M. S. Z. Tienstra
msz.tienstra@gmail.com
Master Thesis

Regularization of Ill-posed Statistical Inverse Problems

A data driven method for choosing the regularization parameters in Tikhonov regularization.

Thesis Supervisors: Dr. Tristan van Leeuwen,
Dr. Evgeny Verbitsky

Mathematisch Instituut, Universiteit Leiden
Date: ??-02-2022

Summary

In this thesis we discuss a data driven way to chose the regularization parameter in Tikhonov regularization of ill-posed inverse problems. We then implement and compare three different kinds of iterative methods for solving the minimization problem. Common parameter choice methods are a-prior rules, discrepancy principle, L-curve method, and cross-validation, but we would like to somehow learn/choose the regularization parameter based alone on the observed data. We show in this thesis that we can choose λ from the data by utilization the Bayesian framework.By posing an inverse problems in the Bayesian framework we gain insight into how the regularization parameter depends on the variance of model. Namely that λ consists of two parts, 1) the variance from the observational noise, and 2) the variance of the underlying ground truth. In the particular case of normal likelihood and additive normal error we get a posterior distribution where solving for the MAP estimate is equivalent to solving the Tikhonov regularization problem with $= \beta/\alpha$. If we bound the variance parameters of the observation noise and x, the resulting minimization problem is well-posed. In the special case above we have closed form solutions x, α, β and see that the minimization problem is bi-convex so coordinate decent method seems a natural choice. We alternatively implement other gradient methods in the case that we don't have closed form solutions. Using the partial solutions we can solve the resulting minimization problem. We also convergence and consistency of this method.

Contents

1 Inverse Problems

In this chapter we give a brief introduction to first functional analytic inverse problems, and why they are usefully and interesting to study. We then give a brief overview of some of the research done in this field and where this paper is situated among.(maybe I should put this in the summary) We then discuss the outline of this thesis.

1.1 Brief Introduction to inverse problems

In inverse problems, our goal is to recover the unknown parameter x from observation y. We want to solve the following equation for x

$$y = Ax \tag{1}$$

where $y \in \mathcal{Y}$ is the observed/measured data, $x \in \mathcal{X}$ is the unknown parameter, and $A : \mathcal{X} \to \mathcal{Y}$ is the forward linear operator that relates how x relates to y. We assume there exists some $\bar{x}$ such (1) holds, and that the forward problem linking the x to y is well-defined. We typically observe only noisy measurements of the x, and express this as

$$y = A(x, e) \tag{2}$$

It is assume through out this paper that e is some additive noise so we can write

$$y = Ax + e \tag{3}$$

Examples of inverse problems abound in many areas in science and mathematics. In mathematical modeling we know the parameters of the model, and their relation and therefore can predict the outcome. In inverse problems we are interested the opposite where we see only the the outcome but not the parameters that cause these the observations or how they are related. Numerous classical examples can be found in [7]. Below we give example of an inverse problem that arises in imagining and signal processing. This demonstrates that finding a solution to (3) is not trivial.

Example 1.1. *(DeConvolution [2])*

Let $\mathcal{X} = \mathcal{Y} = L^2(\mathbb{R})$ be the space of square integrable functions. Let $A : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ such that

$$(Af)(x) = g \circ f = \int_{\mathbb{R}} g(x-y)f(y)dy$$

Let $g(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. The Fourier transform of (Af) is

$$\mathcal{F}(Af)(\xi) = \int_{\mathbb{R}} e^{-i\xi x} Af(x)dx = \hat{g}\hat{f}(\xi)$$

If $Af = 0 \implies \hat{f} = 0 \implies f = 0$ so A is injective. So the solution is unique and exists. The solution to $Af = h$ is given by

$$f(x) = \mathcal{F}^{-1}(\hat{g}^{-1}\hat{h})(x)$$

The solution is not well defined for an arbitrary $h \in L^2(\mathbb{R})$. Suppose we observe small errors in in h. As $\hat{g}^{-1}$ grows exponentially, h may no longer be in the range of A.So the integral does not converge, and no solution exists.

1.2 Outline

The outline to the thesis is as follows. In Chapter 2 we explain one of the key questions in inverse problems. Is the formulated problem of solving for x well-posed. We define what ill-posedness is in a specific setting, and then discus numerous situation in which we encounter ill-posedness. This then natural leads us to the resolution of ill-posed problems where we explain how we can stabilize the solution through regularization. So far everything until then has been in the finite dimensional vector space setting, and we transition to the statistical setting in Chapters 3 and 4. Chapter 3 is a brief over view the statistical and probability notation used to define statistical inverse problems. Then chapter 4 briefly introduces statistical inverse problems, overviews some examples, and how they are related to the functional analytical setting. Chapter 5 is really the first main purpose of this thesis where we explain how from the Bayesian setting of inverse problems we can have a data driven method to infer the regularization parameters from the observations. We prove that the purely empirical Bayesian method is ill-posed in certain cases, and that a hierarchical model resolves this. The chapters 6 and 7 contain the second major half of this thesis. In this chapter we design three different numerical algorithms to numerical solve the resulting minimization problem derived in chapter 5. We show that these converge to a critical point of regularization functional. Then in chapter 7 we implement the methods in python, and explain the results. We also look at the models sensitively to the choice of hyper priors and the convergence and consistency. In chapter 8, we conclude with a discussion of where the thesis could go next and how we can improve on the results seen in chapter 7.

2 Regularization

We have seen that solving for x is more than just inverting the forward operator. In fact in the above example was an ill-posed problem. The direct inverse problem is often ill-posed, so we need a method to solve for x such that resulting problem is well-posed. Even when a unique solution exists or when we can define a unique solution, the solution is not guaranteed to depend on the the data continuously. To resolve these problems will introduce regularization into the direct inverse problem. The resulting problem will be well-posed and the resulting solution will be regularized.

2.1 Ill-posedness

Hadamard defined a well-posed problem as one that met all of the following three conditions are met [7]:

Definition 2.1. *A problem is well posed if the following three condition hold*

1. *Existence; There exists a solution*
2. *Uniqueness: The solution is unique.*
3. *Stability: The solution depends continuously on the observed data.*

Now let $\mathcal{X} = \mathbb{R}^m, \mathcal{Y} = \mathbb{R}^n$, and suppose that we wish to solve the following for x

$$y = Ax$$

where now $y \in \mathbb{R}^n, x \in \mathbb{R}^m, A \in \mathbb{R}^{n \times m}$.

Remark 2.1. *We then consider the following cases:*

1. *If A is a square matrix, and A has full rank, then A is invertible. We then have that $x = A^{-1}y$ is the solution to the above.*
2. *if $n > m$, and $rank(A) = m$, then the system of equations is undetermined and inconsistent. So no solution can exist.*
3. *if $n < m$, and $rank(A) = n$, then the system of equations is overdetermined, and a solution exist but may not be unique.*
4. *If A is not full rank then then by the rank-nullity theorem, the dimension of the null space can be greater than 0. In this case the solution may not exist and/or may not be unique.*

So a unique solution exists if and only if $y \in Ran(A)$ and that $ker(A) = 0$. Then we have that

$$x = A^{-1}y$$

To check if this solution is stable we recall the following two definitions. Since we are in the finite dimensional case, A can be represented by a singular system.

Definition 2.2. *Let $A \in \mathbb{R}^{n \times m}$, then the singular value decomposition of A is a factorization of A,*

$$A = U\Sigma V^*$$

where $U \in \mathbb{R}^{n \times n}$ orthogonal matrix, $\Sigma \in \mathbb{R}^{n \times m}$ rectangular diagonal matrix with non-negative entries, and $V \in \mathbb{R}^{m \times m}$ orthogonal matrix. The diagonal entries of Σ, denoted by σ_i are the singular values of A, and are listed in descending order. That is $\sigma_1 \geq \sigma_2 \geq .... \geq \sigma_m \geq 0$. We sometimes write $U = [u_1, ...u_n]$ and $V = [v_1, ..., v_m]$ where u_i and v_i are orthogonal basis for $\mathbb{R}^n$ and $\mathbb{R}^m$. The s respectively. The singular system of A is then $(u_i, v_i, \sigma_i)_{1 \leq i \leq \min(n,m)}$.

and

Definition 2.3. *Suppose we would like to solve the following equation for x,*

$$y = Ax$$

Let δy be the error in y. Assume that A is inevitable. Then the $A(x + \delta x) = y + \delta y$, so $(x + \delta x) = A^{-1}(y + \delta y) = A^{-1}y + A^{-1}\delta y$. The error in the solution is then $A^{-1}\delta y$. We can compute the ratio of the relative error in the solution compared to the the relative error in y as

$$\frac{||A^{-1}\delta y||}{||\delta y||}\frac{||y||}{||A^{-1}y||}$$

Then

$$\begin{aligned}\kappa(A) &= \max_{\delta y, y\neq 0}\left\{\frac{||A^{-1}\delta y||}{||\delta y||}\frac{||y||}{||A^{-1}y||}\right\} \\ &= \max_{\delta y\neq 0}\left\{\frac{||A^{-1}\delta y||}{||\delta y||}\right\}\max_{y\neq 0}\left\{\frac{||Ay||}{||y||}\right\} \\ &= ||A^{-1}||\ ||A|| \\ &= \frac{\sigma_{max}}{\sigma_{min}}\end{aligned}$$

where $||y||$ is the euclidean norm, $||A||$ is the induced matrix norm, and $\sigma_{max}, \sigma_{min}$ are the maximum and minimum singular values of A respectively.

So then

$$\frac{||x - x_\delta||}{||x||} \leq \kappa(A)\frac{||y - y_\delta||}{||y||}$$

Example 2.1 (Matrix Inversion [4])**.** *Let $y \in \mathbb{C}^n, x \in \mathbb{C}^n, A \in \mathbb{C}^{n\times n}$. Assume that A is symmetric positive definite. From the above, we can write*

$$A = \sum_{i=1}^{n}\sigma_i a_i a_i^T$$

where σ_i are the eigenvalues of A order such that $_1 \geq \sigma_2 \geq \ldots > 0$, and eigenvectors $a_i \in \mathbb{R}^n$ where $a_i \perp a_j$ for $i \neq j$. Assume we observe y_δ where $y_\delta = Ax^\delta$. Then we have that

$$x - x^\delta = \sum_{i=1}^{n}\sigma_i^{-1}a_i a_i^T(y - y^\delta).$$

The error between x and the estimate x^δ is

$$\begin{aligned}||x - x^\delta||_2^2 &= \sum_{i=1}^{n}\sigma_i^{-2}||a_i||^2|a_i^T(y - y^\delta)|^2 \\ &\leq \sigma_n^{-2}||y - y^\delta||_2^2 \\ &\leq \sigma_n^{-1}|||y - y^\delta||_2^2 \\ &\leq \kappa(A)\delta\end{aligned}$$

Suppose that δ is the magnitude of the error and $\kappa(A) \ll \infty$, then then the solution depends on continuously on the data, as the relative error in x is bounded by the relative error in y times a small constant. On the other hand if $\kappa(A)$ is very large, in which case A is ill-conditioned, then the solution is does not depend continuously on the data as a small change in y results in a large change in x. Since $\kappa(A) = \frac{\sigma_{max}}{\sigma_{min}}$ we see that a large $\kappa(A)$ occurs if σ_{min} is very small. If $\sigma_i \to 0$ very quickly then $\kappa(A) \to \infty$. So stability is determine by the decay of the singular values of A. To guarantee a stable solution, we need to bound the singular values of A away from zero.

Suppose that our system is $n > m$ and $y \notin Ran(A)$. Suppose also that A has full rank. The SVD of A is

$$A = U_m \Sigma_m V_m^* \tag{4}$$

Then

$$Ax = U_m U_m^* y$$

since U_m is an orthogonal matrix $U_m U^*$ projects y onto the range of A. The solution is then given by

$$\hat{x} = V_m \Sigma_m^{-1} U_m^* y = A^\dagger y$$

where we define A below.

Definition 2.4. *Let $A \in \mathbb{R}^{n \times n}$ with $rank(A) = r \leq \min\{n, m\}$. Then using SVD of A the Mooroe-Penrose pseudo is defined as*

$$A^\dagger = V_r \Sigma_r U_r^* \tag{5}$$

where U_r, V_r, Σ_r are defined in definition 2.2.

Claim 2.1. *The least squares solution to $y = Ax$ is given by*

$$x_{LS} := \min_x ||Ax - y||_2^2 \equiv V_m \Sigma_m U_m^* y \equiv A^\dagger y \tag{6}$$

where $A = U_m \Sigma_m V_m^*$ is the singular decomposition of the matrix operator A, $U_m = (u_1, ..., u_m), V_m = (v_1, ..., v_m)$, are the left and right m singular vectors and Σ_m is the diagonal matrix with the first m singular values. $A^\dagger$ is the Moore-Penrose pseudo inverse of A.

Suppose now that $n < m$, and A has full rank. The solution exists but is not unique. The solution we would like then is the minimum norm solution. In this case the solution is given by

$$x = x' + \sum_{i=1}^{n} \frac{\langle y, u_i \rangle}{\sigma_i}$$

where $x' \in ker(A)$. Since $n < m$, the $ker(A) = span(v_n + 1, ..., v + m)$. So $x' = Vc$ with $V = [v_n + 1, ..., v + m]$. so the minimum norm solution is when

$x' = 0$, and the solution does not contribute to the $ker(A)$. So the minimum norm solution is

$$\hat{x} = V_n \Sigma_n^{-1} U^* ny = A$$

Is the least squares or minimum norm solution stable? Using the SVD of the pseudo inverse of A we get that

$$\hat{x} = \sum_{i=1}^{r} \frac{\langle x_i, y \rangle}{\sigma_i} v_i$$

where $r = min(m, n)$. We see that the continuity of $\hat{x}$ is depending on the singular values σ_i. If σ_i is small then $\langle u_i, y \rangle$ can be large amplifying the $v_i^t h$ component of y. So it is possible that $v_i^t h$ with small singular values exaggerate the noise of y. So the solution is not continuous. So when is the solution stable?

Definition 2.5. *For $y = Ax$, y satisfies the Picard condition if the Fourier coefficients $\langle u_i, y \rangle$ as derived above decay faster that σ_i, the singular values defined above. That is*

$$\sum_{i=1}^{r} \left| \frac{\langle u_i, y \rangle}{\sigma_i} \right|^2 < \infty$$

2.2 Stabilization

Recall that we can decompose the mean square error of an estimator into it's bias and variance. For x_{LS}, the bias is zero but the variance can be so high that the solution is still ill-posed as we have seen. To lower the variance we can introduce a biased estimator. Ideally we would and up with a lower MSE over all. One way to avoid dividing by large singular values to regularize the $\sigma_i' s$ by with some regularization function $\mathcal{R}_\alpha$, with regularisation parameter α. The regularization solution is then

Definition 2.6.

$$x_\alpha = V_k \mathcal{R}_\alpha(\sigma_k) U_k^* y \tag{7}$$

where $\mathcal{R}_\alpha$ is the regularizing function depending on regularization parameter α.

Possible regularization functions are threasholding functions, such as TVSD, or shifting functions such as Tikhonov regularization.

Definition 2.7. *The Tikhonov regularization solution is*

$$x_\alpha = \sum_{i=1}^{r} \frac{\sigma_i \langle x_i, y \rangle}{\sigma_i^2 + \alpha} v_i \tag{8}$$

where $\mathcal{R}_\alpha(\sigma) = \sigma/(\sigma^2 + \alpha)$, and α is the regularization parameter.

In the above we modify pseudo inverse by adding some weight to the singular values. The denominator is then bounded by α even if $\sigma \to 0$. When $\sigma_i \gg \alpha$ The ratio $\frac{\sigma_i \langle x_i, y \rangle}{\sigma_i^2 + \alpha}$ is relatively unchanged. In the case where $\sigma_i \ll \alpha$, the ratio is decreased, thus overall decreasing the variance $\hat{x}$. The consequence of this is that resulting estimator x_α will be a biased. We can compare the difference between the non-regularized solution $\hat{x}$ and the regularized solutions x_α. This difference displays the bias-variance trade.

Definition 2.8. *Let $\hat{x} = A^\dagger y$ the non-regularized solution with $A^\dagger$ pseudo inverse. Let $\hat{x}_\alpha = A_\alpha^\dagger y$ the regularized solution with $A^\dagger$ regularized pseudo inverse.*

$$||\hat{x} - x_\alpha|| \leq ||(A^\dagger - A_\alpha^\dagger)y|| + ||A^\dagger(y - y^\delta)||$$

The bias is measured as $||(A^\dagger - A_\alpha^\dagger)y||$ and variance is measured as $||A^\dagger(y - y^\delta)||$.

When $\alpha \to 0$, $A^\dagger = A_\alpha^\dagger \implies ||(A^\dagger - A_\alpha^\dagger)|| = 0$. Now that we have a good estimator we would like know how good this estimator is. To do this we can compute the mean squared error as

Definition 2.9. *Let $x = Ay$ be the true parameter. Let $x_\alpha = A_\alpha^\dagger y^\delta$ be the regularized solution. The measure of how close this estimator is to the truth is given by*

$$||x - x_\alpha|| \leq ||x - A_\alpha^\dagger y|| + ||A_\alpha^\dagger(y - y^\delta)||$$

2.3 Tikhonov Regularization Revisited

Above we defined everything in terms of SVD. But there is a variational formulation of Tikhonov regularization that turns solving for x into an optimization problem. Give reason why.

Definition 2.10. *Let $\lambda > 0$ be a a fixed constant. The Tikhonov regularized solution x_λ to (3) $x_\lambda \in \mathcal{X}$ is the minimum of the the functional*

$$\mathcal{R}_\lambda(x) = ||Ax - y||^2 + \lambda ||x||^2 \tag{9}$$

assuming that such a minimizer exists. $\mathcal{R}_\lambda(x) : \mathcal{X} \to \mathcal{Y}$ and λ is called the regularization parameter.

To find the estimate for $\bar{x}$ is now an optimization problem, where we want to minimize $\mathcal{R}_\lambda(x)$ for some fixed λ. We get the following scheme

1. Minimize: $\min_{x \in \mathbb{R}^m} (||Ax - y||_2^2 + \lambda ||Lx||)$. This can also be written as some constrained optimization problem.
2. The solution is $x_e^\lambda = (A^*A + \lambda L^*L)^{-1} A^*y$. Note that if there is no noise in the model, so we need no regularization, then we get back the least-squares solution.

We will denote the functional $\min_{x\in\mathbb{R}^m}(||Ax-y||_2^2+\lambda||Lx||)$ by $\mathcal{J}(x)$, which consist of two portions, the data fidelity term $||Ax-y||_2^2$, and the regularization term $||x||_2^2$. We can check that the problem of minimizing $\mathcal{R}_\lambda(x)$ for a given λ is well-posed problem, by checking that

1. For fixed λ, $\mathcal{R}_\lambda(x)$ is well defined
2. For fixed λ, $\mathcal{R}_\lambda(x)$ is continuous in $\mathcal{Y}$.
3. We can select λ such that if $y \to A(\bar{x})$, then $\mathcal{R}_\lambda(x) \to \bar{x}$.

In this setting, we can check well-posedness by looking at the SVD of the regularized solution. We can find the solution to the minimization problem by writing down the normal equation. We get that

$$x_\alpha = (A^*A+\alpha L)^{-1}A^*y = V_(\Sigma_r^2+\alpha L)^{-1}\Sigma_{U^*y} \tag{10}$$

If the regularization guarantees stability, and $ker\ A \cap ker\ L = \{0\}$, then (12) is a well-posed problem. The estimate x_e^λ, depends on fixed λ, so we must need some method to choose λ such that the solution to the optimization problem is continuous (condition 3 in the above). Common methods to choose are via

1. a-prior rules knowing the noise level
2. Discrepancy principle
3. L-curve
4. Cross validation

Definition 2.11. *Assume we know the noise level, and denote the noise level by e. We can then a-prior choose $\alpha(e)$, the regularization parameter now depending on e. This is called an a-prior rule. This is called convergent if and only if*

$$\lim_{e\to 0}\alpha(e) = 0$$
$$\lim_{e\to 0} e||A^\dagger_{\alpha(e)}|| = 0$$

Claim 2.2. *If the $\alpha(e)$ as defined above is convergent then the total error $||A^\dagger y_e A^\dagger y||_2^2 \to 0$ as $e \to 0$. So we have consistency.*

Definition 2.12. *The discrepancy principle chooses α a-posterior depending on both y_e and e, such that*

$$||AA^\dagger y_e - y_e||_2^2 \le \eta e$$

for $\eta > 1$ fixed. If $y_e \in ker(A^\dagger)$ then no such α can exist.

Definition 2.13. *The L-curve method chooses α heuristically via a minimization problem*

$$\min_{\alpha>0} ||A_\alpha^\dagger||_2^2 ||AA_\alpha^\dagger y_e - y_e||_2^2$$

the optimal α should lie at the corner of the curve $||A_\alpha^\dagger||_2^2 ||AA_\alpha^\dagger y_e - y_e||_2^2$. We do not have consistency with this choice of α.

So far we have seen a a-prior rule, a posterior rule, and a heuristic rule, but we now introduce a more data driven rule, that being choosing α via cross validation. Later on we will see that there is a connection between this rule, and the Bayesian method we describe in this thesis. We would like to somehow choose from the data alone. We also want that as $\sigma \to 0 \to 0$ so $\hat{x}$ converges to x_{LS} should it exist. To do this we will use Bayesian statistics, and frame (3) in the Bayesian framework. With certain choices, we will see that the Bayesian interpretation of (3), is closely related to the Tikhonov regularization we described above.

Finally we remark that

$$\mathcal{R}_\lambda(x) = ||Ax - y||^2 + \lambda||x||^2 \tag{11}$$

for $\in (0, \infty)$ can be written as [13]

$$\mathcal{R}_\lambda(x) = (1-)||Ax - y||^2 + \lambda||x||^2 \tag{12}$$

for $\in (0, 1)$ where here we see that that regularization parameter can be seen as balancing the fit versus the smoothing. At the end points when $= 0$ we get interpolation of the data points, and when $= 1$ we get over-smoothing. We try in the rest of this thesis to make explicit how to choose from the data, and why indeed when is minimized we have also balanced the fit versus the smoothing.

3 Statistical and Probability

Below we review some measure theoretic probability facts and definitions that are necessary for defining Bayes formula for inverse problems. In the second section we recap the definitions of certain distributions as characterized by their probability density functions. We also review some algebra rules for the multivariate normally distributed vectors.

3.1 Probability Theory

Recall that a probability space consists of a sample space Ω, a $\sigma-$algebra $\mathcal{F}$, and a probability measure $\mathbb{P}$. In this thesis we consider only $\sigma-$finite measures, that is measures that are countable unions of finite measure measurable sets. Particularly we use the $\sigma-$ finite Lebesgue measure on $\mathbb{R}^n$. we denote a measurable space as $(S, \mathcal{B}(S))$ where X is some (metric) space and $\mathcal{B}(X)$ is the borel $\sigma-$algebra. In the case that $X = \mathbb{R}$ we know that $\mathcal{B}(X)$ is generated by the

open intervals $(a, b]$ for $a, b \in \mathbb{R}$. We can extend this to $\mathbb{R}^n$. Also recall that random variable,f, is a measurable map $f : \Sigma \to X$ and induces a probability measure X.

Definition 3.1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(X, \mathcal{B}(X))$ a measurable space then the measure μ induced by the random variable $f : \Omega \to X$ is define as*

$$\mu(A) = \mathbb{P}(f^{-1}(A)) = \{\omega \in \Omega \mid f(\omega) \in A\}, A \in \mathcal{B}(X)$$

where μ is the distribution f. We denote this as $f \sim \mu$.

Definition 3.2. *Let μ and ν be measures on a measure space (x, Σ). Then we have the following*

1. *If $\nu(A) = 0 \implies \mu(A) = 0$ for all $A \in \Sigma$, then μ is dominated by ν and we say that μ is absolutely continuous with respect to ν. We denote this as $\mu \ll \nu$*
2. *If $\mu \ll \nu$ and $\nu \ll \mu$ then μ and ν are equivalent.*
3. *Let A and $B \in \mathcal{B}(X)$ be disjoint sets such that $A \cup B = \mathcal{B}(X)$ and $\mu(A) = 0$ while $\nu(B) = 0$, then μ and ν are mutually singular. We denote this as $\mu \perp \nu$.*

We will now also state the Radon–Nikodym theorem which we can use to relate random variables to their probability density functions, as well as proving that the conditional posterior distribution is a unique solution to an Bayesian inverse problem if Bayes rule holds.

Theorem 3.1. *Let (X, Σ) be a measurable space. Then let μ and ν be $\sigma-$finite measures defined on this space. Suppose also that $\nu \ll \mu$. Then there exists a unique up a $\mu-$null set Σ measurable function $f : X \to [0, \infty)$ such that for all $A \subset X$,*

$$\nu(A) = \int_A f d\mu$$

We call f the Radon–Nikodym derivative and denote it as $\frac{d\nu}{d\mu}$.

In the case that the measure space is $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ for some finite n, and $X \sim \nu$ and $\mu = leb(\cdot)$, then by the Radon–Nikodym theorem, f $\in$L$^1(\mathbb{R}^n)$ is the unique is probability density function for $X \sim \nu$.

What follows is a few definitions to define conditional distributions. with these we can precisely write down the posterior distribution as a conditional distribution.

Definition 3.3. *Let $\mathcal{G} \subset \mathcal{F}$ be a sub$-\sigma-$algebra. Let y be a $\mathcal{G}$ measurable function. We call $y : \Omega \to X$ a conditional expectation of a random variable $f : \Omega \to X$ with respect to $\mathcal{G}$ if*

$$\int_{\mathcal{G}} f d\mathbb{P} = \int_{\mathcal{G}} y d\mathbb{P}$$

Definition 3.4. *Let $\mathcal{G} \subset \mathcal{F}$ be a sub$-\sigma-$algebra. The condition probability of $B \in \mathcal{B}(X)$, given $\mathcal{G}$ is*

$$\mathbb{P}(B \mid \mathcal{G}) = \mathbb{E}(1_B \mid \mathcal{G})$$

Definition 3.5. *Let $(\mu(\cdot, \omega))_{\in \Omega}$ be a family of probability distributions on $(X, \mathcal{B}(X))$. Then $(\mu(\cdot, \omega))_{\in \Omega}$ is a regular conditional distribution of f given $\mathcal{G} \subset \mathcal{F}$ if*

$$\mu(B, \cdot) = \mathbb{E}(1_B(f) \mid \mathcal{G}) \ a.s.$$

for all $B \in \mathcal{B}(X)$. If f is as defined about then such a a regular conditional distribution exists.

Remark 3.1. *These theorems are use full when we let $\mathcal{G} = \sigma(y)$ be the sub-$\sigma-$algebra generated by the observations $y = AX + e$ in the Bayesian setting. If we let π_{post} denote the posterior distribution and π_{prior} denote the prior distribution on x, then we have that*

$$\pi_{post}(B, y(\omega) = \mathbb{E}(1_B(x) \mid \sigma(y))(\omega)$$

for $B \in \mathcal{B}(X)$. So then

$$\pi_{post}(B, y) = \pi_{prior}(B \mid y).$$

3.2 Some statistics definitions

Denote random variables by capital letters. For example X,Y, E. Denote realization of these random variables by the corresponding lower case letter for $Y = y$, y is one realization of Y.

Definition 3.6. *Multivariate normal distribution A random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ if its probability density function is give by*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2} \tag{13}$$

for $\mu \in \mathbb{R}$ and $\sigma > 0$.

Now let $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ a non-negative symmetric matrix. Then we can define $X \sim \mathcal{N}(\mu, \Sigma)$ for X a random $n-$ dimensional random vector.

Definition 3.7. *A random vector $X \sim \mathcal{N}(\mu, \Sigma)$ if and only if its probability density function is given by*

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^T (x-\mu)} \tag{14}$$

for parameters $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ symmetric positive definite matrix.

Proof. See Aard van der Vaart chapter 2 lemma 2.3 dictaat. [11] □

Theorem 3.2. *If $c \in \mathbb{R}^n$ and X is an $n-$dimensional random vector such that $X \sim \mathcal{N}(\mu, \Sigma)$. Then $c + X \sim \mathcal{N}(c + \mu, \sigma)$.*

Theorem 3.3. *If X is an $n-$dimensional random vector such that $X \sim \mathcal{N}(\mu, \Sigma)$, and $A \in \mathbb{R}^{m \times m}$ a fixed matrix with rank $m \leq n$, then $AX \sim \mathcal{N}(A\mu, A\Sigma A^T)$ is an m dimensional normally distributed random vector with mean $A\mu$ and covariance $A\Sigma A^T$.*

Remark 3.2. *If $Z \sim \mathcal{N}(0, I)$ and $X \sim \mathcal{N}(\mu, \Sigma)$ then we can write X as and $X = \mu + \Sigma^{1/2} Z$.*

Remark 3.3. *If $X \sim \mathcal{N}(\mu, \Sigma)$ is an $n-$dimensional random vector such that each X_i is independent from X_j for $i \neq j$, then Σ is a diagonal matrix with the variance of each X_i on the diagonal.*

Definition 3.8. *Let $\alpha > 0$, the the Gamma function $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$. A random variable X is Gamma distributed with parameters $\alpha, \beta > 0$ if the pdf is characterized by*

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, x > 0.$$

We denote this as $X \sim Gamma(, \beta)$. [12]

We can extend this definition to a random matrix symmetric X of dimension $p \times p$. In which we have the Wishart distribution.

Definition 3.9. *Let M be a $p \times p$ positive definite matrix. Then the multivariate Gamma function Γ_p for a given α is define as*

$$\Gamma_p(M) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{n}{2} - \frac{j-1}{2}\right)$$

here $|M|$ is the determinate of M and $tr(M)$ is the trace. For X a random variable then, X is Gamma distributed if the pdf is characterized by

$$f_X(x) = \frac{|x|^{(n-p-1)/2} e^{-tr(M^{-1}x)/2}}{2^{\frac{np}{2}} |M|^{n/2} \Gamma_p(\frac{n}{2})}$$

where $n > p - 1$ is the degrees of freedom.

4 Statistical Inverse Problems

We now would like to have a Bayesian interpretation of our model in order to choose the regularization parameter from the data. To do this we will consider y, x, e as random variables and A as some fixed carefully chosen operator. The model (2) can then we written as

$$Y = A(X, E) \tag{15}$$

where X, Y, E are random vectors define on the probability space $\Omega = \Omega_1 \times \Omega_2$ such that $X : \Omega_1 \to \mathbb{R}^m$, $Y : \Omega_2 \to \mathbb{R}^n$, and $E : \Omega \to \mathbb{R}^n$. We want to learn the relation between X, Y, E ie their conditional probability distributions. To do this we can relate X after making observations of Y using Bayes formula. First note again some notation. As in the non-random case Y is the observed/measured data, with $Y = y$ the realization of Y, X is the unknown parameter again with $X = x$ the realization of X, and E is still the noise. We will now consider this noise to be additive and Y, X, E to be random vectors in $\mathbb{R}^n, \mathbb{R}^m$ and $\mathbb{R}^n$ respectively. so we can rewrite (15) as

$$Y = AX + E \tag{16}$$

The solution to the above problem is a conditional posterior distribution for X given $Y = y$. Two things are gained from posing the inverse problem in the Bayesian setting. 1) we can obtain point estimates by computing the most likely value for X which we will describe later, and compute the uncertainty of this estimate by calculating the spread of the posterior distribution. 2) One such estimate, the MAP estimate, connects the non-Bayesian Tikhonov regularization setting to this Bayesian setting where we can then have a better understanding of the regularization parameter and the model's variance. This insight allows us to to choose λ in a data driven way.

4.1 Bayes Formula

On key aspect of Bayesian statistics is that we assume we have prior knowledge of certain parameters in our model. In this setting what we can observe are realizations is Y, and we assume that we have prior knowledge of X, mainly which values of X are occurring and at what frequency. To express this prior knowledge we place a prior distribution on X. We denote this by F_X with Lebesgue's density (since we are in the finite case) π_X. We also assume that $E \sim F_E$ with Lebesgue density π_E and that E is independent of X. Note that we will see later that independence is key here to get the desired likelihood and resulting Tikhonov regularization. With these assumption we can find the likelihood $Y \mid X$ for $X = x$.

Theorem 4.1. *The likelihood* $L(Y = y \mid X = x) = \pi_E(y - Ax)$.

Proof. .... □

Lemma 4.1. $(X, Y) \in \mathbb{R}^m \times \mathbb{R}^n$ *is a random variable with Lebesgue density* $\pi(x, y) = \pi_E(y - Ax)\pi_X(x)$

We now formulate Bayes theorem which will tells us how X is depending on Y.

Theorem 4.2. *Assume that the* $m(y) = \int_{\mathbb{R}}^n \pi_E(y - Ax)\pi_X(x)dx > 0$. *This is called the normalizing constant. Then* $Y = y \mid X = x$ *is a random variable with Lebesgue density* $\pi(x, y) \overset{(rem3.1)}{=} \pi_X(x \mid y) = \frac{1}{m(y)}\pi_E(y - Ax)\pi_X(x)$.

What we can do now is find a conditional probability for $X = x$ given our measurements $Y = y$. We see that the conditional posterior distribution is a product of the likelihood and the prior on X. To understand this we go through some examples.

Example 4.1. *Examples here?*

It is now natural to ask what well-posedness is in this case as the solution is not a point estimator but rather an entire distribution. We can still check if the solution exists,if it is unique and if it is stable. While this is interesting area of research, it is beyond the scope of this thesis, but has been studied in papers [10][2].

Definition 4.1. *Recall that $\pi_E(y - Ax)$ is the likelihood of $Y = y$ given $X = x$. Let $\phi(y; x) = -log(\pi_E(y - Ax))$. Then $\phi(y; x)$ is called the potential function. Note that this is also called the negative log-likelihood.*

We would now somehow like to understand and explore the posterior distribution. We can do this by computing point estimators of the distribution. Such point estimators are explained in the following connection.

Definition 4.2. *The maximum a posterior estimator of x is found by maximizing the posterior distribution if the maximum exists. That is*

$$x_{MAP} = \max_{\mathbb{R}^m} \pi(x \mid y) \tag{17}$$

To compute this point estimator is an optimization problem. Another point estimator is the conditional mean estimator defined as

Definition 4.3. *The conditional mean estimator of x given y is*

$$x_{CM} = \mathbb{E}(x \mid y) = \int_{R^m} x\pi_X(x \mid y)dx$$

To compute this point estimator is an integration problem, which can be very difficult in high dimensional settings.

We can also compute spread estimators by computing Bayesian Credible sets. These are defined as follows:

Definition 4.4. *Let $\alpha \in (0, 1)$, then a $1 - \alpha$ level credible set C_α is given by*

$$\Pi(C_\alpha \mid y) = \int_{C_\alpha} \pi_X(x \mid y)dx = 1 - \alpha$$

To compute this is a root finding problem. We can use these to compute our uncertainty about the true value of $\bar{x}, \alpha$ and β.

4.2 Connection to Tikhonov Regularization

Now that we have seen some examples. We will now explain under which circumstances how we can return to the Tikhonov regularization. We will consider only the independent Gaussian noise model with a normal prior. Suppose that the noise E is additive Gaussian noise with each e_i i.i.d. Suppose also that we assume X is Gaussian, and that X is smooth. We formally write this as

$$E \sim \mathcal{N}(0, \alpha I) \tag{18}$$
$$X \sim \mathcal{N}(0, \beta\Sigma) \tag{19}$$

for fixed α, β, Σ. Using Bayes formula we find that posterior distribution of $X = x \mid Y = y$ to be proportional to

$$\pi(x, \alpha, \beta) \propto \alpha/2||Ax - y||^2 - \beta/2||Lx||^2 + m/2log(\alpha) + n/2log(\beta) \tag{20}$$

and the potential is

$$\mathcal{J}(x, \alpha, \beta)) = \alpha/2||Ax - y||^2 + \beta/2||Lx||^2 - m/2log(\alpha) - n/2log(\beta) \tag{21}$$

The resulting posterior distribution is Gaussian with mean μ and variance Σ. A Gaussian distribution is completely characterized by its mean and variance, so we compute μ and Σ. Do to this we need compute the MAP estimate of $\pi(x \mid \alpha, \beta)$, which is equivalent to minimizing the potential.

$$x_{MAP} = \min_x \mathcal{J}(x, \alpha, \beta)) \tag{22}$$
$$= \min_x \ \alpha/2||Ax - y||^2 + \beta/2||Lx||^2 - m/2log(\alpha) - n/2log(\beta) \tag{23}$$

We can now easily see that this minimization problem is equivalent minimization problem as the Tikhonov regularization with $= \alpha/\beta$. We find that

$$\mu = (\alpha A^*A + \beta L^*L)^{-1}\alpha A^*y \tag{24}$$
$$\Sigma = (\alpha A^*A + \beta L^*L)^{-1} \tag{25}$$

5 Bayesian Regularization

The next natural question is how to choose λ in this setting. We already have some indication for the choice of λ as it is the ratio of the data variance and the observed variance. To chose requires choosing α and β such that the resulting minimization problem is well-posed.

5.1 Empirical Bayesian Method

Our first guess is to use empirical Bayesian method that would allow us to determine α and β from the data only. To do this we minimize the objective

function over all three parameters.

$$(x_{MAP}, \alpha_{MAP}, \beta_{MAP}) = \min_{x,\alpha,\beta} \mathcal{J}(x, \alpha, \beta)) \tag{26}$$

$$= \min_{x,\alpha,\beta} \alpha/2||Ax - y||^2 + \beta/2||Lx||^2 - m/2log(\alpha) - n/2log(\beta) \tag{27}$$

5.2 Well-posedness

We now check that this minimization problem is well-posed.

Claim 5.1. *The empirical Bayesian method is ill-posed.*

Proof. We wish to

$$\min_{x,\alpha,\beta} \alpha/2||Ax - y||^2 + \beta/2||Lx||^2 - m/2log(\alpha) - n/2log(\beta)$$

Suppose that A is invertible. Recall that $y = Ax + e$. Then $||Ax - y|| = 0$, and minimizing $\mathcal{J}$ occurs when $\alpha \to \infty$ and $\beta \to 0$. If $\alpha \to \infty$, and $\beta \to 0$, no regularization occurs. The bias goes to zero, but the variance is high and $\beta \to 0$ then $\widehat{x}$ goes to the zero vector. So the method is not well-posed. On the other hand suppose that A not invertible. Rewrite (27) as

$$||Ax - y||_2^2 + \alpha/\beta||L||_2^2 - m/2log(\alpha) - n/2log(\beta)$$

Then $\mathcal{J}$ is minimized when $\beta \to 8$, and $\alpha \to 0$. Then by recalling that $= \beta/\alpha$, $\to \infty$ so the solution is over smoothed and will have large bias. □

The problem is then that minimization occurs at the bound extreme values for α, β, which we have seen results in ill-posedness.

5.3 Hierarchical Bayesian Method

We have seen that the empirical Bayesian method of regularization is not a well-posed problem. Failure occurred exactly when A is invertable leading to a solution $A^{-1}y = \hat{x}$, so that $\alpha \to \infty$ and $\beta \to 0$. What is need then is bound α and β. To do this place priors on the precision parameters α and β. Since we assume a Gaussian prior on X and E the natural(conjugate) hyper priors are Gamma distributions.

We now add the following hyper priors on the variances

$$\alpha \sim Gamma(a_0, b_0) \tag{28}$$

$$\beta \sim Gamma(a_1, b_1) \tag{29}$$

Then the posterior becomes

$$p(x, \alpha, \beta \mid y) \propto \rho(Ax - y \mid \alpha) \times \pi(\alpha)\pi(x \mid \beta) \times \pi(\beta) \tag{30}$$

$$\propto \alpha^{n/2} e^{-\alpha/2||Ax-y||^2} \alpha^{a_0-1} e^{-b_0\alpha} \beta^{n/2} e^{-\beta/2||Lx||^2} \beta^{a_1-1} e^{-b_1\beta} \tag{31}$$

the objective function is

$$\mathcal{J}(x,\alpha,\beta) = \ell(x,\alpha,\beta \mid y) = \alpha/2||Ax-y||^2 - (n/2+a_0-1)log(\alpha) + b_0\alpha + \\ \beta/2||Lx||^2 - (n/2+a_1-1)log(\beta) + b_1\beta$$

Note that if we let $a_0, a_1 = 1$ and $b_0, b_1 = 0$ the we recover the objective function given no hyper priors.

5.4 Well-posedness

We now check that this addition of hyper priors leads to a well-posed problem. Recall the following two theorems.

Theorem 5.1. *If $\mathcal{J}$ is proper, coercive, bounded from below and lower semi-continuous, then $\mathcal{J}$ has a least one minimizer.*

Example 5.1.

Claim 5.2. *J has a least one minimizer.*

To check if the solution is now stable we again use the optimal solutions. The partial derivatives are

$$\partial/\partial x(J(x,\alpha,\beta)) = (A^*A + \beta/\alpha L^*L)x - A^*y \tag{32}$$

$$\partial/\partial\alpha J(x,\alpha,\beta) = 1/2||Ax-y||^2 - (n/2+a_0-1)/\alpha + b_0 \tag{33}$$

$$\partial/\partial\beta J(x,\alpha,\beta) = 1/2||Lx||^2 - (n/2+a_1-1)/\beta + b_1 \tag{34}$$

so then the optimal solutions are

$$x = (A^*A + \beta/\alpha L^*L)^{-1}A^*y \tag{35}$$

$$\alpha = \frac{(n/2+a_0-1)}{1/2||Ax-y||^2 + b_0} \tag{36}$$

$$\beta = \frac{(n/2+a_1-1)}{1/2||Lx||^2 + b_1} \tag{37}$$

The resulting theorem is by [8], and relies on many intermediate lemmas. We restart the main theorem here and briefly prove the main theorem in the case $L = I$.

Theorem 5.2. *Let σ_0^2 denote the variance. Assume that the random variable η_i is such that $|\eta_i| \leq c_\omega\sigma_0$ (is c_ω the decay rate of the eigenvalues of which matrix?) for $i = 1, ..., n$. Fix b_1 and let $\frac{n}{2} + a_1 - 1 \sim \sigma_0^d$ for $0 < d < 2$. Then*

$$\lim_{\sigma_0 \to 0} ||\hat{x}_- x_{LS}||_2^2 = 0$$

that is as the variance goes to zero the regularization should also go to zero.

Proof. □

6 Numerical Methods

In this section we propose three different iterative methods to numerically solve $\min_{x,\alpha,\beta} \mathcal{J}(x,\alpha,\beta)$. Suppose that we have some function $f(x,y,z)$ and we want to find the minimum of this function. All critical points are then found when $\frac{\partial f}{\partial x} = 0, \frac{\partial f}{\partial y} = 0$, and $\frac{\partial f}{\partial z} = 0$. Suppose f is strictly convex then any local minimum on a convex set is a global minimum and in fact this minimum is unique. A common method to numerically solve for minimum of f is coordinate descent, where we used the fact that in the parameter of interested we are guaranteed to move downward by strict convexity. We can then do coordinate descent to find the minimum.

Example 6.1. *Suppose we would like to minimize*

$$f(x,y) = (x-y)^2$$

The function is minimized along the line $x = y$. We could have also solved this numerically by first noting that $f(x) = x^2 - 2y'x$ and $f(y) = y^2 - 2yx'$ are convex functions with first derivatives equal to $x - y'$ and $y - x'$. Suppose we start at the point $(1,0)$. And suppose that first we want to minimize along x. Then

$$y = argmin_x f(x,0) = argmin_x x^2 = 0$$

Then

$$x = argmin_y f(0,y) = y^2 = 0$$

So the minimum is at $(0,0)$ which is on the line $x = y$.

Now recall that we have computed the partial derivatives to $\mathcal{J}(x,\alpha,\beta)$ in (**??**) and computed the critical points as (**??**). However our function is only bi-(strictly) convex. But we can still implement a similar algorithm to coordinate descent by splitting $\mathcal{J}$ into its two strictly convex parts, and do alternating minimization. The following method was proposed by [8] and we use it as a benchmark to compared to the other two algorithms.

6.1 Method 1

In this method the estimates are found by simultaneous minimize over all three parameters. That is we

$$\min_{x,\alpha,\beta} \mathcal{J}(x,\alpha,\beta) \tag{38}$$

This is done by alternating method, where at each iteration we define either a normal equation for x or a normal equation for α, β. The next best guess for x,

respectively $\alpha\beta$ is then the root of the normal equation. Recall that these roots are the optimal solutions (**??**) to the partial derivatives.

$$\begin{aligned} x(\beta/\alpha) &= (A^*A + \beta/\alpha L^*L)^{-1}A^*y \\ \alpha_(x) &= \frac{(n/2 + a_0 - 1)}{1/2||Ax - y||^2 + b_0} \\ \beta(x) &= \frac{(n/2 + a_1 - 1)}{1/2||Lx||^2 + b_1} \end{aligned}$$

For fixed a_0, b_0, a_1, b_01 , these defined closed form solutions for x, α, β. Suppose, then, that We start with some initial values α_0, β_0. Then using the optimal solution for x we can define an estimate

$$\begin{aligned} x_0 &= x(\beta_0/\alpha_0) \\ &= (A^*A + \beta_0/\alpha_0 L^*L)^{-1}A^*y \end{aligned}$$

Notice that this is the normal Tikhonov regularized solution with regularization parameter β_0/α_0. To find the next estimates for α, β we compute

$$\begin{aligned} \alpha_1 &= \alpha_(x_0) \\ &= \frac{(n/2 + a_0 - 1)}{1/2||Ax_0 - y||^2 + b_0} \\ \beta_1 &= \beta(x_0) \\ &= \frac{(n/2 + a_1 - 1)}{1/2||Lx_0||^2 + b_1} \end{aligned}$$

We can repeatedly alternate between minimizing over x versus minimizing over α, β until some stopping criterion is met. Let I be the number of desired iterations the resulting algorithm is then

Algorithm 1

1. Set $i = 0$. Fix a_0, b_0, a_1, b_1 and choose initial values α_0, β_0
2. Compute the β_0/α_0 regularized Tikhonov solution
$$(A^*A + \beta_k/\alpha_k L^*L)^{-1}A^*y$$
3. Compute
$$\alpha_{i+1} = \frac{(n/2 + a_0 - 1)}{1/2||Ax_i - y||^2 + b_0}$$
$$\beta_{i+1} = \frac{(n/2 + a_1 - 1)}{1/2||Lx_i||^2 + b_1}.$$
Update the regularization parameter by setting $\lambda_{i+1} = \beta_{i+1}/\alpha_{i+1}$
4. If stopping condition is met then x_i is the estimate for x. Else set $i = i+1$ and repeat (2-4).

In [8] the following two theorems are proven for algorithm 1.

Theorem 6.1. *(Theorem 3.1 in [8]) Let $\{x_i, \alpha_i, \beta_i\}_{i\in I}$ be the sequence of estimators generated by Algorithm 1. Then the sequence $\{\mathcal{J}(x_i, \alpha_i, \beta_i)\}_{i\in I}$ converges monotonically.*

Theorem 6.2. *(Theorem 3.2 in [8]) Let $\{x_i, \alpha_i, \beta_i\}_{i\in I}$ be the sequence of estimators generated by Algorithm 1, then this sequence converges to a critical point of $\mathcal{J}(x, \alpha, \beta)$.*

This method works when we can compute the partial derivatives of x, α, β, and have closed form solutions. This is indeed the case when we assume normal prior on x, and Gamma priors on α, β, where by we could recover the Tikhonov regularization with L^2 penalty on x. We now propose two additional methods where we assume we do not have closed form solutions. This way we can still numerically solve $\min_{x,\alpha,\beta} \mathcal{J}(x, \alpha, \beta)$ in case where different prior assumptions lead to different penalties.

6.2 Method 2

Suppose now that we do not have closed form solutions for α, β. Define

$$\mathcal{J}_1(\alpha, \beta) = \mathcal{J}(\alpha, \beta; x) = \alpha/2||A\widehat{x}(\alpha, \beta) - y||^2 - (n/2 + a_0 - 1)log(\alpha) + b_0\alpha + \beta/2||L\widehat{x}(\alpha, \beta)||^2 - (n/2 + a_1 - 1)log(\beta) + b_1\beta$$

for fixed $\widehat{x}(\alpha, \beta)$. First supposing that α, β are fixed we can then minimize $\mathcal{J}$ with

$$x \overset{set}{=} (A^*A + \beta/\alpha L^*L)^{-1}A^*y$$

So the inner minimization fixes x at the Tikhonov estimate for some given initial values of α, β. So want we really want to do is

$$\min_{\alpha,\beta}\left[\min_{x}\mathcal{J}(x,\alpha,\beta)\right]$$

The inner minimization can be done using the closed form solution of x, but to minimize over α, β we must use gradient descent that is

$$\alpha_{i+1} = \alpha_i - \mu\partial_\alpha\mathcal{J}_1(\alpha_i,\beta_i)$$
$$\beta_{i+t} = \beta_i - \mu\partial_\beta\mathcal{J}_1(\alpha_i,\beta_i)$$

Thus to solve the joint minimization problem start with an initial α, β, compute $x(\beta/\alpha)$, then solve for $\widehat{\alpha}, \widehat{\beta}$ by taking one step along the gradient in the direction α, β. Let I be the number of desired iterations the resulting algorithm is then **Algorithm 2**

1. Fix $i = 0$. Fix a_0, b_0, a_1, b_1. Chose step size μ. Choose initial value α_0, β_0.
2. Compute the β_0/α_0 regularized Tikhonov solution x_0.
3. Compute

$$\alpha_{i+1} = \alpha_i - \mu\partial_\alpha\mathcal{J}_1(\alpha_i,\beta_i)$$
$$\beta_{i+t} = \beta_i - \mu\partial_\beta\mathcal{J}_1(\alpha_i,\beta_i)$$

4. Update the regularization parameter by setting $\lambda_{i+1} = \beta_{i+1}/\alpha_{i+1}$
5. If stopping condition is met then x_i is the estimate for x. Else set $i = i+1$ and repeat (2-5).

We now prove the same convergence theorems for algorithm 2.

Theorem 6.3. *Let $\{x_i, \alpha_i, \beta_i\}_{i\in I}$ be the sequence of estimators generated by Algorithm 2. Then the sequence $\{\mathcal{J}(x_i, \alpha_i, \beta_i)\}_{i\in I}$ converges monotonically.*

Theorem 6.4. *Let $\{x_i, \alpha_i, \beta_i\}_{i\in I}$ be the sequence of estimators generated by Algorithm 2, then this sequence converges to a critical point of $\mathcal{J}(x, \alpha, \beta)$.*

6.3 Method 3

Suppose now that we do not have a closed form solution for x. Then define

$$\mathcal{J}_2(x) = \mathcal{J}(x;\widehat{\alpha}(x),\widehat{\beta}(x)) = \widehat{\alpha}(x)||Ax-y||_2^2 + \widehat{\beta}(x)||Lx||_2^2$$

for fixed $\widehat{\alpha}(x), \widehat{\beta}(x)$ and constants a_0, b_0, a_1, b_1 such that $a_0, a_1 \neq 1$ and $b_0, b_1 \neq 0$. Similarly as in method two we compute estimates for x by taking one step along the gradient of $\mathcal{J}$ in the direction of x. Our minimization problem is then

$$\min_x \left[\min_{\alpha,\beta} \mathcal{J}(x, \alpha, \beta) \right]$$

The inner minimization can be solve by using the closed form solution for α, β given some fixed x. For the outer minimization we need to use a gradient method. To do this we compute

$$\nabla_x \mathcal{J}_2(x) = (A^*A + \beta(x)/\alpha(x)L^*L)y - A^T.$$

. And take one step along the gradient there by finding the next best guess for x given α, β

$$x_{i+1} = x_i - \mu \nabla_x \mathcal{J}_2(x_i)$$

where i denotes where we are in the iteration. So to solve the joint minimization problem start with an initial x, then minimize over α, β by computing their optimal solutions, then iterative solve for $\widehat{x}$ and update α, β. Now let I be the number of desired iterations the resulting algorithm is then **Algorithm 3**

1. Fix $i = 0$. Fix a_0, b_0, a_1, b_1. Chose step size μ. Choose initial value x_0.
2. Compute the β_0, α_0 by computing the optimal solution given x_0.
3. Compute
$$x_{i+1} = x_i - \mu \nabla_x \mathcal{J}_1(x_i)$$
4. If stopping condition is met then x_i is the estimate for x. Else set $i = i+1$ and repeat (2-4).

We now prove the same convergence theorems for algorithm 3.

Theorem 6.5. *Let $\{x_i, \alpha_i, \beta_i\}_{i \in I}$ be the sequence of estimators generated by Algorithm 3. Then the sequence $\{\mathcal{J}(x_i, \alpha_i, \beta_i)\}_{i \in I}$ converges monotonically.*

Theorem 6.6. *Let $\{x_i, \alpha_i, \beta_i\}_{i \in I}$ be the sequence of estimators generated by Algorithm 3, then this sequence converges to a critical point of $\mathcal{J}(x, \alpha, \beta)$.*

7 Implementation

7.1 Example

In this section we show the resulting implementation in python and remark on an our observations. The inverse problem we are interested in is to recovering

$sin(x)$ with $x \in [-4\pi, 4\pi]$. We observe

$$y = A(sin(x)) + \epsilon \tag{39}$$

where

$$Af(t) = \int_0^1 \frac{f(y)}{((1 + (t - y)^2)^3/2)} dy \tag{40}$$

for our implementation we discretize the the interval $[-4\pi, 4\pi]$ into n even spaced points, and discretize A by letting the step size be equal to m. The resulting forward problem $A(sin(x))$ is an m system of equations with n variables which is well defined as the number of columns of A equals the number of rows of $sin(x)$. We observe y with random Gaussian noise center at 0, and variance σ. In our experiments we let $n = m = 200$, and the noise be distributed $\mathcal{N}(0, 0.1)$.

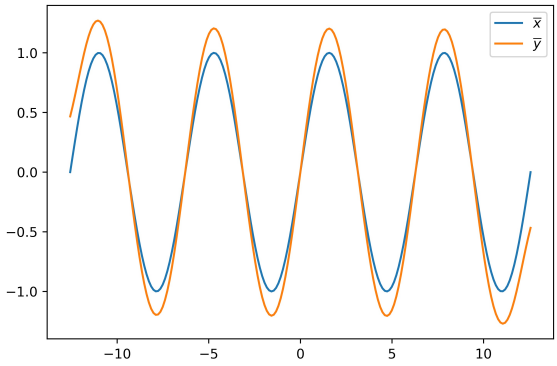

Figure 1: Ground Truth $\bar{x}$, where $\bar{y} = A\bar{x}$

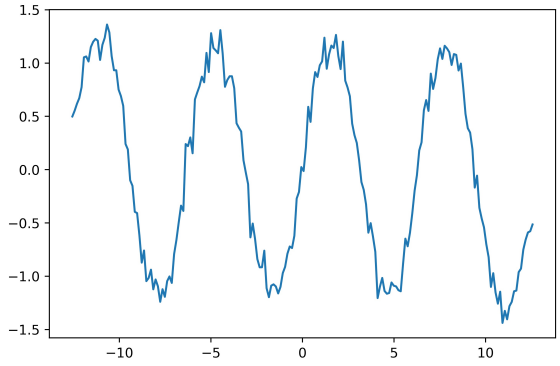

Figure 2: Noisy observation with noise level $\mathcal{N}(0, 0.1)$

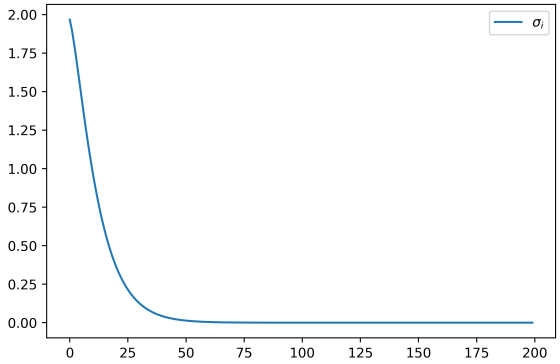

Figure 3: Decay of Singular Values of A

7.2 Ill-posedness

Our method is for regularizing the ill-posed inverse problems. In this section we looked at the ill-posedness of finding the least squares estimate. The conditioning number of A is $4887979232 \approx 10^{9.689} >> 1$ which is much larger than 1. So A is ill-conditioned. We now examine the decay of the singular values of A. We can see that the eigenvalues quickly decay to zero. Now we examine the Picard condition. Where we see that the Picard condition is not met (4), so the problem is ill-posed. If we were to ignore the ill-posedness and directly invert A, which we can do since A is full rank and $det(A) \neq 0$. The resulting solution is the least squares solution which we plot below (5). We see that this solution has high variance and does not accurately recover $\bar{x}$. We note that a majority of the variance is around the peaks (why?).

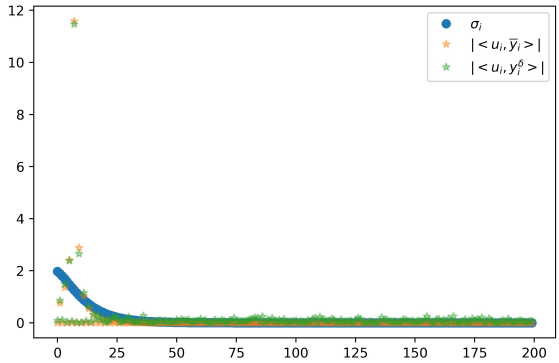

Figure 4: Picard Condition

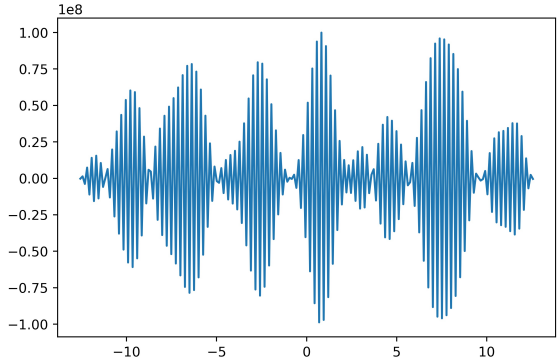

Figure 5: Least Squares solution

7.3 Regularization

Since the problem of recovering $\bar{x}$ directly is ill-posed we instead solve the well-posed problem

$$\min_{x,\alpha,\beta} \mathcal{J}(x,\alpha,\beta) = \ell(x,\alpha,\beta \mid y) = \alpha/2||Ax-y||^2 - (n/2+a_0-1)log(\alpha)+b_0\alpha+ \\ \beta/2||Lx||^2 - (n/2+a_1-1)log(\beta)+b_1\beta$$

using the three proposed methods in section 6.1. The stopping condition is the first order condition

$$||\partial_x \mathcal{J}||_2^2 + ||\partial_\alpha \mathcal{J}||_2^2 + ||\partial_\beta \mathcal{J}||_2^2 < tol$$

The parameters were set to Recall that $\mathcal{J}(x,\alpha,\beta)$ is convex in α,β. Below we fix x and plot the contour plots. We see that the contour lines are slight non-circular from the non-linearity of our problem. The dashed line is the line of

tol	$1e-3$
max iter	100,000
n	200
$\boldsymbol{\alpha_{initial}}$	10
$\boldsymbol{\beta_{initial}}$	1
$\boldsymbol{x_{initial}}$	$\kappa = \frac{\|\|A*y^{\delta}\|\|_2^2}{\|\|AA^*y^{\delta}\|\|_2^2}, x = \kappa A^*y^{\delta}$
$\boldsymbol{a_0 = a_1}$	$1+1e-6$
$\boldsymbol{b_0 = b_1}$	$1e-6$

Table 1: Initial parameters

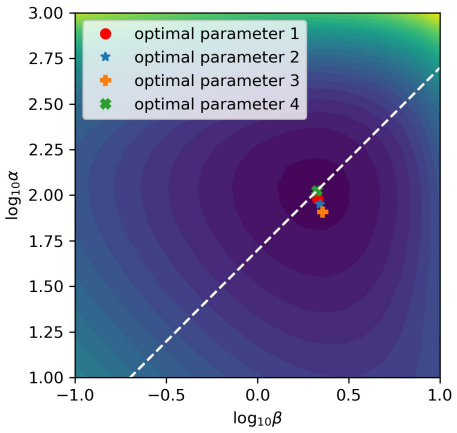

Figure 6: Contour plots of $J(\widehat{x}(\alpha,\beta),\alpha,\beta) = z$. Optimal parameter found by all four algorithms.

steepest decent, and the red dot approximates the minimum of $\mathcal{J}(\widehat{x}(\alpha,\beta),\alpha,\beta)$. Finally, taking advantage of the Bayesian setting, we can see these contour plots as highest posterior density confidence intervals for α,β given fixed x. Below we plot the results of running all 3 Algorithms. We ran a fourth algorithm algorithm that was a modification of method 1. We replaced the close form solution to x in method 1 with that of a gradient method. In all four algorithms we see that the graph of $\mathcal{J}$ is monotonically decreasing (fig 7). In Algorithm 1 and 4, converged based on our stopping criterion, and we can see from the graphs that after 3 iterations the graph at $\mathcal{J}$ is flat. We see a similar result in the right hand side plot of figure 7) for Algorithm 2 and 3. However these did not converge after the maximum number of iterations. We can also see that the slop of $\mathcal{J}$ in Algorithm 1 and 4 is very steep in comparison to that of Algorithm 2 and 3. We note that the graph of objection function from Algorithm 2 seems to reach a lower value than Algorithm 3 will ever reach. In the upper plots of figure 8 we see that all estimates of $\bar{x}$ are pretty close to the ground truth, and

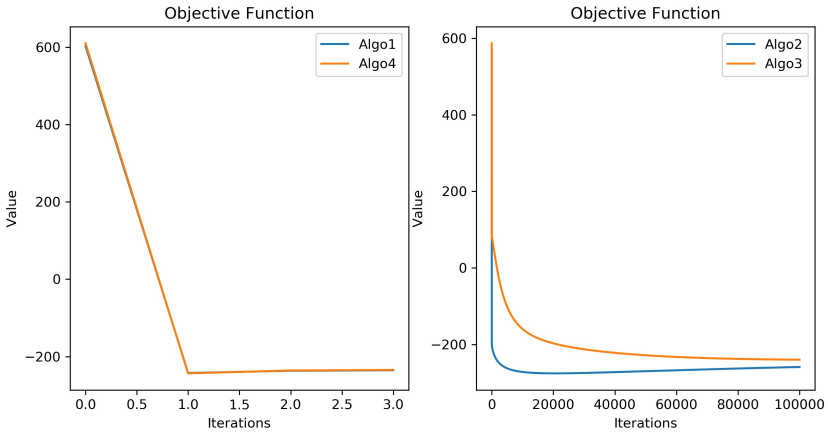

Figure 7: Plot of Objective function over all iterations. On left we plot results from Algorithm 1 and 4. On the right we plot the results of Algorithm 2 and 3.

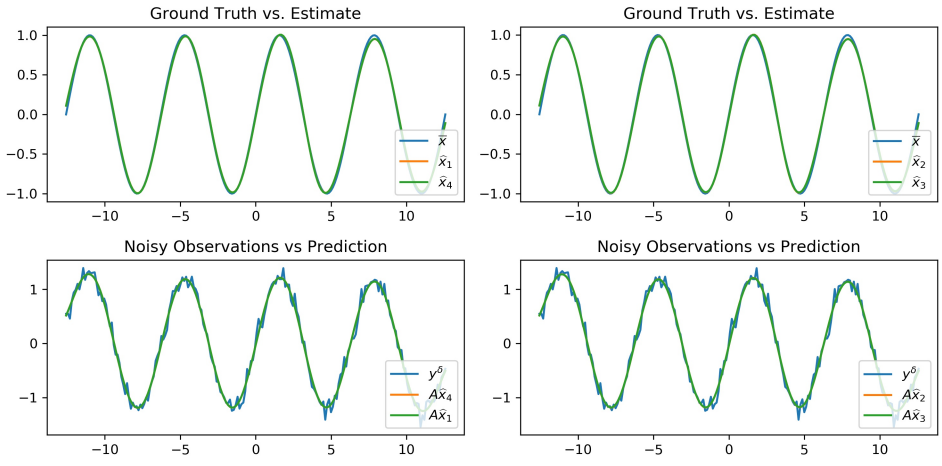

Figure 8: Top plots are of $\widehat{x}$ found by each algorithm versus the ground truth $\bar{x}$. Bottom plots compare the noisy observations versus $A\widehat{x}$ for each Algorithm. On the left we compare Algorithm 1 and 4, and on the right we compare Algorithm 2 and 3.

	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\lambda}$	$\boldsymbol{\mathcal{J}}(x, \alpha, \beta)$	$\boldsymbol{\|\|\overline{x} - \widehat{x}\|\|_2^2}$	**niter**
Algo1	95.61	2.13	0.02	-235.48	0.16	3
Algo2	89.05	2.20	0.02	-258.91	0.21	100000
Algo3	81.06	2.28	0.03	-239.90	0.70	100000
Algo4	105.58	2.12	0.02	-234.57	0.21	3

Table 2: Results

even hard to distinguish among each other. We also see that each estimate has much lower variance that that of the least squares estimate (fig. 5). In table 2 we summarize the all of our results. Overall all algorithms 1,2, and 4 found roughly the same regularization parameter. Algorithm 4 found a sightly higher regularization parameter and resulted in a high error.

7.4 Convergence and Consistency

In the this section we examine consistency and convergence of Algorithm 1, since this was the only one that converged and converged relatively quickly. We proposed that.... We see that as the noise level of the model goes to zero, so

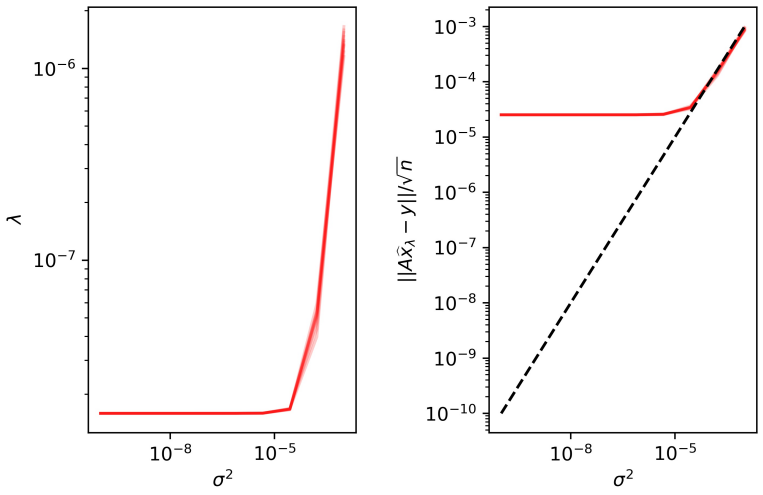

Figure 9: Regularization vs noise and Estimate of residuals vs noise

does the regularization. We also see the effects of bounding β/α in the plot (fig 9) on the right hand side where at some level of noise Algorithm 1 slightly over estimates the noise level as it is bounded away from zero.

$b_0 = b_1$	α	β	λ	$\mathcal{J}(x, \alpha, \beta)$	$\|\|\overline{x} - \widehat{x}\|\|_2^2$	niter
$1e4$	0.01	0.01	1.00	47086.61	14.94	2
$1e2$	0.90	0.84	0.94	156.39	13.81	5
$1e1$	49.98	2.15	0.04	-267.60	0.25	3
$1e-2$	104.43	2.12	0.02	-235.70	0.21	3
$1e-4$	105.57	2.12	0.02	-234.58	0.21	3
$1e-6$	105.58	2.12	0.02	-234.57	0.21	3
$1e-8$	105.58	2.12	0.02	-234.57	0.21	3

Table 3: Results of varying hyper priors $b_0 = b_1$

7.5 Sensitivity

Varying b_0, b_1 the variance parameters to the hyper priors does not seem to effect convergence. We see that too large $b_0 = b_1$ leads to over regularization, but for $b_0 = b_1 < 1e-4$ we see little change. It is strange that in the first two cases $\mathcal{J}$ is minimized at a high positive value [figs 12, 13]. It does seem to be the case that $a_0 = a_1$ can be chosen more freely,([8] pages 16-18), in that small or large values lead to similar regularization and convergence. However the again large values for $a_0 = a_1$ lead to a high value positive of $\mathcal{J}$ at the critical point. It seems we can find another critical point?

8 Discussion

9 Appendix

9.1 Plots of Estimators

They are not all monotonic....

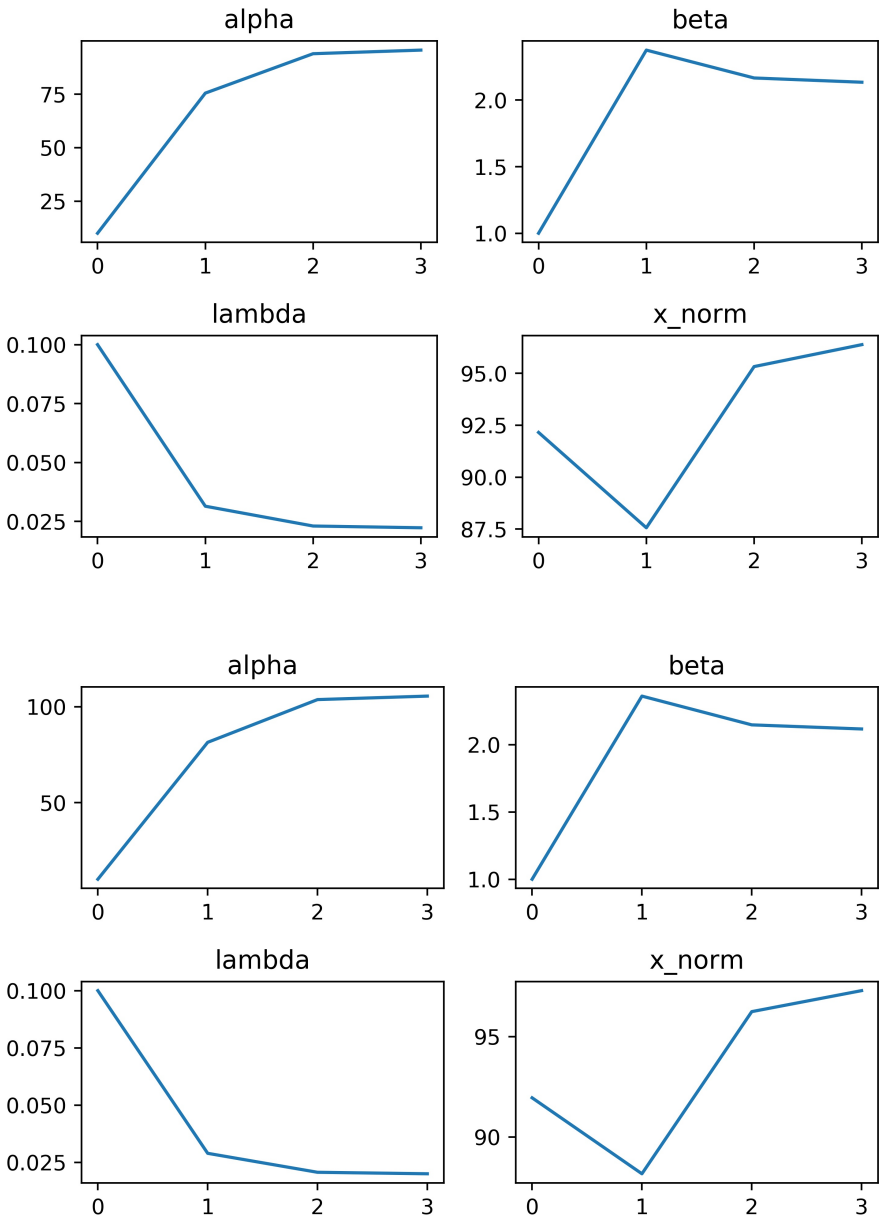

Figure 10: Convergence of estimators Algorithm 1 and 4

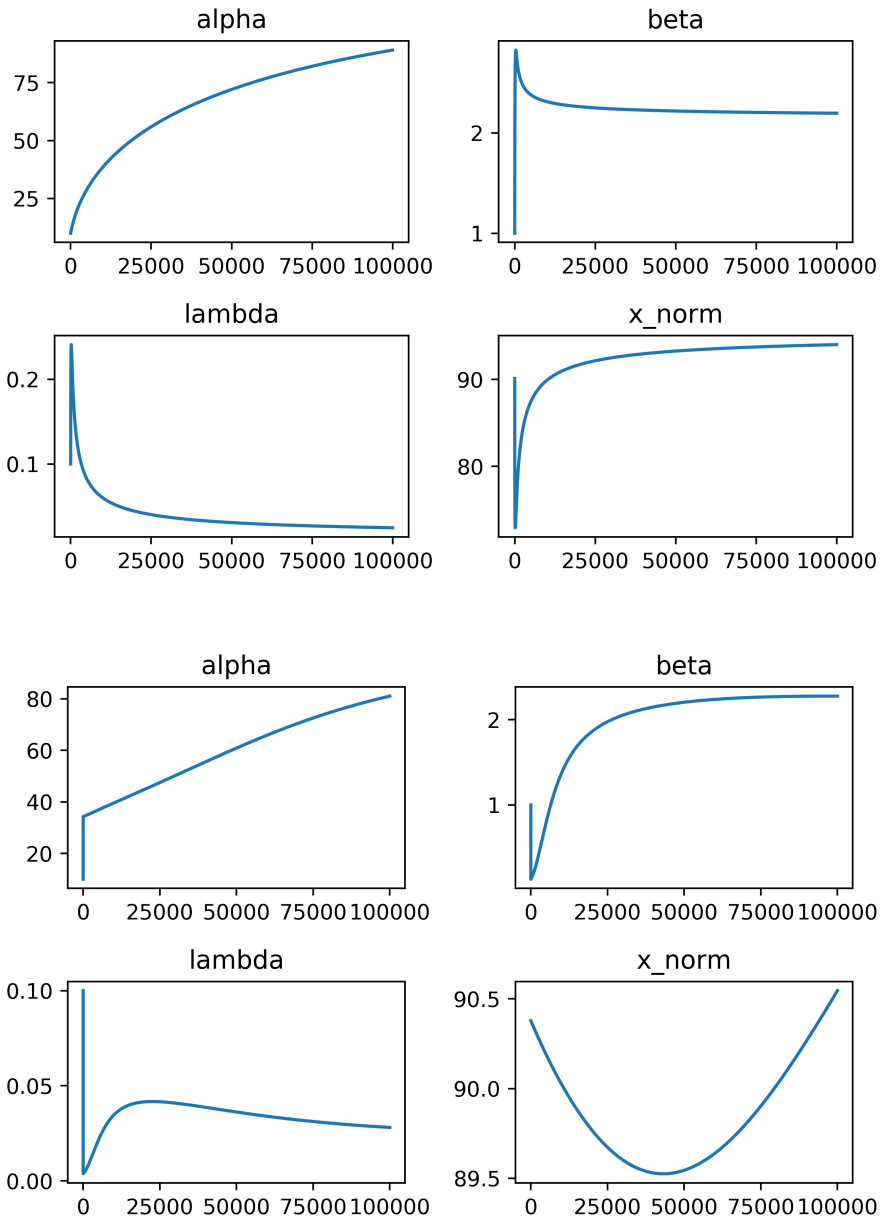

Figure 11: Convergence of estimators Algorithm 2 and 3

9.2 Sensitivity Plots

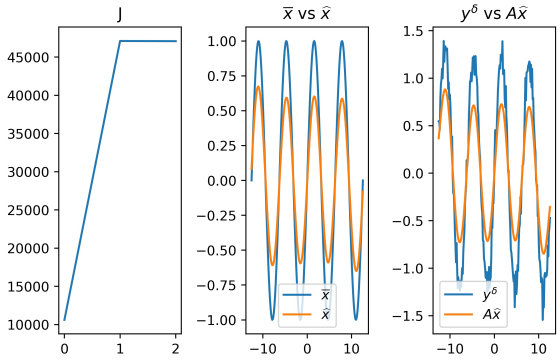

Figure 12: Results when $b_0 = b_1 = 1e4$

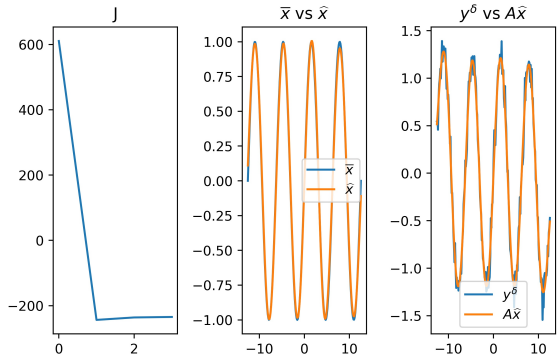

Figure 13: Results when $b_0 = b_1 = 1e - 8$

References

[1] S S Antman et al. *Statistical and Computational Inverse Problems*. Vol. 160. 2004.

[2] Simon Arridge et al. "Solving Inverse Problems Using Data-driven Models". In: *Acta Numerica* 28 (May 2019), pp. 1–174. ISSN: 14740508. DOI: 10.1017/S0962492919000059.

[3] Masoumeh Dashti and Andrew M. Stuart. *The Bayesian Approach to Inverse Problems*. June 2017. DOI: 10.1007/978-3-319-12385-1_7.

[4] Matthias J Ehrhardt and Lukas F Lang. *Inverse Problems*. 2018.

[5] Andrew Gelman et al. *Bayesian Data Analysis CHAPMAN HALL/CRC Texts in Statistical Science Series Series Editors Analysis of Failure and Survival Data*. 2014.

[6] Christian Hansen. "The Discrete Picard Condition for Discrete Ill-Posed Problems". In: *BIT* 30 (1990), pp. 658–072.

[7] Engl Hienz, Hanke Martin, and Andreas Neubauer. *Regularization of Inverse Problems (Mathematics and Its Applications)-Springer (1996)*. Vol. 1. 1996.

[8] Bangti Jin and Jun Zou. "Augmented Tikhonov Regularization". In: *Inverse Problems* 25 (2 2009). ISSN: 02665611. DOI: 10.1088/0266-5611/25/2/025001.

[9] Ali Mohammad-Djafari. *A Full Bayesian Approach for Inverse Problems*. 2001.

[10] A. M. Stuart. "Inverse problems: A Bayesian Perspective". In: *Acta Numerica* 19 (May 2010), pp. 451–459. ISSN: 09624929. DOI: 10.1017/S0962492910000061.

[11] A W Van Der Vaart. *Mathematische Statistiek*. 1997.

[12] Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. 2004.

[13] Wessel N. van Wieringen. *Lecture notes on ridge regression*. 2021. arXiv: 1509.09169 [stat.ME].

[14] Stephen J. Wright. "Coordinate Descent Algorithms". In: *Mathematical Programming* 151 (1 June 2015), pp. 3–34. ISSN: 14364646. DOI: 10.1007/s10107-015-0892-3.