Defensible Metrics and Merit Functions

Making Informative Comparisons of Computer Simulations and Experiments

Kevin R. Vixie and Thomas J. Asaki

Introduction

Comparison of experiments with simulations is at the heart of progress in physics. Experiments are required to validate models, and simulations based on those models can be used to explore new physical regimes and define interesting experiments. Significant progress, however, is intimately tied to the way in which the "comparison" between experiment and simulation is carried out. The procedure has been largely a matter of personal preference, and familiarity with a certain analytical technique has often been the criterion for the selection of that technique. The complexity of the problem suggests that this eclectic approach is here to stay, but the complexity also demands a better understanding of what constitutes a meaningful comparison.

In this article, we discuss and then illustrate the following common-sense criteria for developing and judging comparison metrics: First, a good metric will ignore differences that don't matter and quantify important differences. Second, prior knowledge of the physics details should be incorporated into any analysis. Third, knowledge of the measurement operator, which is essentially a mathematical description of the experiment, should be exploited. Finally, because precise implementation of these criteria is typically too difficult, the comparison metrics should be created under the influence of these ideal requirements, and the ways in which they fall short should be indicated and understood. Our specific approach, illustrated below, is to develop *defensible metrics*[1] based on *merit functions, rigorously understood priors, and sensible measurement models* that, together, capture what is important, what is known, and how measurements relate to reality.

The defensible metric approach does not guarantee an easy solution to the comparison problem. However, because it (1) makes explicit the parameter space of the model, (2) distinguishes between model- and measurement-induced differences resulting from a comparison of simulation and experiment, and (3) measures only differences that matter (via the merit functions), our comparison metrics avoid the pitfalls of many standard approaches. One common approach is the "eyeball" metric (the researcher decides that the simulated results do or do not 'look' like the data); another is some form of difference metric in which the plots or images of the simulated data are subtracted pixel by pixel from corresponding images of the experimental data and various L^p norms of those differences are computed. The various norms are quantitative, yet it is often not clear that they quantify something of interest. For example, equation-of-state parameters are not easily associated with full-image norms, and pixel-to-pixel differences are not easily related to the differences in physics the two different images represent. An example of a slightly more careful approach is the study of how the norms of image differences scale with grid size of the simulation and with experimental parameters. One might also look at regions of high pixel difference and make informed guesses about why the experiment and simulation differ in these regions. Often though, these common methods are poorly connected to any sort of quantifiable conclusion. In contrast to these common methods, the defensible metric/merit function approach, which includes priors based on rigorous knowledge of the relevant state space, makes our comparison metrics quantifiably informative.

The Path to Defensible Metrics and Merit Functions

As stated above, we believe that defensible comparison metrics are objective measures that not only use prior information and knowledge of the measurement operator, but quite importantly, also *quantify only those differences that matter.*

A key concept for quantifying differences of interest, namely, the merit function, is illustrated in Figure 1, along with the state space of a system U, the measurement space D, and the merit function space K. While the concept of a merit function is general to data analysis problems, we will refer to concrete examples in explaining the various abstract functions and spaces in Figure 1. The state space U is the space describing some object or situation we are interested in interpreting. It is typically the very high dimensional description of a physical situation. A given state u could be the space and time concentrations of a multiple-constituent chemical reaction or the material and density description of a human leg. The state space cannot be directly probed. A particular state is given by a simulation or model u_s, or it is inferred through observables. Observables d lie in the measurement space D. This space of observables is typically of lower dimension than the state space. In the chemical reaction example, d may be spectroscopy and temperature measurements. In the human leg example, d may be external geometry measurements and x-ray radiographs. The measurement operator P is the physical description of an experiment (or simulation) that connects states with obervables, $d = Pu$. In our examples, P is the descriptor of the spectroscopy or radiography process. Often, however, we are not interested in knowing the

[1] We use the term *metric* more loosely than the usual mathematically precise definition requires. Typically, we use the term for something that is coercive with respect to some norm (we might also have to go to some appropriate quotient space) and positive. So it does give a sense for how close two inputs are, though it might not, for example, satisfy the triangle inequality. By coercive, we mean that the metric ρ satisfies $\rho\,(u_1,u_2)\rightarrow\infty$ as $\| u_1 - u_2\| \rightarrow \infty$, where $\|\cdot\|$ is some typical norm on U.

Back to Contents

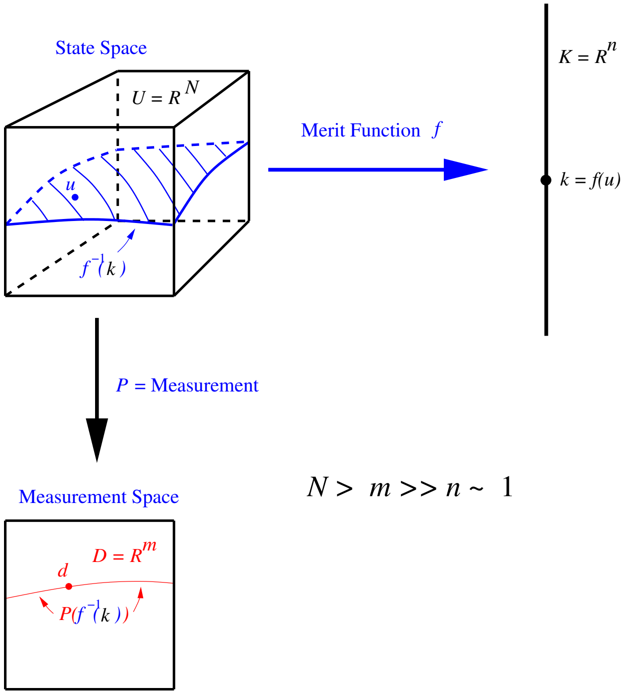

Figure 1. In a general picture, physical descriptions u, data d, and analysis results k are represented as points in state U, measurement D, and merit function K spaces, respectively. The measurement and state spaces are generally thought of as being connected through a measurement operator P (an experiment or its model description). The merit function f seeks to provide a result of interest, k = f(u). Metrics in K are simple to construct. Careful data analysis hinges on constructing metrics in U and D that reflect understanding of metrics on K.

entire state u given the observables d. Instead, we are interested in the answers to a few very specific questions. The answers to these questions lie in the merit function space K. (We will later mathematically identify K as a quotient space of U.) Specific answers k could be the final volume fraction of lithium triflate in our chemically reacting system or the likelihood of a person suffering from leukemia in the next 10 years. The merit function f associates every point (state) u in U with an answer $k = f(u)$ in K. In each space, we also need to know the metrics that quantify "distances" between points. The metric $\rho_K(k)$ in the merit function space quantifies how well we have answered our important questions. Constructing this metric is often the easy part. Now, within this picture, we can restate our goal: *We seek to construct metrics in the state space and/or data space, $\rho_U(u)$ and $\rho_D(d)$, respectively, that accurately reflect $\rho_K(k)$.* Success in achieving this goal is intimately tied to understanding the merit function f, the measurement operator P, and the relationship between ρ_U and ρ_K.

A Probabilistic Approach Gives a Coherent Framework

Our viewpoint permits us to develop good comparison metrics, but in order to more fully motivate and justify this approach, we show how these metrics can be embedded or understood within a standard (Bayesian) probabilistic framework.

We begin by considering how one might measure the quality of a particular simulation in the probabilistic sense (that is, how *typical* the simulation is). Later we will modify our result in a rather nonstandard way to get a measure of how good the simulation is relative to a merit function.

We now assume that we are given a simulated state u_s and an experimental measurement d. We do not directly know the experimental state u_e. The most likely state $u*$ is that which maximizes the posterior, or conditional probability, of u_e given u_s and d:

$$\begin{aligned} u^* &\cong \max_{u_e} p(u_e|d,u_s) \\ &\sim \max_{u_e} p(d|u_e,u_s)p(u_s|u_e)p(u_e) \qquad (1) \\ &\sim \max_{u_e} p(d|u_e)p(u_s|u_e)p(u_e)\ . \end{aligned}$$

Equivalently, we could minimize the negative logarithm of the conditional probability:

$$\begin{aligned} u^* &\cong \min_{u_e}[-\log p(u_e|d,u_s)] \qquad (2) \\ &= \min_{u_e}[-\log p(d|u_e)-\log p(u_s|u_e)-\log p(u_e)]\ . \end{aligned}$$

From the point of view of metrics, we can identify the three negative log likelihood terms as metrics and then find the state that minimizes the sum of those metrics:

$$u^* = \min_{u_e}[\rho_{df}(d,u_e)+ \rho_s(u_s,u_e)+ \rho_{prior}(u_e)]\ , \qquad (3)$$

where $-\log(d \mid u_e) = \rho_{df}(d,u_e)$ is called a data fidelity metric and involves some approximation to the measurement operator P, as well as assumptions about the stochastic nature of data itself, $-\log p(u_e) = \rho_{prior}(u_e)$ is typically a smoothing or regularization term that reflects our prior knowledge about the nature of the experimental state, and $-\log p(u_s \mid u_e) = \rho_s(u_s, u_e)$ is a stochastically inspired metric on the state space of the simulations and experiments (equivalent to the space U in Figure 1).

Our Modification. Since our intent is to evaluate the simulation, we do not want the simulation to influence our choice of the most probable experimental state. Therefore, we first find the state that minimizes the sum of the first and third terms,

$$u^* = \min_{u_e}[\rho_{df}(d,u_e)+ \rho_{prior}(u_e)]\ , \qquad (4)$$

and then use the stochastically inspired state-space metric $-\log p(u_s \mid u_e) = \rho_s(u_s, u_e)$ to understand differences in proposed models since the different models will yield different states u_s.

We now modify this standard probabilistic framework by replacing the stochastically inspired state-space metric, $-\log p(u_s \mid u_e) = \rho_s(u_s, u_e)$, with a merit-function-inspired metric, $\rho_U(u_1, u_2)$. Ideally, the latter is directly related to the merit function metric through the

Back to Contents

relationship $\rho_U(u_1, u_2) = \rho_K(f(u_1), f(u_2))$. The result is a metric that explicitly ignores differences that don't matter according to the merit function.[2] However, because such a metric is often difficult to compute, we must seek approximations.

With this substitution for the state space metric, we are making comparisons that explicitly relate to the goal of our prediction as encoded in the values of the merit function. What do we lose or gain by this substitution? There are two cases to consider. In the first, the instabilities or the stochastic nature of the experiment makes the stochastically based metric $\rho_s(u_s, u_e)$ small when the merit-function-based metric $\rho_U(u_s, u_e)$ is large. That situation means that either the instabilities in nature make predicting u_e difficult or that the model we are using has a very difficult time making the right prediction of *k*. In the second case, $\rho_s(u_s, u_e)$ is large when $\rho_U(u_s, u_e)$ is small, which implies that the model and experiment have serious differences (viewed from the state space), but that, when viewed from the merit function point of view, they are in good agreement. In either case, the size of the metric $\rho_U(u_s, u_e)$ is telling us much of what we wanted to know: How good is this simulation at predicting *k*? The only deficiency is the fact that, in the first case above, we will not know whether only this model or all models have difficulty predicting *k*.

Metrics and Regularization: Illustrative Examples

In the remainder of this article, we illustrate two elements of the picture outlined above, (1) the creation and use of likelihoods or metrics that are based on merit functions and (2) prior models in the form of regularization terms that correctly reflect the state space of interest. In particular, we use a face recognition problem to illustrate the building of a merit function *f*, based on a given metric ρ_K, that attempts to ignore unimportant state-space differences. Second, we use Abel inversion tomography to illustrate the use of a nonstandard regularization term, or prior, in the inversion of the measurement operator *P*.

Face Recognition: Classification Modulo Invariance. The classic face recognition problem is to identify a person from an image of the face by comparison with a database of images of known persons. While simply stated, the task is very difficult to perform accurately. First, images must be standardized for lighting differences, shifts, rotations, and scalings (that is, those features that, in principle, are independent of the subject). Then one must try and account for changes beyond the control of the photographer such as pose, expression, and grooming. It is not hard to imagine why simple metrics can fail to capture the essential features of an image relevant for subject identification. If one were to create a metric based on difference images (for example, L^p norms), identification results would tend toward matching exactly those unimportant features listed above. For example, the image of a laughing man could easily be most closely associated with database images showing a laughing person, rather than images of the same man with a different expression, leading to an incorrect identification. Such a metric is very bad at ignoring differences that do not matter (in this case, facial expression). The concept of a better metric is illustrated in Figure 2. All possible face images of a particular person lie in the same class. All images *u* within a class should project, via the merit function *f*, to the same value *k* (in this example *k* is an identification). Mathematicians call *K* a quotient space. A good state-space metric ρ_U will ignore distances within a class. In the high-dimensional space in which images live, the submanifolds of persons can be expected to be highly convoluted, and metrics must be constructed with care. The task is to construct an image-space metric ρ_U using a properly designed merit function *f* that maps to our well-defined quotient-space metric ρ_K.

Many methods are available for face recognition, and the difficulty of the problem, together with its importance in security applications, guarantees that it will remain important. One popular and intellectually satisfying method is based on eigenfaces. It was introduced by Kirby and Sirovich (1990) and then popularized by Turk and Pentland (1991). We cannot present here an overview of methods. However, it is important for us to compare results of our new techniques with those of several others. For this comparison, we will utilize the Colorado State University (CSU) face image database and 13 algorithms (http://www.cs.colostate.edu/evalfacerec). The particular method that inspired the current work is that of Simard et al. (1998, 2000), in which tangent spaces were used to locally approximate class submanifolds in *U*.

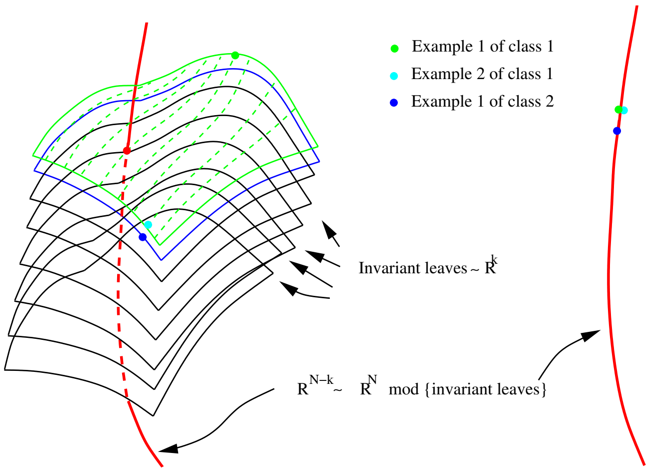

Figure 2. This figure illustrates a typical state space (at left) in which there are submanifolds of equivalent points relative to some merit function. Equivalently, we can think of these as the level sets of the merit function. Differences between points on the same submanifold are not important. Ideally, we would like to build a metric on the quotient space (at right) that replaces each submanifold with a point. In this illustration, we see that, before we take the quotient, it appears that example 1 of class 2 and example 2 of class 1 are much closer together than example 1 of class 1 and example 2 of class 1. After the quotient, we see that this difficulty disappears.

We chose to test our ideas by constructing a metric that attempts to ignore image scaling, rotation, and shifts. The face images equivalent under these transformations form (roughly speaking) five-dimensional,

[2] In mathematical terms, we care only about what our model does in the quotient space *F*, where *F* is the collection of level sets of *f* (the possibly singular foliation given by $f^{-1}(k)$ as *k* ranges over *R*).

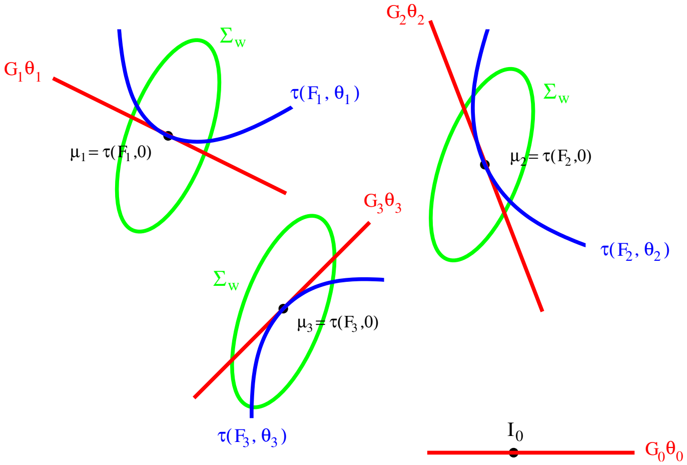

Figure 3. Green ellipses schematically represent level curves of a global metric distance to three individual example face images μ_n. Blue curves represent submanifolds of equivalent face images (same individual), along which a perfect metric would return zero distance. Clearly, the global metric does not capture the submanifold structure. Tangent approximations (red lines) give us local approximations to the submanifolds that we use to locally modify the global metric.

highly curved, nonlinear submanifolds of the space of all images. This feature appears to render impractical a direct computation of the quotient *f*(*u*). So, instead, we computed local tangent approximations that we then used to make local modifications to a global metric, thereby incorporating the quotient information in an approximate way (refer to Figure 3). We used second-order information to adjust the region of modification. The more highly curved the manifold in that region (or the larger the second derivative), the smaller the region of validity of the local approximation.

The CSU face image database consists of 4 images each of 160 individuals. For each classification, we randomly select a test face image of the target person. We then create a comparison image set composed of 159 randomly chosen images from the database, together with one additional image of the target person. The algorithm then chooses the face image in the comparison set closest to the test image according to the algorithm metric ρ_U. This basic test was repeated 160 times in order to calculate an average percent recognition success rate. Then the recognition rate was calculated 10,000 times for each algorithm to provide a recognition rate distribution. These distributions are shown in Figure 4 for each of the 13 CSU algorithms and three variants of our new method. Our new methods outperform all previous methods. They utilize approximate metrics that best ignore image differences that do not matter for face identification. It is important to note that the concepts and algorithms are not specific to face recognition. The algorithms were developed for general data comparisons. Full details can be found in Fraser et al. (2003).

Total Variation Regularized Abel Inversion. Understanding data from sparse radiography of fast events is important to the solution of problems central to the mission of the Laboratory. The sparsity of the radiographic data is a result of the combination of fast measurements, thick objects, and the very high expense of building additional viewing angles into the measurement apparatus. Much of our data is single-angle data. This restriction limits the radiographic reconstruction method to the inversion of the Abel projection, which assumes the object is cylindrically symmetric.[3] While such experiments are designed to make this assumption very nearly correct, rapid time evolution of the objects under study often introduces nonsymmetrical effects.[4] But even when symmetry is not in question, noisy data can make a good inversion difficult. Though the Abel projection is invertible, a typical discretization of the projection yields a condition number on the order of 10^3 (the condition number represents the magnitude of potential noise magnification), which is bad enough for noisy data to yield nonsense results under the Abel inversion. This unwelcome result is addressed by introducing regularization. As shown in Equation (5), the usual method minimizes the sum of two terms: a regularization metric (or prior), consisting of the integral of the squared gradient (first term), which favors smooth states, and a standard data-fidelity metric (second term):

$$u_{recon} = \min_{u}\left(\int|\nabla u|^2 d\Omega + \lambda\int|Pu-d|^2 d\Omega\right) \qquad (5)$$

This standard formulation for the Abel inversion is equivalent to the probabilistic formulation (Equation 3) presented in the last section except that no simulations are involved, and so the state space metric is not needed. Also note that the standard data fidelity met-

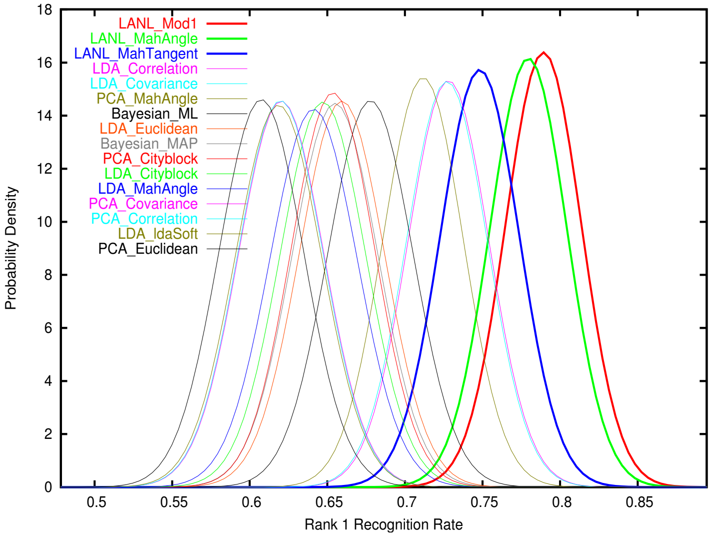

Figure 4. Shown here are face recognition probability results for 13 algorithms in the CSU database (leftmost curves) and our three new methods based on tangent approximations (rightmost three curves). As explained in the text, the three new methods outperform all previous methods.

[3] Actually, there is a whole host of similar assumptions that will yield invertability. All these assumptions boil down to the existence of a two-dimensional (2-D) parameterization of a three-dimensional (3-D) object. The 2-D Abel projection maps a density distribution $\omega(r)$, which is constant on concentric circles (and is therefore a function of the radius *r*), to a function *g*(*r*) whose value equals the line integral along a line that is at a distance *r* from the center.

[4] In some experiments, single-angle data are measured at multiple times (in sequence), and for these we can begin to make up for the data sparsity through variational approaches (Asaki et al. 2004).

Back to Contents

ric assumes that the operator P models the measurement and the noise is Gaussian and additive (a fair approximation for many applications).

It turns out that the total variation (TV) seminorm is a better choice of regularization for many common applications. The use of this seminorm for regularization of image reconstructions was introduced in the Rudin, Osher, and Fatemi (1992) variational model:

$$u_{recon} = \min_{u}\left(\int |\nabla u|\, d\Omega + \lambda \int |Pu-d|^2 d\Omega\right) \qquad (6)$$

In addition to a large amount of intriguing and beautiful theory to be explored, this and similar functionals are big improvements on previously used methods for certain applications. We have recently introduced TV regularization to the Abel inversion, and the results are quite encouraging (Asaki et al., in preparation). At the root of this improvement is the fact that, unlike the $|\nabla u|^2$ gradient prior, the TV prior does not cause bias against objects (states) with discontinuities in density and is therefore a better match to objects with edges (Asaki and Vixie 2004, Evans and Gariepy 1999).

Results on a test simulation can be seen in Figure 5. These results were obtained for a single two-dimensional (2-D) slice of a cylindrically symmetric object, which is therefore parametrically one-dimensional (1-D). The object is exactly described by 200 density rings of equal width. The density profile is shown in Figure 5a. The data shown in Figure 5b is the 1-D mass projection of the 2-D object with additive Gaussian noise (variance 5% of the noiseless data maximum). The remaining subfigures show the results of four reconstruction methods: (c) the nonregularized Abel inversion (an expected disaster); (d) $|\nabla u|^2$ regularized; (e) TV regularized; and (f) adaptive TV regularized. The last method iteratively locates edges using strongly TV-regularized reconstructions and then finalizes the inversion with a stronger smoothing term off these edges. The advantage in using the TV regularizer is clear in these examples. Development of these methods is continuing, and they are being brought to bear on difficult programmatic, mission-critical data analysis problems in two and three dimensions.

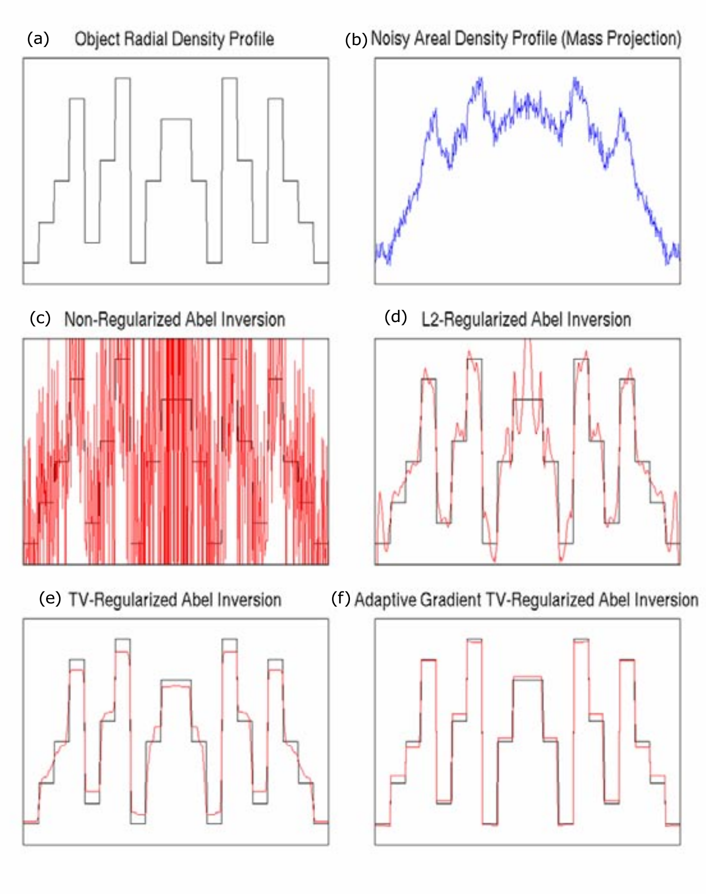

Figure 5. The importance of using an appropriate regularization is illustrated in this example of Abel inversion reconstructions of a cylindrically symmetric object. Individual plots show (a) the initial density profile, (b) projection data with added noise, and four reconstruction results: (c) nonregularized, (d) $|\nabla u|^2$ regularized, (e) $|\nabla u|$ (TV) regularized, and (f) adaptive TV regularized.

Looking Ahead

The path to defensible simulation validation is clearly not a simple one, yet we believe that our goal is achievable. We have set forth a broad picture in which to frame the questions and analyze the answers. And we have shown how the ideas of merit-function-based metrics play out in some simple examples. But the job is far from complete, and there remains much to understand.

We believe that there are many valuable exploration paths. Much can be accomplished quickly using innovative physics-driven regularizations and careful data fidelity metrics. Code parameter optimization might be efficiently managed using cost-function-based image registration methods. Images can be analyzed using detailed statistical projections and information divergences. But all these diverse techniques fall within the picture we have presented here. And all rely on the philosophy that defensible comparisons demand (1) a firm grounding in rigorous (often geometric) analysis and (2) metrics incorporating knowledge of what does and doesn't matter.

Back to Contents

List of Publications

1. T. J. Asaki, R. Chartrand, C. E. Powell, K. R. Vixie, and B. W. Wohlberg, "Total Variation Regularized Abel Inversions: Theory," (in preparation).

2. T. J. Asaki, R. Chartrand, C. E. Powell, D. Sigeti, K. R. Vixie, and B. W. Wohlberg, "Total Variation Regularized Abel Inversions: Applications," (in preparation).

3. T. J. Asaki, E. Bollt, and K. R. Vixie, "Sparse Radiographic Tomography and System Identification Imaging Form Single View, Multiple Time Sample Density Plots," (submitted to *SIAM J. Sci. Comput.*).

4. T. J. Asaki and K. R. Vixie, "SVD Analysis for Radiographic Object Reconstruction: Total Variation Regularization," Los Alamos National Laboratory document LA-UR-04-7076 (2004).

5. R. Beveridge, "Evaluation of Face Recognition Algorithms Website," Colorado State University. http://www.cs.colostate.edu/evalfacerec.

6. L. C. Evans and R. F. Gariepy, "Measure Theory and Fine Properties of Functions," (Boca Raton, FL: CRC Press, 1999).

7. A. M. Fraser, N. W. Hengartner, K. R. Vixie, and B. W. Wohlberg, "Classification Modulo Invariance, with Application to Face Recognition," *J. Comput. Graph. Stat.* **12** (4), 829 (2003).

8. M. Kirby and L. Sirovich, "Application of the Karhunen-Loeve Procedure for the Characterization of Human Faces," *IEEE Trans. Patterns Anal. Machines Intell.* **12** (1), 103 (1990).

9. L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear Total Variation Based Noise Removal Algorithms," *Physica D* **60**, 259 (1992).

10. P. Y. Simard, Y. A. L. Cun, J. S. Denker, and B. Victorri, "Transformation Invariance in Pattern Recognition: Tangent Distance and Tangent Propagation," *Int. J. Imag. Syst. Technol.* **11** (3), 181 (2000).

11. P. Y. Simard, Y. A. L. Cun, J. S. Denker, and B. Victorri, "Transformation invariance in pattern recognition – Tangent distance and tangent propagation. In *Neural Networks: Tricks of the Trade*. G. B. Orr and K. R. Muller, Eds. (Berlin; New York: Springer, 1998).

12. M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cog. Neurosci.* **3** (1), 71 (1991).