SBL models the θ_i as a Gaussian r.v. Positivity is not taken into account in MAP2 either; however, as can be seen, the reconstruction is non-negative. We note that both methods reconstruct the amplitude of the positive non-zero pixels accurately. One possible remedy for SBL is to threshold $\hat{\underline{\theta}}$ at some pre-determined level. The ensuing image when the threshold $t = 10^{-2}\|\underline{\theta}\|_\infty = 10^{-2}$ is used is shown in Fig. 5.12a. Compared to the MAP2 estimator in Fig. 5.11b, the thresholded SBL estimator has more spurious positive values. When a threshold of $t = 0.5$ is chosen, the resulting image, given in Fig. 5.12b, resembles the MAP2 reconstruction. The challenge for a thresholded SBL estimator is to select the appropriate level of the threshold to use. This would require some prior knowledge of θ.

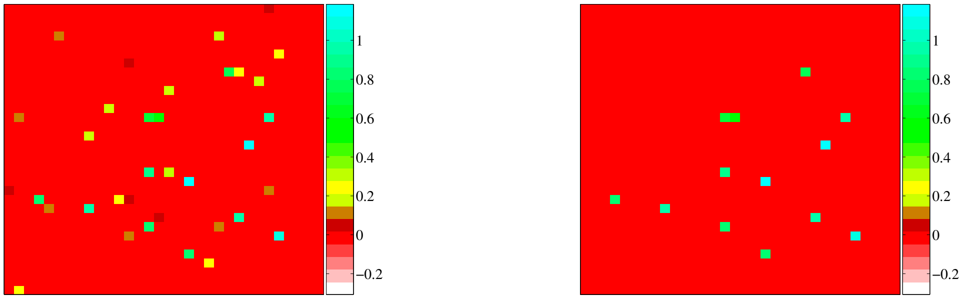


(a) SBL

(b) MAP2, $g^* = (\sqrt{2})^{-1}$

Figure 5.11: Reconstructed images for the binary-valued $\underline{\theta}$ under an SNR of 1.76 dB for SBL and MAP2.

The reconstruction for SBL and HHS-SURE at SNR = 20 dB is given in Figs. 5.13a and 5.13b respectively. The reconstructions are better in the high SNR case than in the lower SNR case, which is to be expected. The performance of SBL improves, and there are less spurious non-zero pixels. However, the HHS-SURE reconstruction

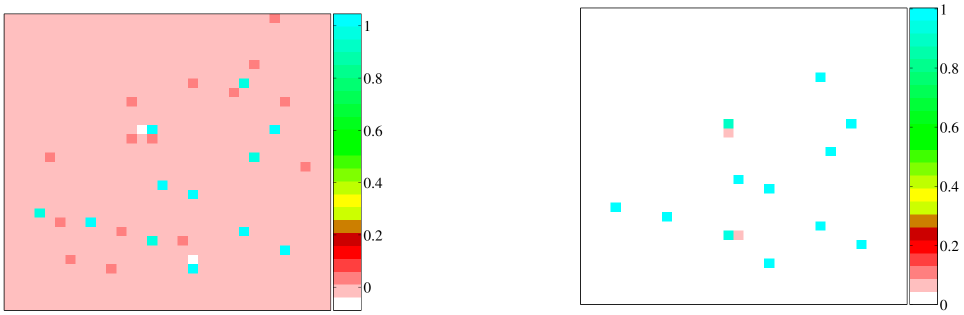


(a) Threshold $t = 10^{-2}$.

(b) Threshold $t = 0.5$.

Figure 5.12: Thresholded $\hat{\underline{\theta}}$ of SBL with two different threshold values.

more closely resembles $\underline{\theta}$. Both methods accurately reconstruct the non-zero pixel values as 1. From Table 5.1, HHS-SURE has better performance than SBL in the low SNR case as well as the high SNR case. The same cannot be said of MAP2. While it has better performance than SBL under the low SNR case, its performance deteriorates in the high SNR case.



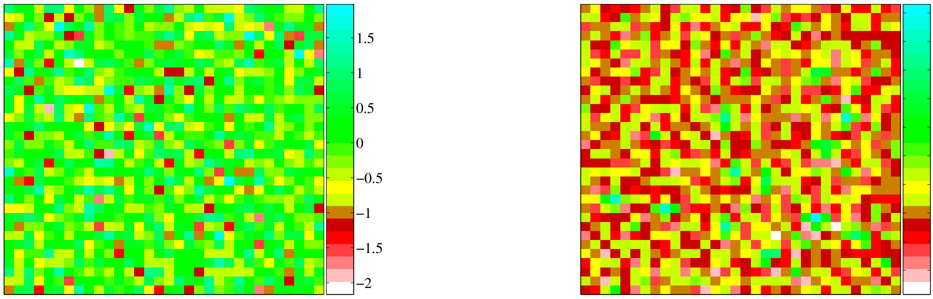
(a) SBL

(b) HHS-SURE

Figure 5.13: Reconstructed images for the binary-valued $\underline{\theta}$ under an SNR of 20 dB for SBL and HHS-SURE.

The Landweber reconstructions for the low and high SNR cases are given in Figs. 5.14a and 5.14b respectively. Under the low SNR case, there are many negative valued $\hat{\underline{\theta}}$. If one were to focus on just the positive-valued pixels, there are a number of spurious pixels incorrectly reconstructed around 1, the non-zero pixel value in $\underline{\theta}$. Under the high SNR case, we see that the higher SNR results in a smaller number of negative-valued pixels. Some of the non-zero pixel locations can be roughly discerned;

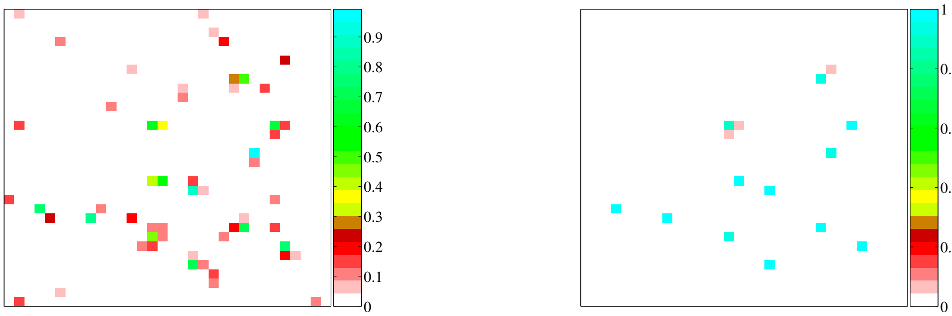
however, they have quite a bit of blurring. In both low and high SNR cases, the amplitudes of the non-zero pixels are not correctly reconstructed. Consider the projected Landweber reconstructions for the same observation $\underline{y}$. They are illustrated in Fig. 5.15a and 5.15b for the low and high SNR case respectively. In the low SNR case, the estimate $\hat{\underline{\theta}}$ has blurring around the non-zero pixel values and many spurious non-zero values. However, in the high SNR case, the reconstruction closely resembles the true $\underline{\theta}$. Despite the fact that the projected Landweber's error metrics are worse than HHS-SURE's, Fig. 5.15b is an acceptable reconstruction.



(a) SNR = 1.76dB

(b) SNR = 20dB

Figure 5.14: Reconstructed Landweber estimates for the binary-valued $\underline{\theta}$ under SNRs of 1.76 dB and 20 dB.

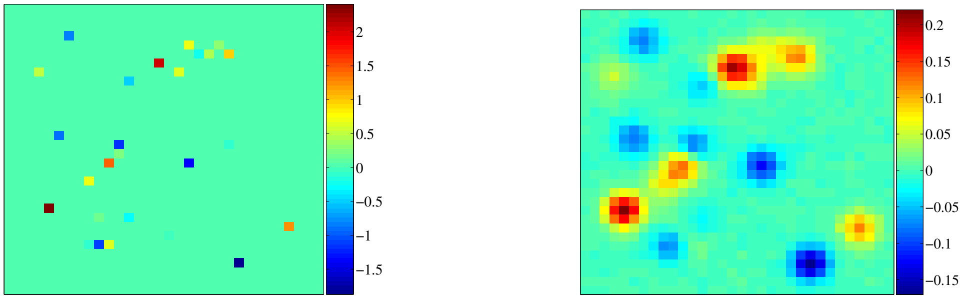


(a) SNR = 1.76dB

(b) SNR = 20dB

Figure 5.15: Reconstructed projected Landweber estimates for the binary-valued $\underline{\theta}$ under SNRs of 1.76 dB and 20 dB.

We shall move on to examine the performance of the reconstruction methods under the LAZE image. We expect that EBD-LAZE, MAP1, and MAP2 would have better performance here than the other methods, as the image $\underline{\theta}$ was generated using the LAZE prior. The LAZE $\underline{\theta}$ is displayed in Fig. 5.16a, and a noisy realization under an SNR of 20 dB is displayed in Fig. 5.16b.



(a) Image $\underline{\theta}$

(b) Noisy observation $\underline{y}$

Figure 5.16: LAZE distributed $\underline{\theta}$ and a realization of $\underline{y}$ under an SNR of 20 dB.

The numbers for the error criteria are given in Table 5.3. Again, the reconstruction method with the best number for each criterion is underlined. For the LAZE $\underline{\theta}$, $||\underline{\theta}_{\text{true}}||_0 = 27$. As in the case of the binary-valued $\underline{\theta}$, the Landweber iteration, which solves for the least-squares estimator, is not competitive in either the low or high SNR case. In the low SNR case, no one estimator dominates in terms of performance. However, it can be said that SBL does not have better performance than the proposed estimators. The estimators that use the LAZE prior for $\underline{\theta}$ appear to be competitive in terms of performance; more will be said on this later. We see that, under low SNR, MAP1 produces the conservative estimate of all zeros, just as with the case of the binary-valued $\underline{\theta}$.

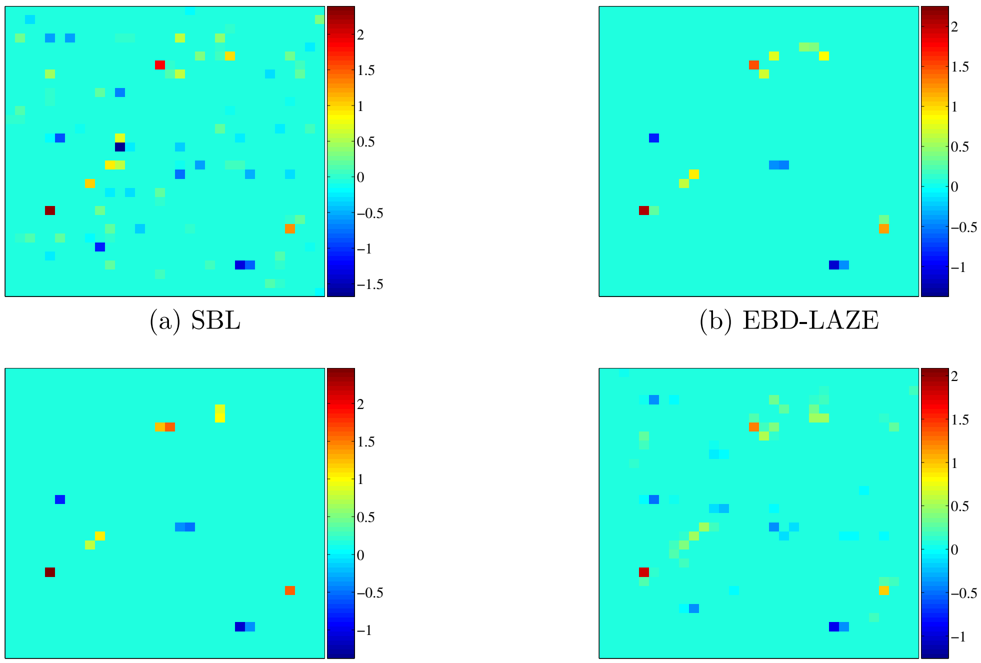
The reconstructions of SBL, EBD-LAZE, MAP2, and HHS-SURE under the SNR of 1.76 dB are given in Fig. 5.17a-d. The numbers for $||\hat{\underline{\theta}}||_0$ EBD-LAZE and MAP2 in Table 5.3 indicate that these two methods are producing estimates that are too sparse. There are in fact 27 non-zero values in $\underline{\theta}$; yet, the mean values of $\hat{\underline{\theta}}_0$ for EBD-LAZE and MAP2 are 13.6 and 9.77 respectively. That is reflected in Figs. 5.17b,c. Both appear to be similar, and do not correctly reconstruct several negative-valued

Table 5.3: Performance of the reconstruction methods for the LAZE $\underline{\theta}$.

Method	$\lVert\underline{e}\rVert_0$	$\lVert\underline{e}\rVert_1$	$\lVert\underline{e}\rVert_2$	$E_d(\underline{\theta},\hat{\underline{\theta}})$	$\lVert\hat{\underline{\theta}}\rVert_0$
SNR = 1.76 dB					
Landweber	1024	807.1	31.6	977.2	1024
Proj. Landweber	80.9	22.8	4.14	63.5	61.7
SBL	1024	28.1	3.99	72.6	1024
EBD-LAZE	33.7	$\underline{17.4}$	3.69	26.8	$\underline{13.6}$
MAP1	$\underline{27}$	21.2	5.21	27	0
MAP2, $g^* = (\sqrt{2})^{-1}$	30.9	17.5	3.98	$\underline{25.1}$	9.77
L1-SURE	92.6	20.3	3.15	69.3	81.9
HHS-SURE	67.2	19.1	$\underline{3.14}$	51.1	54.7
SNR = 20 dB					
Landweber	1024	122.2	5.34	855.5	1024
Proj. Landweber	69.6	17.3	3.85	37.5	52.3
SBL	1024	$\underline{4.32}$	$\underline{0.814}$	33.7	1024
EBD-LAZE	$\underline{62.4}$	8.56	1.62	41.4	$\underline{54.3}$
MAP1	69.7	6.53	1.34	31.9	63.8
MAP2, $g^* = (\sqrt{2})^{-1}$	216.3	10.8	1.44	86.6	211.5
L1-SURE	118.5	6.63	1.32	$\underline{31.1}$	115.8
HHS-SURE	84.4	6.73	1.35	33.0	78.7

pixels. As well, there are some spurious non-zero values. The SBL estimate overestimates the number of non-zero pixels. Although most of the negative values of $\underline{\theta}$ are reconstructed, other spurious negative values are also present. The reconstruction of HHS-SURE is better in the sense that it has less artifacts, although several pixels are blurred. These reconstructions provide an idea of why no one estimator is dominating in performance.

The reconstructions for SBL, EBD-LAZE, L1-SURE, and HHS-SURE under the SNR of 20 dB are illustrated in Figs. 5.18a-d. The amplitude of one of the negative-valued pixels is not properly reconstructed in L1-SURE, and there are spurious non-zero pixels. HHS-SURE reduces the number of non-zero pixels; however, it is not able to correctly reconstruct the amplitude of the negative-valued pixel in the lower-left



(a) SBL

(b) EBD-LAZE

(c) MAP2, $g^* = (\sqrt{2})^{-1}$

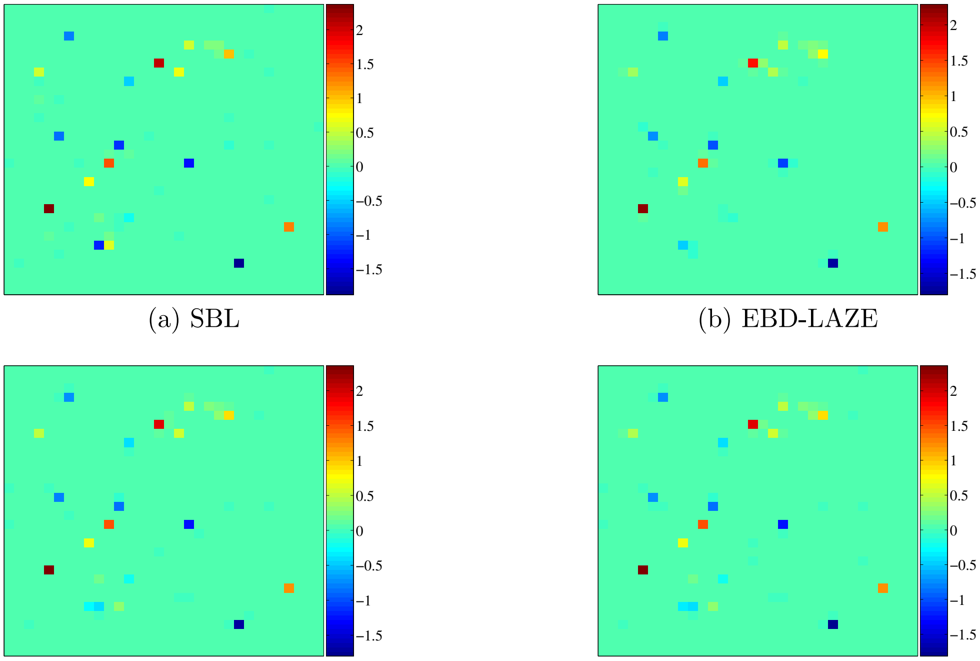
(d) HHS-SURE

Figure 5.17: Reconstructed images for the LAZE $\underline{\theta}$ under an SNR of 1.76 dB for SBL, EBD-LAZE, MAP2 with $g^* = (\sqrt{2})^{-1}$, and HHS-SURE.

corner. Rather, the pixel is blurred. The reconstruction for SBL and EBD-LAZE are similar except in two respects. While the EBD-LAZE reconstruction looks sparser, there is blurring around some of the non-zero pixel locations. In contract, SBL produces an image that has more spurious artifacts, but it correctly reconstructs the amplitude of the non-zero pixels, and does not have any blurring. It is likely that, for this reason, SBL has the lowest $\|\underline{e}\|_1$ and $\|\underline{e}\|_2$ numbers in the high SNR case.

Extrapolating from the simulation results, it appears that SBL has better performance in cases when the non-zero pixels of $\underline{\theta}$ assume both positive and negative values, and when under high SNR. When the non-zero pixels of $\underline{\theta}$ assume only positive or negative values (but not both cases), or when the SNR is low, SBL has poorer performance. The other aspect that makes SBL unappealing is its heavy computational complexity.

Based on the values of the error criteria, the Landweber iteration is not a compet-



(a) SBL

(b) EBD-LAZE

(c) L1-SURE

(d) HHS-SURE

Figure 5.18: Reconstructed images for the LAZE $\underline{\theta}$ under an SNR of 20 dB for SBL, EBD-LAZE, and HHS-SURE.

itive method over the range of SNRs considered. The projected Landweber iteration produces estimates with better error numbers than the standard Landweber iteration under both low and high SNR. The improvement in going from low to high SNR is not as marked as some of the other reconstruction methods. This is undoubtedly due to the model mismatch in the true $\underline{\theta}$ and the projection operator used by the projected Landweber method. The LAZE $\underline{\theta}$ has negative pixel values, whereas the projection used projects the iterative estimates on to the positive orthant.

The fact that EBD-LAZE, MAP1, and MAP2 did not produce superior performance over the other methods in the case of the LAZE image is unintuitive. One would think that, if the prior assumed is the same as the prior that is used, then the reconstruction technique has an advantage. The question to consider is if the estimate of the hyperparameters is accurate. The hyperparameter estimates for the three methods is shown in Table 5.4 below. The number reported is the mean over the simulations along with one standard deviation. Clearly, the hyperparameter estimates are biased. EBD-LAZE's hyperparameter estimate is less biased than MAP1's and MAP2's. The bias of the hyperparameter estimate is perhaps why these three methods do not perform as well as one would have expected.

Table 5.4: Hyperparameter estimates of EBD-LAZE, MAP1, and MAP2 with $g^* = (\sqrt{2})^{-1}$.

Method	$\hat{a}$	$\hat{w}$
	SNR = 1.76 dB	
EBD-LAZE	1.72 ± 0.310	$(18.8 \pm 2.24) \times 10^{-3}$
MAP1	276.3 ± 38.9	$(7.78 \pm 1.48) \times 10^{-3}$
MAP2, $g^* = (\sqrt{2})^{-1}$	0.752 ± 0.0743	$(9.54 \pm 1.44) \times 10^{-3}$
	SNR = 20 dB	
EBD-LAZE	3.65 ± 0.686	$(6.34 \pm 1.28) \times 10^{-2}$
MAP1	52.1 ± 0.329	$(6.23 \pm 0.446) \times 10^{-2}$
MAP2, $g^* = (\sqrt{2})^{-1}$	9.04 ± 2.29	0.207 ± 0.0582

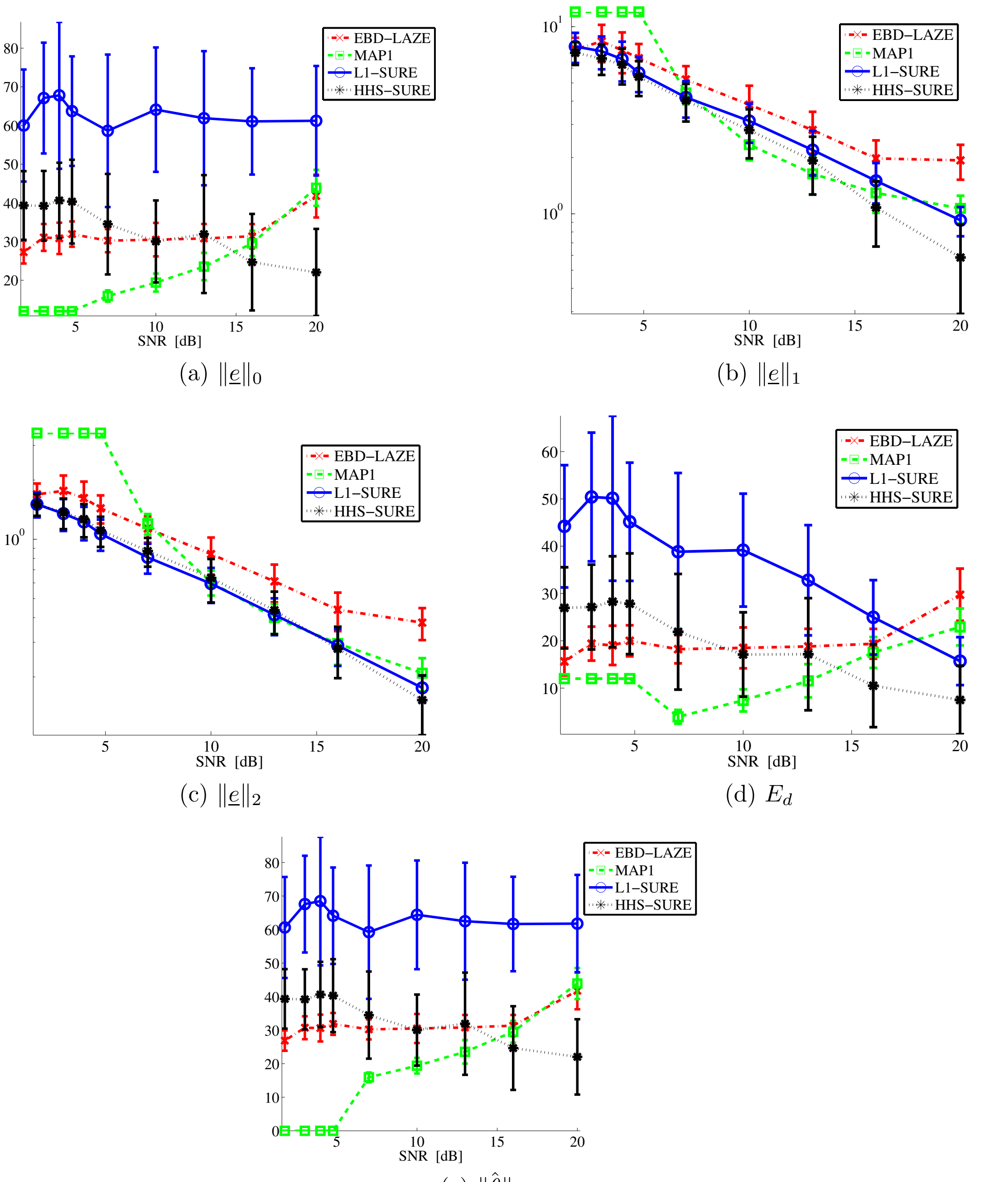
This ends the simulation study of SBL vs. the other reconstruction methods. From this point onwards, only the proposed reconstruction methods and the two Landweber iterations will be studied.

5.7.2 Performance vs. SNR of the proposed reconstruction methods

The performance of the proposed reconstruction methods when applied to the binary-valued $\underline{\theta}$ is examined with respect to SNR. We shall include the standard and projected Landweber method in this section. L1-SURE will be used to benchmark the other estimators. Even though the SURE criterion has not been previously applied to selecting the regularization parameter for the L1 estimator, the l_1 penalty is known to encourage sparsity. Because there are several methods to consider, the presentation will be divided into three sets. The first set contains EBD-LAZE, MAP1, L1-SURE, and HHS-SURE; the second contains MAP2 with the values of $g^* = 10^{-4}, (\sqrt{2})^{-1}, 10^3$, and L1-SURE; the third set contains L1-SURE, Landweber, and the projected Landweber iteration. The number of simulation runs for each estimator is given in Table 5.2. For each estimator, the mean is plotted along with error bars of one standard deviation.

The plots of error criteria for the first set is given in Fig. 5.19. First, consider the $||\underline{e}||_0$, $||\underline{e}||_1$, and $||\underline{e}||_2$ error criteria. L1-SURE has the highest $||\underline{e}||_0$ over the range of SNR considered. However, in terms of $||\underline{e}||_1$ and $||\underline{e}||_2$, it is a competitive estimator. In contrast, MAP1 is unable to distinguish the location of the non-zero pixels in low SNR. Under high SNR conditions, it has performance that is comparable to L1-SURE and HHS-SURE in terms of the $||\underline{e}||_1$ and $||\underline{e}||_2$ errors. The error curve for EBD-LAZE is always higher than L1-SURE's, except in the case of $||\underline{e}||_0$. The value of $||\underline{e}||_0$ increases with respect to increasing SNR for MAP1 and EBD-LAZE. Taken together with the $||\underline{e}||_1$ and $||\underline{e}||_2$ curves, the trend is indicative of small non-zero coefficients appearing in $\hat{\underline{\theta}}$ that are spurious. L1-SURE and HHS-SURE do not exhibit the same behaviour. HHS-SURE's error curve is lower than L1-SURE's for $||\underline{e}||_0$ and $||\underline{e}||_1$, and it is almost identical for $||\underline{e}||_2$.

Next, consider the E_d and $||\hat{\underline{\theta}}||_0$ error criterion. L1-SURE's curve for $||\hat{\underline{\theta}}||_0$ is relatively flat, and its E_d curve decreases for high SNR. This indicates that, while the number of non-zero coefficients in $\hat{\underline{\theta}}$ remains the same, the amplitude at the spurious locations are decreasing. With MAP1 and EBD-LAZE, the opposite trend is true. For low SNR, the number of non-zero coefficients in $\hat{\underline{\theta}}$ is small, but increases with higher SNR. A similar increase can be seen in the E_d curves. One can conclude that the number of spurious non-zero locations is increasing, which is unexpected. This phenomenon is likely due to the bias of the hyperparameter estimates. With HHS-SURE, both the E_d and $||\hat{\underline{\theta}}||_0$ curves decrease slightly as the SNR increases. This behaviour is intuitive, as higher SNR should result in better performance. The



(a) $\|\underline{e}\|_0$

(b) $\|\underline{e}\|_1$

(c) $\|\underline{e}\|_2$

(d) E_d

(e) $\|\underline{\hat{\theta}}\|_0$

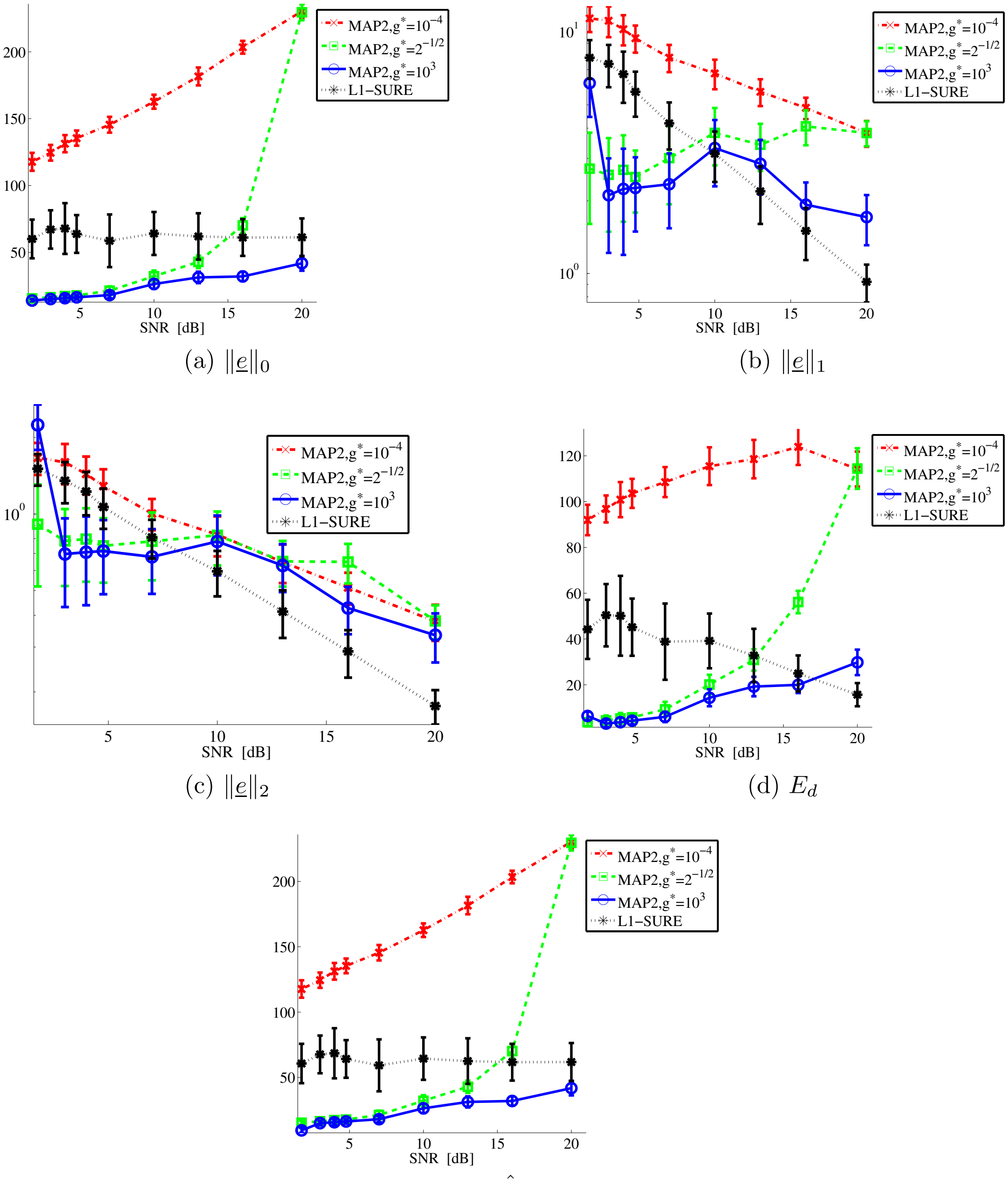
Figure 5.19: Performance vs. SNR for EBD-LAZE, MAP1, L1-SURE, and HHS-SURE when applied to the binary-valued $\underline{\theta}$.

plots Fig. 5.19b,c,d indicate that HHS-SURE produces a sparser estimate that has approximately the same (or slightly lower) $||\underline{e}||_1$ and $||\underline{e}||_2$.

The plots of error criteria for the second set of estimators is given in Fig. 5.20. As with the previous set of estimators, focus first on the error plots of $||\underline{e}||_0$, $||\underline{e}||_1$, and $||\underline{e}||_2$. It is clear that the selection of g^* is vital to the performance of MAP2. The choice $g^* = 10^{-4}$ is bad: its $||\underline{e}||_0$, $||\underline{e}||_1$, and $||\underline{e}||_2$ error curves all lie above L1-SURE's. The error curves for $g^* = (\sqrt{2})^{-1}$ and $g^* = 10^3$ have better performance than L1-SURE for low SNR, but worse performance at high SNR. For $g^* = (\sqrt{2})^{-1}$, $||\underline{e}||_0$ quickly increases as the SNR increases. This indicates that many spurious non-zero components are being added for higher SNR values. We note that all MAP2 versions perform worse than L1-SURE under the $||\underline{e}||_1$ and $||\underline{e}||_2$ criteria at high SNR.

It remains to consider the E_d and $||\hat{\underline{\theta}}||_0$ plots. The MAP2 curves for both criteria generally increases with higher SNR values. The notion of bias does not exactly apply in this scenario, as the binary-valued $\underline{\theta}$ is not from the LAZE prior. Nevertheless, the increasing trend is likely symptomatic of the bias effect of the hyperparameter estimate. As the SNR increases, an unbiased estimator $\hat{\underline{\phi}}$ would, in general, become more accurate. That should result in better performance.

The last group of estimators to consider is L1-SURE, Landweber, and the projected Landweber iteration. In the plots of $||\underline{e}||_1$ and $||\underline{e}||_2$, only the error curves for L1-SURE and the projected Landweber iteration are plotted, as inclusion of the Landweber error curve (it is a lot higher) will obscure the difference between the former two curves. It is interesting that the error curves for $||\underline{e}||_0$ and $||\hat{\underline{\theta}}||_0$ are flat from an SNR of 1.76 dB to 20 dB. The Landweber error curve is also flat for the E_d criterion, whereas L1-SURE's and the projected Landweber's curves decrease with higher SNR. We see from Fig. 5.21a-e that L1-SURE has better performance over



(a) $\|\underline{e}\|_0$

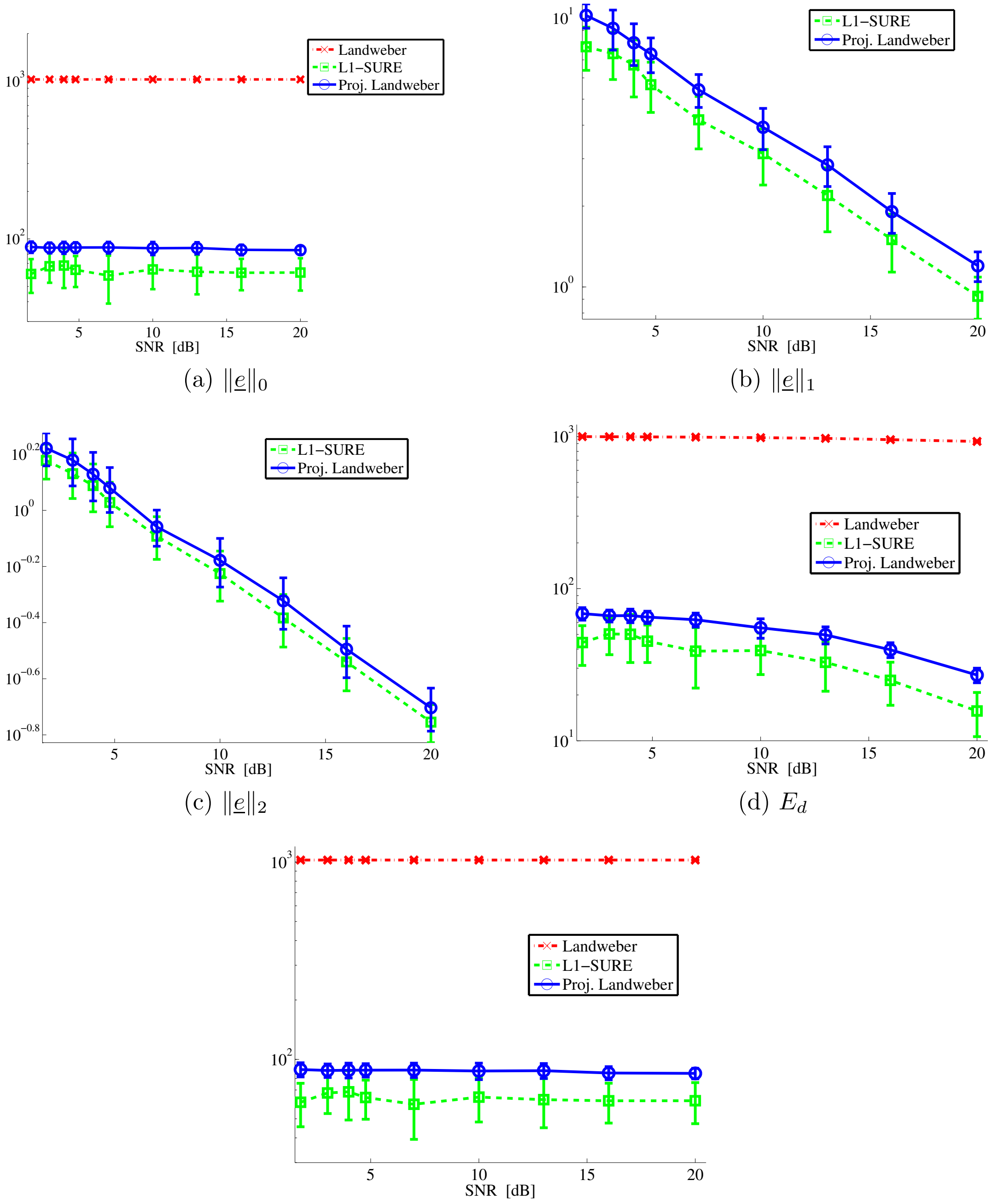
(b) $\|\underline{e}\|_1$

(c) $\|\underline{e}\|_2$

(d) E_d

(e) $\|\hat{\underline{\theta}}\|_0$

Figure 5.20: Performance vs. SNR for MAP2 with $g^* = 10^{-4}, (\sqrt{2})^{-1}, 10^3$, and L1-SURE when applied to the binary-valued $\underline{\theta}$.



(a) $\|\underline{e}\|_0$

(b) $\|\underline{e}\|_1$

(c) $\|\underline{e}\|_2$

(d) E_d

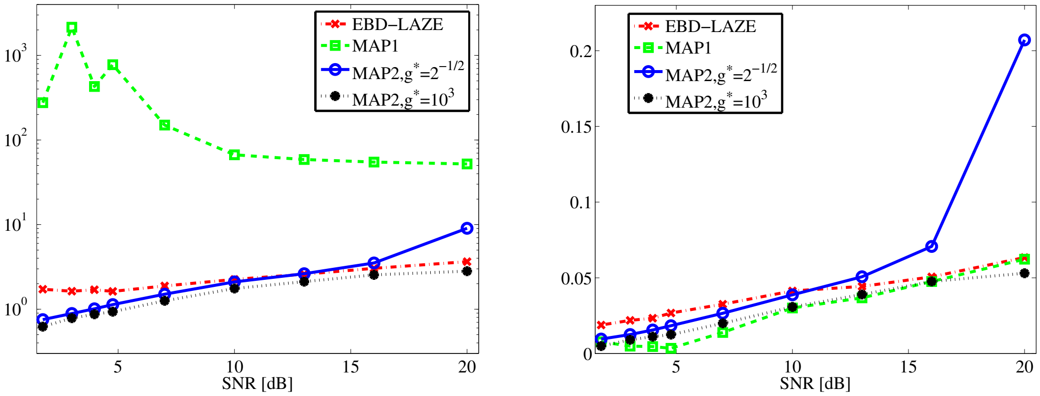
(e) $\|\hat{\underline{\theta}}\|_0$

Figure 5.21: Performance vs. SNR for L1-SURE, Landweber, and projected Landweber iteration when applied to the binary-valued $\underline{\theta}$.

the range of SNRs considered for the different error criteria.

5.7.3 Hyperparameter estimates of EBD-LAZE, MAP1, and MAP2

The quality of the hyperparameter estimates of EBD-LAZE, MAP1, and MAP2 will be examined when applied to the LAZE image. The mean hyperparameter estimate $\underline{\hat{\phi}} = (\hat{a}, \hat{w})$ is plotted vs. SNR in Figs. 5.22a-b. MAP1's $\hat{a}$ appears to converge as the SNR increases, whereas the other estimates slowly increase with higher SNR. However, MAP1's estimate is far from the true value of 1, whereas the other methods' estimates are closer. The estimators of EBD-LAZE and MAP2 are close to 1 when the SNR is in the range of 2–8 dB. The estimator $\hat{w}$ increases with higher SNR for all of the methods considered here. When the SNR is in the range of 2–8 dB, the estimates are close to 0.026.



(a) $\hat{a}$

(b) $\hat{w}$

Figure 5.22: Hyperparameter estimate vs. SNR for EBD-LAZE, MAP1, and MAP2 with $g^* = (\sqrt{2})^{-1}, 10^3$.

The E_d error curves of the reconstruction methods that use the LAZE prior is illustrated in Fig. 5.23. In addition, the L1-SURE error curve is displayed as a basis of comparison. When the hyperparameter estimate is relatively unbiased, we expect the MAP methods to perform well under the E_d criterion. This is because MAP minimizes the error probability $P_e(\underline{\hat{\theta}}) = P(\|\underline{\theta} - \underline{\hat{\theta}}\| > \epsilon)$, and the $E_d(\underline{\theta}, \underline{\hat{\theta}})$

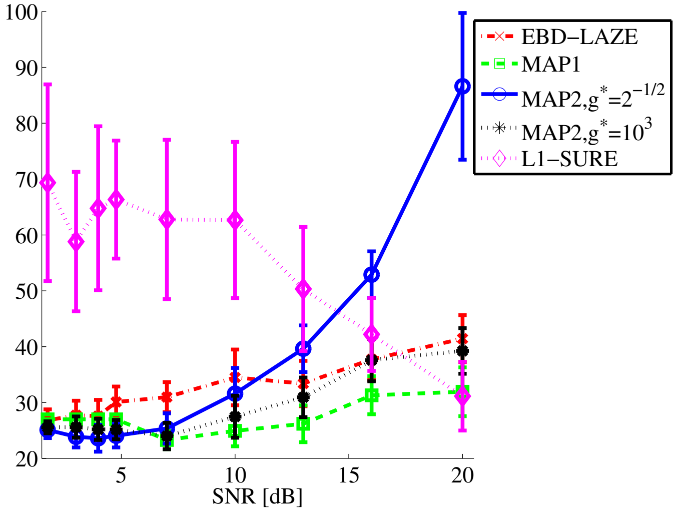


Figure 5.23: E_d error curves for EBD-LAZE, MAP1, MAP2, and L1-SURE with the LAZE image.

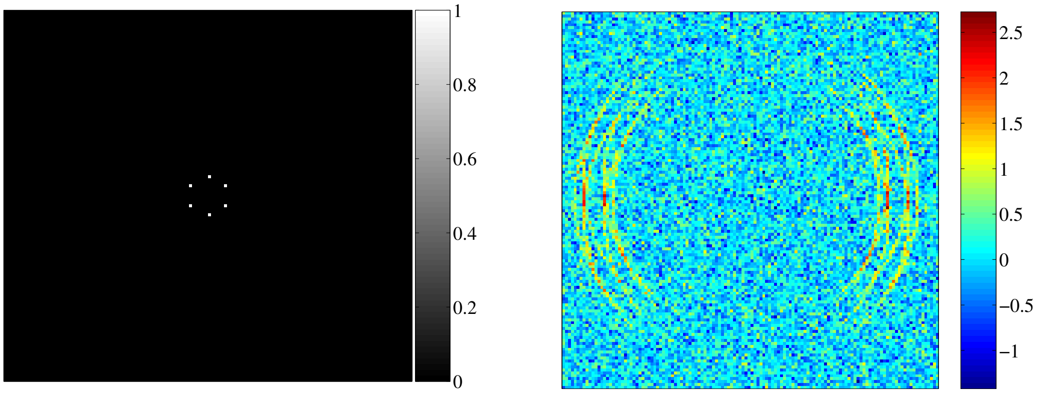
criterion can be regarded as the error probability on the support of $\underline{\theta}$. When the SNR is in the range of 2–8 dB, which is where the hyperparameter estimates are relatively unbiased, the MAP methods and EBD-LAZE have lower error curves than L1-SURE. The increase of MAP2's error curve when $g^* = (\sqrt{2})^{-1}$ closely parallels the increase of its $\hat{w}$. Interestingly enough, the bias of $\hat{a}$ of MAP1 does not result in poor performance in terms of E_d. Since MAP1's $\hat{w} \leq 1/2$, the hybrid thresholding function is used in (5.63). In the simulation, $\alpha < 1$, and because $\hat{w}$ is small over the range of SNRs considered, the expression for $t_1 = \hat{a}\alpha^2 + \sqrt{2\alpha^2 \log((1-\hat{w})/\hat{w})}$ is dominated by the second term.

5.7.4 MRFM reconstruction examples

Benzene example

A two dimensional reconstruction was carried out using the six hydrogen atoms of the benzene molecule as $\underline{\theta}$. Each hydrogen atom location was set to one, and the rest of the image set to zero. The two dimensional image was 128 × 128, and a 2-d slice of the MRFM psf was used for the linear transformation $\mathbf{H}$. The parameters of the psf used are the same as in Table 2.1.

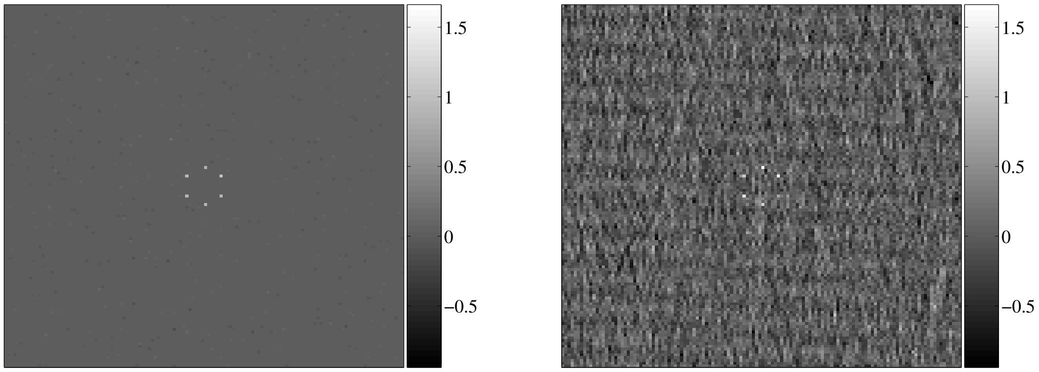
The image $\underline{\theta}$ and the noisy observation $\underline{y}$ are shown in Figs. 5.24a and 5.24b respectively. The SNR was -5 dB, which corresponded to a noise standard deviation of $\sigma = 0.372$. MAP2 with $g^* = (\sqrt{2})^{-1}$ and the Landweber iteration were applied, each with 2×10^4 iterations. The respectively reconstruction results are depicted in Figs. 5.25a and 5.25b respectively. The hexagonal pattern of the hydrogen atoms is clear in the MAP2 reconstruction. No spurious non-zero pixels are visible. In contrast, the Landweber reconstruction contains background noise, and one of the hydrogen locations appears to be "missing". The error criteria for these two methods



(a) Hydrogen atom locations of benzene

(b) Noisy observation

Figure 5.24: Hydrogen atom locations of benzene and noisy observation after convolution with a 2-d slice of the MRFM psf. The hydrogen atoms trace out a hexagon in the plane.



(a) MAP2, $g^* = (\sqrt{2})^{-1}$

(b) Landweber

Figure 5.25: MAP2 and Landweber reconstruction of benzene's hydrogen atoms under an SNR of -5 dB.

are given in Table 5.5 below. The value of $\delta = 10^{-2}\|\underline{\theta}\|_\infty = 10^{-2}$ was used for the $E_d(\underline{\theta}, \hat{\underline{\theta}})$ error criterion. The numbers confirm that MAP2 produces a sparser estimate.

Table 5.5: Error criteria for the benzene reconstruction with an SNR of -5 dB.

Method	$\|\underline{e}\|_0$	$\|\underline{e}\|_1$	$\|\underline{e}\|_2$	$E_d(\underline{\theta}, \hat{\underline{\theta}})$	$\|\hat{\underline{\theta}}\|_0$
MAP2, $g^* = (\sqrt{2})^{-1}$	458	24.5	1.25	447	458
Landweber	16384	3.27×10^3	32.1	1.58×10^4	16384

An interesting point is that the reconstructions of these two algorithms are relatively good under a SNR of -5 dB. Even the Landweber iteration, which does not encourage sparsity in its estimate, produces a reconstruction where the hexagonal pattern is mostly visible. Define the coherence of $\mathbf{H}$ as $\mu(\mathbf{H}) \triangleq \max_{i \neq j} | < \underline{h}_i, \underline{h}_j > |$. The coherence plays a strong role in the amenability of the sparse representation problem $\underline{y} = \mathbf{H}\underline{\theta}$ [22, 69, 13]: a smaller $\mu(\mathbf{H})$ is more desirable. As a result, we expect a smaller $\mu(\mathbf{H})$ to have better sparsity performance in the inverse problem $\underline{y} = \mathbf{H}\underline{\theta} + \underline{w}$.

The 2-d MRFM psf has a support that is almost one dimensional. It is not quite one dimensional because each of the two elliptical curves that make up the 2-d MRFM psf has a non-zero thickness. When the MRFM psf is moved in some direction, there will be very little overlap between the support of the new location and the old location. Consequently, $\mathbf{H}$ will have columns that have very low correlation with each other. The 2-d MRFM psf is a good psf for sparse image reconstruction. In contrast, a psf that has a large area or volume will result in the columns of the corresponding $\mathbf{H}$ being more strongly correlated with each other.

103D molecule (DNA) example

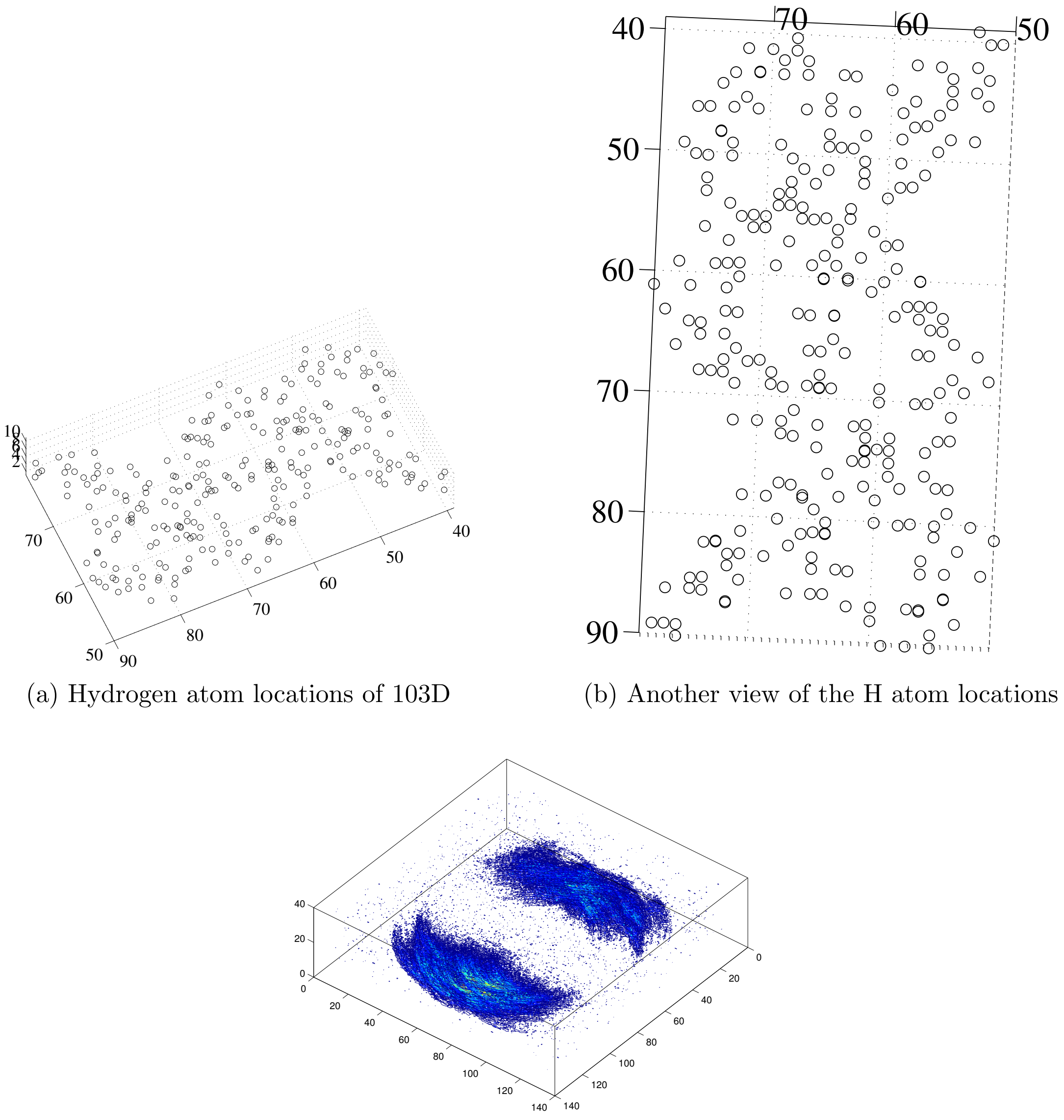
A three dimensional reconstruction was carried out using the 272 hydrogen atoms of the 103D molecule (DNA) as $\underline{\theta}$. As with the benzene example previously, each hydrogen location was set to one, and the rest of the image set to zero. The image size used was $128 \times 128 \times 32 \approx 5.24 \times 10^5$ voxels. The hydrogen atom structure was aligned so that its longer dimension resided in the x-y plane, so as to take advantage of the greater number of grid points in the x and y dimensions. The matrix $\mathbf{H}$ represented convolution with the 3-d MRFM psf. The parameters of the psf are given in Table 5.6 below.

Table 5.6: Psf parameters used for the 3-d MRFM reconstruction example.

Parameter		**Value**
Description	Name	
Amplitude of external magnetic field	B_{ext}	2.8835×10^4 G
Value of B_{mag} in the resonant slice	B_{res}	3×10^4 G
Radius of tip when modelled as a sphere	R_0	3 nm
Distance from tip to sample	d	3 nm
Cantilever tip moment†	m	1.9227×10^5 emu
Peak cantilever swing	x_{pk}	0.049 nm
Maximum magnetic field gradient‡	G_{max}	407 G/nm

† Assuming a spherical tip.
‡ Assuming optimal sample position.

The SNR used was 6.02 dB, which corresponded to a noise $\sigma = 0.380$. The non-zero portion of $\underline{\theta}$ is shown in Fig. 5.26a-b. The noisy observation $\underline{y}$ is shown in Fig. 5.26c. The first four slices of the noisy observation $\underline{y}$ are depicted in Fig. 5.27. MAP2 with $g^* = (\sqrt{2})^{-1}$ was used with 5×10^4 iterations. Several different volume rendering views are illustrated in Fig. 5.28a-d. In Figs. 5.28a-b, the helical structure of $\hat{\underline{\theta}}$ is apparent. There are spurious non-zero voxels present in the reconstruction. In particular, two extraneous Xs are traced out in Fig. 5.28c. The flatness of 103D



(a) Hydrogen atom locations of 103D

(b) Another view of the H atom locations

(c) Noisy observation

Figure 5.26: Hydrogen atom locations of 103D and noisy observation after convolution with the MRFM psf. Note the helical structure traced out by the hydrogen atoms.

can be seen in Fig. 5.28d, where the hydrogen atoms are concentrated in a v-shaped slice.

Landweber iterations were carried out on the same noisy observation using 5×10^4 iterations. The result is rendered at different viewing angles in Fig. 5.29a-d. The helical structure of the hydrogen atoms is also visible. As compared to the MAP2, however, there is more background noise and spurious non-zero voxels. The two

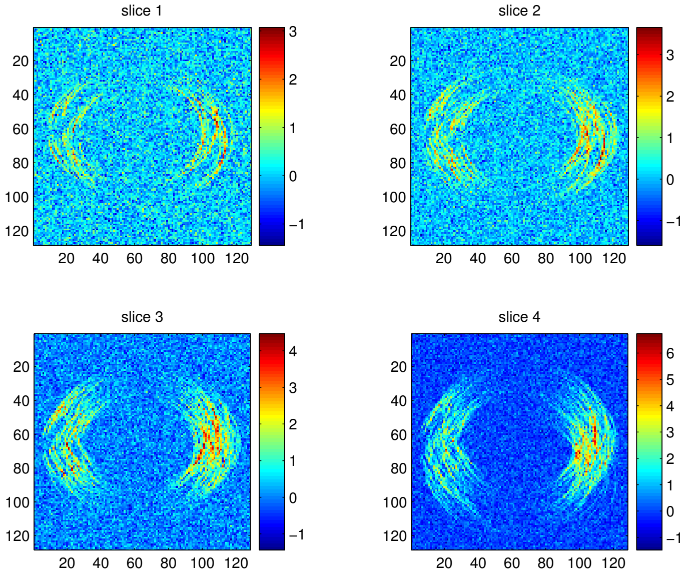
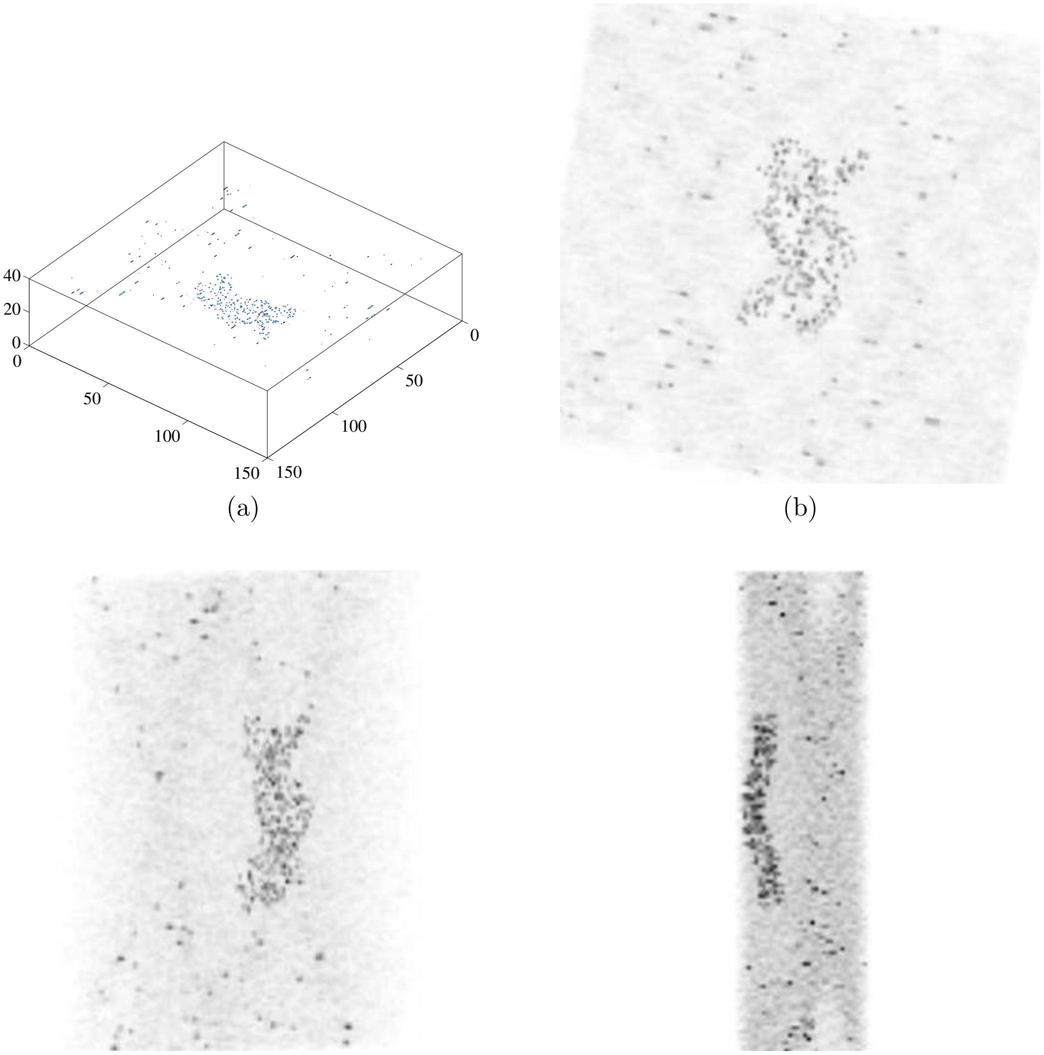


Figure 5.27: First four slices of noisy observation $\underline{y}$ resulting from convolving the MRFM 3-d psf with the H atom locations of 103D.

extraneous Xs are more apparent in Fig. 5.29c. Note that Figs. 5.28 and 5.29 use the same colour map, but not the same colour scale. The error criteria for the two methods are given in Table 5.7. The numbers confirm that the MAP2 reconstruction is sparser. This is also evident from the histogram of $\hat{\underline{\theta}}$ values given in Fig. 5.30. There is a multiplicative reduction in spurious $\hat{\theta}_i$ values that are less than one in magnitude. The reduction of the error criteria for MAP2 vs. Landweber is not as dramatic as for the benzene example. It is likely because of the higher SNR of 6.02 dB used here. In a lower SNR environment, we expect the MAP2 estimator to have a bigger reduction in the error criteria.

Table 5.7: Error criteria for the reconstruction of 103D's hydrogen atoms under an SNR of 6.02 dB.

Method	$\|\underline{e}\|_0$	$\|\underline{e}\|_1$	$\|\underline{e}\|_2$	$E_d(\underline{\theta}, \hat{\underline{\theta}})$	$\|\hat{\underline{\theta}}\|_0$
MAP2, $g^* = (\sqrt{2})^{-1}$	4.605×10^5	6.336×10^4	122.7	4.602×10^5	4.605×10^5
Landweber	5.243×10^5	1.056×10^5	185.0	5.240×10^5	5.243×10^5



(a)

(b)

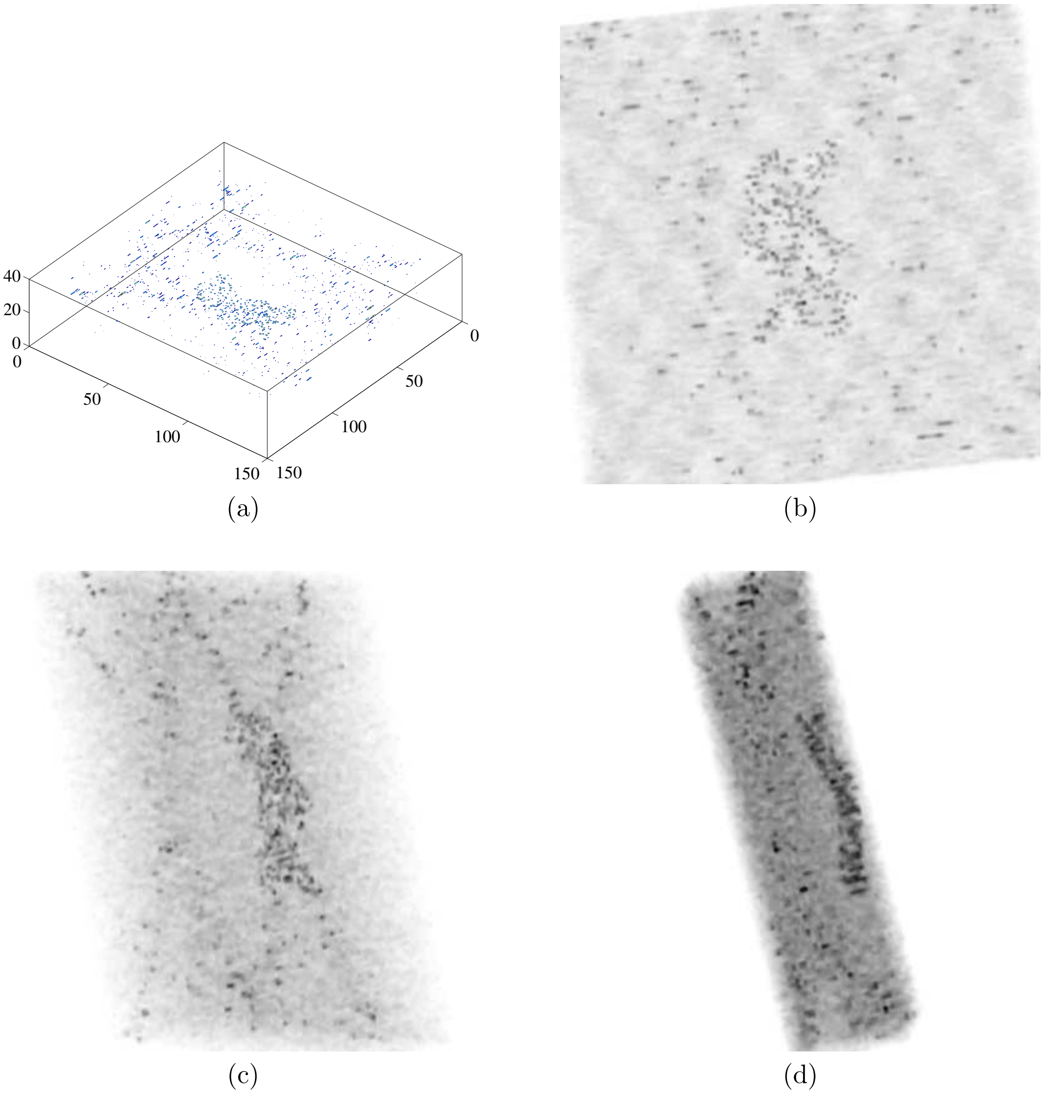
(c)

(d)

Figure 5.28: Three dimensional visualization of the MAP2 reconstruction of 103D's hydrogen atoms with $g^* = (\sqrt{2})^{-1}$ at an SNR of 6.02 dB. Different viewing angles are shown. The helical structure of 103D is apparent.

5.8 Conclusion

This chapter proposed methods of performing simultaneous deconvolution and denoising of sparse images. We wanted methods that estimated the tuning parameters in a data-driven fashion and which were scalable. Two approaches were taken. The first was to impose a sparsifying prior on the image $\underline{\theta}$ that contained unspecified parameters, e.g., the LAZE p.d.f. The unknown parameters of the prior,



(a)

(b)

(c)

(d)

Figure 5.29: Three dimensional visualization of the Landweber reconstruction of 103D's hydrogen atoms at an SNR of 6.02 dB. Different viewing angles are shown. The helical pattern is also visible, but there is more background noise and spurious non-zero voxels.

which were called the hyperparameter, were estimated in an empirical fashion, either through marginal ML or MAP. This first approach gave rise to EBD-LAZE, MAP1, and MAP2. MAP2 requires the specification of a tuning parameter. The second approach taken was to use a MPLE with a penalty that encouraged sparsity. The penalty function contained unspecified parameters which were estimated by minimizing SURE of the l_2 risk between $\mathbf{H}\underline{\theta}$ and $\mathbf{H}\hat{\underline{\theta}}$. The methods L1-SURE and

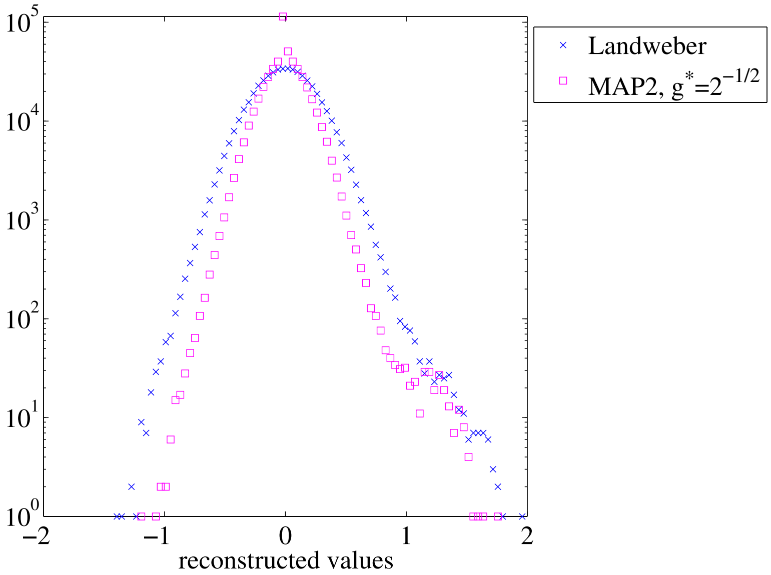


Figure 5.30: Histogram of $\underline{\hat{\theta}}$ for the Landweber iteration and MAP2 when applied to reconstruction of 103D's hydrogen atoms.

HHS-SURE were the result of this approach.

In the simulation study performed, SBL had poor performance with the binary-valued image, but better performance under high SNR with an image that contained positive and negative values (the LAZE image). A possible remedy is to threshold the SBL estimator; however, this requires prior knowledge of $\underline{\theta}$. In addition, SBL is computationally intensive and requires the inversion of a matrix that is of size $P \times P$, where $P = \min(M, N)$. Recall that N is the length of the observation $\underline{y}$ and M is the length of $\underline{\theta}$. The inversion of large matrices is numerically unstable, which affects the scalability of SBL.

Overall, the L1-SURE estimator is a consistent performer under the error criteria examined. The MAP2 estimator has good performance under low SNR, but not high. Under high SNR, L1-SURE and MAP1 have comparable performance in terms of the l_1 and l_2 norm of the reconstruction error. EBD-LAZE has performance that is worse than L1-SURE's for the previous two criteria, although its error detection performance is good over the SNR range of 2 dB to 15 dB. Under higher SNR, the hyperparameter estimate of EBD-LAZE, MAP1 and MAP2 become increasingly biased, which explains why their performance worsens under higher SNR. HHS-SURE achieves a comparable $\|\underline{e}\|_1$ and $\|\underline{e}\|_2$ as L1-SURE, but with a sparser estimate. While the Landweber iteration, which solves for the least-squares solution, produces noisy reconstructions, the projected Landweber has good performance for the non-negative binary-valued image under high SNR.

The MAP1 and MAP2 estimators scale well with the size of the problem. They have low computational complexity: no matrix inversion is required, and the hyperparameter optimizations are computed in closed form. EBD-LAZE, however, requires a 2-d search in order to optimize its hyperparameters. Fortunately, the search remains in two dimensions regardless of the values of M and N. To decrease computational complexity, the hyperparameter search can be performed every nth iteration instead of every iteration. L1-SURE can be efficiently implemented for $\underline{\theta}$ that are highly sparse using the LARS algorithm. A matrix inversion is needed in each step of LARS. The number of LARS steps is approximately proportional to the number of non-zero voxels in $\underline{\theta}$. If this number is fixed, M, the length of $\underline{\theta}$, can vary arbitrarily without affecting the computational complexity of LARS. HHS-SURE's scalability is dependent on the scalability of L1-SURE and of subsequent optimizations for $\underline{t} = (t_1, t_2)$. The subsequent optimizations can be efficiently implemented via parallel processing. As a second option, both L1-SURE and HHS-SURE can be implemented in their entirety as a parallel operation.

CHAPTER VI

Future directions

In this chapter, we will discuss the open issues in Chapters III, IV, and V. A question that concerns all of the work done in this thesis is the noise model that is used. We have assumed throughout the thesis that the noise is AWGN; however, it might not always be appropriate. In the future, we might want to consider coloured Gaussian noise, for example, or a Poisson noise model.

6.1 Detection of the CTC model

While the PLKF was formulated for the detection of the soft nonlinear system given by the CTC model, no convergence or error bound properties were shown for the estimator. As such, while the simulations for a certain parameter set demonstrated that the PLKF outperformed the EKF, it might not be true for other parameter values. The PLKF has not been applied to other softly nonlinear models, and so one does not how it compares against the EKF in general.

The same can be said of the KF/GLR innovations and innovations energy detector. Its formulation was based on heuristic principles, and while the simulations showed that the KF/GLR innovations detector worked well for two parameter sets, one cannot guarantee this in general for all parameter values. It would be desirable to obtain bounds on the false alarm (P_F) and detection (P_D) probabilities.

While the KF/GLR innovations detector had good performance in the simulations, one needed to know the parameter values of the model, e.g., G, k, etc. Some of these values are not precisely known in the experiment, which leads to two questions: how sensitive is the detector to changes in the parameter values, and how can the detector be made more robust to uncertainties in the parameter values? These are both questions that have not been looked into.

Would it be possible to design the rf waveform based on a detectability criterion? This is a question that extends to the detection of the DT single spin-cantilever models as well.

6.2 Detection of the discrete-time models

Applying the FE statistic requires knowledge of α used in the LPF. In practice, a bank of LPFs with different αs are used to perform detection, which amounts to the application of the GLR principle. Nonetheless, one wonders if there is a better way to estimate α. In the general case of the DTRT when the probabilities are not symmetric, i.e., $p \neq q$, the hybrid detector approximation to the LRT given in (4.21) requires knowledge of p, q, A, σ. If these are not available, how are these to be estimated? A sensitivity analysis of these parameters should also be carried out.

It would be interesting to extend the approximation of the LRT for a certain class of DTRWs to a broader category of DT finite-state processes. This will depend on the tractability of the eigendecomposition of the probability transition matrix $\mathbf{P}$ and the matrix $\mathbf{Q}$, among other things.

6.3 Sparse image reconstruction

Even though the MAP1 and MAP2 reconstruction methods monotonically increase their respective objective function, we have not shown that the iterations will converge. In addition, the uniqueness of the maximizer has not been investigated. It would be interesting to see if L1-SURE has optimality properties along the lines of SureShrink.

The proposed reconstruction methods addressed the issue of sparsity, but did not address the issue of non-negativity of $\underline{\theta}$. We did, however, investigate the performance of a method that enforced non-negativity in the estimate: the projected Landweber iteration. In the simulation study, it was noted that positivity seemed to produce a sparsifying effect when the true $\underline{\theta}$ is non-negative. This bears further study. The iterative thresholding framework suggests a possible solution to enforcing non-negativity: apply a thresholding function $T(x)$ such that $T(x) = 0$ for $x < 0$. Indeed, the projected Landweber has this characteristic: it fits the thresholding framework with the following thresholding function $T^{+}(x) = \max(0, x)$.

What if one knew that the non-zero θ_is assumed values in a finite set? This is a discrete constraint on the non-zero θ_is that could be exploited.

The priors of $\underline{\theta}$ that were considered have all been independent, i.e., θ_i and θ_j are independent for $i \neq j$. However, as molecules generally have structure, the independence assumption is not accurate. Modelling the structural dependence might produce better reconstructions. This thesis assumed that there was no spin coupling present; that might not be true when the distance between spins is small. Can the prior or penalty used by a reconstruction method be adapted to model spin coupling effects?

Design of the MRFM point spread function for sparse image reconstruction is another fertile area to be explored. As was noted, a matrix $\mathbf{H}$ that has columns with low correlation is desirable. By optimizing the parameters of the experiment so as to produce this effect, a performance gain can be realized.

APPENDICES

APPENDIX A

Second order approximation of the LRT for the DTRT model

Let $f(y_0, \ldots, y_{N-1})$ denote the log LRT function of the DT random telegraph; this is obtained by taking the log of the left-hand side of (4.18). Let $g(y_0, \ldots, y_{N-1})$ be the filtered energy detector function in (4.17). We want to analyze the two functions f and g under the regime of low SNR ($|A/\sigma| \ll 1$) and long observation time ($N \gg 1$).

The strategy used is to obtain the approximate Taylor series expansion of f about $\underline{y} = \underline{0}$ and compare that with g. Define:

$$\theta_i \triangleq \frac{q_i(A)e^{\frac{A}{\sigma^2}y_i}}{q_i(A)e^{\frac{A}{\sigma^2}y_i} + q_i(-A)e^{-\frac{A}{\sigma^2}y_i}}$$

for $i \geq 0$. From (4.19), a recursive equation for θ_i can be derived. Its approximate solution is

$$\begin{aligned}
\theta_i &\approx \beta_i + \frac{qA}{\sigma^2}\sum_{j=0}^{i} \xi_{ij} y_j, \quad i \geq 0 \quad \text{where:} \\
\beta_i &= \frac{1-q}{1-r} + \left(\frac{1}{2} - \frac{1-q}{1-r}\right) r^i, \quad i \geq 0 \\
\xi_{ij} &= \frac{2(1-q)r^{i-j} + (2q-r-1)r^i}{1-r}, \quad 0 \leq j \leq i-1 \\
\xi_{ii} &= \frac{2(1-q)}{1-r} + \frac{r^i(2q-r-1)}{1-r} = 2\beta_i, \quad i \geq 0
\end{aligned} \tag{A.1}$$

and $r = p + q - 1$. Note that $p, q \in (0,1) \Rightarrow |r| < 1$. Define $s_i \triangleq \frac{A}{\sigma^2} y_i$. Then,

$$f \approx \sum_i \left\{ \left[s_i(2q_i(A) - 1) + \frac{1}{2}s_i^2\right] - \frac{1}{2}\left[s_i(2q_i(A) - 1) + \frac{1}{2}s_i^2\right]^2 \right\} \tag{A.2}$$

By solving for $q_i(A)$ in terms of θ_i and using (A.1), one obtains the approximate Taylor series expansion of f as

$$f \approx L_1 + L_{2a} + L_{2b} + \text{h.o.t.}, \quad \text{with} \tag{A.3}$$

$$L_1 = \frac{A}{\sigma^2} C_m \sum_i (1 - r^i) y_i \tag{A.4}$$

$$L_{2a} = 2q \left(\frac{A}{\sigma^2}\right)^2 \sum_i \sum_{j=0}^{i-1} \left[\frac{2(1-q)}{1-r} r^{i-j} - r^i C_m\right] y_i y_j \tag{A.5}$$

(A.6)

$$L_{2b} = \left(\frac{A}{\sigma^2}\right)^2 \sum_i \left\{4r \left(\frac{1-q}{1-r}\right)^2 + 2\frac{(q-r)(1-q)}{(1-r)^2} - C_m(2q + C_m) r^i + \frac{1}{2} C_m^2 r^{2i}\right\} y_i^2$$

and $C_m \triangleq \frac{p-q}{2-p-q}$. In (A.3), "h.o.t." denotes the higher-order terms; specifically, terms of degree three or higher. C_m is a parameter that indicates the mismatch between the transition probabilities p and q. In the symmetric case, $p = q \Rightarrow C_m = 0$, and one obtains a simpler expression for f. Let f_{sym} be the function f under symmetric transition probabilities, i.e., $p = q$. Then,

$$f_{\text{sym}} \approx 2p \left(\frac{A}{\sigma^2}\right)^2 \left\{\sum_{i=1}^{N-1} \sum_{j=0}^{i-1} (2p-1)^{i-j} y_i y_j + \sum_{i=0}^{N-1} \left(1 - \frac{1}{4p}\right) y_i^2\right\} \tag{A.7}$$

For sufficiently large N, it can be shown that

$$g \approx D \left\{\sum_{i=1}^{N-1} \sum_{j=0}^{i-1} \alpha^{i-j} y_i y_j + \frac{\alpha}{1+\alpha} \sum_{i=0}^{N-1} y_i^2\right\} \tag{A.8}$$

where $D = \frac{1-\alpha^2}{2\alpha}$ is a constant; note that D plays no role in the performance of the test statistic. Let $\tilde{f}_{\text{sym}} \triangleq (2p)^{-1}(A/\sigma^2)^{-2} f_{\text{sym}}$ and $\tilde{g} \triangleq D^{-1} g$. Comparing (A.7) and (A.8), we see that they are nearly identical in form if $\alpha = 2p - 1$. If $\alpha = 2p - 1 \Rightarrow |\tilde{f}_{\text{sym}} - \tilde{g}| \approx \frac{1}{4p} \sum_i y_i^2$. Now, $E_1[\sum_{i=0}^{N-1} y_i^2] - E_0[\sum_{i=0}^{N-1} y_i^2] = A^2 N$. On the other hand, for large N,

$$E_1\left[\sum_{i=1}^{N-1} \sum_{j=0}^{i-1} \alpha^{i-j} y_i y_j\right] - E_0\left[\sum_{i=1}^{N-1} \sum_{j=0}^{i-1} \alpha^{i-j} y_i y_j\right] \approx G A^2 N \tag{A.9}$$

where $G = \frac{\alpha(2p-1)}{1-\alpha(2p-1)}$. When $\alpha = 2p-1$, $G = \frac{(2p-1)^2}{1-(2p-1)^2} = \frac{1}{4(1-p)} + \frac{1}{4p} - 1$. For p close to 1, $G \gg \frac{1}{4p}$, and $GA^2N \gg \frac{1}{4p}A^2N$. So to the first moment, the difference of $\frac{1}{4p}\sum_i y_i^2$ between $\tilde{f}_{\text{sym}}$ and $\tilde{g}$ does not represent a significant difference when $p \approx 1$. Under these conditions, we expect that the performance of the filtered energy detector and the DT random telegraph LRT to be similar.

It is possible to obtain an approximation to the DT random telegraph LRT that holds when we make no assumption about p being equal to q. When $p \neq q$, $C_m \neq 0$, and there are terms of the form $r^i C_m$ and $r^{2i} C_m^2$ in (A.4)-(A.6). Since $|r| < 1$, $r^i \to 0$ in the limit as $i \to \infty$. So drop these terms to get:

$$f \approx C\left\{\frac{(p-q)\sigma^2}{4q(1-r)A}\sum_i y_i + \sum_i \sum_{j<i} r^{i-j} y_i y_j + \left[\frac{1}{2} + \frac{r(1-q)}{2q(1-r)}\right]\sum_i y_i^2\right\} \tag{A.10}$$

where $C = 4q\frac{1-q}{1-r}\left(\frac{A}{\sigma^2}\right)^2$ is a constant. Define $C_a \triangleq \frac{(p-q)\sigma^2}{4q(1-r)A}$ and $C_e \triangleq \frac{r(1-q)}{2q(1-r)}$. In order to equate the coefficients of the cross-terms $y_i y_j$ between (A.10) and g in (A.8), we require $\alpha = r = p + q - 1$. In g, the ratio of the energy terms to the cross-terms is $\frac{\alpha}{1+\alpha}$. For $r = \alpha \approx 1 \Rightarrow \frac{\alpha}{1+\alpha} \approx 1/2$. The idea is to add the energy and amplitude statistics to g so that all three statistics are in the same ratio as in (A.10). Let g_{hyb} be the "extended" version of g, which we shall call the hybrid filtered energy/amplitude/energy detector:

$$\begin{aligned} g_{\text{hyb}} &\triangleq g + \frac{1-\alpha^2}{2\alpha}\left[C_a \sum_i y_i + C_e \sum_i y_i^2\right] \\ &= g + \frac{1-\alpha^2}{2\alpha} C_a \sum_i y_i + \frac{1-\alpha^2}{2\alpha} C_e \sum_i y_i^2 \end{aligned} \tag{A.11}$$

We expect g_{hyb} to have performance that is similar to f under the conditions of large N, low SNR, and $r \approx 1$.

The constants in (A.11) can be further simplified. Let $K_a = C_a(1-\alpha^2)/2\alpha$ and $K_e = C_e(1-\alpha^2)/2\alpha$. As it is required that $\alpha = p + q - 1$, after some algebra, one obtains

$$K_a = \frac{p^2 - q^2}{8q(p+q-1)} \left(\frac{A}{\sigma^2}\right)^{-1} \quad \text{and} \quad K_e = \frac{(p+q)(1-q)}{4q} \tag{A.12}$$

It is interesting that the constants K_a and K_e are not symmetric in p, q.

APPENDIX B

Matrix results

Proposition B.1. *A real tridiagonal matrix* $\mathbf{A} = (a_{ij})$ *of order* n *has only real simple eigenvalues if* $a_{ij}a_{ji} > 0$ *for* $j = i + 1$*.*

See [52].

Proposition B.2. *Suppose* $\mathbf{A}$ *is an* $n \times n$ *matrix and that it has distinct eigenvalues* $\lambda_1, \ldots, \lambda_k$*. Then, the corresponding eigenvectors* $x_1, \ldots, x_k$ *form a linearly independent set.*

This is a well-known result of matrix theory. An immediate corollary is that if $\mathbf{A}$ has n distinct eigenvalues, then $\mathbf{A}$ has a basis of eigenvectors for $\mathbb{C}^n$ (or $\mathbb{R}^n$). We can apply this corollary to real tridiagonal matrices with $a_{ij}a_{ji} > 0$ for $j = i + 1$. By Prop. B.1, such matrices have real simple eigenvalues. Therefore, all of the eigenvalues of such matrices are distinct, and by the corollary, there exists a basis of eigenvectors. Note that the eigenvalues are real, and so the eigenvectors can be chosen to be real-valued as well. The eigenvectors form a basis for $\mathbb{R}^n$ over the reals.

Proposition B.3. *Suppose* $\mathbf{A}$ *is a real tridiagonal matrix. If* $a_{ii} = 0$ *for* $i = 1, \ldots, n$*, then whenever* $\lambda \in \mathbb{R}$ *is an eigenvalue of* $\mathbf{A} \Longrightarrow -\lambda$ *is also an eigenvalue of* $\mathbf{A}$*.*

Proof: Note that by Prop. B.1, $\mathbf{A}$ has only real eigenvalues. Let $p_n(\lambda)$ be the characteristic polynomial of $\mathbf{A}$ when it is of order n. We claim that:

$$p_n(\lambda) = \begin{cases} \text{poly in } \lambda^2 & \text{n even} \\ \lambda(\text{poly in } \lambda^2) & \text{n odd} \end{cases}$$

From this, whenever λ satisfies $p_n(\lambda) = 0$, then $p_n(-\lambda) = 0$ also, i.e. whenever λ is an eigenvalue, so is $(-\lambda)$. We shall prove the claim by applying strong induction on the order of $\mathbf{A}$.

Base cases $n = 1, 2$: For $n = 1$, the only valid $\mathbf{A} = (0)$, and $p_1(\lambda) = -\lambda$. For $n = 2$, $p_2(\lambda) = \lambda^2 - a_{12}a_{21}$. So the claim holds for $n = 1, 2$.

Inductive step: Assume the claim is true for $n = 1, \ldots, k$, where $k \geq 2$. Consider then $n = k + 1$. Now, $p(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}_{k+1})$. Expand the determinant by the last column:

$$p(\lambda) = (-1)^{k+(k+1)} a_{k,k+1} \det \begin{pmatrix} -\lambda & a_{12} & & & & \\ a_{21} & -\lambda & a_{23} & & & \\ & \ddots & \ddots & \ddots & & \\ & & a_{k-1,k-2} & -\lambda & a_{k-1,k} \\ & & & & a_{k+1,k} \end{pmatrix}$$

$$+ (-1)^{k+k}(-\lambda) \det \begin{pmatrix} -\lambda & a_{12} & & & \\ a_{21} & -\lambda & a_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & a_{k-1,k-2} & -\lambda & a_{k-1,k} \\ & & & a_{k,k-1} & -\lambda \end{pmatrix}$$

The second determinant is the characteristic polynomial of a real $k \times k$ matrix with zeros down the diagonal. As such, it falls under the inductive hypothesis. Let us rename it as $q_2(\lambda)$. Expand the first determinant by the last row:

$$p(\lambda) = a_{k,k+1}a_{k+1,k} \det \begin{pmatrix} -\lambda & a_{12} & & & \\ a_{21} & -\lambda & a_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & a_{k-2,k-3} & -\lambda & a_{k-2,k-1} \\ & & & a_{k-1,k-2} & -\lambda \end{pmatrix} - \lambda\, q_2(\lambda)$$

The determinant in the expression above is the characteristic polynomial of a real $(k-1) \times (k-1)$ matrix with zeros down the diagonal. It falls under the inductive hypothesis as well—let us rename it as $q_1(\lambda)$. The net result then is that $p(\lambda) = a_{k,k+1}a_{k+1,k}\, q_1(\lambda) - \lambda\, q_2(\lambda)$. Consider two cases:

Case 1: $(k+1)$ is even. Then, k is odd and $(k-1)$ is even:

$$\begin{aligned} \therefore p(\lambda) &= a_{k,k+1}a_{k+1,k}(\text{poly in } \lambda^2) - \lambda \cdot \lambda \cdot (\text{poly in } \lambda^2) \\ &= \text{poly in } \lambda^2 \end{aligned}$$

Case 2: $(k+1)$ is odd. Then, k is even and $(k-1)$ is odd:

$$\begin{aligned} \therefore p(\lambda) &= a_{k,k+1}a_{k+1,k}\, \lambda(\text{poly in } \lambda^2) - \lambda(\text{poly in } \lambda^2) \\ &= \lambda(\text{poly in } \lambda^2) \end{aligned}$$

Either way, $p(\lambda)$ satisfies the claim, i.e. the claim is true for $n = k+1$. By strong induction, the claim is true for all $n \geq 1$. ■

Proposition B.4. *Let* **A** *be a real matrix, and suppose that all of its rows sum to a constant value* c*. Then,* **A** *has eigenvalue* c*.*

Let $x = (1, \ldots, 1)^T$. Then, one can verify that $\mathbf{A}x = cx$. In particular, if $\mathbf{P}$ is a probability transition matrix, each of its rows sums up to 1. So $\mathbf{P}$ has 1 as an eigenvalue.

Proposition B.5. *Consider a random walk in $\mathcal{R}_0^r$ that has r states $\psi_1, \ldots, \psi_r$. Let* $\mathbf{P}$ *be its probability transition matrix. Then:*

1. *the eigenvalues of* $\mathbf{P}$ *can be ordered as $\lambda_1 > \lambda_2 > \ldots > \lambda_{r-1} > \lambda_r$, with $\lambda_k \in \mathbb{R}$ for $k = 1, \ldots, r$*
2. *$\lambda_k + \lambda_{r+1-k} = 0$ for $k = 1, \ldots, r$.*
3. *$\lambda_1 = 1$*

Proof: $\mathbf{P}$ for a random walk is tridiagonal. Since the random walk is in $\mathcal{R}_0^r$, $p_{ij}p_{ji} > 0$ for $j = i + 1$. So Prop. B.1 applies, and $\mathbf{P}$ has real simple eigenvalues; this accounts for the first statement. Because the random walk has no self-loops, $p_{ii} = 0$ for $1 \leq i \leq r$, and Prop. B.3 can be used. This justifies the second statement. Lastly, we consider the third statement. By Prop. B.4, 1 is an eigenvalue of $\mathbf{P}$. We would like to show that it is in fact the largest positive eigenvalue. Let λ be an eigenvalue of $\mathbf{P}$. By Gerschgorin's theorem [40], there exists some $1 \leq j \leq r$ s.t.

$$|\lambda - p_{jj}| \leq \sum_{k \neq j} |p_{jk}|$$

Since all of the elements of $\mathbf{P}$ are real and non-negative, $\sum_{k \neq j} |p_{jk}| = \sum_{k \neq j} p_{jk} = 1$. Moreover, $p_{jj} = 0$ for all j. Hence, $|\lambda| \leq 1$. ■

Note: There is another way of showing that $|\lambda| \leq 1$. The *spectral radius* of a matrix $\mathbf{A}$ is defined as $\rho(\mathbf{A}) \triangleq \max_i |\lambda_i|, \lambda_i = \lambda_i(\mathbf{A})$. We would like to show that $\rho(\mathbf{P}) \leq 1$. However, one knows that norms dominate the spectral radius, i.e., $\rho(\mathbf{P}) \leq \|\mathbf{P}\|$. In particular, consider the l^∞-norm which induces the norm $\|\mathbf{A}\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$, i.e., it is equal to the max absolute value row sum. In particular, $\|\mathbf{P}\|_\infty = 1 \Longrightarrow \rho(\mathbf{P}) \leq 1$.

Proposition B.6. *Let $\mathbf{D}_\delta = \mathrm{diag}(\delta_1, \ldots, \delta_r)$ be a diagonal matrix with $\delta_j < 1$ for all $1 \leq j \leq r$. Let $\mathbf{P}$ be the probability transition matrix of the random walk that we consider. Then, $\widetilde{\mathbf{P}} = (\mathbf{I} - \mathbf{D}_\delta)\mathbf{P}$ satisfies the first two statements of Prop. B.5*

Proof: Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$. Then,

$$\mathbf{A}\mathrm{diag}(\gamma_1, \ldots, \gamma_n) = (\gamma_1 a_{*1}, \ldots, \gamma_n a_{*n}) \text{ and}$$

$$\mathrm{diag}(\gamma_1, \ldots, \gamma_n)\mathbf{A} = \begin{pmatrix} \gamma_1 a_{1*} \\ \vdots \\ \gamma_n a_{n*} \end{pmatrix},$$

where $\mathbf{A} = (a_{ij})$, and a_{*k} refers to the k-th column of $\mathbf{A}$, whereas a_{k*} refers to the k-th row of $\mathbf{A}$. Let $\mathbf{P} = (p_{ij})$ and $\widetilde{\mathbf{P}} = (\tilde{p}_{ij})$. Note that

$$(\mathbf{I} - \mathbf{D}_\delta)\mathbf{P} = \begin{pmatrix} a_1 p_{1*} \\ \vdots \\ a_r p_{r*} \end{pmatrix},$$

and so $\tilde{p}_{ij}\tilde{p}_{ji} = (1-\delta_i)(1-\delta_j)p_{ij}p_{ji}$. If $\delta_j < 1$ for all j, then $\tilde{p}_{ij}\tilde{p}_{ji} > 0 \iff p_{ij}p_{ji} > 0$. So whenever Prop. B.1 applies to $\mathbf{P}$, it also applies to $\widetilde{\mathbf{P}}$. The second statement follows because $\tilde{p}_{ii} = 0$ for all $1 \leq i \leq r$. ■

Proposition B.7. *Let $\mathbf{A}$ be an $n \times n$ matrix that is diagonalizable, i.e. there exists $\mathbf{U}$ invertible, $\mathbf{\Lambda}$ diagonal s.t. $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$. If either (i): $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ or (ii): $\lambda_1 < \lambda_2 \leq \ldots \leq \lambda_n \leq 0$, then $\mathbf{A}^m \approx \lambda_1^m \mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1}$ for large m, where $\mathbf{M}_{ij}$ is an $n \times n$ matrix with zeros everywhere except for a 1 in the (i,j)-th position.*

Proof: Now,

$$\begin{aligned}\mathbf{A}^m &= \mathbf{U}\mathbf{\Lambda}^m\mathbf{U}^{-1} = \mathbf{U}\,\mathrm{diag}(\lambda_1^m, \ldots, \lambda_n^m)\,\mathbf{U}^{-1} = \lambda_1^m\mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1} + \ldots + \lambda_n^m\mathbf{U}\mathbf{M}_{nn}\mathbf{U}^{-1} \\ &= \lambda_1^m\left[\mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1} + \ldots + \left(\frac{\lambda_n}{\lambda_1}\right)^m \mathbf{U}\mathbf{M}_{nn}\mathbf{U}^{-1}\right]\end{aligned}$$

If either (i) or (ii) applies, then for $j \neq 1$, $0 \leq |\lambda_j/\lambda_1| < 1$. So $\lim_{m\to\infty} |\lambda_j/\lambda_1|^m = 0$ for $j \neq 1$, and

$$\lim_{m\to\infty} \left[\mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1} + \ldots + \left(\frac{\lambda_n}{\lambda_1}\right)^m \mathbf{U}\mathbf{M}_{nn}\mathbf{U}^{-1} \right] = \mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1}$$

We have $\lim_{m\to\infty} \mathbf{A}^m = \lim\,(\lambda_1^m \mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1})$. So for large m, $\mathbf{A}^m \approx \lambda_1^m \mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1}$. ■

Proposition B.8. *Let $\mathbf{A} = (a_{ij})$ be a tridiagonal $n \times n$ matrix with the following properties: (i) $a_{ij}a_{ji} > 0$ for $j = i+1$ and (ii) $a_{ii} = 0$ for $1 \leq i \leq n$. Then, for large m, $\mathbf{A}^m \approx \lambda_1^m \left[\mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1} + (-1)^m \mathbf{U}\mathbf{M}_{nn}\mathbf{U}^{-1}\right]$, where $\mathbf{A}$ can be eigendecomposed as $\mathbf{A} = \mathbf{U}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)\,\mathbf{U}^{-1}$ with $\lambda_1 > \ldots > \lambda_r$.*

Proof: Since $\mathbf{A}$ satisfies properties (i) and (ii), Props. B.1, B.2, and B.3 apply. The eigenvalues of $\mathbf{A}$ can be written as $\lambda_1, \ldots, \lambda_n$ where $\lambda_1 > \lambda_2 > \ldots > \lambda_n$ and $\lambda_j + \lambda_{n+1-j} = 0$ for $j = 1, \ldots, n$. Moreover, $\mathbf{A}$ can be diagonalized as $\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Suppose $n = 2p+1$, $p > 0$, i.e., n is odd. Then, $\lambda_{p+1} = 0$. Define $\mathbf{A}_1 \triangleq \mathbf{U}\,\mathrm{diag}(\lambda_1, \ldots, \lambda_p, 0, \ldots, 0)\,\mathbf{U}^{-1}$ and $\mathbf{A}_2 \triangleq \mathbf{U}\,\mathrm{diag}(0, \ldots, 0, \lambda_{p+2}, \ldots, \lambda_{2p+1})\,\mathbf{U}^{-1}$. So $\mathbf{A}_1$ contains the positive eigenvalues of $\mathbf{A}$ and $\mathbf{A}_2$ contains the negative eigenvalues of $\mathbf{A}$. Now,

$$\begin{aligned}\mathbf{A}^m &= \mathbf{U}\,\mathrm{diag}(\lambda_1^m, \ldots, \lambda_p^m, 0, \ldots, 0)\,\mathbf{U}^{-1} + \mathbf{U}\,\mathrm{diag}(0, \ldots, 0, \lambda_{p+2}^m, \ldots, \lambda_{2p+1}^m)\,\mathbf{U}^{-1} \\ &= \mathbf{A}_1^m + \mathbf{A}_2^m\end{aligned}$$

The eigenvalues of $\mathbf{A}_1$ are $\lambda_1, \ldots, \lambda_p$ and $(n-p)$ zeros s.t. $\lambda_1 > \lambda_2 > \ldots > \lambda_p > 0$. Those of $\mathbf{A}_2$ are $\lambda_{p+2}, \ldots, \lambda_{2p+1}$ and $(n-p)$ zeros s.t. $0 > \lambda_{p+2} > \ldots > \lambda_{2p+1}$. Therefore, apply Prop. B.7 to $\mathbf{A}_1$ and $\mathbf{A}_2$. For large m,

$$\begin{aligned}\mathbf{A}^m &\approx \lambda_1^m \mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1} + \lambda_{2p+1}^m \mathbf{U}\mathbf{M}_{2p+1,2p+1}\mathbf{U}^{-1} \\ &\approx \lambda_1^m \left[\mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1} + (-1)^m \mathbf{U}\mathbf{M}_{2p+1,2p+1}\mathbf{U}^{-1}\right]\end{aligned}$$

since $\lambda_{2p+1} = -\lambda_1$. The case when n is even can be similarly treated. ■

Proposition B.9. *Let* $\mathbf{P}$ *be a square* $r \times r$ *matrix and* $N \geq 3$. *Then,*

$$\begin{aligned}
\sum_{1 \leq j < k \leq N-1} (-1)^k \mathbf{P}^{k-j} &= \frac{(-1)^{N-1}}{2} \sum_{n=1}^{N-2} \mathbf{P}^n - \frac{1}{2} \sum_{n=1}^{N-2} (-\mathbf{P})^n \\
\sum_{1 \leq j < k \leq N-1} (-1)^j \mathbf{P}^{k-j} &= \frac{(-1)^{N-1}}{2} \sum_{n=1}^{N-2} (-\mathbf{P})^n - \frac{1}{2} \sum_{n=1}^{N-2} \mathbf{P}^n \\
\sum_{1 \leq j < k \leq N-1} \mathbf{P}^{k-j} &= \sum_{n=1}^{N-2} (N-1-n) \mathbf{P}^n
\end{aligned}$$

These identities can be verified by rearranging the sum on the LHS.

Proposition B.10. *Let* $\mathbf{P}$ *be the probability transition matrix associated with a random walk in* $\mathcal{R}_0^r$. *Define the functions* $w_{1,m}(\mathbf{P}) = \sum_{n=1}^m \mathbf{P}^n$ *and* $w_{2,m}(\mathbf{P}) = \sum_{n=1}^m n\mathbf{P}^n$. *Let* $\delta \in \mathbb{R}$.

$$\begin{aligned}
w_{1,m}(\delta\mathbf{P}) &= \sum_{i=1}^r \mathbf{P}_i^\star \left(\sum_{n=1}^m (\delta\lambda_i)^n \right) \\
w_{2,m}(\delta\mathbf{P}) &= \sum_{i=1}^r \mathbf{P}_i^\star \left(\sum_{n=1}^m n(\delta\lambda_i)^n \right)
\end{aligned}$$

where $\mathbf{P}$ *can be eigendecomposed as* $\mathbf{P} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$ *and* $\mathbf{P}_j^\star \triangleq \mathbf{U}\mathbf{M}_{jj}\mathbf{U}^{-1}$.

The result follows from applying the definition.

Proposition B.11. *This is a continuation of the previous proposition. Here, we restrict our attention to* $|\delta| \leq 1$. *Using the big-O notation,* $w_{1,m}(\pm\mathbf{P}) \sim \mathcal{O}(m)$ *and* $w_{2,m}(\pm\mathbf{P}) \sim \mathcal{O}(m^2)$. *When* $|\delta| < 1$, $w_{1,m}(\delta\mathbf{P})$ *and* $w_{2,m}(\delta\mathbf{P})$ *are both* $\mathcal{O}(1)$.

Proof: The following identities are useful:

$$\sum_{n=1}^m \delta^n = \begin{cases} \frac{\delta(1-\delta^m)}{1-\delta} & \delta \neq 1 \\ m & \delta = 1 \end{cases} \tag{B.1}$$

and

$$\sum_{n=1}^m n\delta^n = \begin{cases} \frac{\delta(1-\delta^m)}{(1-\delta)^2} - \frac{m\delta^{m+1}}{1-\delta} & \delta \neq 1 \\ \frac{1}{2}m(m+1) & \delta = 1 \end{cases} \tag{B.2}$$

First consider $|\delta| = 1$. Using (B.1), one can evaluate $w_{1,m}(\pm\mathbf{P})$ for large m:

$$w_{1,m}(\mathbf{P}) \approx m\mathbf{P}_1^\star$$
$$w_{1,m}(-\mathbf{P}) \approx m\mathbf{P}_r^\star,$$

and using (B.2), one can evaluate $w_{2,m}(\pm\mathbf{P})$ for large m:

$$\begin{aligned} w_{2,m}(\mathbf{P}) &\approx \frac{1}{2}m(m+1)\mathbf{P}_1^\star \\ w_{2,m}(-\mathbf{P}) &\approx \frac{1}{2}m(m+1)\mathbf{P}_r^\star \end{aligned}$$

From these, we see that $w_{1,m}(\pm\mathbf{P}) \sim \mathcal{O}(m)$ and $w_{2,m}(\pm\mathbf{P}) \sim \mathcal{O}(m^2)$.

Next consider the case when $|\delta| < 1$. As $|\lambda_i| \leq 1$, we have $|\delta\lambda_i| < 1$ for $i = 1, \ldots, r$. So as $m \to \infty$, the terms that depend on m go to zero, as can be seen from (B.1) and (B.2). Hence both are $\mathcal{O}(1)$. ■

Proposition B.12. *For $p, q \in \mathbb{R}$ and $N \geq 3$,*

$$\sum_{1\leq j<k\leq N-1} p^j q^k = \begin{cases} (N^2-3N+2)/2 & p=q=1 \\ \frac{q(1-q^{N-1})}{(1-q)^2} - \frac{(N-2)q^N+q}{1-q} & p=1, q\neq 1 \\ (N-2)\frac{p}{1-p} - \frac{p^2(1-p^{N-2})}{(1-p)^2} & p\neq 1, q=1 \\ \frac{pq^2(1-q^{N-2})}{(1-p)(1-q)} - \frac{(pq)^2(1-(pq)^{N-2})}{(1-p)(1-pq)} & p\neq 1, q\neq 1, pq\neq 1 \\ \frac{q(1-q^{N-2})}{2-p-q} - (N-2)\frac{1}{1-p} & p\neq 1, q\neq 1, pq=1 \end{cases}$$

This can be showed by going through each of the cases and applying the summation for a geometric series where appropriate.

Proposition B.13. *With $\mathbf{Q}$ defined as in (4.26), let $\kappa_1, \ldots, \kappa_r$ be its eigenvalues. Then, if $\psi_2, \psi_{r-1} \neq 0$, $|\kappa_i| < 1$ for all $1 \leq i \leq r$.*

Proof: Apply Gerschgorin's theorem as in Prop. B.4. Let λ be an eigenvalue of $\mathbf{Q}$. Then, $|\lambda| \leq \max_j \sum_{k\neq j} |q_{jk}|$, where $\mathbf{Q} = (q_{jk})$. From (4.26), $\mathbf{Q} = \mathbf{P}\mathbf{M}_{\psi 1}$, where

$\mathbf{M}_{\psi 1} = \mathrm{diag}(e^{-\psi_1^2/2\sigma^2}, \ldots, e^{-\psi_r^2/2\sigma^2})$. So $\mathbf{Q} = (e^{-\psi_1^2/2\sigma^2} p_{*1}, \ldots, e^{-\psi_r^2/2\sigma^2} p_{*r})$, where $\mathbf{P} = (p_{*1}, \ldots, p_{*r})$. Therefore,

$$|\lambda| \leq \max_j \sum_{k \neq j} e^{-\psi_k^2/2\sigma^2} |p_{jk}|$$

Now, $\sum_{k \neq j} |p_{jk}| = 1$, and $e^{-\psi_j^2/2\sigma^2} \leq 1$ for all j. So if $e^{-\psi_k^2/2\sigma^2} < 1$ for some k where $p_{jk} \neq 0$, the RHS of the expression above will be strictly less than one. If this is true for all $1 \leq j \leq r$, then $|\lambda| < 1$. We are assured of this when $\psi_2, \psi_{r-1} \neq 0$. ■

Proposition B.14. *The eigenvalues of* $\mathbf{Q}$ *can be made arbitrarily close to those of* $\mathbf{P}$ *by decreasing the SNR.*

Proof: Under the low SNR assumption, $\mathbf{Q} \approx \mathbf{P} - \frac{1}{2\sigma^2}\mathbf{P}\mathbf{M}_\psi^2$. Let $\mathbf{E} \triangleq -\frac{1}{2\sigma^2}\mathbf{P}\mathbf{M}_\psi^2 = -\frac{1}{2\sigma^2}(\psi_1^2 p_{*1}, \ldots, \psi_r^2 p_{*r})$. The problem comes down to characterizing the eigenvalues of $\mathbf{Q}$, which is a perturbed version of $\mathbf{P}$. By Prop. B.5 and Prop. B.2, $\mathbf{P}$ has a basis of eigenvectors, so that we can write $\mathbf{P} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$. Apply the Bauer-Fike theorem to $\mathbf{Q} = \mathbf{P} + \mathbf{E}$. Let λ be an eigenvalue of $\mathbf{Q}$: then $\min_{1 \leq i \leq r} |\lambda - \lambda_i| \leq \|\mathbf{U}\| \cdot \|\mathbf{U}^{-1}\| \cdot \|\mathbf{E}\| = \kappa(\mathbf{U})\|\mathbf{E}\|$, where $\kappa(\cdot)$ is the condition number of the matrix argument.

Let us evaluate $\|\mathbf{E}\|$: using the infinity-norm, $\|\mathbf{E}\|_\infty = \|\frac{1}{2\sigma^2}(\psi_1^2 p_{*1}, \ldots, \psi_r^2 p_{*r})\|_\infty \leq \frac{1}{2\sigma^2} \max_k \psi_k^2 \Rightarrow \min_{1 \leq i \leq r} |\lambda - \lambda_i| \leq \kappa(\mathbf{U}) \frac{1}{2\sigma^2} \max_k \psi_k^2$. Since $\kappa(\mathbf{U})$ is a bounded constant, we can make the RHS arbitrarily small by decreasing the SNR. ■

Proposition B.15. *Let* $\mathbf{P}$ *be a matrix associated with a random walk in* $\mathcal{R}_0^r$. *Recall the definition of* $\mathbf{M}_\psi$: $\mathbf{M}_\psi = \mathrm{diag}(\psi_1, \ldots, \psi_r)$. *The quantity* $p_{\mathrm{ss}}^T \mathbf{M}_\psi \mathbf{P}_1^\star \mathbf{M}_\psi \underline{1} = (\lim E[X_j])^2$, *where* $\lim E[X_j]$ *is the steady state expected value of the random walk* X_j, *and* p_{ss} *is the steady state probability distribution of* X_j. *The matrix* $\mathbf{P}_1^\star$ *is defined in Prop. B.10.*

Proof: Recall the definition of $\mathbf{P}_1^\star$, which is $\mathbf{P}_1^\star = \mathbf{U}\mathbf{M}_{11}\mathbf{U}^{-1}$, where $\mathbf{U}$ is the matrix s.t. $\mathbf{P} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{-1}$. The diagonal matrix of eigenvalues $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$, where by Prop. B.5, we can order $1 = \lambda_1 > \lambda_2 > \ldots > \lambda_r = -1$.

So let $\mathbf{U} = (u_{ij})$ and $\mathbf{U}^{-1} = (t_{ij})$. First, note that since

$$\mathbf{U}^{-1}\mathbf{U} = \mathbf{I} \Longrightarrow \sum_i t_{1i}u_{i1} = 1 \tag{B.3}$$

Denote by an asterisk the rest of the valid indices, e.g., u_{*1} is the first column of $\mathbf{U}$, t_{1*} is the first row of $\mathbf{U}^{-1}$, etc. Since we have $\mathbf{P}\mathbf{U} = \mathbf{U}\mathbf{\Lambda}$, u_{*1} is a right eigenvector of $\mathbf{P}$ with eigenvalue 1. By Prop. B.4, we can write

$$u_{*1} = c_1(1, \ldots, 1)^T \tag{B.4}$$

for some constant $c_1 \in \mathbb{R}$. Similarly, since $\mathbf{U}^{-1}\mathbf{P} = \mathbf{\Lambda}\mathbf{U}^{-1}$, t_{1*} is a left eigenvector of $\mathbf{P}$ with eigenvalue 1. Therefore, it must be a multiple of p_{ss}, the steady state probability distribution, i.e.,

$$t_{1*} = c_2 p_{\mathrm{ss}}^T \tag{B.5}$$

for some constant $c_2 \in \mathbb{R}$.

Substituting (B.4) and (B.5) into (B.3) results in $c_1c_2 = 1$. Now,

$$p_{\mathrm{ss}}^T \mathbf{M}_\psi \mathbf{P}_1^\star \mathbf{M}_\psi \underline{1} = \sum_{i,j} p_{\mathrm{ss},i}\psi_i \, \mathbf{P}_1^\star[i,j] \, \psi_j, \tag{B.6}$$

where $p_{\mathrm{ss},i}$ denotes the ith value of p_{ss}, and $\mathbf{P}_1^\star[i,j]$ denotes the (i,j)th value of $\mathbf{P}_1^\star$. By going through the definition of $\mathbf{P}_1^\star$, one obtains $\mathbf{P}_1^\star[i,j] = u_{i1}t_{1j}$. Substituting the former relation into (B.6), and applying (B.4) along with (B.5),

$$\begin{aligned} p_{\mathrm{ss}}^T \mathbf{M}_\psi \mathbf{P}_1^\star \mathbf{M}_\psi \underline{1} &= \sum_{i,j} p_{\mathrm{ss},i}\psi_i \, u_{i1}t_{1j} \, \psi_j \\ &= \sum_{i,j} p_{\mathrm{ss},i}\psi_i \, c_1c_2 p_{\mathrm{ss},j} \, \psi_j \\ &= \sum_{i,j} (p_{\mathrm{ss},i}\psi_i)(p_{\mathrm{ss},j}\psi_j) \qquad \text{(B.7)} \end{aligned}$$

where the fact that $c_1 c_2 = 1$ was used in the last step.

However, the last expression in (B.7) is just $\left(\sum_i p_{\mathrm{ss},i}\psi_i\right)^2 = (\lim E[X_j])^2 \geq 0$. ■

APPENDIX C

Second order approximation of the LRT for the DTRW model

We shall start with (4.29) and continue with the second order approximation for the LRT of the DTRW model. Now, $\kappa_1 \leq 1$ (and note that $\kappa_1 > 0$). Divide (4.29) by κ_1^{N-1}. Since this term is independent of the observations, it does not affect the performance of the test statistic. Let

$$\tilde{z}_j \triangleq \frac{z_j}{\sigma}, \;\; \widetilde{\mathbf{\Lambda}}_Q \triangleq \frac{1}{\kappa_1}\mathbf{\Lambda}_Q = \mathrm{diag}(\kappa_1', \ldots, \kappa_r'), \;\; \widetilde{\mathbf{R}} \triangleq \mathbf{U}_Q^{-1}\mathbf{R}\mathbf{U}_Q,$$

so that $\kappa_1' = 1$ and $\kappa_r' = -1$. One can write:

$$\begin{aligned} L_{N-1}(\tilde{z}^{N-1}) \approx{}& \tilde{z}_0\pi^T\mathbf{M}_{\psi 2}\mathbf{U}_Q\widetilde{\mathbf{\Lambda}}_Q^{N-1}\mathbf{U}_Q^{-1}\underline{1} + \frac{1}{\kappa_1}\sum_{j=1}^{N-1}\tilde{z}_j\pi^T\mathbf{M}_{\psi 1}\mathbf{U}_Q\widetilde{\mathbf{\Lambda}}_Q^{j-1}\widetilde{\mathbf{R}}\widetilde{\mathbf{\Lambda}}_Q^{N-1-j}\mathbf{U}_Q^{-1}\underline{1} \\ &+ \frac{1}{\kappa_1}\sum_{j=1}^{N-1}\tilde{z}_0\tilde{z}_j\pi^T\mathbf{M}_{\psi 2}\mathbf{U}_Q\widetilde{\mathbf{\Lambda}}_Q^{j-1}\widetilde{\mathbf{R}}\widetilde{\mathbf{\Lambda}}_Q^{N-1-j}\mathbf{U}_Q^{-1}\underline{1} \\ &+ \frac{1}{\kappa_1^2}\sum_{1\leq j<k\leq N-1}\tilde{z}_j\tilde{z}_k\pi^T\mathbf{M}_{\psi 1}\mathbf{U}_Q\widetilde{\mathbf{\Lambda}}_Q^{j-1}\widetilde{\mathbf{R}}\widetilde{\mathbf{\Lambda}}_Q^{k-1-j}\widetilde{\mathbf{R}}\widetilde{\mathbf{\Lambda}}_Q^{N-1-k}\mathbf{U}_Q^{-1}\underline{1}, \end{aligned} \tag{C.1}$$

ignoring constants and terms of higher order. In order to focus on the important properties of (C.1), define $\pi_\alpha^T \triangleq \pi^T\mathbf{M}_{\psi 1}\mathbf{U}_Q$, $\pi_\beta^T \triangleq \pi^T\mathbf{M}_{\psi 2}\mathbf{U}_Q$, and $d \triangleq \mathbf{U}_Q^{-1}\underline{1}$. Then, the RHS is:

$$\begin{aligned} L_{N-1}(\tilde{z}^{N-1}) \approx\ & \tilde{z}_0 \pi_\beta^T \widetilde{\mathbf{\Lambda}}_Q^{N-1} d + \frac{1}{\kappa_1} \sum_{j=1}^{N-1} \tilde{z}_j \pi_\alpha^T \widetilde{\mathbf{\Lambda}}_Q^{j-1} \widetilde{\mathbf{R}} \widetilde{\mathbf{\Lambda}}_Q^{N-1-j} d \\ & + \frac{1}{\kappa_1} \sum_{j=1}^{N-1} \tilde{z}_0 \tilde{z}_j \pi_\beta^T \widetilde{\mathbf{\Lambda}}_Q^{j-1} \widetilde{\mathbf{R}} \widetilde{\mathbf{\Lambda}}_Q^{N-1-j} d \\ & + \frac{1}{\kappa_1^2} \sum_{1 \leq j < k \leq N-1} \tilde{z}_j \tilde{z}_k \pi_\alpha^T \widetilde{\mathbf{\Lambda}}_Q^{j-1} \widetilde{\mathbf{R}} \widetilde{\mathbf{\Lambda}}_Q^{k-1-j} \widetilde{\mathbf{R}} \widetilde{\mathbf{\Lambda}}_Q^{N-1-k} d \end{aligned} \tag{C.2}$$

We shall analyze (C.2) in parts. Separate the first-order and second-order terms of (C.2) as follows:

$$L_{N-1,1}(\tilde{z}^{N-1}) \triangleq \tilde{z}_0 \pi_\beta^T \widetilde{\mathbf{\Lambda}}_Q^{N-1} d + \frac{1}{\kappa_1} \sum_{j=1}^{N-1} \tilde{z}_j \pi_\alpha^T \widetilde{\mathbf{\Lambda}}_Q^{j-1} \widetilde{\mathbf{R}} \widetilde{\mathbf{\Lambda}}_Q^{N-1-j} d \tag{C.3}$$

$$L_{N-1,2a}(\tilde{z}^{N-1}) \triangleq \frac{1}{\kappa_1} \sum_{j=1}^{N-1} \tilde{z}_0 \tilde{z}_j \pi_\beta^T \widetilde{\mathbf{\Lambda}}_Q^{j-1} \widetilde{\mathbf{R}} \widetilde{\mathbf{\Lambda}}_Q^{N-1-j} d \tag{C.4}$$

$$L_{N-1,2b}(\tilde{z}^{N-1}) \triangleq \frac{1}{\kappa_1^2} \sum_{1 \leq j < k \leq N-1} \tilde{z}_j \tilde{z}_k \pi_\alpha^T \widetilde{\mathbf{\Lambda}}_Q^{j-1} \widetilde{\mathbf{R}} \widetilde{\mathbf{\Lambda}}_Q^{k-1-j} \widetilde{\mathbf{R}} \widetilde{\mathbf{\Lambda}}_Q^{N-1-k} d \tag{C.5}$$

and $L_{N-1,2}(\tilde{z}^{N-1}) = L_{N-1,2a}(\tilde{z}^{N-1}) + L_{N-1,2b}(\tilde{z}^{N-1})$.

The term $L_{N-1,2a}(\tilde{z}^{N-1})$ is the effect of z_0 on the LR. When N is large, we expect that the effect is negligible compared to $L_{N-1,2b}(\tilde{z}^{N-1})$. Define c_{jk} to be the coefficient of $(1/\kappa_1^2)\tilde{z}_j \tilde{z}_k$ in (C.5) above. Let

$$\mathbf{\Upsilon}[j,k] \triangleq \widetilde{\mathbf{\Lambda}}_Q^{j-1} \widetilde{\mathbf{R}} \widetilde{\mathbf{\Lambda}}_Q^{k-1-j} \widetilde{\mathbf{R}} \widetilde{\mathbf{\Lambda}}_Q^{N-1-k} \tag{C.6}$$

so that $c_{jk} = \pi_\alpha^T \mathbf{\Upsilon}[j,k] d$. Recall that $\widetilde{\mathbf{\Lambda}}_Q = \mathrm{diag}(\kappa_1', \ldots, \kappa_r')$, where $1 = \kappa_1' > \kappa_2' > \ldots > \kappa_r' = -1$.

For N large, most of the c_{jk}'s will have j and $(N-k)$ sufficiently large that

$$\widetilde{\mathbf{\Lambda}}_Q^{j-1} \approx \mathrm{diag}(1, 0, \ldots, 0, (-1)^{j-1}) \text{ and } \widetilde{\mathbf{\Lambda}}_Q^{N-1-k} \approx \mathrm{diag}(1, 0, \ldots, 0, (-1)^{N-1-k})$$

Defining M_{ij} to be a $r \times r$ matrix with all zeros except for a 1 in the (i,j)-th position,

$$\begin{aligned}\boldsymbol{\Upsilon}[j,k] &\approx [\mathbf{M}_{11} + (-1)^{j-1}\mathbf{M}_{rr}]\widetilde{\mathbf{R}}\widetilde{\boldsymbol{\Lambda}}_Q^{k-1-j}\widetilde{\mathbf{R}}[\mathbf{M}_{11} + (-1)^{N-1-k}\mathbf{M}_{rr}] \\ &= \mathbf{M}_{11}\widetilde{\mathbf{R}}\widetilde{\boldsymbol{\Lambda}}_Q^{k-1-j}\widetilde{\mathbf{R}}\mathbf{M}_{11} + (-1)^{N-k+j}\mathbf{M}_{rr}\widetilde{\mathbf{R}}\widetilde{\boldsymbol{\Lambda}}_Q^{k-1-j}\widetilde{\mathbf{R}}\mathbf{M}_{rr} + \\ &\qquad (-1)^{j-1}\mathbf{M}_{rr}\widetilde{\mathbf{R}}\widetilde{\boldsymbol{\Lambda}}_Q^{k-1-j}\widetilde{\mathbf{R}}\mathbf{M}_{11} + (-1)^{N-1-k}\mathbf{M}_{11}\widetilde{\mathbf{R}}\widetilde{\boldsymbol{\Lambda}}_Q^{k-1-j}\widetilde{\mathbf{R}}\mathbf{M}_{rr} \end{aligned} \tag{C.7}$$

for j and $(N-k)$ sufficiently large.

The first two terms of (C.7) are functions of $(k-j)$, while the last two are not. One of the defining characteristics of the filtered energy statistic in (A.8) is that the coefficient of $z_j z_k$ is α^{k-j}. NB. the observations in Appendix A are denoted by y_i as compared to z_i in this appendix. In the event that the first two terms of (C.7) are dominant, c_{jk} will consist of a weighted sum of these exponential terms. Indeed, we can see that the exponential terms in c_{jk} will have the form $(\kappa_i')^{k-j}$. Consequently,

$$L_{N-1,2b}(\tilde{z}^{N-1}) \approx \sum_{n=1}^{r}\sum_{j<k} A_n(\kappa_n')^{k-j}\tilde{z}_j\tilde{z}_k \tag{C.8}$$

for some constants A_n, $n = 1,\ldots,r$. The filters for $n = 2,\ldots,(r-1)$ can be approximated by the FE statistic given by (4.17), while the filters for $n = 1, r$ can be generated as second order polynomials in $\tilde{z}_i$. Recall that $|\kappa_1'| = |\kappa_r'| = 1$, and so the FE statistic cannot be used for $n = 1, r$.

If $c_{jk} \approx C\alpha^{k-j}$ for some appropriate $C, \alpha \in \mathbb{R}$, $L_{N-1,2b}(\tilde{z}^{N-1})$ can be realized by a single filtered energy statistic, assuming that $\alpha \neq \pm 1$.

The FE statistic in (A.8) contains terms in the form of z_i^2. A way of ensuring that (C.8) is properly implemented is to subtract out the energy terms from the FE statistic. We shall digress for a moment to investigate the relative importance of the energy terms vs. the cross terms in the FE statistic, just as in the DTRT model.

Now,

$$E_1\left[\frac{\alpha}{1+\alpha}\sum_i z_i^2\right] - E_0\left[\frac{\alpha}{1+\alpha}\sum_i z_i^2\right] \approx N\frac{\alpha}{1+\alpha}p_{\mathrm{ss}}^T\mathbf{M}_\psi^2\underline{1}, \tag{C.9}$$

where we assume that the DTRW is approximately stationary, so that (4.24) applies. Next, as $E_0[\sum_{j<k}\alpha^{k-j}z_jz_k] = 0$, it remains to compute

$$\begin{aligned} E_1\left[\sum_{j<k}\alpha^{k-j}z_jz_k\right] &\approx p_{\mathrm{ss}}^T\mathbf{M}_\psi\left[\sum_{j<k}(\alpha\mathbf{P})^{k-j}\right]\mathbf{M}_\psi\underline{1} \\ &= p_{\mathrm{ss}}^T\mathbf{M}_\psi[Nw_{1,N}(\alpha\mathbf{P}) - w_{2,N}(\alpha\mathbf{P})]\mathbf{M}_\psi\underline{1} \end{aligned} \tag{C.10}$$

where $w_{1,N}(\cdot)$ and $w_{2,N}(\cdot)$ are defined in Prop. B.10 in Appendix B, and Prop. B.9 was applied in going to the second step. Note that we require $|\alpha| < 1$ for stability of the LPF in (4.16). By Prop. B.11, both $w_{1,N}(\alpha\mathbf{P})$ and $w_{2,N}(\alpha\mathbf{P})$ are $\mathcal{O}(1)$ in the limit as $N \to \infty$, since $|\alpha| < 1$. Therefore, the $Nw_{1,N}(\alpha\mathbf{P})$ term is dominant, and applying Prop. B.10

$$\begin{aligned} E_1\left[\sum_{j<k}\alpha^{k-j}z_jz_k\right] &\approx Np_{\mathrm{ss}}^T\mathbf{M}_\psi\left\{\sum_{i=1}^r\mathbf{P}_i^\star\left[\sum_{n=1}^m(\alpha\lambda_i)^n\right]\right\}\mathbf{M}_\psi\underline{1} \\ &\approx N\sum_{i=1}^r\frac{\alpha\lambda_i}{1-\alpha\lambda_i}p_{\mathrm{ss}}^T\mathbf{M}_\psi\mathbf{P}_i^\star\mathbf{M}_\psi\underline{1} \end{aligned} \tag{C.11}$$

as $N \to \infty$, where the λ_is are the eigenvalues of $\mathbf{P}$ defined according to Prop. B.5.

It is not clear if (C.11) is bigger than (C.9). Let

$$w_i \triangleq p_{\mathrm{ss}}^T\mathbf{M}_\psi\mathbf{P}_i^\star\mathbf{M}_\psi\underline{1} \tag{C.12}$$

In Prop. B.15, we show that $w_1 = (\lim E[X_j])^2 \geq 0$; however, it might be the case that $\lim E[X_j] = 0$. Define

$$M(u) \triangleq \sum_{i=1}^r\frac{u\lambda_i}{1-u\lambda_i}w_i \tag{C.13}$$

For the FE statistic to be used in (C.8), we require

$$M(u) \gg \frac{\kappa_n'}{1+\kappa_n'}p_{\mathrm{ss}}^T\mathbf{M}_\psi^2\underline{1} \text{ for all } u \in \{\kappa_n'\}_{n\notin\{1,r\}} \tag{C.14}$$

If (C.14) is not true for a certain value of $n = N$, the energy terms will for the LPF with $\alpha = \kappa'_N$ will have to be subtracted out. Note that $p_{\text{ss}}^T \mathbf{M}_\psi^2 \underline{1} = \lim E[X_j^2]$, i.e., it is the steady state expected energy of the DTRW. For values of $0 \leq \kappa'_n \leq 1$, it is sufficient to show that

$$M(\kappa'_n) \gg p_{\text{ss}}^T \mathbf{M}_\psi^2 \underline{1} = \lim E[X_j^2] \tag{C.15}$$

as $\left|\frac{\kappa'_n}{1+\kappa'_n}\right| \leq \frac{1}{2}$.

Ending the digression and returning to (C.8), let us now investigate conditions under which c_{jk} is approximately a function of $(k-j)$. Let $\widetilde{\mathbf{R}} = (\rho_{ij})$. An asterisk in either the row or column index shall denote all valid values. For example, the notation ρ_{1*} refers to the first row of $\widetilde{\mathbf{R}}$, ρ_{*r} refers to the last column of $\widetilde{\mathbf{R}}$ etc. For $x, y \in \mathbb{R}^r$, define the operator $x \odot y \triangleq (x_1 y_1, \ldots, x_r y_r)^T$. Define $S : \mathbb{R}^r \to \mathbb{R}$ by $S(x) = \sum_{i=1}^r x_i$. NB. $S(x)$ would equal the l_1 norm of x if all of the x_is were positive.

Let $\kappa' \triangleq [\kappa'_1, \ldots, \kappa'_r]^T$ and use the notation that for $x \in \mathbb{R}^r, x^{<i>} = [x_1^i, \ldots, x_r^i]$. Rewrite $\boldsymbol{\Upsilon}[j,k]$ using the newly defined notation as

(C.16)

$$\boldsymbol{\Upsilon}[j,k] = S((\kappa')^{<k-1-j>} \odot \rho_{1*}^T \odot \rho_{*1})\mathbf{M}_{11} + (-1)^{N-k+j} S((\kappa')^{<k-1-j>} \odot \rho_{r*}^T \odot \rho_{*r})\mathbf{M}_{rr} +$$
$$(-1)^{j-1} S((\kappa')^{<k-1-j>} \odot \rho_{r*}^T \odot \rho_{*1})\mathbf{M}_{r1} + (-1)^{N-1-k} S((\kappa')^{<k-1-j>} \odot \rho_{1*}^T \odot \rho_{*r})\mathbf{M}_{1r}$$

so that the dependence on j, k, and N is clear.

We shall say that c_{jk} is approximately a function of $(k-j)$ if the terms of the vector $(\rho_{r*}^T \odot \rho_{*1})$ and $(\rho_{1*}^T \odot \rho_{*r})$ are negligible compared to $(\rho_{1*}^T \odot \rho_{*1})$ and $(\rho_{r*}^T \odot \rho_{*r})$. For example, the l_∞ norm could be used, so that c_{jk} is approximately a function of $(k-j)$ if

$$\|\rho_{1*}^T \odot \rho_{*1}\|_\infty,\ \|\rho_{r*}^T \odot \rho_{*r}\|_\infty \gg \|\rho_{r*}^T \odot \rho_{*1}\|_\infty,\ \|\rho_{1*}^T \odot \rho_{*r}\|_\infty \tag{C.17}$$

If, in addition, there exists $C, \alpha \in \mathbb{R}$ for which

$$S((\kappa')^{<k-1-j>} \odot \rho_{1*}^T \odot \rho_{*1})\pi_\alpha^T \mathbf{M}_{11} d + (-1)^{N-k+j} S((\kappa')^{<k-1-j>} \odot \rho_{r*}^T \odot \rho_{*r})\pi_\alpha^T \mathbf{M}_{rr} d \approx C\alpha^{k-j}, \tag{C.18}$$

then $c_{jk} \approx C\alpha^{k-j}$.

In particular, suppose that for some $1 < i < \lfloor \frac{r}{2} \rfloor$, we have

$$\rho_{1*}^T \odot \rho_{*1} \approx C_1 e_i \quad \text{and} \quad \rho_{r*}^T \odot \rho_{*r} \approx C_2 e_{r+1-i} \tag{C.19}$$

for some $C_1, C_2 \in \mathbb{R}$ and where the e_is are the standard unit vectors in $\mathbb{R}^r$. We rule out the case of $i = 1$; since $\kappa'_1 = 1$, the FE statistic cannot be used. Then, the LHS of (C.18) reduces to

$$[C_1 \pi_\alpha^T \mathbf{M}_{11} d + (-1)^{N-1} C_2 \pi_\alpha^T \mathbf{M}_{rr} d](\kappa'_i)^{k-j-1}$$

since $\kappa'_{r+1-i} = -\kappa'_i$. Therefore,

$$\begin{aligned} C &= [C_1 \pi_\alpha^T \mathbf{M}_{11} d + (-1)^{N-1} C_2 \pi_\alpha^T \mathbf{M}_{rr} d](\kappa'_i)^{-1} \\ \alpha &= \kappa'_i \end{aligned} \tag{C.20}$$

If $\lim E[X_j] = 0$ and (C.15) is satisfied for $\kappa'_n = \kappa'_i$, we have sufficient conditions for a FE statistic to approximate the LRT for the class of random walks considered.

APPENDIX D

Sparse image reconstruction results

D.1 Derivation of Proposition 5.3

Use the following definitions, which appear in [10]:

$$\Xi(\underline{\theta}; \underline{a}) \triangleq C\|\underline{\theta} - \underline{a}\|^2 - \|\mathbf{H}\underline{\theta} - \mathbf{H}\underline{a}\|^2 \tag{D.1}$$

$$\Phi^{\text{SUR}}(\underline{\theta}; \underline{a}) \triangleq \Phi(\underline{\theta}) + \Xi(\underline{\theta}; \underline{a}), \tag{D.2}$$

where C is chosen to ensure that $\Xi(\underline{\theta}; \underline{a})$ is strictly convex in $\underline{\theta}$ for any choice of $\underline{a}$ [10]. By assumption, $\|H\| < 1$, and so we can select $C = 1$. The function $\Phi^{\text{SUR}}(\underline{\theta}; \underline{a})$ is the surrogate function that is minimized in place of $\Phi(\underline{\theta})$. Consider the minimization of $\Phi^{\text{SUR}}(\underline{\theta}; \underline{a})$, which can be simplified as

$$\Phi^{\text{SUR}}(\underline{\theta}; \underline{a}) = \|\underline{\theta}\|^2 - 2(\underline{a} + \mathbf{H}^T(\underline{y} - \mathbf{H}\underline{a}))^T\underline{\theta} + J(\underline{\theta}) + \|\underline{y}\|^2 + \|\underline{a}\|^2 - \|\mathbf{H}\underline{a}\|^2 \tag{D.3}$$

Since $J(\underline{\theta}) = \sum_i J_1(\theta_i)$, we see that the minimization of $\Phi^{\text{SUR}}(\underline{\theta}; \underline{a})$ can be decomposed into M subproblems, where each θ_i is separately minimized. Indeed, each θ_i should minimize

$$\varphi(\theta_i) \triangleq \theta_i^2 - 2s_i\theta_i + J_1(\theta_i), \tag{D.4}$$

where $\underline{s} \triangleq \underline{a} + \mathbf{H}^T(\underline{y} - \mathbf{H}\underline{a})$. Since (D.4) is convex, the minimizing θ_i can be found by solving for $\varphi'(\theta_i) = 0$. This results in $\theta_i = T(s_i)$.

Let $\hat{\underline{\theta}}^{(t)}$ denote the sequence generated by the following:

$$\hat{\underline{\theta}}^{(t+1)} = \text{argmin}_{\underline{\theta}}\, \Phi^{\text{SUR}}(\underline{\theta}; \hat{\underline{\theta}}^{(t)}) \tag{D.5}$$

where $\hat{\underline{\theta}}^{(0)}$ is the initial estimate of $\underline{\theta}$. Then, the sequence $\hat{\underline{\theta}}^{(t)}$ monotonically decreases the cost function $\Phi(\underline{\theta})$.

D.2 Derivation of Stein's unbiased risk estimator for the L1 estimator

Let

$$\Psi_\beta(\underline{y}, \underline{\theta}) \triangleq \|\underline{y} - \mathbf{H}\underline{\theta}\|^2 + \beta\|\underline{\theta}\|_1, \tag{D.6}$$

which is the cost function that the L1 estimator minimizes. In this derivation, let $\hat{\underline{\theta}}(\beta)$ denote the L1 estimator with regularization parameter β. Assume that the columns of $\mathbf{H}$ are linearly independent.

The starting point that we use is [62, (2)], which is

$$\hat{R}(\beta) = N\sigma^2 + \|\underline{e}\|^2 - 2\sigma^2 \text{tr}(\mathbf{H} \cdot (\mathbf{D}_{\underline{\theta}\underline{\theta}}\Psi_\beta)^{-1} \cdot \mathbf{D}_{\underline{\theta}\underline{y}}\Psi_\beta)\Big|_{\underline{\theta}=\hat{\underline{\theta}}(\beta)}, \tag{D.7}$$

where $\underline{e} \triangleq \underline{y} - \mathbf{H}\hat{\underline{\theta}}(\beta)$ and $\mathbf{D}_{\underline{u},\underline{v}}(\cdot) \triangleq \partial^2(\cdot)/\partial\underline{u}\partial\underline{v}^T$. Define $\mathbf{Z}(\underline{\theta}) \triangleq \text{diag}(\delta(\theta_1), \ldots, \delta(\theta_M))$, where $\delta(\cdot)$ is the Dirac delta. One can compute that

$$\mathbf{D}_{\underline{\theta}\underline{\theta}}\Psi_\beta = 2\mathbf{H}^T\mathbf{H} + \mathbf{Z}(\underline{\theta}), \text{ and } \mathbf{D}_{\underline{\theta}\underline{y}}\Psi_\beta = -2\mathbf{H}^T. \tag{D.8}$$

This leads to

$$\hat{R}(\beta) = N\sigma^2 + \|\underline{y} - \mathbf{H}\hat{\underline{\theta}}(\beta)\|^2 + 2\sigma^2 \text{tr}(\mathbf{H}^T\mathbf{H}[\mathbf{H}^T\mathbf{H} + \frac{1}{2}\mathbf{Z}(\hat{\underline{\theta}}(\beta))]^{-1}) \tag{D.9}$$

We would like to evaluate the last term of (D.9). Henceforth, omit the β in $\hat{\underline{\theta}}(\beta)$ for the sake of brevity. Define $\mathbf{Z}_A(\underline{\theta}) \triangleq \text{diag}(A \cdot I(\theta_1 = 0), \ldots, A \cdot I(\theta_M = 0))$, a well-behaved version of $\mathbf{Z}(\underline{\theta})$. Note that $\mathbf{Z}(\underline{\theta}) = \lim_{A\to\infty} \mathbf{Z}_A(\underline{\theta})$.

Suppose that $0 \leq r \leq M$ elements of $\hat{\underline{\theta}}$ are zero, and the remaining elements are non-zero. Let $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{M\times M}$ be permutation matrices such that $\mathbf{P}\mathbf{Z}_A(\hat{\underline{\theta}})\mathbf{Q}$ looks like $\mathrm{diag}(A, \ldots, A, 0, \ldots, 0)$, i.e., a matrix whose diagonal contains r values of A followed by $(M-r)$ values of 0. In particular, we shall select $\mathbf{P}$ and $\mathbf{Q}$ in the way described in the next paragraph.

Denote the zero-valued indices of $\hat{\underline{\theta}}$ by $n_1, \ldots, n_r$, i.e., $\hat{\theta}_{n_i} = 0$ for $1 \leq i \leq r$. The permutation matrices

$$\mathbf{P} = \mathbf{P}_r \cdots \mathbf{P}_1 \text{ and } \mathbf{Q} = \mathbf{Q}_1 \cdots \mathbf{Q}_r, \tag{D.10}$$

where the effect of $\mathbf{P}_i\mathrm{diag}(d_1, \ldots, d_M)\mathbf{Q}_i$ is to exchange the places of d_i and d_{n_i}. So

$$\mathbf{P}_i = \mathbf{I}\Big|_{\underline{e}_i^T \leftrightarrow \underline{e}_{n_i}^T} \text{ and } \mathbf{Q}_i = \mathbf{I}\Big|_{\underline{e}_i \leftrightarrow \underline{e}_{n_i}}, \tag{D.11}$$

and therefore $\mathbf{P}_i = \mathbf{Q}_i^T$ for each i. So $\mathbf{Q}^T = \mathbf{P}$. Since each $\mathbf{P}_i$ and $\mathbf{Q}_i$ are orthogonal matrices, so is $\mathbf{P}$ and $\mathbf{Q}$. With the $\mathbf{P}$ and $\mathbf{Q}$ defined as in (D.10) and (D.11), it is clear that $\mathbf{P}\mathbf{Z}_A(\hat{\underline{\theta}})\mathbf{Q}$ equals $\mathrm{diag}(A, \ldots, A, 0, \ldots, 0)$.

Let $\mathbf{K} \triangleq \mathbf{H}^T\mathbf{H}$, a square matrix. Then, as matrix multiplication is commutative under the trace operator,

$$\mathrm{tr}(\mathbf{K}[\mathbf{K} + \frac{1}{2}\mathbf{Z}_A(\hat{\underline{\theta}})]^{-1}) = \mathrm{tr}(\mathbf{PKQ}[\mathbf{PKQ} + \frac{1}{2}\mathbf{P}\mathbf{Z}_A(\hat{\underline{\theta}})\mathbf{Q}]^{-1}) \tag{D.12}$$

Without loss of generality then, suppose that $\mathbf{Z}_A(\hat{\underline{\theta}})$ has all of its non-zero diagonal entries in the front, followed by zeros. Consider the expression

$$[\mathbf{K} + \frac{1}{2}\mathbf{Z}_A(\hat{\underline{\theta}})]^{-1} = \begin{pmatrix} \frac{1}{2}\mathrm{diag}(A, \ldots, A) + \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{pmatrix}^{-1} \tag{D.13}$$

Inverting the right hand side of (D.13) in parts,

$$[\mathbf{K} + \frac{1}{2}\mathbf{Z}_A(\hat{\underline{\theta}})]^{-1} = \begin{pmatrix} \mathbf{F}_{11}^{-1} & -\widetilde{\mathbf{K}}_{11}^{-1}\mathbf{K}_{12}\mathbf{F}_{22}^{-1} \\ -\mathbf{F}_{22}^{-1}\mathbf{K}_{21}\widetilde{\mathbf{K}}_{11}^{-1} & \mathbf{F}_{22}^{-1} \end{pmatrix} \tag{D.14}$$

where:

$$\widetilde{\mathbf{K}}_{11} \triangleq \frac{1}{2}\mathrm{diag}(A, \ldots, A) + \mathbf{K}_{11} \tag{D.15}$$

$$\mathbf{F}_{11} \triangleq \widetilde{\mathbf{K}}_{11} - \mathbf{K}_{12}\mathbf{K}_{22}^{-1}\mathbf{K}_{21} \tag{D.16}$$

$$\mathbf{F}_{22} \triangleq \mathbf{K}_{22} - \mathbf{K}_{21}\widetilde{\mathbf{K}}_{11}^{-1}\mathbf{K}_{12} \tag{D.17}$$

The equations (D.14)-(D.17) assume that $\widetilde{\mathbf{K}}_{11}$ and $\mathbf{K}_{22}$ are invertible. For sufficiently large A, $\widetilde{\mathbf{K}}_{11}$ will indeed be invertible. The invertibility of $\mathbf{K}_{22}$ is addressed in the following proposition. The subscript $(\cdot)_{22}$ denotes the lower right $(M-r)\times(M-r)$ submatrix of its argument.

Proposition D.1. *For the permutation matrices $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{M\times M}$ defined above, and for $\mathbf{H}$ with linearly independent columns, det$((\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{Q})_{22}) \neq 0$.*

By assumption, exactly r elements of $\underline{\hat{\theta}}$ are zero. Now, $\mathbf{P}$ and $\mathbf{Q}$ are given in (D.10) and (D.11). It is known that $\mathbf{P} = \mathbf{Q}^T$, so

$$\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{Q} = \mathbf{Q}^T\mathbf{H}^T\mathbf{H}\mathbf{Q} = (\mathbf{H}\mathbf{Q})^T(\mathbf{H}\mathbf{Q}) \tag{D.18}$$

Let $\widetilde{\mathbf{H}} \triangleq \mathbf{H}\mathbf{Q}$: it is $\mathbf{H}$ with its columns permuted. By assumption, the columns of $\mathbf{H}$ are linearly independent; then, so are the columns of $\widetilde{\mathbf{H}}$. For a square matrix $\mathbf{X}$, let $\mathrm{gram}(\mathbf{X}) \triangleq \det(\mathbf{X}^T\mathbf{X})$ to be the Grammian of $\mathbf{X}$, which is the determinant of the Gram matrix of $\mathbf{X}$. Denote the Gram matrix of $\mathbf{X}$ by $\mathbf{G}(\mathbf{X}) \triangleq \mathbf{X}^T\mathbf{X}$. Since the columns of $\widetilde{\mathbf{H}}$ are linearly independent, $\mathrm{gram}(\widetilde{\mathbf{H}}) > 0$.

We argue by contradiction. Suppose that

$$0 = \det((\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{Q})_{22}) = \det((\widetilde{\mathbf{H}}^T\widetilde{\mathbf{H}})_{22}) = \det(\mathbf{G}(\widetilde{\mathbf{H}})_{22})$$

Since $\det(\mathbf{G}(\widetilde{\mathbf{H}})_{22})$ is a principal minor of $\mathbf{G}(\widetilde{\mathbf{H}})$, we get that $\mathrm{gram}(\widetilde{\mathbf{H}}) = 0$. This is a contradiction. It must be the case that $\det((\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{Q})_{22}) \neq 0$. ■

As $A \to \infty$, $\widetilde{\mathbf{K}}_{11}^{-1} \to \mathbf{0}$, $\mathbf{F}_{11}^{-1} \to \mathbf{0}$, and $\mathbf{F}_{22} \to \mathbf{K}_{22}$. Therefore,

$$[\mathbf{K} + \frac{1}{2}\mathbf{Z}(\underline{\hat{\theta}})]^{-1} = \lim_{A\to\infty} [\mathbf{K} + \frac{1}{2}\mathbf{Z}_A(\underline{\hat{\theta}})]^{-1} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{22}^{-1} \end{pmatrix}, \tag{D.19}$$

and so as $A \to \infty$,

$$\mathbf{K}[\mathbf{K} + \frac{1}{2}\mathbf{Z}_A(\underline{\hat{\theta}})]^{-1} \to \begin{pmatrix} \mathbf{0} & \mathbf{K}_{12}\mathbf{K}_{22}^{-1} \\ \mathbf{0} & \mathbf{I}_{22} \end{pmatrix}. \tag{D.20}$$

Therefore, $\mathrm{tr}(\mathbf{K}[\mathbf{K} + \frac{1}{2}\mathbf{Z}_A(\underline{\hat{\theta}})]^{-1}) \to \#\{i : \hat{\theta}_i \neq 0\} = \|\underline{\hat{\theta}}\|_0$ as $A \to \infty$. Substitution of this result into (D.9) leads to

$$\hat{R}(\beta) = N\sigma^2 + \|\underline{y} - \mathbf{H}\underline{\hat{\theta}}(\beta)\|^2 + 2\sigma^2\|\underline{\hat{\theta}}(\beta)\|_0 \tag{D.21}$$

D.3 Derivation of cost function for hybrid thresholding function

Let $\underline{t} \triangleq (t_1, t_2)$, and $T(\cdot)$ denote the hybrid hard-soft thresholding function. Assume that $\|\mathbf{H}\| < 1$, so that the results of Prop. 5.3 are applicable. Then,

$$\Psi_{\underline{t}}(\underline{y}, \underline{\theta}) = \|\mathbf{H}\underline{\theta} - \underline{y}\|^2 + \sum_i J_1(\theta_i)$$
$$\text{where: } J_1(x) = 2T^{-1}(x)x - x^2 - 2\int T(\xi)d\xi\bigg|_{\xi=T^{-1}(x)} \tag{D.22}$$

Now,

$$\int T(\xi)d\xi = \begin{cases} (1/2)\xi^2 - |\xi|t_2 + c_1 & |\xi| > t_1 \\ c_2 & \text{o.w.} \end{cases} \tag{D.23}$$

where we shall set the constants $c_1 = c_2 = 0$, as their values do not affect the minimization of $\Psi_{\underline{t}}$. Then

$$\int T(\xi)d\xi = I(|\xi| > t_1)((1/2)\xi^2 - |\xi|t_2). \tag{D.24}$$

Next,

$$T^{-1}(x) = I(|x| < t_1 - t_2)\mathrm{sgn}(x)t_1 + I(|x| \geq t_1 - t_2)(x + \mathrm{sgn}(x)t_2) \tag{D.25}$$

which leads to

$$(\text{D.26}) \qquad (T^{-1}(x))^2 = I(|x| < t_1 - t_2)t_1^2 + I(|x| \geq t_1 - t_2)(x^2 + t_2^2 + 2t_2|x|), \text{ and}$$

$$(\text{D.27}) \qquad |T^{-1}(x)| = I(|x| < t_1 - t_2)t_1 + I(|x| \geq t_1 - t_2)(|x| + t_2).$$

Using (D.24)-(D.27) in (D.22), we obtain after some simplification

$$(\text{D.28}) \quad J_1(x) = I(|x| < t_1 - t_2)[-(x - \operatorname{sgn}(x)t_1)^2 + 2t_1t_2] + I(|x| \geq t_1 - t_2)(2t_2|x| + t_2^2)$$

Since

$$\lim_{x \uparrow 0}[-(x - \operatorname{sgn}(x)t_1)^2 + 2t_1t_2] = 2t_1t_2 - t_1^2 = \lim_{x \downarrow 0}[-(x - \operatorname{sgn}(x)t_1)^2 + 2t_1t_2]$$

and

$$\lim_{x \uparrow (t_1 - t_2)} J_1(x) = 2t_1t_2 - t_2^2 = \lim_{x \downarrow (t_1 - t_2)} J_1(x),$$

$J_1(x)$ is a continuous function.

D.4 Derivation of Stein's unbiased risk estimator for the HHS estimator

D.4.1 Preliminaries

Assume that the columns of $\mathbf{H}$ are linearly independent, and that the Gram matrix $\mathbf{G}(\mathbf{H})$ does not have an eigenvalue of 1/2.

We shall use [62, (2)], just as in the case of the L1 estimator, to derive an unbiased risk estimator. That is,

$$(\text{D.29}) \qquad \hat{R}(\underline{t}) = N\sigma^2 + \|\underline{e}\|^2 - 2\sigma^2 \operatorname{tr}(\mathbf{H} \cdot (\mathbf{D}_{\underline{\theta}\underline{\theta}}\Psi_{\underline{t}})^{-1} \cdot \mathbf{D}_{\underline{\theta}\underline{y}}\Psi_{\underline{t}})\Big|_{\underline{\theta} = \hat{\underline{\theta}}(\underline{t})},$$

where the quantities $\underline{e} = \underline{y} - \mathbf{H}\hat{\underline{\theta}}(\underline{t})$ and $\mathbf{D}_{\underline{u},\underline{v}}$ retain their original meaning in App. D.2. The cost function $\Psi_{\underline{t}}$ is given in (D.22). Now $\mathbf{D}_{\underline{\theta}\underline{y}}\Psi_{\underline{t}} = -2\mathbf{H}^T$, just as for the L1 estimator. However, to evaluate $\mathbf{D}_{\underline{\theta}\underline{\theta}}\Psi_{\underline{t}}$, we have to compute $J_1'(x)$ and $J_1''(x)$ for the $J_1(x)$ given in (D.22).

Let $\text{rect}(x) \triangleq 1$, $|x| \leq 1/2$ and 0 otherwise. It is known that $J_1'(x) = 2(T^{-1}(x) - x)$, which leads to

$$J_1''(x) = \delta(x) - \text{rect}\left(\frac{x}{2(t_1 - t_2)}\right). \tag{D.30}$$

Define

$$\mathbf{U}(\underline{\theta}) \triangleq \text{diag}\left(\text{rect}\left(\frac{\theta_1}{2(t_1 - t_2)}\right), \ldots, \text{rect}\left(\frac{\theta_M}{2(t_1 - t_2)}\right)\right), \tag{D.31}$$

for $t_2 < t_1$, and $\mathbf{0}$ when $t_2 = t_1$. The Hessian of $\Psi_{\underline{t}}$ w.r.t. $\underline{\theta}$ is

$$\mathbf{D}_{\underline{\theta}\underline{\theta}}\Psi_{\underline{t}} = 2\mathbf{H}^T\mathbf{H} + \mathbf{Z}(\underline{\theta}) - \mathbf{U}(\underline{\theta}) \tag{D.32}$$

We retain the meanings of $\mathbf{Z}(\underline{\theta})$ and $\mathbf{Z}_A(\underline{\theta})$ that were defined in App. D.2. Substituting the results into (D.29),

$$\hat{R}(\underline{t}) = N\sigma^2 + \|\underline{e}\|^2 + 2\sigma^2\text{tr}\left(\mathbf{H}^T\mathbf{H}\left[\mathbf{H}^T\mathbf{H} - \frac{1}{2}\mathbf{U}(\underline{\hat{\theta}}(\underline{t})) + \frac{1}{2}\mathbf{Z}(\underline{\hat{\theta}}(\underline{t}))\right]^{-1}\right). \tag{D.33}$$

For the rest of this section, assume that $t_2 < t_1$, so that $\mathbf{U}(\underline{\hat{\theta}}) \neq \mathbf{0}$. If $t_2 = t_1$, then $\hat{R}(\underline{t})$ will equal SURE of the L1 estimator, which has already been derived. Emulating the development of SURE for the L1 estimator, suppose that $0 \leq r \leq M$ elements of $\underline{\hat{\theta}}$ are zero, and the remaining elements are non-zero. Let $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{M\times M}$ be permutation matrices defined by (D.10)-(D.11) such that $\mathbf{P}\mathbf{Z}_A(\underline{\hat{\theta}})\mathbf{Q} = \text{diag}(A, \ldots, A, 0, \ldots, 0)$, where there are r values of A followed by $(M - r)$ values of 0. Recall that the subscript $(\cdot)_{22}$ denote the lower right $(M - r) \times (M - r)$ submatrix of the argument, e.g., (D.13).

Proposition D.2. *Suppose* $\mathbf{H}$ *has linearly independent columns. If* $\det[\mathbf{G}(\mathbf{H}) - \frac{1}{2}\mathbf{U}(\underline{\hat{\theta}})] = 0$*, then* $\mathbf{G}(\mathbf{H})$ *has an eigenvalue of* $\frac{1}{2}$*.*

If $\det[\mathbf{G}(\mathbf{H}) - \frac{1}{2}\mathbf{U}(\underline{\hat{\theta}})] = 0$, then the nullspace of $\mathbf{G}(\mathbf{H}) - \frac{1}{2}\mathbf{U}(\underline{\hat{\theta}})$ is non-trivial. There exists $x \in \mathbb{R}^M, x \neq 0$ such that

$$\mathbf{G}(\mathbf{H})x = \frac{1}{2}\mathbf{U}(\underline{\hat{\theta}})x. \tag{D.34}$$

Suppose that the diagonal of $\mathbf{U}(\underline{\hat{\theta}})$ has p ones that are indexed by $k_1, \ldots, k_p$ where $1 \leq p \leq M$. Define $\mathcal{K} \triangleq \{k_1, \ldots, k_p\}$. If $p = 0$, then $\mathbf{G}(\mathbf{H})x = \underline{0}$. But that is impossible since, by assumption, the columns of $\mathbf{H}$ are linearly independent, so $\mathbf{G}(\mathbf{H})$ is strictly positive definite. Denote the columns of $\mathbf{H}$ by $\underline{h}_i$, i.e., $\mathbf{H} = (\underline{h}_1 | \ldots | \underline{h}_M)$. With regard to (D.34),

$$\text{RHS} = \frac{1}{2}\sum_{j=1}^{p} x_{k_j}\underline{e}_{k_j}, \text{and the} \tag{D.35}$$

$$\text{LHS} = \begin{pmatrix} <\underline{h}_1, \underline{h}_1> x_1 + <\underline{h}_1, \underline{h}_2> x_2 + \ldots + <\underline{h}_1, \underline{h}_M> x_M \\ \vdots \\ <\underline{h}_M, \underline{h}_1> x_1 + \ldots + <\underline{h}_M, \underline{h}_M> x_M \end{pmatrix}$$

$$= \begin{pmatrix} <\underline{h}_1, \sum_j \underline{h}_j x_j> \\ \vdots \\ <\underline{h}_M, \sum_j \underline{h}_j x_j> \end{pmatrix} \tag{D.36}$$

Equating both sides, we have

$$<\underline{h}_i, \sum_j \underline{h}_j x_j> = \begin{cases} \frac{1}{2}x_{k_m} & i = k_m \text{ for some } 1 \leq m \leq p \\ 0 & \text{o.w.} \end{cases} \tag{D.37}$$

If $x_j = 0$ for all $j \in \mathcal{K}$, then (D.37) implies that $\sum_j \underline{h}_j x_j \perp \text{span}(\underline{h}_1, \ldots, \underline{h}_M)$. The only way that can happen is if $\sum_j \underline{h}_j x_j = 0$, which means that the $\underline{h}_j$s are linearly dependent since $x \neq 0$, a contradiction. So not all of the x_{k_j}s can be zero.

Define the subspaces $\mathcal{H}_1 \triangleq \text{span}(\{\underline{h}_j : j \in \mathcal{K}\})$ and $\mathcal{H}_2 \triangleq \text{span}(\{\underline{h}_j : j \notin \mathcal{K}\})$. As $\{\underline{h}_1, \ldots, \underline{h}_M\}$ is linearly independent by assumption, $\mathcal{H}_1 \perp \mathcal{H}_2$. Let $\underline{c} \triangleq \sum_j \underline{h}_j x_j$.

From (D.37), $\underline{c} \perp \mathcal{H}_2$. Therefore, $x_j = 0$ for $j \notin \mathcal{K}$. The LHS of (D.34) is

(D.38)
$$\begin{pmatrix} < \underline{h}_1, \sum_j \underline{h}_{k_j} x_{k_j} > \\ \vdots \\ < \underline{h}_M, \sum_j \underline{h}_{k_j} x_{k_j} > \end{pmatrix} = \begin{pmatrix} \sum_j < \underline{h}_1, \underline{h}_{k_j} > x_{k_j} \\ \vdots \\ \sum_j < \underline{h}_M, \underline{h}_{k_j} > x_{k_j} \end{pmatrix} = \mathbf{G}(\mathbf{H}) \left(\sum_{j=1}^{p} x_{k_j} \underline{e}_{k_j} \right)$$

Since not all of the x_{k_j}s can be zero, $\sum_j x_{k_j} \underline{e}_{k_j} \neq \underline{0}$. From (D.35) and (D.38), this means that 1/2 is an eigenvalue of $\mathbf{G}(\mathbf{H})$. ■

Proposition D.3. *Suppose that* $\mathbf{H}$ *has linearly independent columns. If det*$[(\mathbf{G}(\mathbf{H}) - \frac{1}{2}\mathbf{U}(\hat{\underline{\theta}}))_{22}] = 0$*, then* $\mathbf{G}(\mathbf{H})$ *has an eigenvalue of* $\frac{1}{2}$*.*

If $\det[(\mathbf{G}(\mathbf{H}) - \frac{1}{2}\mathbf{U}(\hat{\underline{\theta}}))_{22}] = 0$, the nullspace of $\mathbf{G}(\mathbf{H})_{22} - \frac{1}{2}\mathbf{U}(\hat{\underline{\theta}})_{22}$ is non-trivial. So there exists $\underline{v} \in \mathbb{R}^{M-r}, \underline{v} \neq \underline{0}$ such that

$$\mathbf{G}(\mathbf{H})_{22}\underline{v} = \frac{1}{2}\mathbf{U}(\hat{\underline{\theta}})_{22}\underline{v} \tag{D.39}$$

Now, $\mathbf{G}(\mathbf{H})_{22}$ is the Gram matrix of the columns $\underline{h}_{r+1}, \ldots, \underline{h}_M$ and looks like

$$\mathbf{G}(\mathbf{H})_{22} = \begin{pmatrix} < \underline{h}_{r+1}, \underline{h}_{r+1} > & \ldots & < \underline{h}_{r+1}, \underline{h}_M > \\ \vdots & \vdots & \vdots \\ < \underline{h}_M, \underline{h}_{r+1} > & \ldots & < \underline{h}_M, \underline{h}_M > \end{pmatrix} \tag{D.40}$$

The results of the previous proposition can be applied, with M substituted with $M - r$. Retain the notation used therein, so that the diagonal of $\mathbf{U}(\hat{\underline{\theta}})$ has exactly p ones indexed by $k_1, \ldots, k_p$ where $1 \leq p \leq M - r$. Then, $\mathbf{G}(\mathbf{H})_{22}$ has eigenvalue 1/2 with eigenvector $\sum_j v_{k_j} \underline{e}_{k_j} \in \mathbb{R}^{M-r}$, i.e.,

$$\mathbf{G}(\mathbf{H})_{22}\underline{v} = \begin{pmatrix} < \underline{h}_{r+1}, \sum_j \underline{h}_{r+k_j} v_{k_j} > \\ \vdots \\ < \underline{h}_M, \sum_j \underline{h}_{r+k_j} v_{k_j} > \end{pmatrix} = \frac{1}{2} \sum_j v_{k_j} \underline{e}_{k_j} \tag{D.41}$$

Let $\underline{\tilde{v}} = \sum_{j=1}^{p} v_{k_j}\underline{e}_{r+k_j} \in \mathbb{R}^M$. Consider

$$\text{(D.42)} \qquad \mathbf{G}(\mathbf{H})\underline{\tilde{v}} = \begin{pmatrix} <\underline{h}_1, \sum_j v_{k_j}\underline{h}_{r+k_j}> \\ \vdots \\ <\underline{h}_M, \sum_j v_{k_j}\underline{h}_{r+k_j}> \end{pmatrix}$$

As in Prop. D.2, let $\mathcal{K} = \{k_1, \ldots, k_p\}$. Define the subspaces $\mathcal{H}_1 \triangleq \text{span}(\{\underline{h}_j : j \in r + \mathcal{K}\})$ and $\mathcal{H}_2 \triangleq \text{span}(\{\underline{h}_j : j \notin r + \mathcal{K}\})$. The set $\{\underline{h}_1, \ldots, \underline{h}_M\}$ is linearly independent; therefore, $\mathcal{H}_1 \perp \mathcal{H}_2$. Let $\underline{c} = \sum_j v_{k_j}\underline{h}_{r+k_j}$. Since $\underline{c} \in \mathcal{H}_1 \Longrightarrow \underline{c} \perp \mathcal{H}_2$. Combining this with (D.41),

$$\text{(D.43)} \qquad <\underline{h}_i, \sum_j v_{k_j}\underline{h}_{r+k_j} >= \begin{cases} \frac{1}{2}v_{k_m} & i = r + k_m \text{ for some } 1 \leq m \leq p \\ 0 & \text{o.w.} \end{cases}$$

Using (D.43) in (D.42),

$$\text{(D.44)} \qquad \mathbf{G}(\mathbf{H})\underline{\tilde{v}} = \frac{1}{2}\sum_j v_{k_j}\underline{e}_{r+k_j} = \frac{1}{2}\underline{\tilde{v}}$$

Since $\sum_j v_{k_j}\underline{e}_{k_j}$ is an eigenvector of $\mathbf{G}(\mathbf{H})_{22}$, not all of the v_{k_j}s are zero. This implies that $\underline{\tilde{v}} \neq 0$. Therefore, from (D.44), we get that 1/2 is an eigenvalue of $\mathbf{G}(\mathbf{H})$, as required. ■

Proposition D.4. *Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be the permutation matrices defined by (*D.*10)-(*D.*11), and suppose that the columns of* $\mathbf{H}$ *are linearly independent. If* $det[(\mathbf{P}(\mathbf{G}(\mathbf{H}) - \frac{1}{2}\mathbf{U}(\underline{\hat{\theta}}))\mathbf{Q})_{22}] = 0$*, then* $\mathbf{G}(\mathbf{H})$ *has an eigenvalue of* $\frac{1}{2}$*.*

Using $\mathbf{P} = \mathbf{Q}^T$,

$$\begin{aligned} \mathbf{P}(\mathbf{G}(\mathbf{H}) - \frac{1}{2}\mathbf{U}(\underline{\hat{\theta}}))\mathbf{Q} &= \mathbf{Q}^T\mathbf{H}^T\mathbf{H}\mathbf{Q} - \frac{1}{2}\mathbf{Q}^T\mathbf{U}(\underline{\hat{\theta}})\mathbf{Q} \\ \text{(D.45)} \qquad &= \mathbf{G}(\widetilde{\mathbf{H}}) - \frac{1}{2}\mathbf{U}(\underline{\theta}^*) \end{aligned}$$

where $\widetilde{\mathbf{H}} \triangleq \mathbf{H}\mathbf{Q}$, i.e., $\mathbf{H}$ with its columns permuted, and $\underline{\theta}^*$ is a permutation of $\underline{\hat{\theta}}$ so that $\mathbf{U}(\underline{\theta}^*) = \mathbf{Q}^T\mathbf{U}(\underline{\hat{\theta}})\mathbf{Q}$. Since the columns of $\widetilde{\mathbf{H}}$ are linearly independent, we can apply Prop. D.3 to get that $\mathbf{G}(\widetilde{\mathbf{H}})$ has an eigenvalue of $1/2$. But since $\mathbf{Q}^T\mathbf{G}(\mathbf{H})\mathbf{Q} = \mathbf{G}(\widetilde{\mathbf{H}})$ and $\mathbf{Q}$ is orthonormal, $\mathbf{G}(\mathbf{H})$ is similar to $\mathbf{G}(\widetilde{\mathbf{H}})$. So $\mathbf{G}(\mathbf{H})$ has an eigenvalue of $1/2$. ■

D.4.2 Derivation of SURE

By substituting $\mathbf{Z}_A(\underline{\hat{\theta}})$ in place of $\mathbf{Z}(\underline{\hat{\theta}})$ and taking the limit as $A \to \infty$, the trace expression in (D.33) can be shown to equal

$$\text{tr}\left((\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{Q})_{22}\left[(\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{Q})_{22} - \frac{1}{2}(\mathbf{P}\mathbf{U}(\underline{\hat{\theta}})\mathbf{Q})_{22}\right]^{-1}\right) \tag{D.46}$$

The steps used to obtain (D.46) are similar to those used to derive SURE for the L1 estimator, and will be omitted. We are assured that the inverse in the trace expression exists because of Prop. D.4.

Let $\mathbf{K} = (\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{Q})_{22}$ and $\mathbf{J} = -(1/2)(\mathbf{P}\mathbf{U}(\underline{\hat{\theta}})\mathbf{Q})_{22}$. Stein's URE for the risk of the HHS estimator is given by

$$\hat{R}(\underline{t}) = N\sigma^2 + \|\underline{e}\|^2 + 2\sigma^2\text{tr}(\mathbf{K}[\mathbf{K}+\mathbf{J}]^{-1}). \tag{D.47}$$

The computation of (D.47) requires the inversion of a $(M-r) \times (M-r)$ matrix. When $t_2 = t_1 \Longrightarrow \mathbf{J} = 0$, and (D.47) reduces to (D.21), SURE of the L1 estimator.

D.4.3 Approximation

Let us try to find another expression for (D.47). Define $0 \le s \le M-r$ to be the number of elements of in the diagonal of $\mathbf{J}$ that equals zero. Let $\widetilde{\mathbf{P}}, \widetilde{\mathbf{Q}} \in \mathbb{R}^{(M-r)\times(M-r)}$ be permutation matrices defined along the lines of (D.10)-(D.11) that re-arrange $\mathbf{J}$ in the following order:

$$\widetilde{\mathbf{P}}\mathbf{J}\widetilde{\mathbf{Q}} = \text{diag}\left(0, \ldots, 0, -\frac{1}{2}, \ldots, -\frac{1}{2}\right) \tag{D.48}$$

Denote $\widetilde{\mathbf{K}} \triangleq \widetilde{\mathbf{P}}\mathbf{K}\widetilde{\mathbf{Q}}$ and $\widetilde{\mathbf{J}} \triangleq \widetilde{\mathbf{P}}\mathbf{J}\widetilde{\mathbf{Q}}$. Then,

$$\text{(D.49)} \qquad \operatorname{tr}(\mathbf{K}[\mathbf{K}+\mathbf{J}]^{-1}) = \operatorname{tr}(\widetilde{\mathbf{P}}\mathbf{K}\widetilde{\mathbf{Q}}[\widetilde{\mathbf{P}}\mathbf{K}\widetilde{\mathbf{Q}}+\widetilde{\mathbf{P}}\mathbf{J}\widetilde{\mathbf{Q}}]^{-1}) = \operatorname{tr}(\widetilde{\mathbf{K}}[\widetilde{\mathbf{K}}+\widetilde{\mathbf{J}}]^{-1})$$

If $s = M - r$, then $\widetilde{\mathbf{J}} = \mathbf{0}$, and $\operatorname{tr}(\mathbf{K}[\mathbf{K}+\mathbf{J}]^{-1}) = M - r$. Suppose that $s < M - r$. By the matrix inversion lemma,

$$\text{(D.50)} \qquad (\widetilde{\mathbf{K}}+\widetilde{\mathbf{J}})^{-1} = \widetilde{\mathbf{K}}^{-1} - \widetilde{\mathbf{K}}^{-1}(\widetilde{\mathbf{J}}^{-1}+\widetilde{\mathbf{K}}^{-1})\widetilde{\mathbf{K}}^{-1}$$

$$\text{(D.51)} \qquad \Longrightarrow \widetilde{\mathbf{K}}(\widetilde{\mathbf{K}}+\widetilde{\mathbf{J}})^{-1} = \mathbf{I}_{M-r} - (\widetilde{\mathbf{J}}^{-1}+\widetilde{\mathbf{K}}^{-1})^{-1}\widetilde{\mathbf{K}}^{-1}$$

Since $\widetilde{\mathbf{J}}$ is not invertible, a limiting argument along the lines of (D.19) has to be made in order to compute $(\widetilde{\mathbf{J}}^{-1}+\widetilde{\mathbf{K}}^{-1})^{-1}$. We note that $\widetilde{\mathbf{K}}$ is invertible, as is show in the following proposition.

Proposition D.5. *The matrix* $\widetilde{\mathbf{K}}$ *is invertible.*

Since $\widetilde{\mathbf{P}} = \widetilde{\mathbf{Q}}^T$ and $\widetilde{\mathbf{Q}}$ is orthogonal,

$$\text{(D.52)} \qquad \det\widetilde{\mathbf{K}} \neq 0 \iff \det\mathbf{K} \neq 0 \iff \det((\mathbf{P}\mathbf{H}^T\mathbf{H}\mathbf{Q})_{22}) \neq 0.$$

Since the columns of $\mathbf{H}$ are linearly independent, we can apply Prop. D.2. ■

The end result is that

$$\text{(D.53)} \qquad (\widetilde{\mathbf{J}}^{-1}+\widetilde{\mathbf{K}}^{-1})^{-1} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\widetilde{\mathbf{K}}^{-1}_{2'2'} - 2\mathbf{I})^{-1} \end{pmatrix}$$

where the subscript $(\cdot)_{2'2'}$ denotes the lower right $(M-r-s)\times(M-r-s)$ submatrix of the argument. NB. $\widetilde{\mathbf{K}}^{-1}_{2'2'}$ is the lower right $(M-r-s)\times(M-r-s)$ submatrix of $\widetilde{\mathbf{K}}^{-1}$. Substituting (D.53) into (D.51) and evaluating the trace produces

$$\text{(D.54)} \qquad \operatorname{tr}(\widetilde{\mathbf{K}}[\widetilde{\mathbf{K}}+\widetilde{\mathbf{J}}]^{-1}) = s - \operatorname{tr}\left(\mathbf{I}_{M-r-s} - (\widetilde{\mathbf{K}}^{-1}_{2'2'} - 2\mathbf{I}_{M-r-s})^{-1}\widetilde{\mathbf{K}}^{-1}_{2'2'}\right)$$

The previous result works for $s = M - r$ if the convention that $\mathbf{I}_0 = \mathbf{0}$ is used. Recognize that $s = \#\{i : \hat{\theta}_i \neq 0 \wedge |\hat{\theta}_i| > t_1 - t_2\}$. So $s = \|\mathcal{D}_{T_\mathrm{s}}(\underline{\hat{\theta}}; t_1 - t_2)\|_0$.

Recall that $\mathcal{D}_{T_\mathrm{s}}(x;t) = \sum_i T_\mathrm{s}(x_i;t)\underline{e}_i$, where $T_s(\cdot;t)$ is the soft-thresholding rule with threshold $t > 0$. Applying this to (D.54), SURE for the HHS estimator can be written as

(D.55)

$$\hat{R}(\underline{t}) = N\sigma^2+\|\underline{e}\|^2+2\sigma^2\|\mathcal{D}_{T_\mathrm{s}}(\underline{\hat{\theta}};t_1-t_2)\|_0-2\sigma^2\mathrm{tr}\left(\mathbf{I}_{M-r-s}-(\widetilde{\mathbf{K}}^{-1}_{2'2'}-2\mathbf{I}_{M-r-s})^{-1}\widetilde{\mathbf{K}}^{-1}_{2'2'}\right).$$

The computation of (D.55) requires the inversion of a $(M-r)\times(M-r)$ matrix, just like (D.47). Thus, it does not seem like anything has been gained.

Consider the approximation to $\hat{R}(\underline{t})$ made by dropping the last term of (D.55), so that

$$\hat{R}(\underline{t}) \approx N\sigma^2 + \|\underline{e}\|^2 + 2\sigma^2\|\mathcal{D}_{T_\mathrm{s}}(\underline{\hat{\theta}};t_1-t_2)\|_0, \tag{D.56}$$

This is easier to compute. When $t_1 = t_2$, (D.56) is equal to SURE of the L1 estimator, i.e., (D.21). But when $t_1 = t_2$, the HHS estimator is the L1 estimator. This shows that, when $t_1 = t_2$, the last term in (D.55) equals 0. For values of (t_1, t_2) such that $t_1 \approx t_2$, we expect the approximation (D.56) to be relatively accurate.

APPENDIX E

Pseudocode for the image reconstruction methods

E.1 Standard and Projected Landweber iteration

The pseudocode for the Landweber iteration is given below.

```
Require: $\mathbf{H}, \underline{y}, \hat{\underline{\theta}}^{(0)}, \epsilon > 0, \tau \in \left(0, \frac{2}{\rho(\mathbf{H}^T\mathbf{H})}\right)$
$\underline{\theta}_{\text{prev}} := \hat{\underline{\theta}}^{(0)}$
repeat
    $\underline{\theta}_{\text{next}} := \underline{\theta}_{\text{prev}} + \tau\mathbf{H}^T(\underline{y} - \mathbf{H}\underline{\theta}_{\text{prev}})$
    $d := \|\underline{\theta}_{\text{next}} - \underline{\theta}_{\text{prev}}\|$
    $\underline{\theta}_{\text{prev}} := \underline{\theta}_{\text{next}}$ { Update the "prev" variable }
until $d < \epsilon$
$\hat{\underline{\theta}} := \underline{\theta}_{\text{next}}$
return $\hat{\underline{\theta}}$
```

The projected Landweber iteration incorporates a projection to the positive orthant after each normal Landweber step. The positive orthant is a closed, convex set. The pseudocode for the projected Landweber iteration is as follows.

```
Require: $\mathbf{H}, \underline{y}, \hat{\underline{\theta}}^{(0)}, \epsilon > 0, \tau \in \left(0, \frac{2}{\rho(\mathbf{H}^T\mathbf{H})}\right)$
$\underline{\theta}_{\text{prev}} := \hat{\underline{\theta}}^{(0)}$
repeat
```

```
3:      $\underline{\theta}_{\text{tmp}} := \underline{\theta}_{\text{prev}} + \tau \mathbf{H}^T(\underline{y} - \mathbf{H}\underline{\theta}_{\text{prev}})$
4:      for  $i = 1$ to $M$  do { Project on to positive orthant }
5:          $\theta_{\text{next},i} := \max(0, \theta_{\text{tmp},i})$
6:      end for
7:      $d := \|\underline{\theta}_{\text{next}} - \underline{\theta}_{\text{prev}}\|$
8:      $\underline{\theta}_{\text{prev}} := \underline{\theta}_{\text{next}}$ { Update the "prev" variable }
9:  until  $d < \epsilon$
10: $\hat{\underline{\theta}} := \underline{\theta}_{\text{next}}$
11: return  $\hat{\underline{\theta}}$
```

The only difference between the standard and projected Landweber iteration is the additional lines 4–6 in the latter.

E.2 EBD

Recall that EBD is a sparse denoising method; without loss of generality, take $\mathbf{H} = \mathbf{I}$, and so $M = N$. The hyperparameter for the LAZE prior is $\underline{\phi} = (a, w)$. The marginalized p.d.f. $p(\underline{y}|\underline{\phi})$ is

$$
\begin{aligned}
p(\underline{y}|\underline{\phi}) &= \int p(\underline{y}|\underline{\theta}) p(\underline{\theta}|\underline{\phi}) \, d\underline{\theta} \\
&= \int (2\pi)^{-N/2} \sigma^{-N} e^{-\|\underline{y}-\underline{\theta}\|^2/2\sigma^2} \prod_{i=1}^{N} \left[(1-w)\delta(\theta_i) + w\frac{1}{2} a e^{-a|\theta_i|} \right] d\underline{\theta} \\
&= \prod_{i=1}^{N} \int \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i-\theta_i)^2/2\sigma^2} \left[(1-w)\delta(\theta_i) + w\frac{1}{2} a e^{-a|\theta_i|} \right] d\theta_i. \qquad \text{(E.1)}
\end{aligned}
$$

The marginalized density of $\underline{y}$ is equal to the product of the marginalized densities of each y_i. This is to be expected, as both the noise and prior p.d.f.s are i.i.d. Now, the integral of the first product is just $(1-w) \cdot \mathcal{N}(y_i; 0, \sigma^2)$. It remains to compute the integral of the second product, which is w times the convolution of a Gaussian and Laplacian p.d.f. The convolution can be shown to equal

$$
\begin{aligned}
g(x; a, \sigma) &\triangleq \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2} * \frac{1}{2} a e^{-a|x|} \\
&= \frac{1}{4} a e^{(a\sigma)^2/2} \left\{ e^{ax} + e^{-ax} + e^{-ax} \mathrm{erf}\left(\frac{x - a\sigma^2}{\sqrt{2}\sigma} \right) - e^{ax} \mathrm{erf}\left(\frac{x + a\sigma^2}{\sqrt{2}\sigma} \right) \right\} \qquad \text{(E.2)}
\end{aligned}
$$

where $\mathrm{erf}(x) \triangleq \frac{2}{\sqrt{\pi}} \int_0^x e^{-s^2} ds$. The marginalized density of each observation $p(y|\underline{\phi})$ is given by

$$
p_{Y|\underline{\Phi}}(y|\underline{\phi}) = (1 - w)\mathcal{N}(y; 0, \sigma^2) + wg(y; a, \sigma) \qquad \text{(E.3)}
$$

and so $p(\underline{y}|\underline{\phi}) = \prod_{i=1}^{N} p_{Y|\underline{\Phi}}(y_i|\underline{\phi})$.

Given an observation y, let us derive the posterior median u. Using the notation that $F(\theta|y) = \int_{-\infty}^{\theta} p(\theta|y)\, d\theta$ is the posterior c.d.f., u is the solution to the equation $F(\theta|y) = 1/2$. This can be solved in closed form to give $u = T_{\text{laze,pmed}}(y; \underline{\phi}, \sigma)$. The details are as follows. The posterior median is a thresholding function [33] with threshold $t \geq 0$ that is the solution to

$$
1 - \mathrm{erf}\left(\frac{-t + a\sigma^2}{\sqrt{2}\sigma} \right) = 2p_{Y|\underline{\Phi}}(-t|\underline{\phi}) \left[aw \cdot \exp\left(-at + \frac{1}{2}(a\sigma)^2 \right) \right]^{-1} \qquad \text{(E.4)}
$$

Then,

$$
T_{\text{laze,pmed}}(y; \underline{\phi}, \sigma) = \begin{cases} y + a\sigma^2 - \sqrt{2}\sigma \mathrm{erf}^{-1}\left(1 - \frac{2p_{Y|\underline{\Phi}}(y;\phi)}{aw \cdot \exp(ay + \frac{1}{2}(a\sigma)^2)} \right) & y < -t \\ 0 & |y| \leq t \\ y - a\sigma^2 + \sqrt{2}\sigma \mathrm{erf}^{-1}\left(1 - \frac{2p_{Y|\underline{\Phi}}(-y;\phi)}{aw \cdot \exp(-ay + \frac{1}{2}(a\sigma)^2)} \right) & y > t \end{cases} \qquad \text{(E.5)}
$$

The pseudocode for EBD is as follows.

```
Require: $\underline{y}, \sigma$ { Recall that $\mathbf{H} = \mathbf{I}$ }
$\hat{\underline{\phi}} := \mathrm{argmax}_{\underline{\phi}} \log p(\underline{y}|\underline{\phi})$ { Recall that $\underline{\phi} = (a, w)$ }
for $i = 1$ to $N$ do
    $\hat{\theta}_i := T_{\text{laze,pmed}}(y_i; \hat{\underline{\phi}}, \sigma)$
```

```
4: end for
5: return  $\underline{\hat{\theta}}$
```

In the pseudocode above, the noise variance σ^2 is assumed to be known. However, it can also be estimated along with $\underline{\phi}$ in the MML framework.

E.3 SBL

SBL relies on MML to estimate its hyperparameters. As was noted in section 5.5.2, $\underline{\phi} = (\zeta_1, \ldots, \zeta_M)$, where ζ_i is the prior variance of θ_i, i.e., $\theta_i \sim \mathcal{N}(0, \zeta_i)$. Here, we assume that σ^2 is given; however, its estimation can also be incorporated in the estimation of $\underline{\phi}$. See [76] for more details. Fortuitously, the marginalized likelihood can be computed in closed-form. Let $\mathbf{Z} \triangleq \operatorname{diag}(\zeta_1, \ldots, \zeta_M) \in \mathbb{R}^{M \times M}$. Then

$$p(\underline{y}|\underline{\phi}) = (2\pi)^{-N/2}(\det \mathbf{\Sigma}_y)^{-1/2} \exp\left(-\frac{1}{2}\underline{y}^T \mathbf{\Sigma}_y^{-1} \underline{y}\right) \tag{E.6}$$

$$\text{where:} \quad \mathbf{\Sigma}_y \triangleq \sigma^2 \mathbf{I} + \mathbf{H}\mathbf{Z}\mathbf{H}^T \tag{E.7}$$

This leads to

$$\underline{\hat{\phi}} = \underset{\underline{\phi}}{\operatorname{argmax}} \log p(\underline{y}|\underline{\phi}) = \underset{\underline{\phi}}{\operatorname{argmin}} \left[\log(\det \mathbf{\Sigma}_y) + \underline{y}^T \mathbf{\Sigma}_y^{-1} \underline{y}\right] \tag{E.8}$$

Before addressing the estimation of $\underline{\hat{\phi}}$, the posterior p.d.f. $p(\underline{\theta}|\underline{y}, \underline{\phi})$ will be mentioned.

$$\begin{aligned} & p(\underline{\theta}|\underline{y}, \underline{\phi}) \sim \mathcal{N}(\underline{\mu}, \mathbf{\Sigma}_\theta) \\ & \text{where:} \quad \underline{\mu} \triangleq \sigma^{-2} \mathbf{\Sigma}_\theta \mathbf{H}^T \underline{y} \text{ and } \mathbf{\Sigma}_\theta \triangleq (\sigma^{-2}\mathbf{H}^T\mathbf{H} + \mathbf{Z}^{-1})^{-1} \end{aligned} \tag{E.9}$$

Note that a $M \times M$ matrix inversion required to compute $\mathbf{\Sigma}_\theta$. If $M > N$, which occurs when the columns of $\mathbf{H}$ are an overcomplete basis, the complexity of the inversion can be reduced by applying the matrix inversion lemma to get [76, (17)]

$$\mathbf{\Sigma}_\theta = \mathbf{Z} - \mathbf{Z}\mathbf{H}^T \mathbf{\Sigma}_y^{-1} \mathbf{H}\mathbf{Z} \tag{E.10}$$

A matrix inversion is still required, i.e., $\mathbf{\Sigma}_y^{-1}$. However, $\mathbf{\Sigma}_y$ is a $N \times N$ matrix, and so its inversion is less computationally complex.

The EM algorithm can be applied to find $\hat{\underline{\phi}}$ using $\underline{\theta}$ as the complete data. In the E-step, the Q function can be computed as follows

$$\begin{aligned} Q(\underline{\phi}; \underline{\phi}^{(n)}) &= E_{\underline{\Theta}|\underline{Y},\underline{\Phi}}[\log p(\underline{\theta}|\underline{\phi})|\underline{y}, \underline{\phi}^{(n)}] \\ &= -\frac{M}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{M}\left(\log \zeta_i + \frac{E[\theta_i^2|\underline{y}, \underline{\phi}^{(n)}]}{\zeta_i}\right) \end{aligned} \tag{E.11}$$

$$\text{where:}\quad E_{\underline{\Theta}|\underline{Y},\underline{\Phi}}[\theta_i^2|\underline{y}, \underline{\phi}^{(n)}] = (\mathbf{\Sigma}_\theta)_{i,i}(\underline{\phi}^{(n)}) + \mu_i^2(\underline{\phi}^{(n)}) \tag{E.12}$$

The maximization of $\underline{\zeta} = (\zeta_1, \ldots, \zeta_M)$ in the M-step can be done on a coordinate-wise basis, as we notice that the Q function in (E.11) can be written as a sum of identical functions in ζ_i. So

$$\zeta_i^{(n+1)} = \underset{\zeta_i \geq 0}{\operatorname{argmax}} = E[\theta_i^2|\underline{y}, \underline{\phi}^{(n)}] \tag{E.13}$$

Once the hyperparameter estimate $\hat{\underline{\phi}}$ is obtained, the posterior mean $\underline{\mu}$ given in (E.9) is used as $\hat{\underline{\theta}}$ with $\underline{\phi} = \hat{\underline{\phi}}_{\text{final}}$. In the pseudocode given below, $\underline{\zeta}$ will be used instead of $\underline{\phi}$ for the sake of clarity. In any case, they are interchangeable, as $\underline{\phi} = (\zeta_1, \ldots, \zeta_M) = \underline{\zeta}$. We shall employ (E.10) in the pseudocode.

```
Require: $\mathbf{H}, \underline{\zeta}^{(0)}, \sigma, \epsilon > 0$
$\underline{\zeta}_{\text{prev}} := \underline{\zeta}^{(0)}$
repeat { Compute estimate of hyperparameter }
    $\mathbf{Z} := \text{diag}(\underline{\zeta}_{\text{prev}})$
    $\mathbf{\Sigma}_y := \sigma^2\mathbf{I} + \mathbf{H}\mathbf{Z}\mathbf{H}^T$ { Compute covariance of marginalized distribution }
    $\mathbf{\Sigma}_\theta := \mathbf{Z} - \mathbf{Z}\mathbf{H}^T\mathbf{\Sigma}_y^{-1}\mathbf{H}\mathbf{Z}$ { Compute covariance of posterior distribution }
    $\underline{\theta}_{\text{prev}} := \sigma^{-2}\mathbf{\Sigma}_\theta\mathbf{H}^T\underline{y}$
    for $i = 1$ to $M$ do
```

```
        $\zeta_{\text{next},i} := (\mathbf{\Sigma}_w)_{i,i} + \mu^2_{\text{prev},i}$ { Compute $\underline{\zeta}_{\text{next}}$ }
    **end for**
    $d := \|\underline{\zeta}_{\text{next}} - \underline{\zeta}_{\text{prev}}\|$
    $\underline{\zeta}_{\text{prev}} := \underline{\zeta}_{\text{next}}$ { Update the "prev" variable }
**until** $d < \epsilon$
$\mathbf{Z} := \text{diag}(\underline{\zeta}_{\text{next}})$ { Compute $\hat{\underline{\theta}}$ using $\underline{\zeta}_{\text{next}}$ }
$\mathbf{\Sigma}_y := \sigma^2\mathbf{I} + \mathbf{H}\mathbf{Z}\mathbf{H}^T$
$\mathbf{\Sigma}_\theta := \mathbf{Z} - \mathbf{Z}\mathbf{H}^T\mathbf{\Sigma}_y^{-1}\mathbf{H}\mathbf{Z}$
$\hat{\underline{\theta}} := \sigma^{-2}\mathbf{\Sigma}_\theta\mathbf{H}^T\underline{y}$
**return** $\hat{\underline{\theta}}$
```

E.4 EBD-LAZE

The algorithm for EBD-LAZE makes use of the EBD denoising method. Denote the latter by $\hat{\underline{\theta}} := \text{EBD}(\underline{y}, \sigma)$.

```
**Require:** $\mathbf{H}, \underline{y}, \hat{\underline{\theta}}^{(0)}, \sigma, c \in (0,1), \epsilon > 0$
$\alpha := c \cdot \sigma\sqrt{\rho(\mathbf{H}\mathbf{H}^T)^{-1}}$
$\underline{\theta}_{\text{prev}} := \hat{\underline{\theta}}^{(0)}$
**repeat**
    $\underline{z} := \underline{\theta}_{\text{prev}} + \left(\frac{\alpha}{\sigma}\right)^2\mathbf{H}^T(\underline{y} - \mathbf{H}\underline{\theta}_{\text{prev}})$
    $\underline{\theta}_{\text{next}} := \text{EBD}(\underline{z}, \alpha)$
    $d := \|\underline{\theta}_{\text{next}} - \underline{\theta}_{\text{prev}}\|$
    $\underline{\theta}_{\text{prev}} := \underline{\theta}_{\text{next}}$ { Update the "prev" variable }
**until** $d < \epsilon$
$\hat{\underline{\theta}} := \underline{\theta}_{\text{next}}$
**return** $\hat{\underline{\theta}}$
```

The pseudocode above performs a search for the hyperparameter $\underline{\phi} = (a, w)$ in each iteration of the repeat–until loop. The computational complexity can be decreased by modifying the EBD procedure so that the search is performed only every nth iteration. In between searches, the most recent $\hat{\underline{\phi}}$ is used.

E.5 MAP1 and MAP2

Only the pseudocode for MAP1 will be given here, as the MAP2 reconstruction methods is similar. They have the same input parameters with the exception of an extra tuning parameter that must be supplied for MAP2.

```
**Require:** $\mathbf{H}, \hat{\tilde{\underline{\theta}}}^{(0)}, \hat{\underline{I}}^{(0)}, \hat{\underline{\phi}}^{(0)}, \sigma, c \in (0,1)$
**Require:** $\epsilon_1 > 0$ { $\epsilon_1$ dictates when convergence is achieved for the optimization in step (ii) }
**Require:** $\epsilon_2 > 0$ { $\epsilon_2$ dictates when convergence is achieved for the overall method }
$\alpha := c \cdot \sigma \sqrt{\rho(\mathbf{H}\mathbf{H}^T)^{-1}}$
$\tilde{\underline{\theta}}_{\text{prev}} := \hat{\tilde{\underline{\theta}}}^{(0)}$
$\underline{I}_{\text{prev}} := \hat{\underline{I}}^{(0)}$
$\underline{\phi}_{\text{prev}} := \hat{\underline{\phi}}^{(0)}$ { Recall that $\underline{\phi} = (a, w)$ }
**repeat**
  $\tilde{\underline{\xi}}_{\text{prev}} := \tilde{\underline{\theta}}_{\text{prev}}$ { $\tilde{\underline{\xi}}$ is a temporary variable corresponding to $\tilde{\underline{\theta}}$ to be used in the inner loop }
  $\underline{J}_{\text{prev}} := \underline{I}_{\text{prev}}$ { Similarly, $\underline{J}$ is a temporary variable corresponding to $\underline{I}$ }
  $(a, w) := \underline{\phi}_{\text{prev}}$
  **repeat** { In this loop, perform step (ii), i.e., hold $\underline{\phi}$ fixed and find the maximizing $\tilde{\underline{\theta}}$ and $\underline{I}$ }
  $\underline{z} := \underline{\xi}_{\text{prev}} + \left(\frac{\alpha}{\sigma}\right)^2 \mathbf{H}^T(\underline{y} - \mathbf{H}\underline{\xi}_{\text{prev}})$ { Using $\xi_{\text{prev},i} = \tilde{\xi}_{\text{prev},i} J_{\text{prev},i}$ }
  if $0 < w \leq 1/2$ then { Set $\underline{J}_{\text{next}}$ according to (5.62) }
    for $i = 1$ to $M$ do
      $J_{\text{next},i} := I\left(|z_i| > a\alpha^2 + \sqrt{2\alpha^2 \log(\frac{1-w}{w})}\right)$
    end for
  else { $1/2 < w \leq 1$ }
    for $i = 1$ to $M$ do
      $J_{\text{next},i} := 1$
    end for
  end if
  for $i = 1$ to $M$ do { Set $\tilde{\underline{\xi}}_{\text{next}}$ according to (5.61) }
    if $J_{\text{next},i} = 0$ then
      $\tilde{\xi}_{\text{next},i} := 0$
    else { $J_{\text{next},i} = 1$ }
      $\tilde{\xi}_{\text{next},i} := T_s(z_i; a\alpha^2)$
    end if
  end for
  $d_1 := \|\underline{\xi}_{\text{next}} - \underline{\xi}_{\text{prev}}\|$ { $\underline{\xi}_{\text{next}}$ is obtained from $\tilde{\underline{\xi}}_{\text{next}}$ and $\underline{J}_{\text{next}}$; same goes for $\underline{\xi}_{\text{prev}}$ }
  $\tilde{\underline{\xi}}_{\text{prev}} := \tilde{\underline{\xi}}_{\text{next}}$ { Update the "prev" variables }
  $\underline{J}_{\text{prev}} := \underline{J}_{\text{next}}$
until $d_1 < \epsilon_1$ { Check for convergence of $\underline{\xi}$ sequence }
$\tilde{\underline{\theta}}_{\text{next}} := \tilde{\underline{\xi}}_{\text{next}}$
$\underline{I}_{\text{next}} := \underline{J}_{\text{next}}$
$a_{\text{new}} := M/\|\tilde{\underline{\theta}}_{\text{next}}\|_1$ { Do step (i) here }
```

```
    $w_{\text{new}} := \|\underline{I}_{\text{next}}\|_0 / M$
    $\underline{\phi}_{\text{next}} := (a_{\text{new}}, w_{\text{new}})$
    $d_2 := \|\underline{\theta}_{\text{next}} - \underline{\theta}_{\text{prev}}\|$
    $\underline{\tilde{\theta}}_{\text{prev}} := \underline{\tilde{\theta}}_{\text{next}}$ { Update the "prev" variables }
    $\underline{I}_{\text{prev}} := \underline{I}_{\text{next}}$
    $\underline{\phi}_{\text{prev}} := \underline{\phi}_{\text{next}}$
until $d_2 < \epsilon_2$ { Check for convergence of $\underline{\theta}$ sequence }
for $i = 1$ to $M$ do { Form the estimate $\underline{\hat{\theta}}$ }
    $\hat{\theta}_i := \tilde{\theta}_{\text{next},i} \cdot I_{\text{next},i}$
end for
return $\underline{\hat{\theta}}$
```

E.6 L1-SURE and HHS-SURE

The pseudocode for L1-SURE is given here without explicitly specifying the method of solving for the L1 estimator. Assume that we are given a list of different βs to evaluate. Recall that β is the regularization parameter of the L1 estimator. The L1 estimator can be solved via the iterative thresholding framework for a general **H** or via the LARS algorithm if the columns of **H** are linearly independent. Note that by using the SURE expression (5.49) in the pseudocode below, we are already assuming that the columns of **H** are linearly independent.

```
Require: $\mathbf{H}, \underline{y}, \sigma$
Require: $\underline{\beta}$ { List of different $\beta$ values to consider; assume there are $p$ of them }
$\underline{\hat{\theta}}_{\text{best}} := \underline{0}$
$C_{\text{best}} := \infty$
$N := \text{length}(\underline{y})$
```

```
for $i = 1$ to $p$ do
    $\underline{\hat{\theta}} := \text{argmin}_{\underline{\theta}} \left( \|\mathbf{H}\underline{\theta} - \underline{y}\|^2 + \beta_i \|\underline{\theta}\|_1 \right)$
    $C := N\sigma^2 + \|\mathbf{H}\underline{\hat{\theta}} - \underline{y}\|^2 + 2\sigma^2 \|\underline{\hat{\theta}}\|_0$
    if $C < C_{\text{best}}$ then { We have a better candidate }
        $\underline{\hat{\theta}}_{\text{best}} := \underline{\hat{\theta}}$
        $C_{\text{best}} := C$
    end if
end for
return $\underline{\hat{\theta}}_{\text{best}}$
```

The pseudocode for HHS-SURE is similar and will be omitted.

APPENDIX F

Miscellaneous details on sparse image reconstruction simulations

F.1 Specification of Gaussian blur psf

The Matlab code used to generate the Gaussian blur psf used in the simulations is given below.

```
propPsf = struct( ’fwhm’, 3, ’nk_half’, 5 );

psf = gaussian_kernel( propPsf.fwhm, propPsf.nk_half );

psf = psf * psf’;
```

The Matlab code for `gaussian_kernel.m` is included here for the sake of completeness.

```
 function kern = gaussian_kernel(fwhm, nk_half)

%function kern = gaussian_kernel(fwhm, nk_half)

% samples of a gaussian kernel at [-nk_half:nk_half]

% with given FWHM in pixels

% uses integral over each sample bin so that sum is

% very close to unity

% Copyright 2001-9-18, Jeff Fessler, The University of Michigan

if nargin < 1, help(mfilename), return, end
if nargin < 2, nk_half = 2 * ceil(fwhm); end

if fwhm == 0

kern = zeros(nk_half*2+1, 1);

kern(nk_half+1) = 1;

else

sig = fwhm / sqrt(log(256));

x = [-nk_half:nk_half]’;

kern = normcdf(x+1/2, 0, sig) - normcdf(x-1/2, 0, sig);

end
```

F.2 Noise values used in the simulations

Table F.1: Table of noise standard deviation used in the simulations.

SNR	σ	**SNR**	σ	**SNR**	σ
Binary-valued image					
1.76 dB	1.99×10^{-2}	3.01 dB	1.72×10^{-2}	3.98 dB	1.54×10^{-2}
4.77 dB	1.41×10^{-2}	7 dB	1.09×10^{-2}	10 dB	7.70×10^{-3}
13 dB	5.45×10^{-3}	16 dB	3.86×10^{-3}	20 dB	2.43×10^{-3}
LAZE image					
1.76 dB	2.90×10^{-2}	3.01 dB	2.51×10^{-2}	3.98 dB	2.25×10^{-2}
4.77 dB	2.05×10^{-2}	7 dB	1.59×10^{-2}	10 dB	1.12×10^{-2}
13 dB	7.96×10^{-3}	16 dB	5.63×10^{-3}	20 dB	3.55×10^{-3}

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] S. Alliney and S. A. Ruzinsky, *An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian estimation*, IEEE Trans. Signal Processing **42** (1994), no. 3, 618–627.

[2] M. J. Beal and Z. Ghahramani, *The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures*, Bayesian Statistics **7** (2003), 453–464.

[3] G. P. Berman, G. D. Doolen, P. C. Hammel, and V. I. Tsifrinovich, *Solid-state nuclear-spin quantum computer based on magnetic resonance force microscopy*, Physical Review B **61** (2000), no. 21, 14694–14699.

[4] G. P. Berman, D. I. Kamenev, and V. I. Tsifrinovich, *Stationary cantilever vibrations in oscillating-cantilever-driven adiabatic reversals: Magnetic-resonance-force-microscopy technique*, Physical Review A **66** (2002), no. 2, 023405/1–6.

[5] R. Budakian, H. J. Mamin, B. W. Chui, and D. Rugar, *Creating order from random fluctuations in small spin ensembles*, Science **307** (2005), no. 5708, 408–411.

[6] D. M. Chickering and D. Heckerman, *Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables*, Machine Learning **29** (1997), 181–212.

[7] K. Chun and N. O. Birge, *Dissipative quantum tunneling of a single defect in a disordered metal*, Physical Review B **54** (1996), no. 7, 4629–4637.

[8] M. A. Clyde and E. I. George, *Empirical Bayes estimation in wavelet nonparametric regression*, Springer-Verlag, New York, 1999.

[9] C. Cohen-Tannoudji, B. Diu, and F. Laloë, *Quantum Mechanics*, Wiley, New York, 1977.

[10] I. Daubechies, M. Defrise, and C. de Mol, *An Iterative Thresholding Algorithm for Linear Inverse Problems with a Sparsity Constraint*, Communications on Pure and Applied Mathematics **57** (2004), no. 11, 1413–1457.

[11] C. M. Dobson, *Chemical space and biology*, Nature **432** (2004), 824–828.

[12] D. L. Donoho, *Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition*, Applied and Computational Harmonic Analysis **2** (1995), 101–126.

[13] D. L. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization*, Proceedings of the National Academy of Sciences of the United States of America **100** (2003), no. 5, 2197–2202.

[14] D. L. Donoho and I. M. Johnstone, *Adapting to unknown smoothness via wavelet shrinkage*, Journal of the American Statistical Association **90** (1995), no. 423, 1200–1224.

[15] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, The Annals of Statistics **32** (2004), no. 2, 407–499.

[16] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models*, Springer, New York, 1995.

[17] J. A. Fessler, *Image Reconstruction: Algorithms and Analysis*, Draft of book.

[18] M. A. T. Figueiredo and R. D. Nowak, *Wavelet-Based Image Estimation: An Empirical Bayes Approach Using Jeffreys' Noninformative Prior*, IEEE Trans. Image Processing **10** (2001), no. 9, 1322–1331.

[19] ———, *An EM Algorithm for Wavelet-Based Image Restoration*, IEEE Trans. Image Processing **12** (2003), no. 8, 906–916.

[20] P. D. Gader, M. Mystkowski, and Y. Zhao, *Landmine detection with ground penetrating radar using Hidden Markov Models*, IEEE Trans. Geosci. Remote Sensing **39** (2001), no. 6, 1231–1244.

[21] M. Girolami, *A variational method for learning sparse and overcomplete representations*, Neural Computation **13** (2001), 2517–2532.

[22] R. Gribonval and M. Nielsen, *Sparse representations in unions of bases*, IEEE Trans. Inform. Theory **49** (2003), no. 12, 3320–3325.

[23] J. Gunther, R. Beard, J. Wilson, T. Oliphant, and W. Stirling, *Fast Nonlinear Filtering via Galerkin's Method*, Proceedings of the 1997 American Control Conference, vol. 5, 1997, pp. 2815–2819.

[24] P. C. Hammel, D. V. Pelekhov, P. E. Wigen, T. R. Gosnell, M. M. Midzor, and M. L. Roukes, *The Magnetic-Resonance Force Microscope: A New Tool for High-Resolution, 3-D, Subsurface Scanned Probe Imaging*, Proc. IEEE **91** (2003), no. 5, 789–798.

[25] A. O. Hero, *Statistical methods for signal processing*, Course notes for EECS 564: Estimation, filtering and detection.

[26] K. K. Herrity, A. C. Gilbert, and J. A. Tropp, *Sparse approximation via iterative thresholding*, Proceedings of the IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, 2006, to appear.

[27] A. Hyvärinen, *Sparse Code Shrinkage: Denoising of Nongaussian Data by Maximum Likelihood Estimation*, Neural Computation **11** (1999), no. 7, 1739–1768.

[28] A. Isidori, *Nonlinear Control Systems: An Introduction*, Springer-Verlag, Berlin, 1985.

[29] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970.

[30] N. L. Johnson and S. Kotz, *Continuous univariate distributions–1*, John Wiley, New York, 1970.

[31] ———, *Continuous univariate distributions–2*, John Wiley, New York, 1970.

[32] I. M. Johnstone and B. W. Silverman, *Empirical Bayes approaches to mixture problems and wavelet regression*, Tech. report, Stanford University, 1998.

[33] ———, *Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences*, The Annals of Statistics **32** (2004), no. 4, 1594–1649.

[34] T. Kailath, *A general likelihood-ratio formula for random signals in Gaussian noise*, IEEE Trans. Inform. Theory **IT-15** (1969), no. 3, 350–361.

[35] T. Kailath and M. Zakai, *Absolute continuity and Radon-Nikodym derivatives for certain measures relative to Wiener measure*, The Annals of Mathematical Statistics **42** (1971), no. 1, 130–140.

[36] N. J. Kasdin, *Modeling and Runge-Kutta Numerical Simulations of Nonlinear Stochastic Differential Equations*, Submitted for publication.

[37] ———, *Discrete Simulation of Colored Noise and Stochastic Processes and $1/f^\alpha$ Power Law Noise Generation*, Proc. IEEE **83** (1995), no. 5, 802–827.

[38] ———, *Runge-Kutta Algorithm for the Numerical Integration of Stochastic Differential Equations*, Journal of Guidance, Control, and Dynamics **18** (1995), no. 1, 114–120.

[39] J. Kay, *The EM algorithm in medical imaging*, Statistical Methods in Medical Research **6** (1997), 55–75.

[40] E. Kreyszig, *Advanced Engineering Mathematics*, seventh ed., Wiley, New York, 1993.

[41] K. Lange, D. R. Hunter, and I. Yang, *Optimization transfer using surrogate objective functions*, Journal of Computational and Graphical Statistics **9** (2000), no. 1, 1–20.

[42] S. V. Lototsky and B. L. Rozovskii, *Recursive Nonlinear Filter for a Continuous-Discrete Time Model: Separation of Parameters and Observations*, IEEE Trans. Automat. Contr. **43** (1998), no. 8, 1154–1158.

[43] H. J. Mamin, R. Budakian, B. W. Chui, and D. Rugar, *Detection and manipulation of statistical polarization in small spin ensembles*, Physical Review Letters **91** (2003), no. 20, 207604/1–4.

[44] H. J. Mamin, R. Budakian, and D. Rugar, *Point Response Function of an MRFM Tip*, Tech. report, IBM Almaden, 2003.

[45] P. Maybeck, *Stochastic models, estimation and control*, vol. 1, Academic Press, Orlando, Florida, 1979.

[46] K. S. Miller and D. M. Leskiw, *An Introduction to Kalman Filtering with Applications*, Krieger, Maralabar, Florida, 1987.

[47] S. K. Mitra, *Digital Signal Processing: A Computer-Based Approach*, second ed., McGraw-Hill, New York, 2001.

[48] R. Molina and A. K. Katsaggelos, *On the hierarchical Bayesian approach to image restoration and the iterative evaluation of the regularization parameter*, Visual communications and image processing, Proceedings of the SPIE, vol. 2308, 1994, pp. 244–251.

[49] R. Molina, A. K. Katsaggelos, and J. Mateos, *Bayesian and regularization methods for hyperparameter estimation in image restoration*, IEEE Trans. Image Processing **8** (1999), no. 2, 231–246.

[50] L. Ng and V. Solo, *Optical flow estimation using adaptive wavelet zeroing*, Proceedings of the IEEE Intl. Conf. on Image Processing, vol. 3, 1999, pp. 722–726.

[51] ———, *Selecting the neighbourhood size, shape, weights and model order in optical flow estimation*, Proceedings of the IEEE Intl. Conf. on Image Processing, vol. 3, 2000, pp. 600–603.

[52] D. K. Ross, *Eigenvalues of a tri-diagonal matrix: problem 80-4*, SIAM Review **23** (1981), no. 1, 112–113.

[53] D. Rugar and R. Budakian, *Classical dynamics of a spin interacting with a MRFM cantilever*, Tech. report, IBM Almaden, 2002.

[54] D. Rugar, R. Budakian, H. J. Mamin, and B. W. Chui, *Single spin detection by magnetic resonance force microscopy*, Nature **430** (2004), no. 6997, 329–332.

[55] D. Rugar, C. S. Yannoni, and J. A. Sidles, *Mechanical detection of magnetic resonance*, Nature **360** (1992), no. 6404, 563–565.

[56] D. Rugar, O. Züger, S. Hoen, C. S. Yannoni, H. M. Vieth, and R. D. Kendrick, *Force detection of nuclear magnetic resonance*, Science **264** (1994), no. 5165, 1560–1563.

[57] L. L. Scharf and L. W. Nolte, *Likelihood ratios for sequential hypothesis testing on Markov sequences*, IEEE Trans. Inform. Theory **IT-23** (1977), no. 1, 101–109.

[58] S. C. Schwartz, *The estimator-correlator for discrete-time problems*, IEEE Trans. Inform. Theory **IT-23** (1977), no. 1, 93–100.

[59] J. A. Sidles, *Nondestructive detection of single-proton magnetic resonance*, Applied Physics Letters **58** (1991), no. 24, 2854–2856.

[60] J. A. Sidles, J. L. Garbini, K. J. Bruland, D. Rugar, O. Züger, S. Hoen, and C. S. Yannoni, *Magnetic resonance force microscopy*, Review of Modern Physics **67** (1995), no. 1, 249–265.

[61] J. A. Sidles, J. L. Garbini, and G. P. Drobny, *The theory of oscillator-coupled magnetic resonance with potential applications to molecular imaging*, Review of Scientific Instruments **63** (1992), no. 8, 3881–3899.

[62] V. Solo, *A sure-fired way to choose smoothing parameters in ill-conditioned inverse problems*, Proceedings of the IEEE Intl. Conf. on Image Processing, vol. 3, 1996, pp. 89–92.

[63] H. Stark and J. W. Woods, *Probability, random processes, and estimation theory for engineers*, second ed., Prentice-Hall, New Jersey, 1994.

[64] C. M. Stein, *Estimation of the mean of a multivariate normal distribution*, The Annals of Statistics **9** (1981), no. 6, 1135–1151.

[65] B. C. Stipe, H. J. Mamin, C. S. Yannoni, T. D. Stowe, T. W. Kenny, and D. Rugar, *Electron spin relaxation near a micron-size ferromagnet*, Physical Review Letters **87** (2001), no. 27, 277602/1–4.

[66] T. D. Stowe, K. Yasumura, T. W. Kenny, D. Botkin, K. Wago, and D. Rugar, *Attonewton force detection using ultrathin silicon cantilevers*, Applied Physics Letters **71** (1997), no. 2, 288–290.

[67] A. M. Thompson, J. C. Brown, J. W. Kay, and D. M. Titterington, *A study of methods of choosing the smoothing parameter in image restoration by regularization*, IEEE Trans. Pattern Anal. Machine Intell. **13** (1991), no. 4, 326–339.

[68] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, Series B **58** (1996), no. 1, 267–288.

[69] J. A. Tropp, *Greed is good: algorithmic results for sparse approximation*, IEEE Trans. Inform. Theory **50** (2004), no. 10, 2231–2241.

[70] H. L. Van Trees, *Detection, estimation, and modulation theory*, vol. 1, Wiley, New York, 1968.

[71] P. Vettiger, G. Cross, M. Despont, U. Drechsler, U. Dürig, B. Gotsmann, W. Häberle, M. A. Lantz, H. E. Rothuizen, R. Stutz, and G. K. Binnig, *The "Millipede"—nanotechnology entering data storage*, IEEE Trans. Nanotechnol. **1** (2002), no. 1, 39–55.

[72] M. Vidyasagar, *Nonlinear Systems Analysis*, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[73] K. Wago, D. Botkin, C. S. Yannoni, and D. Rugar, *Force-detected electron-spin resonance: adiabatic inversion, nutation, and spin echo*, Physical Review B **57** (1998), no. 2, 1108–1114.

[74] Y.-M. Wang and N. R. Sheeley Jr., *On the fluctuating component of the sun's large-scale magnetic field*, The Astrophysical Journal **590** (2003), 1111–1120.

[75] D. P. Wipf, J. A. Palmer, and B. D. Rao, *Perspectives on sparse Bayesian learning*, Advances in Neural Information Processing Systems, vol. 16, 2004.

[76] D. P. Wipf and B. D. Rao, *Sparse Bayesian learning for basis selection*, IEEE Trans. Signal Processing **52** (2004), no. 8, 2153–2164.

[77] ———, *l_0-norm minimization for basis selection*, Advances in Neural Information Processing Systems, vol. 17, 2005.

[78] W. M. Wonham, *Some applications of stochastic differential equations to optimal nonlinear filtering*, J. SIAM Control **2** (1965), no. 3, 347–369.

[79] C. F. J. Wu, *On the convergence properties of the EM algorithm*, The Annals of Statistics **11** (1983), no. 1, 95–103.

[80] Y.-C. Yao, *Estimation of noisy telegraph processes: nonlinear filtering versus nonlinear smoothing*, IEEE Trans. Inform. Theory **IT-31** (1985), no. 3, 444–446.

[81] C.-Y. Yip, A. O. Hero, D. Rugar, and J. A. Fessler, *Baseband detection of bistatic electron spin signals in Magnetic Resonance Force Microscopy*, Asilomar Conference on Signals, Systems, and Computers, 2003, pp. 1309–1313.

[82] ———, *Baseband Detection of Bistatic Electron Spin Signals in Magnetic Resonance Force Microscopy (MRFM)*, ArXiv:Quantum Physics **0307** (2003).

[83] Z. Zhou, R. M. Leahy, and J. Qi, *Approximate maximum likelihood hyperparameter estimation for Gibbs priors*, IEEE Trans. Image Processing **6** (1997), no. 6, 844–861.

[84] O. Züger, S. T. Heon, C. S. Yannoni, and D. Rugar, *Three-dimensional imaging with a nuclear magnetic resonance force microscope*, Journal of Applied Physics **79** (1996), no. 4, 1881–1884.

[85] O. Züger and D. Rugar, *First images from a magnetic resonance force microscope*, Applied Physics Letters **63** (1993), no. 18, 2496–2498.

ABSTRACT

Signal processing for magnetic resonance force microscopy

by

Michael Y. J. Ting

Chair: Alfred O. Hero III

Magnetic resonance force microscopy (MRFM) is an emergent technology that has the potential for three-dimensional, non-destructive, and in-situ imaging of biological molecules with atomic resolution. Experiments at IBM have shown that MRFM is capable of detecting and localizing individual electron spins associated with subsurface atomic defects in silicon dioxide. In principle, detection of single nuclear spins is possible as well. MRFM detects the spins by measuring the small spin-induced forces on a micromachined cantilever.

Detection of a single electron spin was studied in additive white Gaussian noise (AWGN). Four models of the single spin-cantilever interaction were proposed. We investigated three of these models. A heuristic argument was used to formulate a detector for the continuous-time classical model. Approximate forms of the optimal likelihood ratio test (LRT) for the discrete-time (DT) random telegraph and DT random walk models were derived which hold under certain conditions. It was shown

that, under low signal to noise ratio (SNR), the LRT for a DT finite state Markov process in AWGN reduces to the matched filter statistic with the one-step minimum mean-squared error predictor used in place of the known signal values.

The next challenge for MRFM is to demonstrate the technology's applicability as an imaging modality with advantages over those already in existence. We therefore considered the problem of image reconstruction in the MRFM setting, which is reconstructing sparse images from noisy projections. The goal here is to perform sparse reconstruction with the tuning parameters selected in a data-driven fashion. The empirical Bayes framework was investigated, and several sparse image reconstruction methods were proposed that are more scalable and have lower computational complexity than sparse Bayesian learning (SBL). In a simulation study, the proposed methods demonstrate benefits over SBL, Landweber, and the projected Landweber method. Under low SNR, a MAP-based solution produced low l_1 and l_2 reconstruction error. We found that the maximum penalized likelihood estimator using a l_1 norm penalty and with its regularization parameter estimated by minimizing Stein's unbiased risk estimate produced consistently good results across a wide range of SNRs.