Advances in Reliability, Safety and Security, Part 3 – Kołowrocki, Dąbrowska (eds.)
 ISBN 978-83-68136-15-9 (printed), ISBN 978-83-68136-02-9 (electronic)

Advances in Reliability, Safety and Security

ESREL 2024
Monograph Book Series

Regularized Differentiation For Bioburden Density Estimation In Planetary Protection

Andrei Gribok[a], Michael DiNicola[b], Lisa Guan[b]

[a]*Idaho National Laboratory, Idaho Falls, ID, USA*
[b]*California Institute of Technology, Jet Propulsion Laboratory, Pasadena, CA, USA*

Abstract

In this paper, we propose and investigate the performance of two novel shrinkage estimators for bioburden density estimation in planetary protection. The estimators are based on the regularized differentiation of a cumulative count of colony forming units collected throughout the data collecting session or the life cycle of the entire mission. The regularized differentiation recasts the problem of bioburden density estimation as a linear least squares problem. The least squares problem is then solved through regularization techniques, such as truncated singular value decomposition and penalized least squares. The regularization is necessary to avoid noise amplification during the differentiation of noisy data.
It is shown through computer-simulated data that the regularized differentiation based on ridge regression has the smallest mean-squared error among all estimators. The analysis of shrinkage mechanism implemented by regularized differentiation is performed, and it is shown that the regularized differentiation amounts to performing a weighted averaging of all the samples. The weights are determined by the regularization parameter automatically selected by the L-curve technique. Since the method of least squares makes no distributional assumptions about the data, it presents an attractive technique for bioburden density estimation when there are concerns about the misspecification of the distributional model. The paper concludes with the analysis of the bioburden data collected during InSight mission and directions for future work.

Keywords: bioburden density, Bayesian inference, shrinkage estimators, least squares, regularization

1. Introduction

The primary objective of forward planetary protection (PP) is to minimize the inadvertent microbial contamination of other planetary bodies via hitchhiking microbes on robotic spacecraft sent to these planetary bodies. To manage and track the bioburden density and total bioburden throughout the life cycle of the entire mission, PP engineers maintain a PP equipment list, which tracks the surface area, related bioburden, and overall assembly hierarchy of each spacecraft component, subsystem, and system.

Traditionally, an estimation of bioburden density and total microbial bioburden for biologically sensitive interplanetary missions has been performed with statistical tools such as maximum likelihood estimation (Beaudet, 2013). Recently, the Bayesian approach has been suggested as an alternative to deal with zero-inflated datasets and high variance of maximum likelihood estimates (MLE) for small sampling areas (Gribok et al., 2019, 2020, 2022; Benardini et al., 2020). The Bayesian approach allows the user to deal effectively with zero-inflated data and reduce the variance of the estimates by the virtue of shrinking it toward a prespecified target value, for example, the average value of the sampled data. Despite these advantages over the maximum likelihood estimator, the Bayesian approach nevertheless relies on an assumed model of a data-generating mechanism. Having an estimator which does not assume any specific data-generating model is advantageous as it can be used for different distributions of the data and different noise models. The maximum likelihood and Bayesian estimators are effectively parametric methods while model-independent techniques are non-parametric.

In this paper, the problem of bioburden density estimation is reformulated as a linear least squares (LS) problem of differentiating cumulative counts of colony forming units (CFUs) collected during a data sampling session. The LS approach makes minimal assumptions about the data and does not depend on a particular

probabilistic model. Hence, it presents an attractive alternative to the parametric techniques. We propose and validate performance of two least-squares-based estimators.

2. Data collection, processing, and simulation

To validate the performance of the proposed techniques, both computer-simulated and InSight mission PP verification datasets were used. For the InSight data, a single component #261 has been selected for this paper. The component has been sampled 24 times with cotton swabs during two separate data collection sessions. A single swab data sampling covered the area of 0.0025 m^2. Each individual swab was considered a sample. Having been processed in the microbiology laboratory, the samples were deposited in petri dishes and covered with tryptic soy agar. For swabs, only 80% of the total sample solution was deposited into the dishes, thus producing a pour fraction of 0.8 that was taken into account by reducing the surface area sampled which is referred to as exposure in this paper. For the purpose of the analysis presented in this paper, the raw data for each sample were represented by pairs (x_i, E_i), i = 1,2,...,N, where x_i is the number of CFU counts for the i-th swab sample, E_i is the exposure calculated as the area covered with a swab multiplied by the corresponding pour fraction, and N is the number of samples collected for the component. The computer-simulated dataset was generated using the gamma-Poisson compound distribution model described in detail in Gribok et al. (2019, 2020, 2022) and Benardini et al. (2020).

3. Frequentists and Bayesian inference for bioburden density estimation

Prior to finding the bioburden density, a few assumptions about the CFU counts are commonly made when sampling a component:

- the probability of finding a CFU on any specified small exposure area is proportional to the exposure area and does not depend on where that exposure area is located. In other words, the bioburden density does not depend on location;
- the probability of finding more than one CFU on a given small exposure area is negligible in comparison with the probability of finding exactly one CFU on that area;
- finding CFUs on disjoint exposure areas is a statistically independent event.

If the above assumptions hold, the probabilistic model applied to the number of CFUs counts is a Poisson distribution with the probability mass function,

$$P(X = x|\lambda) = \frac{(\lambda \cdot E)^x}{x!} \cdot e^{-\lambda \cdot E}, \lambda \geq 0, x = 0,1, \dots, \quad (1)$$

where X is the random variable describing CFU counts, x is the actual number of CFUs found on the exposure area E, and λ is the bioburden density or expected number of CFUs per unit of exposure, which is unknown and the subject of the statistical inference. If the observed CFU count is x_i for a given exposure E_i, λ_i can be estimated as

$$\widehat{\lambda_i} = \frac{x_i}{E_i}, i = 1, \dots N, \quad (2)$$

where N is the number of samples.

This estimate is the MLE (Atwood et al., 2003) currently used by NASA to evaluate the bioburden density and total CFU counts for biologically sensitive missions (Beaudet, 2013). The MLE allows the bioburden density for each sample to be examined separately, and it has several desirable statistical properties. For example, MLE is unbiased in a frequentist sense and has minimum variance among unbiased estimators. However, it also has some shortcomings, such as large variance, and most importantly, for a small number of observed CFUs, it can overfit the data. It is also known that the MLE is inadmissible in case of Poisson distribution for $N \geq 3$ under squared error function (Clevenson et al., 1975). Inadmissibility of the MLE estimator means there are estimators that are uniformly better, i.e., have lower mean-squared error (MSE) over the entire range of the parameter space.

These shortcomings motivated the search for other estimators to calculate bioburden density, such as Bayesian estimators. Bayesian inference using the gamma-Poisson conjugate model will produce an estimator through (Martz et al., 1991)

$$\hat{\lambda}_i^{Bayes} = \frac{x_i+\alpha_{prior}}{E_i+\beta_{prior}} = \left(\frac{x_i}{E_i}\right) - [B]\cdot\left(\frac{x_i}{E_i} - \underbrace{\frac{\alpha_{prior}}{\beta_{prior}}}_{E_{prior}(\lambda)}\right),\ B = \frac{\beta_{\text{prior}}}{E_i+\beta_{\text{prior}}} \leq 1, \tag{3}$$

which is the mean of the posterior gamma distribution. In (3), α_{prior} and β_{prior} are the parameters of the prior gamma distribution of λ with mean value $E_{prior}(\lambda)$, and B is called shrinkage factor. It is easy to see from (3) that for $B = 0$, the Bayes estimate is MLE, while for $B = 1$ the estimates are reset to the mean value of the prior distribution $E_{prior}(\lambda)$. Thus, the Bayesian inference through the gamma-Poisson model shrinks the MLE estimate toward the mean value of the gamma distribution. The gamma-Poisson models assume the Poisson distribution for the number of CFU counts once bioburden density λ is selected from a prior gamma distribution. While a biased estimator in a frequentist's sense, the Bayes estimator has several advantages, such as not producing zero bioburden density estimates for components with zero CFU counts and often having a lower MSE with respect to the true bioburden density values.

4. Differentiation of noisy data as a least squares problem

Since the numerical differentiation of a noisy function is a well-known, ill-posed problem (Tikhonov et al., 1977; Hansen, 1998, 2010; Engle et al., 2000), it cannot be solved by simply applying the operation of differencing to the cumulative CFU count. To illustrate the technical problems involved in the operation of differentiation, we recall an alternative definition of a function's derivative through the second fundamental theorem of calculus (Tikhonov et al., 1977). Specifically, for a given function $f(x)$ with initial condition $f(0)=0$, its derivative $v(x)$ can be defined as

$$f(x) = \int_0^x v(t)dt \tag{4}$$

because

$$\frac{df(x)}{dx} = \frac{d}{dx}\int_0^x v(t)dt = v(x). \tag{5}$$

For example, the derivative of sin (x), which is a well-known cos (x), can be written as

$$sin(x) = \int_0^x cos(t)dt \tag{6}$$

as it immediately follows that

$$\frac{d(sin(x))}{dx} = \frac{d}{dx}\int_0^x cos(t)dt = cos(x). \tag{7}$$

The innocuous (4) reveals an often-overlooked fact that finding a derivative, by definition, always requires (implicitly or explicitly) solving an integral equation. In this equation, the derivative $v(t)$ occurs under a definite integral with variable upper limit. While this is never a problem for analytically defined functions and experimentally measured functions, the solution of an integral equation presents a problem because its left-hand side is always contaminated with measurement noise as (4) becomes

$$f(x) + \varepsilon = \int_0^x v(t)dt \tag{8}$$

where ε is measurements noise. The function in the left-hand side of (8) is a noisy version of $f(x)$ and $v(t)$ is its derivative. In the framework of our problem, $f(x)+\varepsilon$ is measured cumulative CFU count, and $v(x)$ is bioburden density. For the sake of simplicity, assume that the noise is represented by a single high-frequency sinusoid with amplitude "a" and frequency "ω," i.e., $\varepsilon = a\cdot\sin(\omega\cdot x)$. Finding derivative $v(x)$ requires differentiation of the left-hand side of (8),

$$\frac{d(f(x)+a\cdot sin(\omega\cdot x))}{dx} = f'(x) + a\cdot\omega\cdot\cos(\omega\cdot x) = v(x) + a\cdot\omega\cdot\cos(\omega\cdot x). \tag{9}$$

Equation (9) shows that the derivative of the noise-contaminated $f(x)$ can be arbitrarily far away from the true derivative, $v(x)$ because the term $a\cdot\omega\cdot\cos(\omega\cdot x)$ can be very large for large "ω" even if the amplitude of the noise "a" is small. Notice that the measurement noise is usually a broadband noise containing frequency bands covering the whole spectrum of the signal.

The previous derivations demonstrate the problem of differentiating a noisy function in continuous domain. However, all experimental data are collected in discrete form; hence, it is preferable to conduct our future analysis in the discrete domain. The bioburden data are collected in the form of pairs of x_i and E_i for $i = 1,\ldots,N$,

where N is the number of samples in a data collection session. Having CFU counts x_i for a number of samples N, we can form the monotonically nondecreasing cumulative count

$$\tilde{C}(i) = \sum_{j=1}^{i} x_j, i = 1, \dots, N\ ;\ \tilde{C}(0) = 0. \tag{10}$$

Having obtained the measured cumulative count $\tilde{C}$ and available exposure for each sample, the bioburden density can be represented through first-order differencing of the cumulative count

$$\hat{\lambda}_i = \frac{\tilde{C}(i) - \tilde{C}(i-1)}{E_i} = \frac{x_i}{E_i}. \tag{11}$$

Notice that $\hat{\lambda}_i$ obtained through (11) is mathematically identical to MLE shown in (2) as $\tilde{C}(i) - \tilde{C}(i-1) = x_i, i = 1, \dots, N$ which demonstrates that the MLE estimate is the first-order derivative of the cumulative count. Further, (4) can be rewritten in discrete form as a system of linear equations,

$$\tilde{C} = A \cdot \lambda \tag{12}$$

where A is an (N x N) lower-unitriangular Toeplitz (has constant values along all its diagonals and all ones on the principal diagonal) integration matrix with all nonzero elements equal to one, i.e.,

$$\mathrm{A} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}. \tag{13}$$

$\tilde{C} = C_{true} + \varepsilon$ is the (Nx1) vector of measured CFU counts containing C_{true} and noise ε, λ is an (Nx1) vector of unknown derivatives, i.e., bioburden densities, and N is the number of measurements. The integration matrix A amounts to application of the right end-point rectangle integration rule to bioburden density λ. System of (12) can either be solved with respect to λ, by direct inversion of matrix A, provided it is of full rank, or by the method of LS if A is rank deficient. In general, for a square matrix A, there are three possible options in resolving (12) with respect to λ: unique solution, infinite number of solutions, or no solution. The rank of matrix A determines which of these options is realized. If the rank(A) = r = N, then matrix A has full rank, and the system of equations is consistent.
In this case, direct inversion of matrix A will produce a unique solution,

$$\hat{\lambda} = \left[A^{-1} \cdot \tilde{C}\right]/E, \tag{14}$$

where E is the vector of exposures and $\hat{\lambda}$ is a vector of bioburden densities. The second option is when the system of equations (12) is consistent; however, rank(A) = $r < N$. In this case, matrix A is rank deficient, and system 12 has an infinite number of solutions. In terms of augmented matrix, this situation can be described as: rank [A $\tilde{C}$] = rank(A) where [A $\tilde{C}$] is the augmented matrix obtained by adding column of $\tilde{C}$ to matrix A. Since matrix A cannot be inverted and no exact unique solution can be obtained, we must look for an approximate solution which can be obtained through the method of the LS which minimizes the following functional with respect to λ

$$\left\{\left\|\tilde{C} - A \cdot \lambda\right\|^2\right\} \to min \tag{15}$$

with the symbol $\|\cdot\|$ denoting the Euclidean norm. The solution is obtained through Moore-Penrose pseudoinverse matrix and can be written as

$$\hat{\lambda}_{LS} = [A^T \cdot A]^{-1} \cdot A^T \cdot \tilde{C} = [A^+ \cdot \tilde{C}]/E \tag{16}$$

where $A^+ = [A^T \cdot A]^{-1} \cdot A^T$ is Moore-Penrose pseudoinverse matrix, and E is the vector of exposures. For the case of infinite number of solutions, the LS method is used to narrow down a single solution with the smallest norm called the minimum-norm solution. The last case when system 12 has no solution also arises for matrix A being rank deficient; however, in this case, rank [A $\tilde{C}$] > rank(A), and the system is inconsistent. Nevertheless, the method of LS still can be applied to find an approximate solution which is also the minimum-norm solution. Notice that in the case of full rank, the direct inversion solution is identical to the LS solution since $[A^T \cdot A]^{-1} \cdot A^T = A^{-1}$.

Integration matrix A is a lower unitriangular, and hence, it is invertible by the virtue of its determinant being equal to one and as such is different from zero. It also follows that the matrix is technically full rank, and the system of equations (12) can be solved exactly. However, in practice, there are two problems with applying direct inversion to the system of equations (12). The first problem is that the response variable $\tilde{C}$ on the left-hand side of system 12, which represents the measured cumulative count, is noisy, and the exact solution will fit all the noise in $\tilde{C}$ thus proving an overfitted solution for $\hat{\lambda}$. This problem emphasizes the problem of overfitting for

the MLE from a different perspective of the LS approach. Second, while A is theoretically full rank, it still can be numerically rank deficient or ill-conditioned due to its small singular values. In this respect, singular value decomposition (SVD) is a valuable tool in matrix algebra and theory of LS (Björck, 1996). For a square matrix A, its SVD can be written as

$$\underbrace{A}_{NxN} = \underbrace{U}_{NxN} \cdot \underbrace{\Sigma}_{NxN} \cdot \underbrace{V}_{NxN}^{T} \tag{17}$$

where $U^T U = V^T V = I$ are the orthonormal matrices of left and right singular vectors respectively, while Σ is a diagonal matrix with singular values arranged in a decreasing order. Using the SVD, the LS solution can be written as

$$\hat{\lambda}_{LS} = V \cdot \Sigma^{-1} \cdot U^T \cdot \tilde{C} = \sum_{i=1}^{N} \frac{u_i^T \cdot \tilde{c}}{\sigma_i} \cdot v_i \tag{18}$$

where σ_i are the singular values of matrix A. Equation (18) demonstrates the problem which may arise in the case of ill-conditioned matrix A. First, the solution is a linear combination of the right singular vectors v_i of matrix A with expansion coefficients produced by the dot product of left singular vectors u_i and a noisy cumulative count $\tilde{C}$ divided by corresponding singular values σ_i. A plot of five right singular vectors v_i for integration matrix A ($N = 24$) is presented in Figure 1.

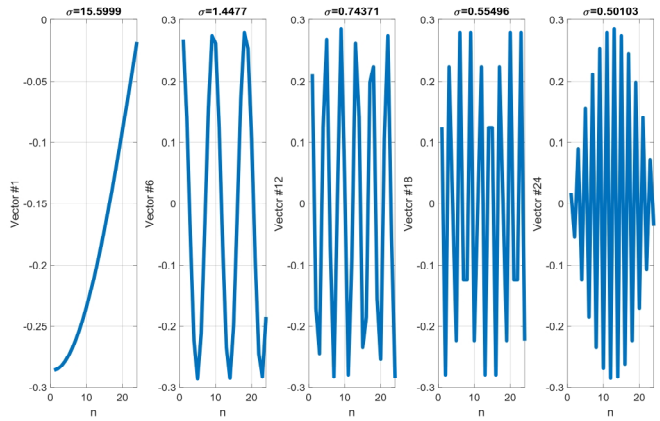

Fig. 1. The right singular vectors # 1, 6, 12, 18, and 24 for integration matrix A ($N = 24$). The corresponding singular values are shown on top of each plot.

As can be seen in Figure 1, the right singular vectors become progressively more oscillatory as the corresponding singular values are decreasing. As a result, the LS solution becomes more and more contaminated with high-frequency components causing data overfitting. However, even more important is the second observation that the expansion coefficient in (18) for each right singular vector is inversely proportional to the corresponding singular value. It means that the high-frequency components are more amplified than the low-frequency components resulting in a highly oscillatory and unstable solution. The singular values spectrum for matrix A ($N = 24$) is shown in Figure 2.

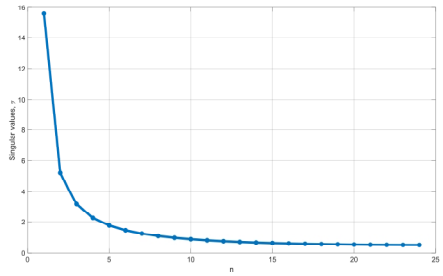

Fig. 2. Singular values spectrum of integration matrix A ($N = 24$).

As can be seen in Figure 2, the singular values are gradually decaying, and while they never reach zero, the condition number which is defined as cond(A)= σ_{max} / σ_{min} can be high with σ_{max} and σ_{min} being the maximum and minimum eigenvalues of matrix A.

The condition number is a very important matrix parameter as according to the classical perturbation theory (Björck, 1996),

$$\frac{\|\lambda_{exact}-\hat{\lambda}_{LS}\|_2}{\|\lambda_{exact}\|_2} \leq cond(A) \cdot \frac{\|\varepsilon\|_2}{\|\lambda_{exact}\|_2}, \tag{19}$$

where λ_{exact} is the theoretically correct derivative and ε is the measurement noise in the cumulative count.

Inequality (19) demonstrates that the relative error in the LS solution, $\frac{\|\lambda_{exact}-\hat{\lambda}_{LS}\|_2}{\|\lambda_{exact}\|_2}$, is bounded from above by the product of the condition number of matrix A – cond(A) and the relative measurement noise in the cumulative count, $\frac{\|\varepsilon\|_2}{\|\lambda_{exact}\|_2}$. Thus, the condition number determines the possible error in the solution of a LS problem, and for the large condition numbers, such an error can be significant. In case of N = 24, for example, the condition number of matrix A is just over 31, meaning that the error in the cumulative count is amplified more than 31 times while calculating the LS solution or MLE. It is important to point out that the straightforward or naive differentiation, by taking the difference of two subsequent measurements and dividing by the exposure between them, is ***mathematically identical*** to the LS solution. For this reason, we restrict ourselves to the LS analysis of the problem of differentiation.

4.1. Regularized differentiation

To contain noise amplification during differentiation of noisy data, several approaches have been proposed which fall under the category of regularized differentiation (Tikhonov et al., 1977; Hansen, 1998, 2010; Engle et al., 2000). Since the problem of noise amplification is caused by small singular values of matrix A, the idea of regularized differentiation is to prevent the small singular values and corresponding right singular vectors from entering the solution presented in (18). The most straightforward approach is to restrict the sum in (18) to the first k components corresponding to matrix's numerical rank. So, the truncated SVD solution (TSVD) is obtained by

$$\hat{\lambda}_{TSVD} = \sum_{i=1}^{k} \frac{u_i^T \cdot \tilde{C}}{\sigma_i} \cdot v_i \tag{20}$$

where $k < N$. The truncation parameter k can be selected by the method of L-curve which will be discussed below. Since not all singular values are used in the TSVD solution, the resulting matrix is obviously better conditioned thus producing a more stable approximate solution. The TSVD solution is a shrinkage estimator as it is evident that the norm of the TSVD solution is smaller than the norm of the LS solution simply due to the truncation of the last terms in the LS solution. While TSVD will produce a more stable and accurate solution, the fact that it simply disregards minor SVD components led to the search of a regularized solution which will "soft" threshold the components instead of "hard" thresholding performed by the TSVD. The "soft" thresholding can be obtained by replacing the LS functional in formula 15 with augmented functional which explicitly incorporates constrains on the solution

$$\left\{\left\|\tilde{C} - A \cdot \lambda\right\|^2 + \mu \cdot \|\lambda - \lambda_0\|^2\right\} \rightarrow min \tag{21}$$

where the first term measures the goodness of fit, while the second term measures the norm of the difference between solution λ and the initial guess λ_0. The initial guess can be set to zero or any other values reflecting prior knowledge about the solution. The balance between the two terms is controlled by regularization parameter μ. The functional 21 is known as ridge regression (Hoerl et al., 1970), and its minimizer for a given μ is

$$\hat{\lambda}_{Ridge} = [A^T \cdot A + \mu \cdot I]^{-1} \cdot A^T \cdot \tilde{C} = V \cdot (\Sigma^2 + \mu \cdot I)^{-1} \cdot \Sigma \cdot U^T \cdot \tilde{C} = \sum_{i=1}^{N} f_i \frac{u_i^T \cdot \tilde{C}}{\sigma_i} \cdot v_i \tag{22}$$

with

$$f_i = \frac{\sigma_i^2}{\sigma_i^2 + \mu}, i = 1, \dots, N \tag{23}$$

called filter factors. In (22) we used the following properties of orthogonal matrixes: $V \cdot V^T = V^T \cdot V = I = V \cdot V = V^{-1} \cdot V;\ V^T = V^{-1} = V$.

The filter factors are guaranteed to be no bigger than one and are functions of singular values σ_i and regularization parameter μ. For μ equal to zero, the ridge solution is reduced to the LS solution. For μ set to

infinity, the ridge solution is either zero or set to the initial guess λ_0. As can be seen from (22), each component of the LS solution represented by $\frac{u_i^T \cdot \vec{c}}{\sigma_i} \cdot v_i$ is multiplied by a filter factor f_i which is smaller than one. However, for large singular values, the filter factors are very close to one; while for smaller components, the filter factors are getting smaller, and thus those components have a smaller influence on the regularized solution. Since each LS component is multiplied by a value smaller than one, the ridge solution is also a shrinkage estimator with respect to the LS and hence with respect to MLE. For TSVD, the filter factors are just ones for i = 1:k and zero otherwise.

A very valuable tool to analyze the shrinkage mechanism of the ridge solution is the resolution matrix which can be expressed as

$$A^{\#} = [A^T \cdot A + \mu \cdot I]^{-1} \cdot A^T \cdot A. \tag{24}$$

The importance of the resolution matrix lies in the fact that it maps the LS solution to the ridge solution because

$$\hat{\lambda}_{Ridge} = [A^T \cdot A + \mu \cdot I]^{-1} \cdot A^T \cdot \tilde{C} = [A^T \cdot A + \mu \cdot I]^{-1} \cdot A^T \cdot \underbrace{A \cdot \hat{\lambda}_{LS}}_{\tilde{C}} = A^{\#} \cdot \hat{\lambda}_{LS}. \tag{25}$$

Notice that for μ = 0, the ridge solution becomes exactly the LS solution. The resolution matrix can be used to quantify the bias or regularization error of the ridge solution with respect to the LS solution as

$$\hat{\lambda}_{Ridge} - \hat{\lambda}_{LS} = A^{\#} \cdot \hat{\lambda}_{LS} - \hat{\lambda}_{LS} = \hat{\lambda}_{LS} \cdot (A^{\#} - I), \tag{26}$$

where I is the identity matrix. From (26), it follows that the bias of the ridge solution is the deviation of the resolution matrix from identity. However, even more importantly is that the resolution matrix sheds light on the shrinkage mechanism of the ridge solution with respect to the LS solution. From (25), it can be seen the ridge solution is obtained by multiplying a square (NxN) matrix $A^{\#}$ by a vector $\hat{\lambda}_{LS}$ of LS solution. It means that each component of the ridge solution is a dot product of the vector of LS solution and the corresponding row of the resolution matrix. Figure 3 shows rows of the resolution matrix for different values of regularization parameter μ corresponding to component 12 of the LS solution.

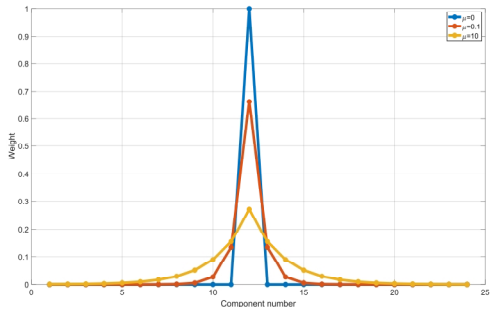

Fig. 3. The row of the resolution matrix $A^{\#}$ (N = 24) corresponding to component 12 and different values of the regularization parameter.

As can be seen in Figure 3, for μ = 0, only single LS component # 12 contributes to the ridge solution since in this case the resolution matrix is an identity matrix. However, as μ starts to increase, more and more components of the LS solution start contributing to the ridge solution thus making each component of the ridge solution a weighted sum of all LS components. Also notice that the samples closest to the sample in question are weighted more heavily than the more distant samples. This reveals the shrinkage mechanism of the ridge solution as a weighted pooling when a single LS estimate is replaced by a pooled estimate from all LS components with different weights. Since component 12 is the middle component for the 24-components solution, the weighting is symmetric, and components that are closer to component 12 are weighted more heavily than the distant components.

5. Performance of different shrinkage estimators

The performance of the regularized differentiation has been evaluated with respect to other shrinkage estimators proposed in the literature as well as against MLE. A recent review of such estimators used for simultaneous estimation of Poisson parameters can be found in Gribok et al. (2022). For this paper, we selected

three of the most known shrinkage estimators used for Poisson parameter inference: Empirical Bayes estimator with Jeffreys prior (Jeffreys, 1946; Martz et al., 1991), Empirical Bayes estimator with gamma prior parameters selected by the method of moments (MOM) (Berger, 1985; Martz et al., 1991), and the Clevenson-Zidek (CZ) estimator (Clevenson et al., 1975). The MLE and CZ estimators are frequentist estimators. For regularized differentiation, we used both TSVD and ridge solution with truncation parameter and regularization parameter selected by the method of L-curve (Hansen, 1998, 2010). The method of L-curve is a plot of the solution norm $\|\lambda\|$ against residual norm $\left\|\tilde{C} - A \cdot \lambda\right\|$as a function of regularization parameter μ or truncation parameter k. Such a plot often exhibits a distinct corner at parameter value μ that provides an optimal balance between fitting the data and shrinking the solution.

The performance of the shrinkage methods was first investigated through computer-generated data using gamma-Poisson data-generating model. As has been reported in an earlier publication (Gribok et al., 2022), eight components from InSight mission data were fitted with a gamma distribution that reflects the sample-to-sample variability of the data. The parameters of the fitted gamma distribution were $\alpha = 0.0447$ and $\beta = 2.5463 \cdot 10^{-4}$, with a λ average equal to 176 CFUs/m^2 and λ variance equal to $7 \cdot 10^5$ CFUs^2/m^4. Having fit the data, the gamma distribution was used to generate vectors of true λ of length 24, and those true λs have been used to generate Poisson variables for a given exposure. In this paper, the swab exposure has been used for all generated samples. The MSE between estimates of bioburden densities produced by each estimator and true bioburden density generated from gamma distribution was then calculated to quantify the accuracy of the estimators. This process has been repeated 100 times to quantify the uncertainty of each estimate.

The MLE has been calculated according to (2) while the Bayes estimator with Jeffreys prior and MOM prior were calculated according to (3) with improper Jeffreys prior proportional to $\lambda^{-0.5}$ and prior gamma distribution parameters selected by MOM for Empirical Bayes.

One of the most popular simultaneous estimators for Poisson means is the CZ estimator which is expressed component-wise as

$$\hat{\lambda}_i^{CZ} = \frac{x_i}{E_i} - \frac{\gamma + N - 1}{\sum \frac{x_i}{E_i} + \gamma + N - 1} \cdot \left(\frac{x_i}{E_i}\right), i = 1, \ldots N. \tag{27}$$

It shrinks the MLE estimate toward zero with parameter $0 \leq \gamma \leq N - 1$. The parameter γ is selected empirically. Fortunately, the performance of the CZ estimator shows weak dependence on the parameter. In this study, it was set to 12. Notice, for large values of $\frac{x_i}{E_i}$, the estimator is close to the MLE, and for a single sample, it is reduced to MLE. The MSE for different estimators along with 95% confidence intervals (CI) is shown in Figure 4.

The estimators were also tested using InSight data from component 261. Twenty-four swab samples were collected during two separate sessions on May 20, 2014 and July 10, 2017. Since in this case, the true bioburden density is unknown, the quality of the estimation can only be inspected qualitatively. Figure 5 shows the original raw cumulative count along with smooth cumulative counts obtained with the shrinkage estimators. The smooth counts have been obtained by integrating back the bioburden densities obtained with different estimators.

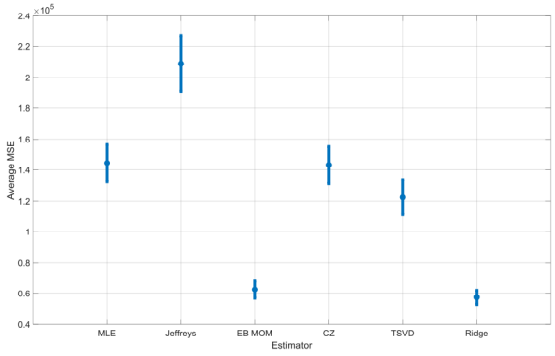

Fig. 4. MSE and 95% confidence intervals for six estimators.

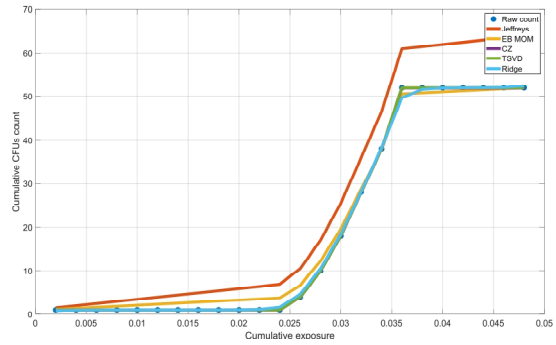

Fig. 5. Raw cumulative count and smooth cumulative counts produced by five different estimators.

6. Discussion and conclusions

The results for computer-generated data shown in Figure 4 demonstrate that Bayesian inference with Jeffreys noninformative prior has the largest MSE which is even larger than MLE. This can be explained by the fact that Jeffreys prior adds half a count to the measured data leaving exposure unchanged. With small exposure, similar to our computer-generated data with E = 0.002, half a count makes a difference driving the estimate up and for small values of bioburden density increasing the MSE. The performance of MLE and CZ are practically identical since, as can be seen from (27), the CZ is very close to the MLE for large values of $\sum\frac{x_i}{E_i}$. In this case, the shrinkage parameter $\frac{\gamma+N-1}{\sum\frac{x_i}{E_i}+\gamma+N-1}$ becomes small, and there is practically no shrinkage. For our simulated data, some $\frac{x_i}{E_i}$ were indeed large due to the small exposure value. The TSVD estimator performed better than MLE and Jeffreys prior estimator and slightly better that CZ. The truncation parameter was selected for each simulation run, and the average parameter over 100 runs was 13. It means that on average more than half of the SVD components have been truncated and not used in the TSVD solution. The difference between TSVD and CZ estimators is statistically significant at the standard 5% significance level. Both estimators are shrinking the MLE toward zero; however, the shrinkage seems to be stronger for the TSVD solution. The two lowest MSEs are achieved for EB-MOM and the ridge estimators with ridge having the smallest MSE at 5% significance level. For both methods, the shrinkage factor B and regularization parameter μ have been selected for each trial. The initial guess of λ_0, used for the ridge method, was set to the mean value of all CFUs counts collected for 24 samples. Thus, in contrast to TSVD and CZ, both EB-MOM and ridge are shrinking the MLE estimates toward the mean value of CFU's count of all samples. On the other hand, the shrinkage mechanisms of these two techniques are different. While EB-MOM pulls each estimate toward the mean value of all measurements, the ridge does a more subtle weighted averaging when measurements that are closer to the current measurement contributes more to the current estimate. This makes it important prior to applying regularized differentiation to a set of CFU samples, to arrange the samples either in chronological order or in the order of their physical proximity. For a set of measurements, it is expected the samples that chronologically or spatially close to each other will have higher intercorrelation.

The results shown in Figure 5 confirm some of the observations made for the simulated data. For example, using Jeffreys noninformative prior does lead to overcounting since the Jeffreys cumulative count is clearly biased upward with respect to the raw count. The EB-MOM fit seems to slightly over smooth the raw count while TSVD and CZ clearly overfit the raw count as they fit every data sample. This is the result of small truncation parameter and shrinkage parameter that are selected for the two methods. Ridge seems to provide the most reasonable fit to the raw cumulative count neither over smoothing nor undersmoothing it.

Replacing the norm of the solution in (20) with a seminorm as $\|L \cdot \lambda\|$ where L is a matrix approximating n-order derivative, we can shrink the solution to different functions such as linear, quadratic, etc., if such a priory information is available.

Acknowledgements

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004). This work was supported

by U.S. DOE-NASA Strategic Partnership Project (SPP) #19701. This manuscript has been authored by Battelle Energy Alliance, LLC under Contract No. DE-AC07-05ID14517 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for U.S. Government purposes. The authors are grateful to Dr. J. Nick Benardini, Dr. Elaine E. Seasly, and Arman Seuylemezian for their contributions to this project.

References

Atwood, C. L., LaChance, J. L., Martz, H. F., Anderson, D. J., Englehardt, M., Whitehead, D., Wheeler, T. 2003. Handbook of Parameter Estimation for Probabilistic Risk Assessment. NUREG/CR-6823, SAND2003-3348P, U.S. Nuclear Regulatory Commission.

Beaudet, R. A. 2013. The Statistical Treatment Implemented to Obtain the Planetary Protection Bioburdens for the Mars Science Laboratory Mission. Advances in Space Research 51, 2261–2268.

Benardini, J. N., Seuylemezian, A., Gribok. A. 2020. Predicting biological cleanliness: an empirical Bayes approach for spacecraft bioburden accounting, presented at the 2020 IEEE Aerospace Conference, pp. 1-12, https://www.doi.org/10.1109/AERO47225.2020.9172725.

Berger, J. O. 1985. Statistical Decision Theory and Bayesian Analysis. Springer, New York.

Björck, A. 1996. Numerical Methods for Least Squares Problems, SIAM.

Clevenson, L., Zidek. J. V. 1975. Simultaneous estimation of the means of independent Poisson laws. Journal of the American Statistical Association 70, 698-705,

Engl, H. W., Hanke, M., Neubauer, A. 2000. Regularization of Inverse Problems. New York, NY: Springer.

Gribok A., Seuylemezian, A., Benardini, J. N. 2019. A Bayesian approach for estimating spacecraft microbial bioburden and managing the risk of biological contamination. presented at PSAM 2019, the Topical Conference Practical Use of Probabilistic Safety Assessment in Operations, December 2019, The Haymarket Hotel, Stockholm, Sweden.

Gribok A., Benardini, J. N., Seuylemezian, A. 2020. Bayesian Framework for Bioburden Density Calculations to Perform Planetary Protection Probabilistic Risk Assessment. Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference. Edited by Piero Baraldi, Francesco Di Maio and Enrico Zio. doi: 10.3850/981-973-0000-00-0 esrel2020p.

Gribok, A., Seuylemezian, A. 2022. Performance of Shrinkage Estimators for Bioburden Density Calculations in Planetary Protection Probabilistic Risk Assessment. PSAM 16, June 26–July 1, 2022. Honolulu, Hawaii.

Hansen, P.C. 1998. Rank-Deficient and Discrete Ill-Posed Problems Numerical Aspects of Linear Inversion. SIAM.

Hansen, P.C. 2010. Discrete Inverse Problems: Insight and Algorithms SIAM Monographs on Mathematical Modeling and Computation (Fundamentals of Algorithms series). Philadelphia.

Hendrickson, R., Kazarians, G., Benardini, J. N. 2020. Planetary Protection Implementation on the Interior Exploration Using Seismic Investigations, Geodesy and Heat Transport Mission. Astrobiology 20, 1151–1157.

Hoerl, A.E., Kennard, R. W. 1970. Ridge regression: Bia sed estimation for nonorthogonal problems. Technometrics 12(1), 55-82.

Jeffreys, H. 1946. An Invariant Form for the Prior Probability in Estimation Problems. Proceedings of the Royal Society of London Series A, Mathematical and Physical Sciences 186, 453–461.

Martz, H. F. Waller, R. A. 1991. Bayesian Reliability Analysis. Reprinted with corrections. Krieger Publishing Co.

Tikhonov, A.N., Arsenin, V.Y. 1977. Solution of Ill-Posed Problems. Washington: Winston & Sons.