Optimal experimental design for machine learning using the Fisher information

Tracianne B. Neilsen, David F. Van Komen, Mark K. Transtrum, Makenzie B. Allen, and David P. Knobles

Citation: Proc. Mtgs. Acoust. **35**, 055004 (2018); doi: 10.1121/2.0000953
View online: https://doi.org/10.1121/2.0000953
View Table of Contents: https://asa.scitation.org/toc/pma/35/1
Published by the Acoustical Society of America

Volume 35 http://acousticalsociety.org/

176th Meeting of Acoustical Society of America
2018 Acoustics Week in Canada
Victoria, Canada
5-9 Nov 2018

Signal Processing in Acoustics: Paper 1pSP13

Optimal experimental design for machine learning using the Fisher information

Tracianne B. Neilsen
Department of Physics and Astronomy, Brigham Young University, Provo, Utah, 84602; tbn@byu.edu; tbnbyu@gmail.com

David F. Van Komen, Mark K. Transtrum and Makenzie B. Allen
Brigham Young University, Provo, Utah, 84602; david.vankomen@gmail.com; mktranstrum@byu.edu; allenmakenzie1427@gmail.com

David P. Knobles
KSA, LLC, Austin, TX; dpknobles@kphysics.org

Optimal experimental design focuses on selecting experiments that minimize the statistical uncertainty in inferred parameter or predictions. In traditional optimizations, the experiment consists of input data, model parameters, and cost function. For machine learning and deep learning, the features, labels, and loss function define the experiment. One tool for optimal experimental design is the Fisher information, which gives an estimate of the relative uncertainty in and correlation among the model parameters based on the local curvature of the cost function. Using the Fisher information allows for rapid assessment of many different experimental conditions. In machine learning, the Fisher information can provide guidance as to which types of input features and labels maximize the gradients in the search space. This approach has been applied, for example, to systems biology models of biochemical reaction networks [Transtrum and Qiu, BMC Bioinformatics 13(1), 181 (2012)]. Preliminary application of the Fisher information to optimize experimental design for source localization in an uncertain ocean environment is a step towards finding an efficient machine learning algorithm that produces results with the least uncertainty in the quantities of interest.

Published by the Acoustical Society of America

 https://doi.org/10.1121/2.0000953

1. INTRODUCTION

In ocean acoustics research, common goals are localizing acoustic sources and estimating the ocean environment, often with optimization techniques. Some of the many challenges of these inverse problems include high-dimensional search spaces, nonlinear relationships between the unknowns, and large uncertainty in many of the inferred parameters due to ill-conditioning. Particularly in this era of machine learning (ML) and deep learning (DL), determining the optimal experiment (combination of input data/features, parameterization/labels, and cost/loss function) is of particular importance. Often many of the possible labels are "sloppy" parameters, i.e., do not affect the input data being used. Ideally, time should not be wasted trying to estimate sloppy parameters. Instead, an optimal experiment should only use "stiff" parameters as labels—those that are strongly linked to the features. The stiffness/sloppiness of the parameters for a particular experiment is given by the Fisher information. In turn, the Fisher information can guide optimal experimental design by revealing which experiments (features + labels + loss) provide large gradients for ML/DL and, thus, give target estimates with less uncertainty.

The relative importance of the source and ocean parameters has been addressed previously using ocean acoustics optimizations. One technique is to constrain the bounds on certain parameters while allowing other parameters to vary more widely. Another approach is to use a set of rotated coordinates—eigenvectors of the covariance matrix—to more effectively navigate the search space. (Collins and Fishman, 1995; Neilsen, 2003) This form of parameter compression was also implemented in iterative schemes in which the stiffer parameters are found, then their bounds reduced, and subsequent parameters are found. (Neilsen, 2005) These approaches, however, required sampling over the entire parameter search space to obtain the rotated coordinates and gave a global measure of the parameter importance and couplings. More recently, a trans-dimensional Bayesian approach has been implemented in which the number of parameters, specifically the number of sediment layers, is determined as part of the optimization and rotated coordinates are used to efficiently conduct Markov Chain Monte Carlo sampling. (Dettmer and Dosso, 2012)

The Fisher information has been used for optimal experimental design in other fields, including in biological systems (Transtrum and Qiu, 2012; Machta *et al.*, 2013; Mannakee *et al.*, 2016; White *et al.*, 2016), condensed matter physics (Machta *et al.*, 2013), and acoustic array output (Rousseau *et al.*, 2003). The Fisher information is a measure of the information content in data, y_i, in the search space near parameters $\boldsymbol{\theta}_0$. Because it is a local metric, calculation of the Fisher information does not require sampling of a posterior distribution. The Fisher Information works near local or global minimum and can be generalized for the case when the accuracy of prediction matters more than estimation of parameters. (Transtrum and Qiu, 2012)

The goal of this paper is to illustrate with a numerical example how the Fisher information can guide experimental design such that parameter uncertainty is reduced for the stiff parameters and no time is wasted trying to find sloppy parameters. Specifically, the Fisher information can guide design of ML/DL experiments by informing selection of features, labels, and loss functions.

2. BACKGROUND

The Fisher information can be used to determine what experiment has the information content necessary to obtain parameter/label estimates or data predictions with low uncertainty. A brief explanation of how the Fisher information is obtained and used is now provided, following the explanation in Transtrum and Qiu (2012). The Fisher information is the inverse of the covariance matrix of the N parameters, θ_μ, in a quadratic Taylor series expansion of the cost function about $\boldsymbol{\theta}_0 = [\theta_1, \theta_2, \ldots, \theta_N]$. Thus, minimizing the variance is equivalent to maximizing the information. The inverse of the Fisher Information for an unbiased estimator forms a lower Cramer-Rao bound.

The Fisher information, $\boldsymbol{I}$, is closely related to the Jacobians of a cost function and is obtained as follows. For input data y_i^{data} and modeled values $y_i^{\text{model}}(\boldsymbol{\theta})$, a least-squares cost function is defined: $C(\boldsymbol{\theta}) = \sum_i e_i^2(\boldsymbol{\theta})$, where $e_i(\boldsymbol{\theta}) = y_i^{\text{data}} - y_i^{\text{model}}(\boldsymbol{\theta})$ is the error or residual for the i^{th} data sample. The Fisher information can be calculated from analytical functions when available or approximated with simulated data. For numerical

simulations, a parameter vector $\boldsymbol{\theta}_0$ is selected near a local or global minimum, and the Fisher information is calculated using numerical derivatives as

$$I_{\mu\nu} = \sum_i \frac{\partial e_i}{\partial \theta_\mu} \frac{\partial e_i}{\partial \theta_\nu}\bigg|_{\boldsymbol{\theta}_0}.$$

The Fisher information is approximately equal to the quadratic term in a Taylor series expansion, i.e., the Hessian, of the cost function around θ_0. (The Fisher information equals the Hessian only is $C(\boldsymbol{\theta}_0) = 0$.) A matrix containing the Fisher information for all combinations μ, ν can be computed as $\boldsymbol{I} = \boldsymbol{J}^T \boldsymbol{J}$, where $\boldsymbol{J}$ is the Jacobian, with $J_{i\mu} = \frac{\partial e_i}{\partial \theta_\mu}$.

When the parameter values in $\boldsymbol{\theta}$ are all positive and vary over several orders of magnitude, the Jacobians, and hence Fisher information, may be expressed in terms of the log of the parameters: $J_{i\mu} = \frac{\partial e_i}{\partial \log \theta_\mu}$. Via the chain rule this becomes $J_{i\mu} = \frac{\partial e_i}{\partial \theta_\mu} \frac{\partial \theta_\mu}{\partial \log \theta_\mu}$. When $e_i(\boldsymbol{\theta}) = y_i^{\text{data}} - y_i^{\text{model}}(\boldsymbol{\theta})$, the partial derivatives of e_i with respect to θ_μ reduce to partial derivatives of the modeled values because the data is independent of θ_μ: $\frac{\partial e_i}{\partial \theta_\mu} = \frac{\partial y_i^{\text{model}}}{\partial \theta_\mu}$. Using $\frac{\partial \theta_\mu}{\partial \log \theta_\mu} = \theta_\mu$, the elements of the Jacobian in terms of log of the parameters are $J_{i\mu} = \frac{\partial y_i}{\partial \theta_\mu} \theta_\mu$.

The Fisher information matrix is the inverse of the covariance matrix of the parameters in $\boldsymbol{\theta}$ in the region of the search space near $\boldsymbol{\theta}_0$. The diagonal terms, $I_{\mu\mu}$, indicate stiffness/sloppiness of the parameters, the influence of θ_μ on the match expressed in the cost function, and the information content about θ_μ in the input data and cost function. The off-diagonal terms, $I_{\mu\nu}$ tell the interdependence of θ_μ and θ_ν: $I_{\mu\nu}$= 0 indicates θ_μ and θ_ν are independent. The relationship between the Fisher information and covariance matrices means that maximum information leads to minimum variance, which is why the Fisher information can be used for optimal experimental design.

The key to understanding the geometry of the search space near $\boldsymbol{\theta}_0$ comes from the eigenvectors of the Fisher information matrix, which are also the right singular vectors of the Jacobian. Prior work in in ocean acoustics has shown that the eigenvectors of the inverse of the covariance matrix integrated over the search space constitute a rotated coordinate system that can be used to more efficiently navigate a large search space in optimizations (Collins and Fishman, 1995; Neilsen, 2003) and in Markov Chain Monte Carlo sampling (Dettmer and Dosso, 2012). Similarly, the eigenvectors, $\boldsymbol{V}$, of the Fisher information matrix provide a new basis for more efficiently navigating the search space about $\boldsymbol{\theta}_0$: EVD$(\boldsymbol{I}) = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^H$. In practice, the singular value decomposition of the Jacobian is often used. The eigenvectors $\boldsymbol{V}$ are also the right singular vectors from SVD$(\boldsymbol{J}) = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^H$:

$$\boldsymbol{I} = \boldsymbol{J}^T\boldsymbol{J} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^H\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^H = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^H.$$

The eigenvectors, $\boldsymbol{V}$, show combinations of θ_μ that make largest change in cost function near $\boldsymbol{\theta}_0$. The matrix $\boldsymbol{S}$ contains the singular values s_j along the diagonal, and $\boldsymbol{\Sigma}$ has diagonal elements s_j^2. Large singular values correspond to singular vectors $\boldsymbol{v}_j$ aligned with steep slopes in the search space and stiff parameter combinations with large information content. Small eigenvalues denote a cost function search space with winding canyons due to the sloppy parameters.

For a given experiment, an estimate of the information content—the inverse of the variance—is given by the diagonal of the Fisher information matrix. When the goal is low variance in parameter estimations, the overall performance of the experiment can be predicted by

$$D_{\text{parm}} \approx \frac{1}{N} \text{tr}(\boldsymbol{I}^{-1}),$$

where tr indicates the trace. The diagonal elements, $(\boldsymbol{I}^{-1})_{\mu\mu}$ give the estimated variance for θ_μ, as shown in Fig. 1. The performance parameter can be rewritten in terms of the singular values of the Jacobian, s_j, as

$$D_{\text{parm}} \approx \frac{1}{N}\sum_{\text{j}} \frac{1}{\text{s}_\text{j}^2}.$$

The individual terms of the summation, $1/s_j$, indicate the estimated variance in the combination of parameters coupled in singular vector $\boldsymbol{v}_j$.

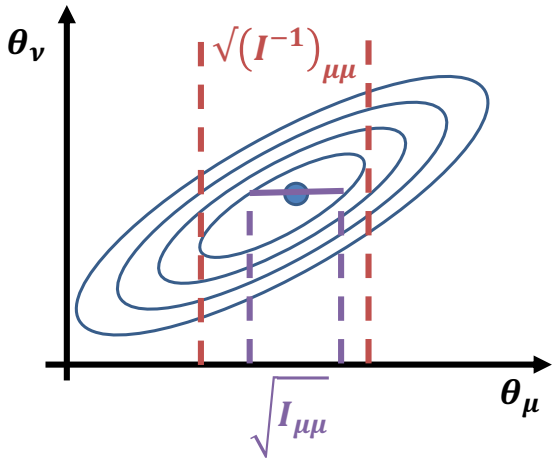

Figure 1. Schematic of the interpretation of the diagonal elements of the Fisher information, $\boldsymbol{I}_{\mu\mu}$, and its inverse, the covariance matrix, $(\boldsymbol{I}^{-1})_{\mu\mu}$, around the point $\boldsymbol{\theta}_0$ (blue filled circle). The blue ellipses represent contours in a cost function space.

The Fisher information can guide optimal experimental design for parameter estimation, as described above, and for the case where prediction accuracy is more important than parameter estimation. The uncertainty of a prediction experiment is related to both the Fisher Information at input data location/conditions and at the M locations/conditions of the predictions:

$$D_{\text{pred}} \approx \frac{1}{\text{M}}\sum_{\mu,\nu} I_{\mu\nu}^{\text{pred}} (\boldsymbol{I}^{-1})_{\mu\nu},$$

where

$$I_{\mu\nu}^{\text{pred}} = \sum_{m} \frac{\partial y_m^{\text{pred}}}{\sigma_\text{m} \partial \log \theta_\mu} \frac{\partial y_m^{\text{pred}}}{\sigma_\text{m} \partial \log \theta_\nu}$$

is a measure of the information content at the prediction locations or conditions.

The Fisher information can be used to estimate the uncertainty in an experiment for optimizations and for machine and deep learning. The steps in this process are as follows. 1) Simulate potential experiments: input data + model parameters + cost function for optimizations and features + labels + loss for machine/deep learning. 2) Calculate $\boldsymbol{I}$ or $\boldsymbol{J}$ near θ_0, then estimate D_{parm} or D_{pred}. 3) Add an appropriate type of noise to the simulations and repeat. 4) Repeat for different types of input data (features), parameterization (labels), and cost (loss)

functions. The first two steps for parameter estimation are now presented as an example of using Fisher information for experimental design.

3. EXAMPLE

An example is provided to illustrate how the Fisher information yields insights into the uncertainty expected in an experiment designed for parameter estimations. This example comes from ocean acoustics. The input data/features are transmission loss (TL) in an ocean waveguide, computed for a range-independent ocean with the ORCA normal-mode model. (Westwood *et al.*, 1996) TL is the sound level reduction, in decibels, due to propagation and can be considered the transfer function of the ocean waveguide between the source and receiver positions. The labels are 13 parameters that describe the water sound speed and depth and properties of the sediment layers. The loss function is a simple difference. The Jacobians and the Fisher information of the simulated TL are calculated to reveal the stiffness/sloppiness of the environmental parameters, identify parameter couplings, and estimate the uncertainty in parameter estimations from this experiment. The eigenvectors of the Fisher information matrix form a new basis—a set of rotated coordinates that is more aligned with the gradients of the search space.

A. EXPERIMENT

Synthetic data in a basic ocean environment and a least-squares cost function is used to show the usefulness of the Fisher Information. The forward model, ORCA (Westwood *et al.*, 1996), is a normal mode model for range-independent environments at 200 Hz. The simple ocean waveguide selected is shown in the left plot of Figure 2. The parameter vector is

$$\boldsymbol{\theta}_0 = [h_w, c_w(0), c_w(h_w), h_1, c_{1\text{top}}, c_{1\text{bot}}, \rho_{1,\text{top}}, \rho_{1,\text{bot}}, \alpha_{1\text{top}}, \alpha_{1\text{bot}}, c_{\text{hsp}}, \rho_{\text{hsp}}, \alpha_{\text{hsp}}],$$

The water column is labeled with the water depth h_w and the sound speed $c_w(z)$ at the top and bottom of the water column. The sediment layer is characterized by its thickness h_1, the sound speed c_1, density ρ, and attenuation α, at the top and bottom of the sediment layer and for the basement half-space (hsp). The $\{y_i^{\text{model}}\}$ are the magnitude of the transmission loss (TL) over the water column (z = 0 – 99 m) depths and ranges (r = 0.1 – 10 km). A least-squares cost function $C(\boldsymbol{\theta}) = \sum_i e_i^2(\boldsymbol{\theta})$ quantifies the match. The forward model, ORCA, is used to calculate the Jacobians of the cost function over this large spatial aperture.

A. JACOBIAN

The Jacobians for the example described in Sec. 3.A illustrate the relative stiffness of different parameters. The elements of the Jacobian are defined as $J_{i\mu} = \frac{\partial e_i}{\partial \theta_\mu}$. In practice, $J_{i\mu}$ is estimated via numerical derivatives using a forward model. For residuals $e_i(\boldsymbol{\theta}) = y_i^{\text{data}} - y_i^{\text{model}}(\boldsymbol{\theta})$, $J_{i\mu} = \frac{\partial e_i}{\partial \theta_\mu} = \frac{\partial y_i^{\text{model}}}{\partial \theta_\mu}$ as the $\{y_i^{\text{data}}\}$ do not vary with $\boldsymbol{\theta}$. For the example described in Sec. 3A, the forward model, ORCA, is used to calculate the TL over the same spatial aperture with a change in just one parameter θ_μ. The difference between the modified TL and the original TL is used to obtain a numerical derivative for $J_{i\mu} = \left.\frac{\partial y_i^{\text{model}}}{\partial \theta_\mu}\right|_{\theta_0}$. Examples of $J_{i\mu}$ are shown in Figure 3 for $\theta_\mu = h_w$ (upper left), $c_w(h_w)$ (upper right), c_1 (lower left), and α_1 (lower right). The maximum value of $J_{i\mu}$varies greatly across these four θ_μ: from 1.5e4 for $\theta_\mu = h_w$ to 1e2 for $\theta_\mu = \alpha_1$. The magnitudes of $J_{i\mu}$ indicate the sensitivity of y_i^{model} to θ_μ.

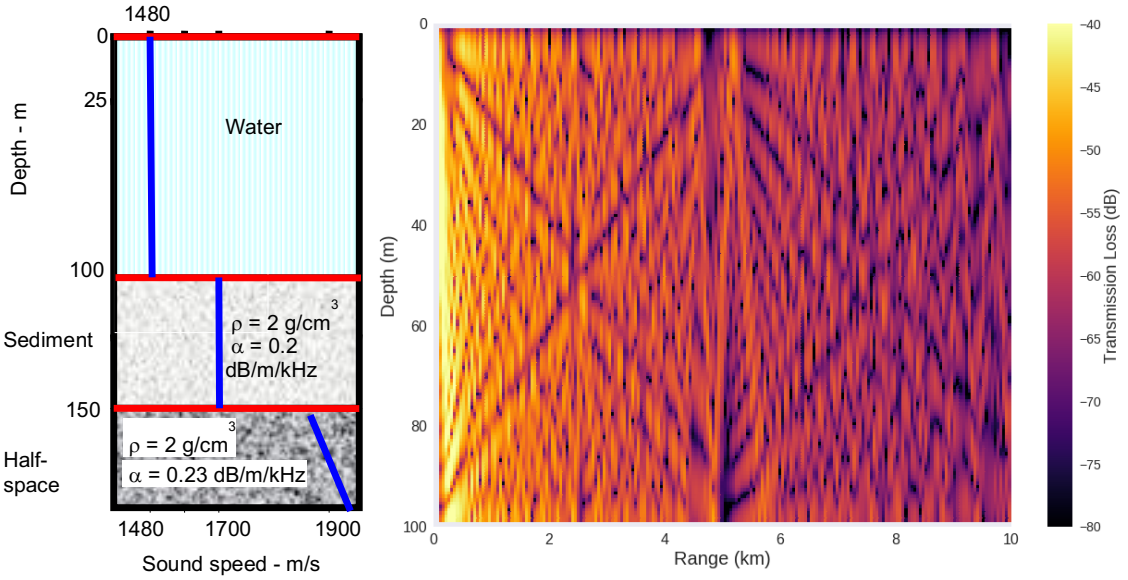

Figure 2. (Left) Schematic of the simulated ocean environment. (Right) Transmission loss at 200 Hz from a source at z_s= 6 m.

B. FISHER INFORMATION MATRIX

The Jacobians of the residuals is used to calculate the Fisher information matrix (FIM), which is computed from the Jacobians as $\boldsymbol{I} = \boldsymbol{J}^T \boldsymbol{J}$. For the experiment described in Sec 3.A, the log (base 10) of the absolute value of the FIM is shown in the left plot of Figure 4. The diagonal elements $I_{\mu\mu}$ indicate the information content in the $\{y_i^{\text{model}}\}$ for each θ_μ. For this experiment, all the $\{y_i^{\text{model}}\}$—over the large two-dimensional plane—contain significant information about the water depth and sound speed in the water column, no information about the thickness of the (deep) sediment layer, varying degrees of information about the other sediment properties, and no information about the half-space. The off-diagonal terms show the coupling between the parameters, although a clearer representation of this is shown in the eigenvectors.

The inverse of the FIM is the covariance matrix. For the example experiment, the covariance matrix on a log (base 10) scale is displayed in the right plot of Figure 4. The diagonal elements of the covariance matrix give the expected variance for an optimization/learned estimate of θ_μ. Different sets of input data yield different variances, which is why the covariance matrix and the FIM are useful in designing an optimal experiment.

To better appreciate the coupling between the parameters, the eigenvalues and eigenvectors of the FIM are calculated. Because the FIM often has many small eigenvalues, the singular value decomposition of the Jacobian matrix is performed instead: $\text{SVD}(\boldsymbol{J}) = \boldsymbol{USV}^H$. As described in Sec..2, the right singular vectors in $\boldsymbol{V}$ are the eigenvectors of $\boldsymbol{I}$, and the square of the singular values s_j in $\boldsymbol{S}$ are the eigenvalues of $\boldsymbol{I}$. Examples of s_j and $\boldsymbol{V}$ are given in Figure 5, with the rows of the right plot showing $\boldsymbol{v}_j$, the basis vectors for a coordinate system rotated to align with the gradients of the cost/loss function space. The $\boldsymbol{v}_j$ with the largest singular values show combinations of θ_μ that make largest change in residuals near θ_0. The s_j indicate the stiffness/sloppiness of the parameter combination contained in $\boldsymbol{v}_j$.

4. CONCLUSION

The Fisher Information can inform optimal experimental design. For optimizations, the experiment consists of the input data, forward model, parameters, and the cost function. For machine learning, the experiment is defined by the selection of the input features, labels, and loss function. The Fisher information guides selection experiments using the optimal design parameter D_{parm}. This estimate of the overall uncertainty informs the design of experiments that maximally increase the curvature of the search space around a local or global minimum. The use of the Fisher Information in optimal experimental design can increase efficiency and reduce uncertainty. The design parameter D_{parm} is not a rigorous estimate of uncertainty or confidence intervals but is sufficient to guide experimental design based on information content. Fisher information can also be used to

design experiment to minimize uncertainty in predictions instead of parameter estimations. (Transtrum and Qiu, 2012).

For machine learning/deep learning, the Fisher information can inform the selection of features, labels, and loss functions. Machine/deep learning algorithms perform best when the search space has large gradients. The Fisher information identifies which parameters correspond to large gradients, and perhaps more importantly which parameters have negligible gradients. It is postulated that when an algorithm is trained only on labels with large Fisher information, the gradients of the search space about are maximized. Increasing the gradients likely reduces the amount of training data needed. Thus, the Fisher information has potential to inform experimental design for machine and deep learning and provide rough estimates of the uncertainty in the learned labels.

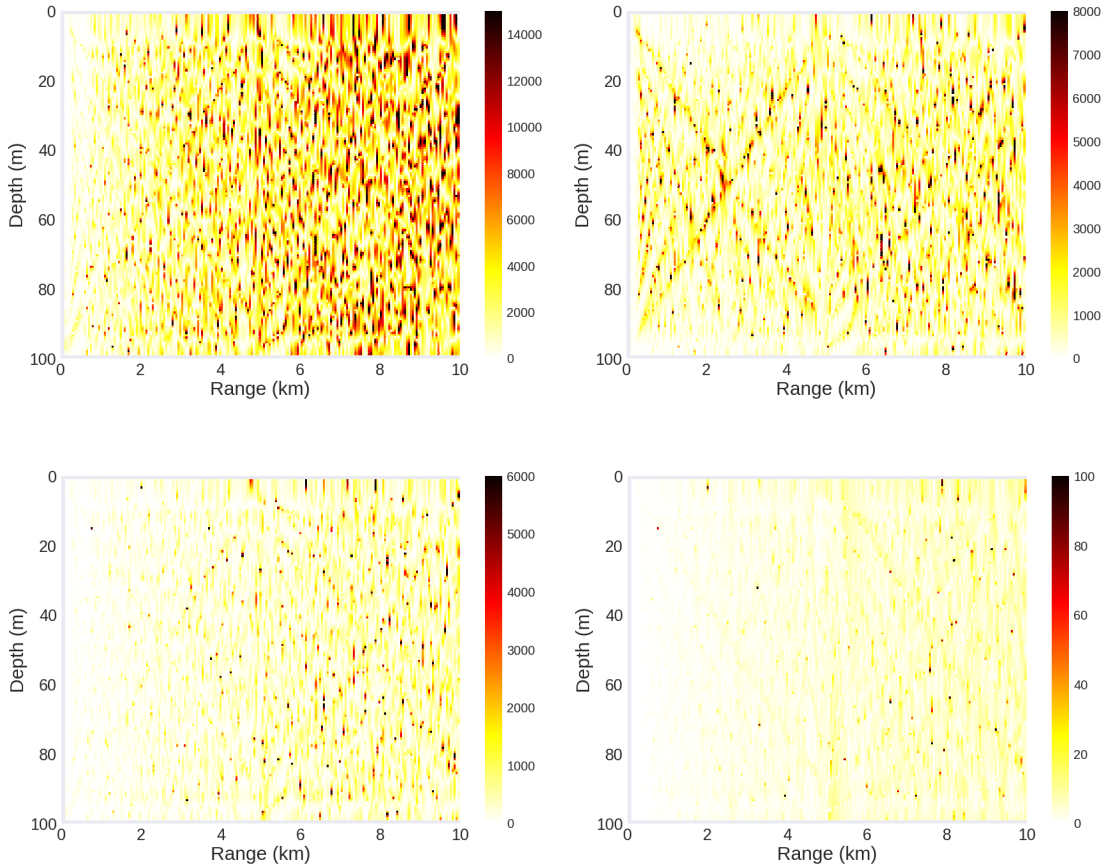

Figure 3. Jacobians $(\partial e_i/\partial \theta_\mu)$ for water depth (h_w, upper left), sound speed at the bottom of the water ($c_w(h_w)$, upper right), compressional sound speed at the top of the sediment (c_1, lower left), and compressional attenuation in the sediment (α_1, lower right).

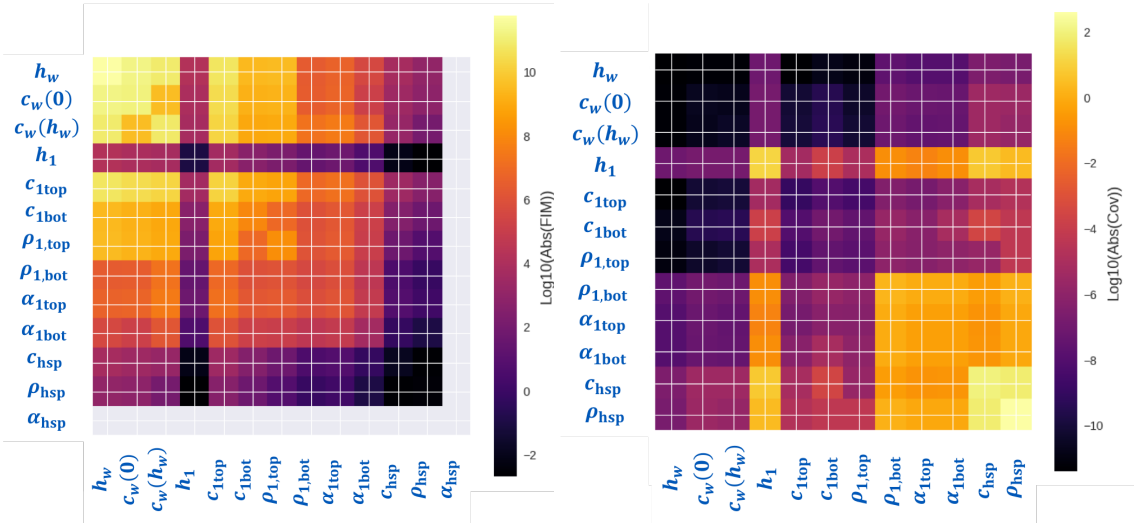

Figure 4. (Left) Fisher information matrix and (Right) its inverse, the covariance matrix, on a log scale, for the TL in the right plot of Figure 2.

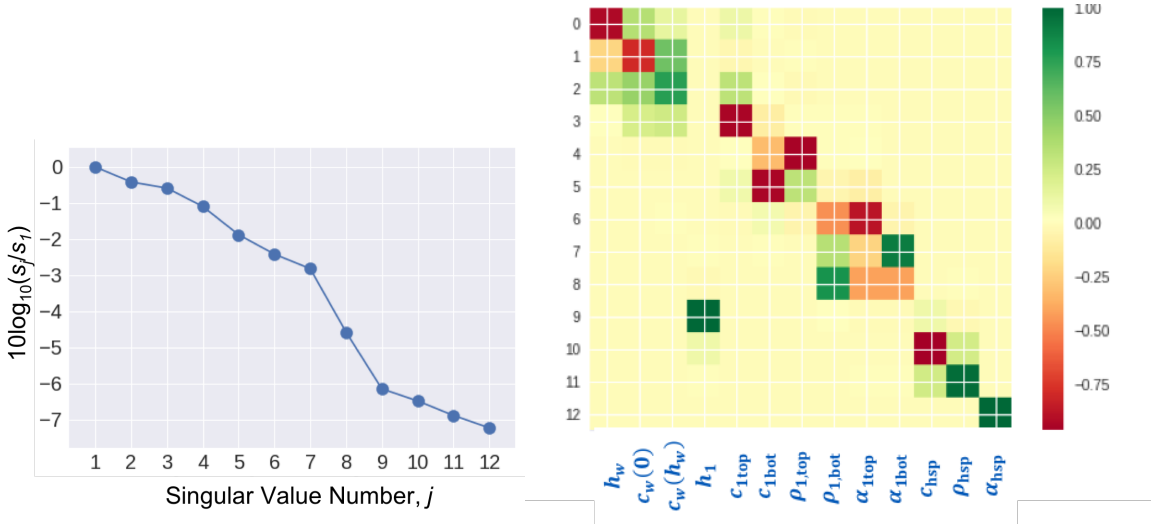

Figure 5. (Left) Singular values s_j and (Right) singular vectors v_j of the Jacobian, from which the Fisher information matrix in Figure 4 was calculated.

ACKNOWLEDGMENTS

This work was funded in part by the Office of Naval Research, SBIR Grant # N68335-18-C-0806.

REFERENCES

Collins, M. D., and Fishman, L. (**1995**). "Efficient navigation of parameter landscapes," J. Acoust. Soc. Am. **98**, 1637-1644.

Dettmer, J., and Dosso, S. E. (**2012**). "Trans-dimensional matched-field geoacoustic inversion with hierarchical error models and interacting Markov chains," J. Acoust. Soc. Am. **132**, 2239-2250.

Machta, B. B., Chachra, R., Transtrum, M. K., and Sethna, J. P. (**2013**). "Parameter space compression underlies emergent theories and predictive models," Science **342**, 604-607.

Mannakee, B. K., Ragsdale, A. P., Transtrum, M. K., and Gutenkunst, R. N. (**2016**). "Sloppiness and the geometry of parameter space," in *Uncertainty in Biology* (Springer), pp. 271-299.

Neilsen, T. B. (**2003**). "An iterative implementation of rotated coordinates for inverse problems," J. Acoust. Soc. Am. **113**, 2574-2586.

Neilsen, T. B. (**2005**). "Localization of multiple acoustic sources in the shallow ocean," J. Acoust. Soc. Am. **118**, 2944-2953.

Rousseau, D., Duan, F., and Chapeau-Blondeau, F. (**2003**). "Suprathreshold stochastic resonance and noise-enhanced Fisher information in arrays of threshold devices," Physical Review E **68**, 031107.

Transtrum, M. K., and Qiu, P. (**2012**). "Optimal experiment selection for parameter estimation in biological differential equation models," BMC bioinformatics **13**, 181.

Westwood, E. K., Tindle, C. T., and Chapman, N. R. (**1996**). "A normal mode model for acousto-elastic ocean environments," J. Acoust. Soc. Am. **100**, 3631-3645.

White, A., Tolman, M., Thames, H. D., Withers, H. R., Mason, K. A., and Transtrum, M. K. (**2016**). "The limitations of model-based experimental design and parameter estimation in sloppy systems," PLoS Comp. Bio. **12**, e1005227.