Proposal

Darren W. Homrighausen
Department of Statistics, Carnegie Mellon University

November 29, 2010

Abstract

Any problem where the observed data are only recorded indirectly is called an inverse problem (IP). A classic example is errors in variables where one wants to estimate regression coefficients but only observes the covariates after corruption with noise.

I propose to address three main questions in inverse problems (IPs). The first is, given a sequence of observations, how well can these observations be combined to form an estimate of an underlying function. If the function is observed directly there are many well-used techniques available. In the context of IPs the situation is less clear. I explore a new regime in IPs where the data are repeated observations of a function, but the function is blurred by different operators and observed under different noise conditions. This situation involves rethinking IPs in new ways.

The second question builds off the previous goal. Suppose an estimate of a function based off of a sequence of observations exists and a new observation is made. One commonly asked question is if there is any difference in the underlying function in the new observation that didn't exist before. This is known as the "transient detection" problem as these differences are referred to as transients in the astrophysical community where this could be, for instance, a supernova, asteroid, or gamma burst. Existing methods that perform this task rely on estimating a smoothing operator. I present a possible way of doing this estimation. Moreover, this smoothing operator is actually a nuisance parameter. I additionally introduce a hierarchical model that may allow for estimating the probability of a transient without estimating the smoothing operator.

Lastly, as in the nonparametric estimation case, there are tuning parameters in IP estimation as well. However, choosing tuning parameters in IPs is much less developed and more difficult than in nonparametric estimation. For instance, the default choice of minimizing generalized cross validation does not work in IPs as it estimates the wrong risk. In effect, unbiased risk estimation is not possible. I present a new method that introduces bias into the risk estimate in a controlled manner that allows for tuning parameter selection for a large number of estimators and distributions.

1 Introduction

The overarching theme of this document is to examine instances or implications of ill-posed problems, which we refer to generally as inverse problems (IPs). A non-informative definition of ill-posedness is that the problem is not well-posed. In work on mathematical physics, Hadamard gave three conditions for well-posedness. For an operator $K : A \to B$ between two metric spaces, suppose we wish to find the $f \in A$ such that $Kf = g$, where $g \in B$.

Definition 1. *A problem is well-posed if*

(a) K *is surjective.*

(b) $\forall f_1, f_2 \in A$, $K(f_1) = K(f_2)$ *i.f.f.* $f_1 = f_2$. *This corresponds to* K *being injective.*

(c) The operator K^{-1} *is continuous on* B.

In practice, the violation of condition (c) is what makes most problems ill-posed. For instance, suppose A and B are both subsets of $L^2(\mathbb{R})$ and K is a convolution operator with kernel k (ie: $Kf(t) = \int_R k(t-u)f(u)du =: k*f$). Define the Fourier transform operator as $\mathcal{F}$. Then under mild conditions on k, $\mathcal{F}k(\omega) = 0$ on at most a set of measure zero in $\mathcal{F}(L^2(\mathbb{R}))$. It follows by the isometry property of $\mathcal{F}$ that for $f_1, f_2 \in A$

$$||Kf_1 - Kf_2||_2 = ||\mathcal{F}k\mathcal{F}f_1 - \mathcal{F}k\mathcal{F}f_2||_2 = 0$$

only if $\mathcal{F}f_1(\omega) - \mathcal{F}f_2(\omega) = 0$ almost everywhere. Therefore

$$0 = ||\mathcal{F}f_1 - \mathcal{F}f_2||_2 = ||f_1 - f_2||_2$$

which implies $f_1 = f_2$ in $L^2(\mathbb{R})$ and hence K is injective. However, K^{-1} is not bounded as a linear operator. This can be shown by direct construction of a sequence of functions $(f_i)_{i=1}^{\infty}$ such that $||f_i||_2 \to \infty$ while $||Kf_i||_2 \to 0$. Note that this implies that B isn't closed in $L^2(\mathbb{R})$ by the Open Mapping Theorem.

Inverse problems themselves have been the focus of intense research over the last century, with many successes. It wasn't until the work in [30] that noisy versions of IPs were considered. Since then, a large amount of work has been done analyzing specific problems and, to a lesser extent, developing general methodology and theory. Some examples of inverse problems are

estimating a derivative, errors in variables, boundary value problems for heat equations, and tomography (which includes MRI and fMRI).

In this thesis, I propose to look at three related problems in IPs. First, in section 2, I examine a realistic instance of IPs where many related observations are made of an underlying function. The motivating application is contemporary and proposed large astronomical sky surveys, such as the LSST. These surveys are operational for many years and end up re-imaging the same piece of sky many, many times. A major objective of these surveys is to use this sequential data aquisition to produce a good summary image for visible celestial structures and use this image to detect when new objects, such as asteroids or supernovae, appear. Note that sequential IPs appear in other signal processing applications such as medical imaging.

This detection problem motivates the next part, found in section 3.1. In subsection 3.2.1 I propose a method for matching new images to archived summary image for the purposes of detecting transients. Also, in subsection 3.2.2 I investigate a Bayesian method for doing the detection without estimating several nuisance parameters directly.

Lastly, in section 4, I consider a method for choosing smoothing parameters in the context of IPs and present shortcomings of contemporary methods.

2 Sequential Inverse Problem

To help understand the sequential problem I first introduce IPs as they are commonly formalized, at least implicitly, in the literature.

Let $\mathcal{L}$ be a separable Banach space of functions[1] and $\Theta \subseteq \mathcal{L}$ be the *model space*. This corresponds to $\theta \in \Theta$ being the true, unobserved scene in imaging problems. Define $\gamma : \Theta \to \mathcal{B}$ to be the function of the scene θ we wish to estimate. Call either the mapping γ or the image $\gamma(\theta)$ the *parameter*. Usually $\mathcal{B} = \mathbb{R}^n$ for some n. In applied mathematics $\mathcal{B} = \Theta$ and g is the identity mapping. Following the seminal work of [3, 4], γ could be some sequence of functionals of θ. This idea, known as the Backus-Gilbert method, has been revisited periodically in various fields; theory: [22], astophysics: [23], and fMRI: [14].

[1] Note that separability is necessary as many solution methods require a nested sequence of subsets of $\mathcal{L}$ that approximate that space asymptotically. Separability is needed for this to be true. Classic examples are the Galerkin methods such as boundary elements or Krylov subspace methods. See [11] for a recent take on using wavelets as a foundation for the approximation spaces.

What makes it an inverse problem is the introduction of an operator K. Assume there is a known[2] linear[3] operator K such that observations of θ can only be made through the function $g(t) = K\theta(t)$. Define a random variable (process) W on $\text{ran}(K)$ that doesn't depend on θ. Then the random process under consideration is

$$Y = K\theta + \sigma W \tag{1}$$

for some $\sigma > 0$. Let $\mathbb{P}_{\theta,K,\sigma}$ be its distribution. Then the entire mapping $(\theta, K, \sigma) \mapsto \mathbb{P}_{\theta,K,\sigma}$ is the *forward problem*.

Some of the IP literature (eg [15]) consider (1) the data. Alternately, as recorded observations are necessarily discretized, the model is phrased as

$$Y_i = K\theta(u_i) + \sigma W(u_i) \tag{2}$$

for some sequence $(u_i)_{i=1}^n$. See [10, 11, 20, 9] for examples. More generally, fix a sequence $(\phi_i)_{i=1}^n \subseteq \mathcal{L}^*$ where $\mathcal{L}^*$ is the continuous dual of $\mathcal{L}$. This is the same as defining $K : \mathcal{L} \to \mathbb{R}^n$, where $K\theta$ is vector-valued with components $\phi_i\theta$. Then suppose our observations are

$$Y_i = \phi_i\theta + \sigma Z_i \tag{3}$$

See e.g. [25, 31, 19, 17, 27] for examples.

Essentially all linear inverse problems can be expressed in this manner. For instance, we can suppose $L^2([0,1]) = \mathcal{L} = \Theta = \mathcal{B}$, $g(\theta) = \theta$, and we observe $\theta \in \Theta$ under the action of an inhomogeneous Fredholm integral equation of the first kind at a finite number of points $0 \leq u_1 \leq \ldots \leq u_n \leq 1$. Then our observations are $Y_i = \int_{[0,1]} k(u_i, v)\theta(v)dv$ for some function $k \in \mathcal{C}([0,1])$ where $\mathcal{C}$ is some space of suitably nice functions such that elements of $\mathcal{C}$ are defined under pointwise evaluation in the first argument and elements of $\mathcal{L}$ in the second argument.

Here we can see the divergence of these approaches. In (1), observations are of $g(u) = \int k(u,v)\theta(v)dv$. In (2), the functions $k_i(\cdot) := k(u_i, \cdot)$ are presumed to be known. In (3), however, I may only know the outcome of the functionals (ϕ_i) generated by integration against k_i. Notice that this is closely related to Galerkin methods; the difference being that here the approximation spaces are given by the problem (span of the $k_i's$), instead of being chosen by the analyst.

[2]While the unknown operator case is very interesting, we don't address it here.

[3]Nonlinear inverse problems are different altogether and are given by a nonlinear operator. For certain special cases, the linearized version behaves nice enough that the following can still be used as a solution format.

2.1 How is this different?

Note that when K is the identity [7, 24] show the asymptotic equivalence of (1) and (2) in nonparametric regression and density estimation, respectively. However, results of this sort for inverse problems are still open questions. This leaves open the choice of statistical model as inferences based on (1) need not align, even asymptotically, with inferences based on (2). Recently, [26] established some conditions for the sequence $(u_i)_{i=1}^n$ where asymptotic equivalence can and cannot be established in the functional deconvolutional setting.

Sequential IPs diverge from the formalization in the previous section in at least two ways. First, we observe many instances of different but related IPs. This can be represented in the white noise model as

$$Y_t = K_t\theta + \sigma_t W_t \quad \text{for} \quad t = 1, \ldots, T. \tag{4}$$

The relationship between (1) and (2) is already murky. When there is another asymptotic regime to consider, the relationship becomes all the more unclear.

Second, and less obviously, the goals are possibly different. For instance, in Astronomical applications, a major goal of template creation is to create a high quality image that can be used to detect transients. The first part, create a good template, matches the classic inverse problem goal: create an estimator $\hat{\theta}$ that is 'close' to θ in some sense. However, finding a $\hat{\theta}$ that can do transient detection is a possibly different objective altogether. In fact, detection seems to suggest a classification-type loss function, while θ estimation suggests an L^p-type loss function.

Proposed Work: Develop and exploit a theoretical platform to answer the questions: Is an estimator that makes a good template also good at transient detection? What estimators are good under these criteria? How good are commonly used approaches at either goal? Note that if the answer to the first question is affirmative then the subsequent questions become more concise.

2.2 Possible Approach

I want to add an extra layer to this formalization to allow for slightly more flexibility. This added flexibility allows for the machinery sufficient for analysis. Suppose there exists unobserved random variables Y as in (1). I refer to this transformation of θ as the *Distortion Step*. Let $D(N) := (D_i)_{i=1}^N$ be

a sequence in L^* that we call *detector functionals*. The data is then

$$Y_i = D_i(K\theta) + \sigma D_i W. \tag{5}$$

We call this the *Detector Step*. These functionals mimic the data aquisition that occurs during analog to digital conversion and can be thought of as integration over disjoint intervals representing pixels. However, more elaborate detector signitures can be expressed such as unequal sensitivity in a detector. This could be important, particularly in modeling anti-aliasing or diffraction in telescoping imaging.

Let $\mathcal{P} = \{\mathbb{P}_{\theta,K,\sigma} : \theta \in \Theta, K \in B(\mathcal{L},\mathcal{L}), \sigma > 0\}$ be defined on the same σ-algebra. Then $(\theta, K, D(N), \sigma) \mapsto \mathbb{P}_{\theta,K,D(N),\sigma}$ is the forward problem. Note that $D(N)$ can usually be expressed as N without confusion.

As T increases, N remains fixed as the properties of the detector won't change over time. This is an instance of having a *fixed resolution detector* where in-fill asymptotics no longer make sense. Overall, for each t, we have a new *distortion step* operator $K_t : \mathcal{L} \to \mathcal{L}$. Now, the forward operator becomes

$$(\theta, K(T), N, \sigma) \mapsto \mathbb{P}_{\theta,K(T),N,\sigma} \tag{6}$$

where $K(T) := (K_t)_{t=1}^T$. Also, I augment the notation to make the time t observation at the i^{th} detector Y_{ti} with mean $\theta_{ti} := D_i K_t \theta$ and look at asymptotics as $T \to \infty$.

As long as the D_i are orthogonal as functions in $\mathcal{L}$, then $(D_i W) =: (Z_i)$ becomes a mutually independent sequence. If W is a zero mean brownian sheet, then Z_i is normal. Combining the distortion and detections steps, the data is

$$Y_{ti} = \theta_{ti} + \sigma Z_{ti}$$

which can be recognized as a normal means problem. The difference between this and the classic one-way ANOVA normal means problem is that there is no asymptotics in n. Hence, we don't necessarily want to form an estimator on the vector of sums $(\sum_{t=1}^T Y_{ti})_{i=1}^n$. Rather, some model selection should possibly be used to pick among the included terms in the sum. This corresponds to lucky image, which we discuss in the next section.

Over time, some K_{t_1} will offer better resolution of the model θ than some other K_{t_2}. In other words, K_{t_1} will be less bandlimiting than K_{t_2}. The question, in time space, becomes: does it make sense to not add the additional bias of including $Y_{t_2 i}$ in an estimator. Note that in the formalization, a somewhat strange inversion has happened from the usual bias-variance trade-off. Adding $Y_{t_2 i}$ to the sum *increases* the bias and *decreases* the variance.

Suppose, as usual, the risk is $\rho : \mathcal{B} \times \mathcal{L} \to \mathbb{R}^+$ with image $\rho(\hat{\theta}, \theta)$. Then, unless $\Theta \subseteq \text{span}((D_i)_{i=1}^N)$, the risk can't go to zero pointwise over Θ as $T \to \infty$. Even if we redefine the risk to be $\rho : \mathcal{B} \times \mathbb{R}^N \to \mathbb{R}^+$ with image $\rho(\hat{\theta}, \theta_N)$, where θ_N is the projection of θ onto $\text{span}((D_i)_{i=1}^N)$ along Θ, the risk will generally not go to zero.

To address these problems, I propose to use a relative efficiency notion to analyze estimators. Specifically, let $\epsilon > 0$ be given. Then, find

$$T_\epsilon(\hat{\theta}) := \arg\min\{T_* : \mathbb{E}[\rho_T(\hat{\theta}, \theta_N)|(K_t)_{t=1}^T] < \epsilon,\ \forall T \geq T_*\},$$

where ρ_T indicates the expectation is with respect to the forward operator defined in (6).

For the purposes of this analysis,

$$\rho(\hat{\theta})_T := \mathbb{E}||\hat{\theta} - \theta_N||^2. \tag{7}$$

Lucky Imaging

In lucky imaging (LI), a large number of images are observed and only the 'best,' according to some criterion, are retained. [18, 29] describe such implementations in detail. An advantage of this approach is that the reconstruction of the true scene is based entirely on high quality data. A disadvantage is that the method requires storing many images to determine which are best. Moreover, the images that are discarded can contain useful information about the scene that is, in effect, wasted.

One approach would be to suppose for $t = 1, \ldots, T$ we have without loss of generality ordered our observations from best to worst[4]. What we mean by best needs to be formalized, but will be encapsulated by $||\cdot||$ on $\mathcal{B}(\mathcal{L})$, although it may not be a norm. Suppose we have the permutation of $\{1, \ldots, T\}$ such that $(||K_t||)$ is ordered from smallest to largest. Let Also, let $T(\alpha)$ be the cardinality of $\mathcal{T}(\alpha)$. Then

$$\widehat{\theta(\alpha)}_i = \frac{1}{T(\alpha)} \sum_{t \in \mathcal{T}(\alpha)} Y_{ti}$$

defines a hyperplane of LI estimators, indexed by α. For example, $\alpha = 1$ indicates taking all images into a pixel-wise mean. Observe that the risk is

[4]We have been exploring related but improved approaches that avoid some of the optimization pitfalls.

$$\rho(\alpha) := ||\widehat{\theta}(\alpha) - \theta_N||^2 = \sum_{i=1}^{n} \left(\frac{1}{T(\alpha)} \sum_{t \in \mathcal{T}(\alpha)} \theta_{ti} - \theta_i \right)^2 + \frac{\sigma^2}{T(\alpha)n}$$

$$=: \beta(\alpha) + \frac{\sigma^2}{T(\alpha)n} \quad (8)$$

Thus, we have an integer programming problem:

$$\begin{aligned} \min \quad & R(\alpha) \\ \text{subject to} \quad & \\ & T(\alpha) \in \mathbb{N} \\ & \alpha \in [0,1] \end{aligned}$$

3 Kernel Matching

3.1 Overview

The kernel matching problem is an interesting instance of an inverse problem. One major goal in many imaging problems is to detect changes, sometimes known as transients, in a particular location, called a scene, over time. To this end, a good quality, low-noise reference image, R, is usually generated by combining many images together. Now the problem is: given a noisy, lower quality science image S, can we detect any changes between S and R? To formalize this approach, suppose there is a (linear) operator K such that

$$\mathbb{E}[S|RT] = KR + KT$$

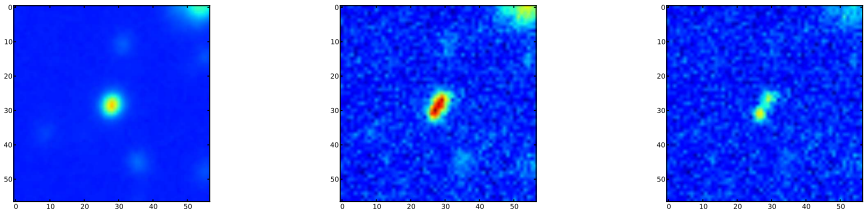
where T corresponds to an image of transients.

The image S is very large, on the order of ten million pixels when recorded. Also, the operator K can be very complicated, varying substantially across the image. Hence, modern Astronomical surveys take the following approach. Break up the image S into disjoint, very small parts $L_1, \ldots, L_p$ such that L_i contains only one source. Also, they assume that K is locally a convolution on L_1, which we call K_i for location L_i and that the noise is a spatially stationary normal with variance nugget σ_i^2. Also, assume that at each location, T is zero . This is usually accomplished by searching a known catelogue of sources for each image S that are isolated and making a small 'postage stamp' for each source. This results in a panel of images

$$S_i = K_i R + \sigma_i Z.$$

Now, with the use of a chosen basis and regularization method, $\hat{K}_i$ is formed for each i. See [2] for an approach using finite sums of Gaussians as a basis with with tuning parameter set at a non-data dependent level and [5] for a step function basis and tuning parameter set by risk estimation. In section 4 we provide details of this risk estimation proceedure. A functional interpolation is performed across the image to get $\hat{K}$, which is brute force applied to R via quadrature. The functional interpolation is always done via fitting a low order polynomial trend to the first few right singular vectors of $(\hat{K}_i)_{i=1}^p$. Lastly, $\Delta_{\hat{K}} := S - \hat{K}R$ is formed, and detection of transients is performed in various ways, usually involving manual human inspection in some fashion. Notice that the conclusion that all relevant inference can be done by examining Δ is an implicit assumption made in all approaches of which we are aware.

See figure 1 for an example of such a 'postage stamp' and the outcome of naively forming $\Delta_i := S_i - R$ without estimating K_i. Notice that Δ_i contains both remnants of the main source in the center and and scattered noise from S_I . Both of these artifacts would likely be labeled as transients.



(a) Reference Image R (b) Science Image S (c) Difference Image Δ

Figure 1: Example of a naive attempt at transient detection on a subset of an Astronomical image of a star.

3.2 Proposed Work:

We propose two new approaches to this problem. First, we introduce a new method for estimating the local convolution K_i, which more closely aligns with the goal of forming an estimate $\hat{K}_i$ such that $S_i - \hat{K}_i R$ has no left-over structure on a variety of scales. We call this the Multiresoluton Operator Estimator (MOE) and discuss it in section 3.2.1.

Second, we re-examine the problem of transient detection from first principals. In reality, neither $\Delta_{\hat{K}}$ nor $\hat{K}$ are of interest in their own right. Hence,

we should treat the difference and K as nuisance parameters and attempt to estimate the probability that there is a transient in S. We call this method BEST Detect, for BayESian Transient Detection. See section 3.2.2 for some preliminary details.

3.2.1 Multiresolution Operator Estimator (MOE)

Suppose we have a method for estimating K_j which results in a suite of estimators $(\hat{K}_j(\lambda))_{\lambda\in\Lambda}$. Both [5, 2], mentioned above, are examples of this. It is tempting to apply one of the methods mentioned in section 4 to choose $\hat{\lambda} \in \Lambda$. However, the extra structure in this problem of having two samples allows for possibly better approaches.

Specifically, we propose a multiresolution, noise-like statistic that attempts to quantify the goal of forming a 'good' subtraction image Δ_i. If we choose the correct operator $\hat{K}_i(\hat{\lambda})$, then $\Delta_i(\hat{\lambda})$ will look like noise only and no additional structure.

For each 'postage stamp' j, identify a mutiresolution system of subsets $\mathcal{I}$. This could be a wavelet system such as the Haar basis, for instance. Then, compute the following

$$NL(\hat{K}_j, \mathcal{I}) := \sup_{I\in\mathcal{I}} \frac{1}{\sqrt{|I|}} \left| \sum_{j\in I} (S_j - \hat{K}_j R)_i \right|. \tag{9}$$

See [13] for application of this idea to nonparametric regression over atypical function spaces. We will equivalently index NL by the tuning parameter λ when $\hat{K}_j$ is otherwise fixed. In other words, (9) takes the form

$$NL(\lambda, \mathcal{I}) := NL(\lambda) = \sup_{I\in\mathcal{I}} \frac{1}{\sqrt{|I|}} \left| \sum_{j\in I} (S_j - \hat{K}_j(\lambda) R)_i \right|. \tag{10}$$

We define $\hat{\lambda} := NL(\lambda)$.

The main idea behind MOE is that if K_j is properly estimated then the difference image Δ_j should be just noise. MOE looks at the difference at many different scales and locations, and minimizes NL over λ in (10). In most problems this would result in overfitting as we would just be interpolating the noisy observations. Having two samples prevents this by having a test set. This test set is not identically distributed as the training set, however.

We propose to undertake the following steps to analyze the NL statistic for estimating K_i. In what follows, we will drop the subscript i with the understanding that we are taking about one of the aforementioned L_i.

Non-random, Known σ Posit a sequence of non-random operators $(K_\lambda)_{\lambda\in\Lambda}$ and $\sigma > 0$ known such that there exists a K_{λ_0} where $\mathbb{E}[S|R] = \mathbb{E}[K_{\lambda_0}R + \sigma Z|R] = K_{\lambda_0}$. In this case, show that we have a consistent test with some rate for finding $\lambda_0 \in \Lambda$ in general and in some specific cases of families. This has largely been accomplished below.

Non-random, unknown σ Either relax the assumption of known variance or look at estimating the kernels from another iid copy of the science images.

Estimated Kernel Finally look at the case were the kernels are estimated from the science image.

Data Analysis Show the method on simulations and data from Andy Connolly.

So far, we have accomplished **Non-random, Known** σ from the list above. This is an outline.

Observe that we can write this as

$$NL(\lambda, \mathcal{I}) := \sup_{I\in\mathcal{I}} \left| \sum_{i\in I} (K_\lambda - K_{\lambda_0})R + \sigma Z_i \right| \tag{11}$$

by adding an subtracting $K_{\lambda_0}R$. Note that the summation of i is supressed in the first term for notational clarity.

Now, one quality this statistic could have would be to asymptotically distinguish between competing hypotheses in Λ. In this case, low-noise asymptotics makes more sense than large sample, so we choose this regime.

Our goal is to look at the power of this statistic to determine amongst hypothesis asymptotically. It is known [12] that asymptotics for fixed alternative hypothesis leads to trivial results, such as power always tending toward 1.

Hence, we wish to look at an analogy to the Pittman slope. This can be phrased as follows. Let $\tau > 0$ be given. Then we want to look at

$$\lim_{\sigma\to 0} \mathbb{P}\left(\frac{NL(\lambda_0 + \Delta C_\sigma)}{NL(\lambda_0)} > \tau \right) \tag{12}$$

where $C(\sigma)$ is a function going to zero with σ and Δ is a constant. We look at the ratio of the test under the alternate and null hypothesis as a way of rescaling. Alternatively, we can make τ a function of σ. We see in what follows the ratio in effect chooses that function.

Note first that we can rewrite (12) as

$$\lim_{\sigma\to 0} \mathbb{P}\left(\frac{\sup_{I\in\mathcal{I}} \left|\sum_{i\in I} (K_{\lambda_0+\Delta C(\sigma)} - K_{\lambda_0})R + \sigma Z_i\right|}{\sup_{I\in\mathcal{I}} \left|\sum_{i\in I} Z_i\right|} > \sigma\tau \right) \tag{13}$$

by using (11) and multiplying by σ.

We use (13) to show that σ is the correct reference rate for the RHS to decay. Hence, we consider

$$\lim_{\sigma\to 0}\mathbb{P}\left(\sup_{I\in\mathcal{I}}\left|\sum_{i\in I}(K_{\lambda_0+\Delta C(\sigma)}-K_{\lambda_0})R+\sigma Z_i\right|>\sigma\tau\right)$$

Before continuing, we need a result for exchanging sup and $\mathbb{P}$:

Lemma 1. *Let $(X_t)_T$ be a sequence of random variables over some index T such that $\sup_t X_t = X_{t_*}$ for some $t_* \in T$. Then for any $\tau > 0$*

$$\mathbb{P}(\sup_t X_t > \tau) \geq \sup_t \mathbb{P}(X_t > \tau)$$

Proof. Write $\mathbb{P}(\sup_t X_t > \tau) = \mathbb{E}\mathbf{1}(\sup_t X_t > \tau)$. Now, since $\mathbf{1}(\sup_t X_t > \tau) = \sup_t \mathbf{1}(X_t > \tau)$, we see that

$$\mathbb{P}(\sup_t X_t > \tau) = \mathbb{E}\sup_t \mathbf{1}(X_t > \tau) \geq \sup_t \mathbb{E}\mathbf{1}(X_t > \tau)$$

where for the last inequality we use that $\sup_t \int f_t \leq \int \sup_t f_t$ for the necessary kinds of sequences of functions and measures. □

Using Lemma 1, we can write

$$\begin{aligned}
&\mathbb{P}\left(\sup_{I\in\mathcal{I}}\left|\sum_{i\in I}(K_{\lambda_0+\Delta C(\sigma)}-K_{\lambda_0})R+\sigma Z_i\right|>\sigma\tau\right)\geq\\
\geq&\sup_{I\in\mathcal{I}}\mathbb{P}\left(\left|\mu_{I,\sigma}+\sum_{i\in I}\sigma Z_i\right|>\sigma\tau\right)=\\
=&1+\Phi\left(-\sqrt{|I|}\left(\tau+\frac{\mu_{I,\sigma}}{\sigma}\right)\right)-\Phi\left(\sqrt{|I|}\left(\tau-\frac{\mu_{I,\sigma}}{\sigma}\right)\right) \qquad (14)
\end{aligned}$$

where we define $\mu_{I,\sigma} := \sum_{i\in I}(K_{\lambda_0+\Delta C(\sigma)} - K_{\lambda_0})R$ and notice that

$$\mu_{I,\sigma}+\sum_{i\in I}\sigma Z_i \sim N\left(\mu_{I,\sigma},\frac{\sigma^2}{|I|}\right).$$

We would like to examine the $C(\sigma)$ such that the RHS of (14) $\overset{\sigma\to 0}{\to} 1$.

Using $\liminf_m \sup_n x_{m,n} \geq \sup_n \liminf_m x_{m,n}$ for any doubly indexed sequence $x_{m,n}$ we see that under (14)

$$\lim_{\sigma\to 0}\mathbb{P}\left(\sup_{I\in\mathcal{I}}\left|\sum_{i\in I}(K_{\lambda_0+\Delta C(\sigma)}-K_{\lambda_0})R+\sigma Z_i\right|>\sigma\tau\right)\geq$$

$$\geq\sup_{I\in\mathcal{I}}\lim_{\sigma\to 0}\left[1+\Phi\left(-\sqrt{|I|}\left(\tau+\frac{\mu_{I,\sigma}}{\sigma}\right)\right)-\Phi\left(\sqrt{|I|}\left(\tau-\frac{\mu_{I,\sigma}}{\sigma}\right)\right)\right]. \qquad (15)$$

This probability goes to 1 when $\mu_{I,\sigma}/\sigma\to\infty$.

We did this calculation for the case where K_λ a non-normalized Gaussian kernel with variance λ for all $\lambda\in\Lambda$. We find

$$\frac{\mu_{I,\sigma}}{\sigma}\to\infty \quad \text{if} \quad C'(\sigma)\to\infty$$

and

$$\frac{\mu_{I,\sigma}}{\sigma}\to 0 \quad \text{if} \quad C'(\sigma)\to 0$$

The second of the two results is not informative. If we additionally assume that $C(\sigma)=\sigma^\alpha$ for $\alpha>0$ we get

$$\frac{\mu_{I,\sigma}}{\sigma}\to\infty \quad \text{if} \quad \alpha\in(0,1)$$

and

$$\frac{\mu_{I,\sigma}}{\sigma}\to 0 \quad \text{if} \quad \alpha>1.$$

The $\alpha=1$ case is R dependent.

3.2.2 BEST Detect

The idea behind BEST Detect is to formulate a hierchical framework that explicitly models the actual parameter of interest; namely if there is a transient or not. Using the notation from section 3.1, we specify the following hierarchy:

$$S|K,T,R\sim GP(K(R+T),C)$$

$$T|M=M\delta$$

$$K|\alpha\sim p(K;\alpha)$$

$$M|\pi\sim \text{Bernoulli}(\pi)$$

where δ is a random field characterizing possible transients and $p(K;\alpha)$ is a distribution parameterized by α. We have a lot of prior information as to what transients of interest look like and how K looks and varies over the image. The transients will either be long streaks corresponding to asteroids and comets, or very bright objects corresponding to supernovae or gamma ray bursts. Hence, we can frame them as a Markov random field over an appropriate basis of shapes. Then, for any pair of images, we can report either $p(M = 1|S, R)$ or a Bayes factor $p(S|M = 1)/p(S|M = 0)$. We have results from computing the Bayes factor, but not the marginal posterior of M.

4 Tuning Parameter Selection In Inverse Problems

Risk estimation is of course a very broad subject. However, not nearly as much work has been done in the context of inverse problems. The issue at hand is that in well-posed problems, regularization is introduced to get better risk performance. Usually, both a tuning parameter and a corresponding risk estimate is introduced with the hope that this will reduce the risk.

On the other hand, in ill-posed problems some form of regularization is required for estimation. However, the same mechanism that makes regularization a requirement also makes risk estimation difficult. This situation is explored below.

There are two main approaches to risk estimation in inverse problems. They correspond loosely to two separate solution methods. Loosely speaking, one centers on expanding f into an appropriate basis, which generally leads to choosing GCV. The other method relies on diagonalizing K, which leads to penalized empirical risk. Some methods, such as [15, 1, 21], attempt to do both simultaneously. While this is attractive and theoretically well justified, there are substantial restrictions to the classes of operators and functions one can consider. For instance, for the wavelet-vaguelette method, K being convolutional with a Gaussian kernel does not qualify.

To fix ideas, suppose that we make observations

$$Y_i = Kf(x_i) + \sigma\epsilon_i, \qquad x_i = i/n, \qquad i = 1, \ldots, n. \tag{16}$$

Here, $f \in L^2$ is an unknown function, K is a known operator, and $\sigma\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We can rewrite $Kf(x_i) = \langle k_{x_i}, f\rangle$, where $k_{x_i} = k(x_i, \cdot)$ is the

kernel of K evaluated at x_i. Also, define the risk as

$$R(f, \hat{f}) = \mathbb{E} \int (f - \hat{f})^2. \tag{17}$$

First, we outline commonly used approaches to risk estimation in inverse problems. The most used method is generalized cross validation (GCV). GCV does not need an estimate of σ and can be calculated with information that is usually obtained during the estimation proceedure.

For example, under some assumptions about the kernel of K and f we can rewrite (16) using a basis (ϕ_ν) as

$$Y = \Phi\theta + \sigma\epsilon$$

where $f = \sum_\nu \theta_\nu \phi_\nu$ and $\Phi_{ij} = \langle k_{x_i}, \phi_j \rangle$. Now, we can get a regularized estimate of θ by specifying a matrix V and forming

$$\theta(\lambda) := (\Phi_N^\top \Phi_N + \lambda V)^{-1} \Phi_N^\top Y$$

where Φ_N is the first N columns of Φ. Note that for different choices of V we can recover different methods such as splines or ridge. Define $L_\lambda := \Phi_N(\Phi_N^\top \Phi_N + \lambda V)^{-1}\Phi_N^\top$. Then we find the GCV estimate of λ by computing

$$\hat{\lambda} := \arg\min_\lambda \frac{\sum_{i=1}^n (Y_i - \Phi_N \theta(\lambda)_i)^2}{tr(I - L_\lambda)^2}.$$

The main, and some may say fatal, flaw of GCV in IPs is $\hat{\lambda}$ is a good approximation to the minimizer of the prediction risk

$$\frac{1}{n}\sum_{i=1}^n (Kf(x_i) - [L_\lambda Y]_i)^2 \approx \int (K(f - f_\lambda))^2$$

where $f_\lambda(x) = \sum_{i=1}^N \phi_i(x)\theta(\lambda)_i$. Therefore, we aren't approximately minimizing R in (17); rather a smoothed squared difference between f and f_λ. This has very real implications for the quality of the estimate as extreme 'wiggiliness' in f_λ can be masked after being smoothed by K. See [25, 31] for comments. In particular, [25] suggests a modified GCV. However, if Φ_N is ill-conditioned, as it is likely to be, the modified GCV is ill-conditioned as well.

Other approaches center around penalized empirical risk [10, 20, 9]. The approach is to specify biorthogonal bases (ϕ_ν) and (ψ_ν) with a coefficient vector (b_ν) such that

$$K\phi_\nu = b_\nu \psi_\nu. \tag{18}$$

These bases exist if K admits a singular value decomposition[5], SVD. However, there are other examples of operators satisfying (18) for bases other than the singular system.

This biorthogonal system allows us to rewrite (16) as

$$Z_\nu = \theta_\nu + \sigma \xi_\nu \epsilon_\nu.$$

Here, $Z_\nu := \langle Y, \psi_\nu \rangle$, $\theta_\nu := \langle f, \phi \rangle$, $\xi_\nu := b_\nu^{-1}$, and $\epsilon_\nu \sim N(0,1)$. Suppose, for example, we wish to choose the tuning parameter N in the estimator $\hat{\theta}_\nu = Z_\nu \mathbf{1}(\nu \leq N)$. Then the penalized empirical risk would be

$$R_{pen} = -\sum_{\nu=1}^{N} Z_\nu^2 + \sum_{\nu=1}^{N} \xi_\nu^2 + pen(N) \tag{19}$$

for some penalty functional *pen*. For unbiased risk estimation, $pen(N) = \sigma^2 \sum_{\nu=1}^{N} \xi_\nu^2$. This corresponds to a plug-in estimator of R after being transformed into sequence space and decomposed into bias and variance. Unsurprisingly, [9] finds that this penality functional has poor properties in practice. We revisit this phenomena later when we propose a new method.

A second proposed penalty is referred to as the Risk Hull Method. As it is somewhat involved, we will only mention that exists. The main downfall of penalized risk estimation, as mentioned in [8], is that it severely limits the possible choices of basis. In this way, no matter the underlying function f you are trying to recover, the chosen basis is determined by your operator K. As noted in [6], optimal risk performance in linear smoothers is intimately connected to choosing a basis that sparsely represents f.

4.1 Proposed Work:

4.1.1 Introduction

Returning to the GCV example, define θ_N to be first N entries in θ. Then, we would like to find a λ such that

$$\rho(\lambda) := \mathbb{E}||\theta_N(\lambda) - \theta_N||^2 = ||\theta_N||^2 + \mathbb{E}||\theta_N(\lambda)||^2 - 2\mathbb{E}\langle \theta_N(\lambda), \theta_N \rangle \tag{20}$$

where the expectation is taken with respect to the distribution of $(\sigma\epsilon_i)_{i=1}^n$. When doing risk estimation we can disregard $||\theta_N||^2$ as it is unknown but doesn't depend on λ. Also, we already have an unbiased estimate of $\mathbb{E}||\theta_N(\lambda)||^2$, namely $||\theta_N(\lambda)||^2$. Hence, we wish to get an estimate of $\mathbb{E}\langle \theta_N(\lambda), \theta_N \rangle$. [16] introduces an unbiased risk estimate that generalizes the work in [28]

[5] A sufficient condition is for K to be compact, which is generally the case.

Theorem 1. *If Y is from an exponential family with parameter vector $\theta_N \in \mathbb{R}^N$ and S is a sufficient statistic, then under some mild assumptions (integrability and almost sure differentiability) about an estimator $h(S)$, it follows that*

$$\mathbb{E}\langle h(S), \theta_N \rangle = -\mathbb{E}\left[tr\left(\frac{\partial h(S)}{\partial S} \right) + h(S)^\top \frac{\partial \log q(S)}{\partial S} \right]$$

where tr is the trace function and q is the normalizing constant in the pdf of S. Specifically, $f_S(s) = q(s)\exp\{\theta^\top s - g(\theta)\}$.

Taking this, for each estimator $\hat{\theta}$ we define

$$\hat{\rho}_{UB}(\hat{\theta}) := ||\hat{\theta}||^2 + 2\left[tr\left(\frac{\partial h(s)}{\partial s} \right) + \hat{\theta}^\top \frac{\partial \log q(s)}{\partial s} \right] \tag{21}$$

Corollary 1. *Under the hypothesis of Theorem 1, $\hat{\rho}_{UB}$ is an unbiased estimate of ρ up to a constant that depends only on θ_N.*

4.1.2 Proposed Goals

Well-conditioned Examine the quality of tuning parameter selection based on this criteria in a linear model when the design matrix is well-conditioned.

Ill-conditioned Introduce an adaptation of the unbiased risk estimation when the design matrix is ill-conditioned and examine its ability to asymptotically find correct tuning parameters

Finite n Comparison Also, show some results on finite sample usage of this method vs. GCV.

4.1.3 Well Conditioned

Suppose we begin with the following linear model

$$Y = \Phi_N \theta_N + \sigma\epsilon$$

Then

$$\hat{\rho}(\theta_N(\lambda)) := \hat{\rho}(\lambda) = ||\theta_N(\lambda)||^2 + 2\left[tr\left(W_\lambda\right) - \theta_N(\lambda)^\top (W_0 s) \right]$$

where $W_\lambda := (\Phi_N^\top \Phi_N + \lambda V)^{-1}$ and W_0 is $W_\lambda\Big|_{\lambda=0}$. I want to show that picking λ based on minimizing $\hat{\rho}$ makes sense asymptotically. In general, an approach would look something like the following. Fix $\theta \in \Theta$ and define

$$\lambda_0 := \arg\min \rho(\lambda, \theta) \;\text{ and }\; \lambda_n := \arg\min \hat{\rho}(\lambda).$$

Then we are interested in the asymptotic behavior of

$$\Delta(\lambda_0, \lambda_n) := \rho(\lambda_0, \theta) - \rho(\lambda_n, \theta).$$

Note that by definition $\Delta(\lambda_0, \lambda_n) \geq 0$.

Now, Δ admits the following decomposition

$$\begin{aligned}\Delta(\lambda_0, \lambda_n) = \\ &=(\rho(\lambda_0, \theta) - \hat{\rho}(\lambda_0, \theta)) + (\hat{\rho}(\lambda_0, \theta) - \hat{\rho}(\lambda_n, \theta)) + (\hat{\rho}(\lambda_n, \theta) - \rho(\lambda_n, \theta)) \\ &=: (a) + (b) + (c).\end{aligned}$$

By the Strong Law of Large Numbers, $(a) \overset{a.s.}{\to} 0$. Also, $(b) \leq 0$ as λ_n minimizes $\hat{\rho}$. So,

$$\lim_{n\to\infty} \Delta(\lambda_0, \lambda_n) \overset{a.s.}{=} 0$$

if (c) converges almost surely to 0.

4.2 Ill Conditioned

In many cases, the matrix X is very ill-conditioned. In statistics, this is sometimes referred to as multicollinarity. It can be defined rigorously by appealing to a SVD of X. Specifically, any matrix can be written as $X = UDV^\top$. Here, U is an orthogonal matrix that forms a basis for $ran(X)$, V is an orthogonal matrix such that $V^\top$ forms a basis for $null(X)^\perp$, and $D = diag([s_1, s_2, \ldots, s_q])$, $q = \min(n, N)$. We can assume that $(s_i)_{i=1}^q$, called the singular values, are ordered from greatest to least and nonnegative.

We say that a matrix is ill-conditioned (in the l^2 norm) if $s_1/s_q := C >> 1$. In particular, $C = 1$ means perfect conditioning (like, for example, a unitary matrix), and $C = \infty$ means X is singular. The condition number C of X is roughly the derivative of X, thought of as a linear function, at a point. In fact, it says how different $X(a + \delta)$ is from Xa when δ is a small perturbation. C large means $||X(a+\delta) - Xa||_2$ can be large even when $||\delta||^2$ is small.

There are interesting connections between ill-conditioning and regularization. Generally, regularization solves a nearby least squares solution that is better conditioned. The specifics are wrapped up in the norms involved. But, we can apply this idea to our current situation and regularize our risk estimate, as the unbiased version developed in Theorem 1 becomes unstable as $C \to \infty$.

There are several ways to do this regularization. We choose to approach it by truncating the spectrum of X. Specifically, choose some number $r \leq q$.

Define a new diagonal matrix $D_r := diag([s_1, \ldots, s_r, 0, \ldots, 0])$. Now, we can form a new matrix $X_r := UD_rV^\top$ that has condition number C_r that is as small as we want. X_r has many properties such as

$$||X_r - X||_2 = \min_{A:rank(A)=r} ||A - X||_2$$

Implicit in what follows, we assume that X is full rank, but that it might be very ill-conditioned. This is a very reasonable assumption, and in fact is treated as almost equivalent in some literature to inverse problems in general.

Define a new function

$$\hat{\rho}_r(\theta_N(\lambda)) = ||\theta_N(\lambda)||_2^2 + 2\left(tr\big(W_\lambda\big) - \hat{\theta}_\lambda^\top \hat{\theta}_{MLE,r}\right)$$

where $\hat{\theta}_{MLS,r} := (X_r^\top X_r)^{-1} S$.

We are still thinking about a more rigorous method for picking r. However, for now, pick a large condition number tolerance for X. This is going to correspond to some some maximal r. Then, get $\hat{\theta}_{MLE,r}$ and compute $\hat{\rho}_r(\theta_N(\lambda))$.

References

[1] Abramovich, F. and Silverman, B. W. (1997), "The vaguelette-wavelet decomposition approach to statistical inverse problems," *Biometrika*, 85, 115–129.

[2] Alard, C. and Lupton, R. (1998), "A Method for Optimal Image Subtraction," *The Astrophysical Journal*, 503, 325–331.

[3] Backus, G. and Gilbert, F. (1968), "The resolving power of gross Earth data," *Geophysical Journal of the Royal Astronomical Society*, 16, 169–205.

[4] Backus, G. and Gilbert, F. (1970), "Uniqueness in the inversion of inaccurate gross Earth data," *Philosphical Transactions of the Royal Society of London Series A*, 266, 123–192.

[5] Becker, A., Homrighausen, D., Genovese, C., Connolly, A., Owen, R., Bickerton, S., and Lupton, R. (2010), "Choice of Basis in PSF-Matching," *In Preparation.*

[6] Beran, R. (2000), "Scatterplot smoothers: superefficiency through basis economy," *Journal of the American Statistical Association*, 95, 155–171.

[7] Brown, L. D. and Low, M. G. (1996), "Asymptotic equivalence of nonparametric regression and white noise," *The Annals of Statistics*, 24, 2384–2398.

[8] Cavalier, L. (2008), "Nonparametric statistical inverse problems," *Inverse Problems*, 24.

[9] Cavalier, L. and Golubev, G. (2006), "Risk hull method and regularization by projections of ill-posed inverse problems," *Annals of Statistics*, 34, 1653–1677.

[10] Cavalier, L., Golubev, G., Picard, D., and Tsybakov, A. (2002), "Oracle inequlities for inverse problems," *Annals of Statistics*, 30, 843–874.

[11] Cohen, A., Hoffman, M., and Reiss, M. (2004), "Adaptive wavelet Galerkin method for linear inverse problems," *SIAM Journal of Numerical Analysis*, 42.

[12] DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, Springer Science, New York, New York.

[13] Davies, P. and Kovac, A. (2001), "Local extremes, runs, strings, and mutliresolution," *The Annals of Statistics*, 29, 1–65.

[14] de Peralta Menendez, R. G. and Andino, S. G. (1999), "Backus and Gilbert method for vector fields," *Human Brain Mapping*, 7.

[15] Donoho, D. L. (1995), "Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition," *Applied and Computational Harmonic Analysis*, pp. 101–126.

[16] Eldar, Y. (2009), "Generalized SURE for exponential families: applications to regularization," *IEEE Transactions on Signal Processing*, 57, 471–481.

[17] Evans, S. and Stark, P. (2002), "Inverse problems as statistics Inverse Problems," *Inverse Problems*, 18.

[18] Fried, D. (1978), "Probability of getting a lucky short-exposure image through turbulence," *Journal of the Optical Society of America*, 68, 1651–1658.

[19] Genovese, C. and Stark, P. (1996), "Data reduction and statistical inconsistency in linear inverse problems," *Department of Statistics, U.C. Berkeley*, 443.

[20] Golubev, Y. (2004), "The principle of penalized empirical risk in severely ill-posed problems," *Probability Theory and Related Fields*, 130, 18–38.

[21] Johnstone, I. M., Kerkyacharian, G., Picard, D., and Raimondo, M. (2004), "Wavelet deconvolution in a periodic setting," *Journal of the Royal Statistical Society, Series B*, 66, 547–573.

[22] Kirsch, A., Schomburg, B., and Berendt, G. (1988), "The Backus-Gilbert method," *Inverse Problems*, 4, 771–783.

[23] Loredo, T. and Epstein, R. (1989), "Analyzing Gamma-Ray Burst Spectral Data," *The Astrophysical Journal*, 336, 896–919.

[24] Nussbaum, M. (1996), "Asymptotic Equivalence of density estimation and Gaussian white noise," *The Annals of Statistics*, 24, 2399–2430.

[25] O'Sullivan, F. (1986), "A statistical perspective on ill-posed inverse problems," *Statistical Science*, 1, 502–527.

[26] Pensky, M. and Sapatinas, T. (2010), "On convergence rates equivalency and sampling strategies in in functional deconvolution models," *The Annals of Statistics*, 38, 1793–1844.

[27] Stark, P. (2008), "Generalizing resolution," *Inverse Problems*, 24, 1–17.

[28] Stein, C. (1981), "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, 90, 1247–1256.

[29] Tubbs, R. (2004), "Lucky exposures: diffraction limited Astronomical imaging through the atmosphere," *PhD Thesis*, Cambridge University, UK.

[30] Wahba, G. (1969), "On the numerical solution of Fredholm integral equations of the first kind," *Technical Report, University of Wisconsin*, 990.

[31] Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadelphia, PA.