On Non-Uniqueness of the Student's t-formulation for Linear Inverse Problems

Aleksandr Y. Aravkin[*], *Tristan van Leeuwen*[*], *Kenneth Bube*[†], *and Felix J. Herrmann*[*]
[*]*University of British Columbia, Earth and Ocean Sciences,* [†]*University of Washington, Mathematics*

SUMMARY

We review the statistical interpretation of inverse problem formulations, and the motivations for selecting non-convex penalties for robust behaviour with respect to measurement outliers or artifacts in the data. An important downside of using non-convex formulations such as the Student's t is the potential for non-uniqueness, and we present a simple example where the Student's t penalty can be made to have many local minima by appropriately selecting the degrees of freedom parameter.
On the other hand, the non-convexity of the Student's t is precisely what gives it the ability to ignore artifacts in the data. We explain this idea, and present a stylized imaging experiment, where the Student's t is able to recover a velocity perturbation from data contaminated by a very peculiar artifact — data from a different velocity perturbation. The performance of Student's t inversion is investigated empirically for different values of the degrees of freedom parameter, and different initial conditions.

INTRODUCTION

Many geophysical inverse problems can be formulated using an optimization approach, where a misfit measure between predicted and observed data is minimized to obtain a set of parameters of interest that explains the data. Time travel tomography and full waveform inversion are two important examples. Both of these problems can be written as

$$\min_x h(d - F(x)) \,, \tag{1}$$

where x are the parameters of interest, d is observed data, F is the forward model, and h is the misfit measure.

Many convex misfit measures h have been proposed and compared, including the classic 2-norm (Tarantola and Valette, 1982), as well as 1-norm, Huber (Guitton and Symes, 2003; Brossier et al., 2010), and hybrid (Bube and Langan, 1997; Bube and Nemeth, 2007) measures. The non-convex Cauchy penalty has also been proposed and tested by (Crase et al., 1990; Amundsen, 1991), and more recently, the penalty derived from the Student's t distribution has been successfully used in the context of full waveform inversion (Aravkin et al., 2011c,b).
The major motivation for alternate measures to the 2-norm is robustness to outliers (Claerbout and Muir, 1973; Tarantola, 2002; Huber and Ronchetti, 2009). Recall that the inverse problem (1) corresponds to the statistical model

$$d = F(x) + \varepsilon \,, \tag{2}$$

where ε is a random vector with density $p(\cdot) \propto \exp(-h(\cdot))$. In this way, the standard 2-norm approach corresponds to an assumption of Gaussian noise on ε, while using the 1-norm penalty corresponds to exponentially distributed ε. The statistical interpretation helps to understand how the solution of (1) behaves for different h. When $h(x) = \|x\|^2$, the tails of the model error distribution ε decay to 0 at the rate $\exp(-\|x\|^2)$, forcing the solution of (1) to avoid large residuals at all costs. The 1-norm, Huber, and Hybrid penalties all decay to 0 at the asymptotic rate $\exp(-\|x\|_1)$, allowing occasional residuals to stay large.

The natural extension to this methodology is to assume that ε has a heavy tailed distribution, and solve the resulting optimization problem. Cauchy and Student's t distributions, for example, have tails that decay to 0 at a polynomial rate, which means they may allow more and larger residuals than the convex penalties, making them attractive for cases with large unexplained artifacts in the data (Aravkin et al., 2011a).

However, it is often a concern from an optimization point of view to move from a convex formulation to a non-convex one. To be precise, problem 1 is guaranteed to be convex when F is linear, and h is convex (e.g. h is the Huber, Hybrid, 2-norm, or 1-norm). For a nonlinear F, such as in the FWI case, 1 is nonconvex even with convex h, but one may still wonder about introducing further nonconvexity in h.

In this note, we consider the issues of non-uniqueness for the Student's t, focusing on a linear model F. We review the Student's t-formulation, discuss the role of the *degrees of freedom* parameter of the Student's t, and show that examples can easily be constructed for linear models F which exhibit multiple local (or global) minima. We then explore the issue of non-uniqueness in practice with a stylized imaging experiment for a perturbation of a constant background velocity. In the experiment, a significant portion of the data is not only wrong, but is in fact consistent with a *different* perturbation. We investigate the role of the degrees of freedom in the behavior of the Student's t-solution, and compare to the least squares solution.

STUDENT'S T-FORMULATION

The basis of the Student's t-formulation is simply to assume that each component of the error ε in Eq. 2 has the Student's t-distribution, with scalar density given by

$$p(\varepsilon_i) = \frac{\Gamma((k+1)/2)}{\sigma\sqrt{\pi k}\Gamma(k/2)} \left(\frac{1}{k\sigma^2}\left(k\sigma^2 + \varepsilon_i^2\right)\right)^{-(k+1)/2} \,, \tag{3}$$

where k is the *degrees of freedom* parameter, and σ is a scale parameter. Note that Eq. 2 is an extension of the classical variant of the Student's t, which is recovered by setting $\sigma = 1$. The inclusion of the scale parameter is particularly important for geophysical applications, where the scale can vary significantly. In the stylized imaging experiments we present, σ is on the order of 3×10^5, for example. In this paper, we treat the scale as fixed and known, and focus on the variation in the degrees of freedom parameter k. We discuss how we obtained the value of σ when we present the experiments.

Taking the negative likelihood, and dropping the part of the objective that does not depend on ε_i, we get a penalty

$$\log(k\sigma^2 + \varepsilon_i^2)\,. \tag{4}$$

This penalty is not convex; in fact the second derivative is given by

$$\frac{2(k\sigma^2 - x^2)}{(k\sigma^2 + x^2)^2}\,,$$

so the function is convex on the interval $(-\sigma\sqrt{k}, \sigma\sqrt{k})$, and concave outside of this interval. The choice of k is clearly important to the behavior of the penalty. For example, if k is large enough so that every initial residual falls within the convex range, the resulting inverse problem is strictly convex. This example makes the Student's t useless from a modeling perspective, since the entire benefit of the approach comes from diminishing the effect of 'outliers' on the parameters x; if every residual is in the convex range, this means there are no outliers. On the other hand, if k is taken so small that most or all residuals fall outside of the convex range, the problem is likely to be highly nonconvex with many local minima. This situation corresponds to the unintuitive case where every residual is an 'outlier'.

Intuitively, one would like to pick a k that would distinguish between inliers and outliers, keeping the inliers in the convex range $(-\sigma\sqrt{k}, \sigma\sqrt{k})$, and pushing the outliers out. This suggests automated approaches for fitting for k, or continuation methods in k. While we do not present these strategies here, we provide examples illustrating non-uniqueness of the Student's t for simple examples, and practical effect of k on the solution of a linear inverse problem.

NON-UNIQUENESS OF THE STUDENT'S T

To better understand the effect of the degrees of freedom parameter k on the behaviour of the model, consider a simple stylized example where we try to recover the mean parameter, which we take to be 0 for simplicity, from a set of observations using only 6 measurements $b = [\pm 1, \pm .6, \pm .2]$, and a scale parameter $\sigma = 1$. In Figure 1, we simply plot the objective function

$$f(x) = \sum_{i=1}^{6} \log(k + (x - b_i))$$

for three values of k: $0.35, 0.15, 0.025$. The effect is clear - for high enough k, the function is convex, and the global minimum is at 0; if we bring the k down, the function can become extremely non-convex, with many local minima. Note that something very interesting happens - all of the local minima lie close to the available measurements, and 0 is a local maximum! When you drive the k down far enough, the objective 'believes' that the right answer is near available data, ignoring the data that is far away. This understanding clarifies both the advantages and the disadvantages of the Student's t formulation — we have the capability to 'ignore' bad data, but also the potential to get stuck at local minima. The degrees of freedom parameter k provides a tuning nob, which ideally would be adjusted so that the 'good' data residuals are captured by the convex region, while the 'bad' residuals, i.e. those corresponding to artifacts in the data we want to ignore, are in the tails.

ADVANTAGES OF THE STUDENT'S T

So far we have focused on the problems of the Student's t formulation that arise because of the non-convexity. Before we present the numerical experiments, we present a simple explanation of the advantage of non-convex formulations.

Consider a situation where the inverse problem 1 with n measurements has been solved, and a 'good' solution $\bar{x}$ has been found, with

$$\bar{x} = \arg\min \sum_{i=1}^{n} \rho(b_i - F_i x)\,.$$

where we assumed that $F(x) = Fx$ with rows F_i, and the objective h in Eq. 1 is differentiable and can be written $h(b - Fx) = \sum \rho(b_i - F_i x)$, which is the case for all penalties we consider.

If the problem has been solved, the gradient at the solution must be 0. Now, suppose a bad measurement b^{n+1} is added to the set of measurements b, together with the corresponding measurement row F_{n+1}. This simply adds another term to the sum above, and at the current solution $\bar{x}$, the gradient of the objective becomes

$$-F_{n+1}^{*} \rho'(r_{n+1})\,,$$

where $r_{n+1} = b_{n+1} - F_{n+1}\bar{x}$. The new measurement therefore tries to pull the solution away from $\bar{x}$ in the direction $-F_{n+1}^{*}$, and the strength of the pull is determined by $\rho'(r_{n+1})$. This explanation is closely related to the notion of *influence functions* in statistics. For the least squares penalty, $\rho'(r_{n+1}) = r_{n+1}$, so the larger the residual, the stronger its pull. In contrast, for the Student's t penalty, $\rho'(r_{n+1}) = \frac{2r_{n+1}}{k\sigma^2 + r_{n+1}^2}$, so the strength of the pull *decreases* with the size of the residual. As follows from (Aravkin et al., 2011a,Theorem 2.1), the least effect any convex penalty can achieve is a constant level of pull for large residuals. This property motivates the use of non-convex methods in cases where there are many large residuals, e.g. when there are large unexplained artifacts in the data.

NUMERICAL EXPERIMENT

To illustrate the ability of the Students t penalty to ignore bad data, and to investigate the effect of the degrees of freedom parameter on inversion performance and non-uniqueness issues, we consider the following stylized experiment on a domain of 1 km by 1 km. The forward model F is the linear Born scattering operator for a constant velocity of 2000 m/s, with 10 equispaced sources and receivers on the left and right edges of the domain, respectively. We model data for two circular perturbations, A and B, with a radius of 100 m, centered around z=8 km, x=4 km and z=2 km, x=6 km, respectively (shown in figure 2 (a,b)). Our goal is to image perturbation A from contaminated data of which 75 % of the data points correspond to perturbation A and 25 % of the data points correspond to perturbation B. The modeling operator is implemented in the frequency-domain and we use an L-BFGS method to solve the

corresponding optimization problem. All results are obtained with 30 iterations. The reconstruction using the least-squares penalty is shown in figure 2 (c). The reconstruction is very noisy and an artifact appears at the location of perturbation B.

Figures 3 (a-c) show the Students t reconstruction for various degrees of freedom. The scale parameter σ was taken to be the infinity norm of the initial residual (largest absolute residual). As a result of this choice, $k = 1$ is the largest degrees of freedom parameter that we tried, because given the choice of scale, it corresponds to *all* of the initial residuals lying in the convex region of the penalty. We see that for low values of k, we get a good reconstruction of perturbation A. For higher values of k the penalty becomes effectively convex and the reconstruction tends to look like the least-squares reconstruction. This is as expected, since one can show using simple calculus that the Students' t density approaches the Gaussian density pointwise as k goes to infinity.

It is worth noting that we did not encounter non-uniqueness issues for any of the k values we used when starting our inversion from the background velocity estimate. This does not mean the issue cannot occur - it is likely that taking k many orders of magnitude smaller would give behaviour similar to that illustrated in Figure 1.

To better investigate non-uniqueness issues, we also performed the inversion starting the optimization from model A. The results are shown in figure 4 (a-c). For $k = 10^{-4}$, the optimization stalls at the initial model. In this case, the data for model A are treated as the outliers! For a higher value of k, we image perturbation A and the optimization tries to get rid of perturbation B by putting lower velocities around it. In fact, it looks to be converging to the least squares solution one gets when starting from the wrong model (see figure 3 (d)). Note that the forward model F in this formulation is known to have a null space, and so it is not surprising that the starting point has such a strong effect on the solution. The non-uniqueness of the inverse problem formulation plays a stronger role than the non-uniqueness of the Student's t-formulation!

The experiments demonstrate the advantage of the Student's t—it is possible to obtain good reconstructions from noisy data, given that we have a reasonable value of k. The local minima inherent in the formulation do not seem to be a real issue in the actual performance of the imaging example. We had to start from the wrong model in order to find the 'wrong' solution, which was consistent with the particular type of artifact that we generated.

The natural question is, how do we estimate k? Ideally, k would be able to separate the outliers from the inliers at the *true* solution, but in this example, it appears possible to obtain a guess of k using even the residuals at the *initial* model. To get some intuition, we consider the cumulative distribution of the scaled initial residuals, shown in figure 5 (a). Based on these plots we can choose a k for which a certain percentage of the residuals lie within the convex region of the penalty function. The Student's t penalty function as well as its derivative (the influence function) for various k are depicted in figure 5 (b,c). These figures clearly illustrate the behaviour of the penalty. The convex region of the penalty is the region where the derivative is strictly increasing. For the zero-guess initial residual, we see that roughly 50 % of the residuals are smaller than 0.1. Choosing $k = 10^{-2}$ ensures that roughly 50 % of the residuals lie within the convex region of the penalty. If too much of the residuals lie within this region, the reconstruction tends to the the least-squares reconstruction. On the other hand, if too little of the data lies within the convex region, we might get stuck in a local minimum. In practical applications we might have some idea about the percentage of contaminated data and we can use this to choose k.

DISCUSSION

We have reviewed statistical motivation for using robust penalties for inverse formulations, and discussed the advantages and disadvantages of non-convex formulations. In particular, we have clearly demonstrated that the Student's t-formulation can have many local minima, and that the 'bad' behaviour depends on the selected degrees of freedom parameter k — ideally, there is a majority of 'good' data at the final solution in the convex region of the selected penalty.

We have also demonstrated, using a stylized imaging experiment based on a linear Born scattering model, that it takes some work to observe the effects of non-uniqueness in practice. In particular, we had to start close to a synthetically created wrong solution that was consistent with the outlier data in our model, and pick a very small degrees of freedom parameter k. For a wide range of k, the solutions obtained by the Student's t, even starting from the wrong solution (perturbation A), look a lot like the least squares solution obtained when starting from A. When starting from the background model only, the Student's t has a clear advantage over least squares, and recovers a better model.

The relationship of the performance of the Student's t inversion to the chosen degrees of freedom parameter k lead to an interesting line of inquiry. We saw that it is possible to obtain a reasonable value by looking at the cumulative distribution of the scaled residual (even at the initial point!) and choosing k such that a large portion of the data lies within the convex region of the penalty. This suggests that automated schemes for selecting k, or continuation methods in k, may improve performance and keep the inversion in a desired regime. The development of such schemes is left to future work.

ACKNOWLEDGEMENTS

This work was in part financially supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant (22R81254) and the Collaborative Research and Development Grant DNOISE II (375142-08). This research was carried out as part of the SINBAD II project with support from the following organizations: BG Group, BGP, BP, Chevron, ConocoPhillips, Petrobras, PGS, Total SA, and WesternGeco.

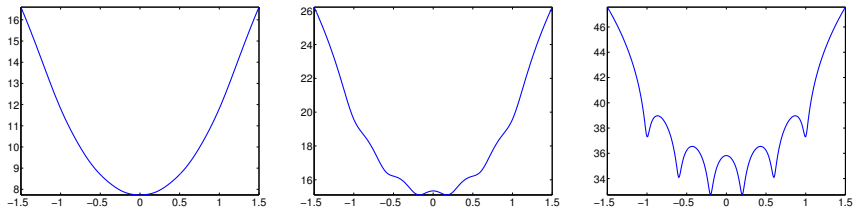
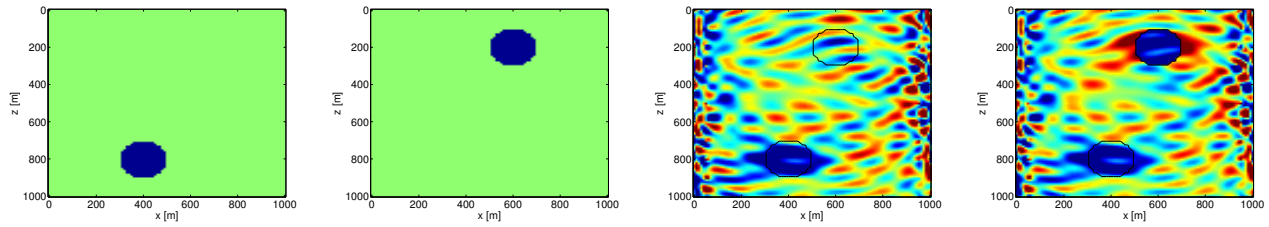
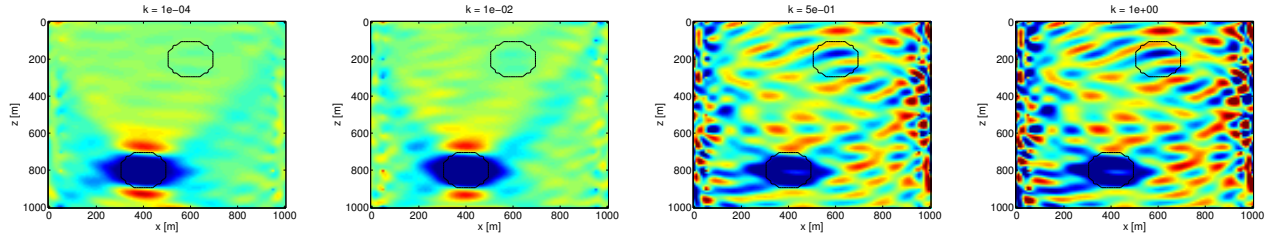


Figure 1: Student's t-penalty for with 6 measurements for different k values. Left to right: $k = 0.35$, $k = 0.15$, and $k = 0.025$.



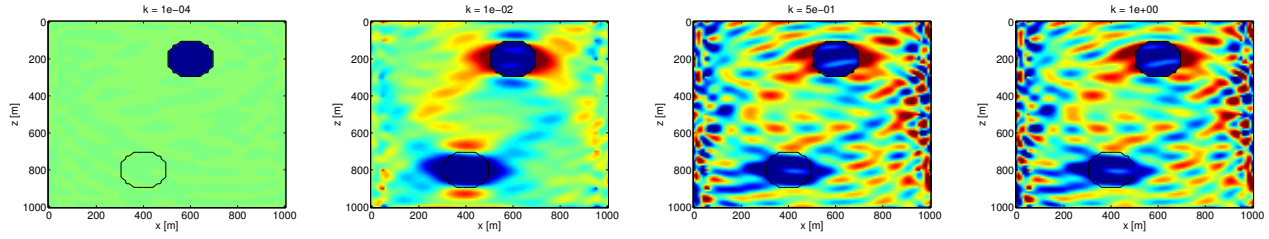
(a) (b) (c) (d)

Figure 2: Model perturbations A and B and least-squares reconstruction from contaminated data; (c) starting from smooth background and (d) starting from A.



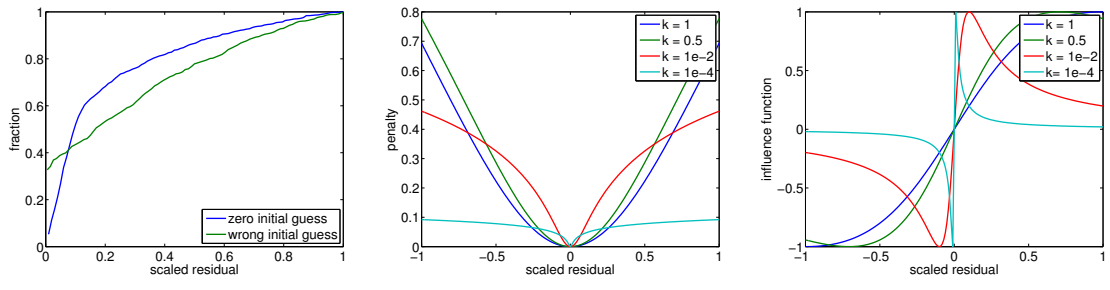
(a) (b) (c) (d)

Figure 3: Students t reconstructions from contaminated data starting from a zero initial guess.



(a) (b) (c) (d)

Figure 4: Students t reconstructions from contaminated data starting using perturbation B as initial guess.



(a) (b) (c)

Figure 5: (a) cumulative distribution of the scaled initial residuals, (b) penaly function for various k and (c) influence function (derivative of the penalty) for various k.