DATA-INFORMED REGULARIZATION FOR INVERSE AND IMAGING PROBLEMS*

JONATHAN WITTMER† AND TAN BUI-THANH‡

1. Introduction. Regularization is a technique often employed to facilitate the well-posedness of inverse (and imaging) problems. An inverse solution is thus a trade-off between the data misfit and the regularization. Due to noise and limited availability, the data typically informs limited directions in the parameter space where the inverse solution resides. A desired regularization, we argue, should minimally interfere with these data-informed directions. However, most regularization techniques regularize all parameter directions, including the data-informed ones, thus polluting the resulting inverse solution. This chapter presents a new regularization method for inverse and imaging problems, called data-informed regularization, that implicitly avoids regularizing the data-informed directions. The DI method combines advantages of the classical truncated SVD and Tikhonov regularization. In particular, it does not pollute the data-informed modes with regularization, but regularizes only the less data-informed ones.

Compared to existing approaches, our method has many distinct and advantageous features: 1) it automatically determines the directions equally informed by the data and any Tikhonov regularization while leaving the most informative directions untouched. In fact, we will show that, similar to the balanced truncation idea in control theory (see, e.g., [4, 1] and the references therein), this is done implicitly by seeking directions in parameter space that balance the information from regularization and data, and removing the regularization on them; 2) We will show that our approach has an intuitive statistical interpretation, namely, it transforms both the data distribution (i.e. the likelihood) and prior distribution (induced by Tikhonov regularization) to the same Gaussian distribution whose covariance matrix is diagonal and the diagonal elements are exactly the singular values of a composition of the prior covariance matrix, the forward map, and the noise covariance matrix. 3) Though constructively derived and its insights obtained from the truncated Singular Value Decomposition (SVD), the inverse solution resulting from our approach does not necessarily require an SVD decomposition, which may not be feasible for large-scale applications. Indeed, we will present a nested matrix-free approach to obtain an approximate inverse solution. 4) By construction, features in our inverse solution dictated by the data-informed directions are insensitive to the regularization parameter. For many inverse and imaging problems, these features dominate the solution and thus the inverse solution resulting from our regularization technique is robust with respect to regularization parameter values. These findings will be demonstrated and supported by various numerical results from deblurring, denoising, and X-ray tomography problems.

*Submitted to the editors DATE.

Funding: This work was partially funded by the National Science Foundation awards NSF-1808576 and NSF-CAREER-1845799; by the Defense Thread Reduction Agency award DTRA-M1802962; by the Department of Energy award DE-SC0018147; by KAUST; by 2018 ConTex award; and by 2018 UT-Portugal CoLab award. The authors are grateful to the supports.

†Department of Aerospace Engineering and Engineering Mechanics, UT Austin, Austin, Texas

‡Department of Aerospace Engineering and Engineering Mechanics, The Oden Institute for Computational Engineering and Sciences, UT Austin, Austin, Texas (tanbui@ices.utexas.edu, https://users.oden.utexas.edu/~tanbui/).

2. A data-informed regularization (DI) approach.

2.1. Data-informed regularization derivation. In this section we review the key ideas behind regularization by truncation using the Singular Value Decomposition (SVD). This provides the basic insights into the data-informed regularization technique. A statistical interpretation of the data-informed inverse framework will be discussed in Section 2.2. To begin, let us consider a linear inverse problem to determine $\boldsymbol{x} \in \mathbb{R}^p$ given

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}, \tag{2.1}$$

where $\boldsymbol{A} \in \mathbb{R}^{d\times p}$, $\boldsymbol{e} \sim \mathcal{N}\left(\boldsymbol{0}, \lambda^2 \boldsymbol{I}\right)$, $\boldsymbol{I} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{y} \in \mathbb{R}^d$. In the following, the identity matrix $\boldsymbol{I}$ may have different size at different places and the actual size should be clear from the context. The simplest approach to attempt to solve this inverse problem is perhaps the least squares approach:

$$\min_{\boldsymbol{x}} \frac{1}{2} \left\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right\|^2, \tag{2.2}$$

where $\|\cdot\|$ denotes the standard Euclidean norm. The least-squares solution is given by

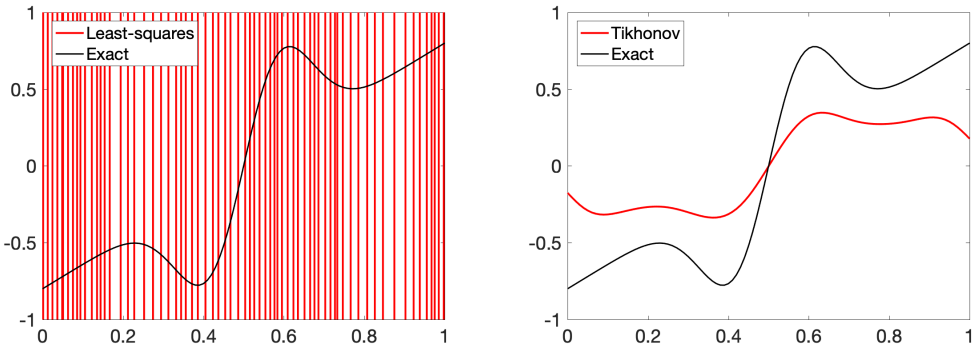
$$\boldsymbol{x}_{LS} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{y}.$$

Figure 1(a) plots the exact synthetic solution (black curve) against the least-squares solution (red curve) for a deconvolution problem with $d = p = 101$ and $\lambda = 0.05$. As can be seen, the least-squares solution blows up (or is unstable), which is not surprising since the inverse problem is ill-posed.

A standard Tikhonov regularization approach casts the above inverse problem into

$$\min_{\boldsymbol{x}} \frac{1}{2} \left\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right\|^2 + \frac{\alpha}{2} \left\|\boldsymbol{x} - \boldsymbol{x}_0\right\|^2,$$

where $\boldsymbol{x}_0$ is given. A Tikhonov solution is presented in Figure 1(b) for $\alpha = 1$ and $\boldsymbol{x}_0 = 0$. Though this approach stabilizes the solution, it also smooths out the solution everywhere.



(a) Least-squares solution (b) Tikhonov solution

FIG. 1. *Deconvolution using a) the least-squares approach, and b) a Tikhonov regularization with regularization parameter* $\alpha = 1$ *and* $\boldsymbol{x}_0 = \boldsymbol{0}$.

Regularization by truncation does not require an explicit introduction of regularization term as in Tikhonov regularization. For example, the truncated SVD starts with the SVD decomposition of $\boldsymbol{A}$ and then truncates all the singular vectors $\boldsymbol{U}_j$ and $\boldsymbol{V}_j$ corresponding to zero (or relatively small) singular values, i.e.,

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$$

$$= \begin{pmatrix} \boldsymbol{U}_1 & \cdots & \boldsymbol{U}_n & \boldsymbol{U}_{n+1} & \cdots & \boldsymbol{U}_d \\ | & & | & | & & | \end{pmatrix} \begin{pmatrix} \sigma_1 & & & & \\ & \ddots & & & \\ & & \sigma_n & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & & 0 \end{pmatrix} \begin{pmatrix} \text{---} & \boldsymbol{V}_1^T & \text{---} \\ & \vdots & \\ \text{---} & \boldsymbol{V}_r^T & \text{---} \\ \text{---} & \boldsymbol{V}_{n+1}^T & \text{---} \\ & \vdots & \\ \text{---} & \boldsymbol{V}_p^T & \text{---} \end{pmatrix}$$

$$= \boldsymbol{U}^n \boldsymbol{\Sigma}^n \left(\boldsymbol{V}^n\right)^T,$$

where $\boldsymbol{U}^n := [\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n]$ (the first n columns of $\boldsymbol{U}$ corresponding to n nonzero singular values or the range of the column space of $\boldsymbol{A}$), $\boldsymbol{\Sigma}^n := \operatorname{diag}[\sigma_1, \ldots, \sigma_n]$ ($\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n$), $n \leq \min\{d, p\}$, and $\boldsymbol{V}^n := [\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n]$ (the first n columns of $\boldsymbol{V}$ corresponding to n nonzero singular values the row space of $\boldsymbol{A}$).

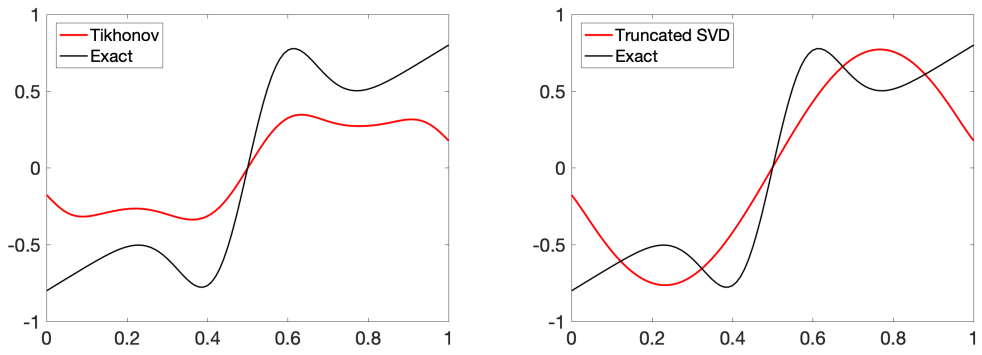
The solution of the least-squares problem (2.2) using the truncated SVD together with pseudo-inverse reads

$$\boldsymbol{x}_{\text{SVD}}^n = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^\dagger \boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{V}^n \left(\boldsymbol{\Sigma}^n\right)^{-1} \left(\boldsymbol{U}^n\right)^T \boldsymbol{y} = \sum_{i=1}^{n} \frac{\boldsymbol{U}_i^T\boldsymbol{y}}{\sigma_i}\boldsymbol{V}_i,$$

and

$$\boldsymbol{x}_{\text{SVD}}^r := \sum_{i=1}^{r} \frac{\boldsymbol{U}_i^T\boldsymbol{y}}{\sigma_i}\boldsymbol{V}_i, \tag{2.3}$$

which is a truncated SVD solution using r largest singular values. Figure 2 applies the truncated SVD approach to the deconvolution problem and compares the results with the Tikhonov regularization. As can be seen, truncated SVD solutions are stable and do not seem over-regularize the solution. However, as r increases, truncated SVD solutions tend to be more oscillatory (more unstable) because more "high-frequency" singular modes (corresponding to smaller singular values) are added to the truncated solutions. In other words, *the truncated SVD approach truncates the most unstable modes out of the solution. Why are these high-frequency modes undesirable?* The answer lies on the fact that the jth column of $\boldsymbol{A}$ is the observational vector when the parameter $\boldsymbol{x}$ is the jth canonical basis vector in $\mathbb{R}^p$. Thus the range space (column space) of $\boldsymbol{A}$ is the observable subspace in $\mathbb{R}^d$. Within this observable subspace, the subspace spanned by the singular vector $\boldsymbol{U}_j$ corresponding to larger (largest) singular value is more (most) observable whereas those corresponding to smaller (zero) singular values are less (not) observable. The truncated solution (2.3) clearly shows that the component of the data vector $\boldsymbol{y}$ in the less (or non-) observable subspace, i.e. the product $\boldsymbol{U}_j^T\boldsymbol{y}$ (due to noise or error in the measurements), is amplified by $1/\sigma_j$, thus promoting an oscillatory or unstable solution.



(a) Tikhonov solution (b) Truncated SVD, $r = 3$

(c) Truncated SVD, $r = 7$ (d) Truncated SVD, $r = 15$

FIG. 2. *Deconvolution using a) a Tikhonov regularization with regularization parameter* $\alpha = 1$ *and* $\boldsymbol{x}_0 = \mathbf{0}$*; b) truncated SVD with* $r = 3$*; c) truncated SVD with* $r = 10$*; and d) truncated SVD with* $r = 15$*.*

In this chapter, we refer to row subspaces spanned by $\boldsymbol{V}_j$, corresponding to the **observable** subspaces spanned by $\boldsymbol{U}_j$, as **data-informed** parameter subspaces. Similarly, row subspaces spanned by $\boldsymbol{V}_j$, corresponding to the least observable subspaces spanned by $\boldsymbol{U}_j$, as **data-uninformed** parameter subspaces. The truncated SVD solution (2.3) clearly resides in the **data-informed** parameter subspaces for small r. The question is *where to truncate so that the solution is data-informed?*. The above discussion suggests that one should take r neither too large nor too small. Popular methods for picking r include the Morozov discrepancy principle, L-curve, and cross-validation, to name a few.

A closer look at the truncated SVD solution (2.3) shows that the truncated SVD approach zeros out the data-uninformed modes $\boldsymbol{V}_j$ for $j \geq r+1$. We next show that this is equivalent to infinitely regularizing data-uninformed directions. To see this, let us now consider a regularziation scheme where the data-uninformed modes are penalized infinitely, i.e.,

$$\min \frac{1}{2} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2 + \frac{1}{2} \|L(\boldsymbol{x} - \boldsymbol{x}_0)\|^2, \tag{2.4}$$

where

$$L^TL := \infty\left[\boldsymbol{I}-\boldsymbol{V}^r\left(\boldsymbol{V}^r\right)^T\right] = \infty\left(\boldsymbol{V}^r\right)^\perp\left(\left(\boldsymbol{V}^r\right)^\perp\right)^T$$
$$= \left[\boldsymbol{V}^r,\left(\boldsymbol{V}^r\right)^\perp\right]\left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & \infty\boldsymbol{I}\end{array}\right]\left[\boldsymbol{V}^r,\left(\boldsymbol{V}^r\right)^\perp\right]^T,$$

and $\left[\boldsymbol{I}-\boldsymbol{V}^r\left(\boldsymbol{V}^r\right)^T\right]$ is the orthogonal projection into the data-uninformed subspace spanned by $\{\boldsymbol{V}_j\}_{j=r+1}^d$. The solution of (2.4) is given by

$$\begin{aligned}\boldsymbol{x}_{Inf} &= \left\{\boldsymbol{A}^T\boldsymbol{A}+\infty\left(\boldsymbol{I}-\boldsymbol{V}^r\left(\boldsymbol{V}^r\right)^T\right)\right\}^{-1}\boldsymbol{A}^T\boldsymbol{y}\\ &= \left\{\left[\boldsymbol{V}^r,\left(\boldsymbol{V}^r\right)^\perp\right]\left(\left[\begin{array}{c|c}\boldsymbol{\Sigma}_r^2 & 0\\ \hline 0 & \boldsymbol{D}\end{array}\right]+\left[\begin{array}{c|c}0 & 0\\ \hline 0 & \infty\boldsymbol{I}\end{array}\right]\right)\left[\boldsymbol{V}^r,\left(\boldsymbol{V}^r\right)^\perp\right]^T\right\}^{-1}\boldsymbol{A}^T\boldsymbol{y}\\ &= \boldsymbol{V}^r\left(\boldsymbol{\Sigma}^r\right)^{-2}\left(\boldsymbol{V}^r\right)^T\boldsymbol{A}^T\boldsymbol{y} = \boldsymbol{V}^r\left(\boldsymbol{\Sigma}^r\right)^{-1}\left(\boldsymbol{U}^r\right)^T\boldsymbol{y} =: \boldsymbol{x}_{SVD}^r,\end{aligned}$$

where $\left(\boldsymbol{V}^r\right)^\perp$ is the orthogonal complement of $\boldsymbol{V}^r$ in $\mathbb{R}^p$ and $\boldsymbol{D} := \operatorname{diag}\left[\sigma_{r+1},\ldots,\sigma_p\right]$. The second equality clearly shows that the regularization scheme adds infinity to all singular values that correspond to data-uninformed modes. The last equality proves that infinite regularization on data-uninformed parameter subspace is the same as the truncated SVD approach.

The beauty of the truncated SVD approach is that it avoids putting any regularization on data-informed parameter directions, and hence avoids polluting inverse solutions in these directions, while annihilating data-uninformed directions. However, it is often the case that there is no clear-cut (i.e. $\sigma_k = 0$ for $k \geq r+1$) between the data-informed and data-uninformed ones, but gradual (sometimes exponential) decay of singular values of $\boldsymbol{A}$. In that case, completely removing less data-informed directions may not be ideal, as they may still contain valuable parameter directions seen by the data. Then we may want to impose finite regularization in less data-informed directions, i.e.,

$$\min \frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\right\|^2 + \frac{1}{2}\left\|L\left(\boldsymbol{x}-\boldsymbol{x}_0\right)\right\|^2, \tag{2.5}$$

where

$$L^TL := \alpha\left(\boldsymbol{I}-\boldsymbol{V}^r\left(\boldsymbol{V}^r\right)^T\right) = \alpha\left(\boldsymbol{V}^r\right)^\perp\left(\left(\boldsymbol{V}^r\right)^\perp\right)^T = \left[\begin{array}{c|c}0 & 0\\ \hline 0 & \alpha\boldsymbol{I}\end{array}\right]\left[\boldsymbol{V}^r,\left(\boldsymbol{V}^r\right)^\perp\right]^T.$$

Let us call this approach the **data-informed (DI) regularization method**. The inverse solution in this case reads

$$\begin{aligned}\boldsymbol{x}_{DI} &= \left\{\boldsymbol{A}^T\boldsymbol{A}+\alpha\left(\boldsymbol{I}-\boldsymbol{V}^r\left(\boldsymbol{V}^r\right)^T\right)\right\}^{-1}\boldsymbol{A}^T\boldsymbol{y}\\ &= \left\{\left[\boldsymbol{V}^r,\left(\boldsymbol{V}^r\right)^\perp\right]\left(\left[\begin{array}{c|c}\boldsymbol{\Sigma}_r^2 & 0\\ \hline 0 & \boldsymbol{D}\end{array}\right]+\left[\begin{array}{c|c}0 & 0\\ \hline 0 & \alpha\boldsymbol{I}\end{array}\right]\right)\left[\boldsymbol{V}^r,\left(\boldsymbol{V}^r\right)^\perp\right]^T\right\}^{-1}\boldsymbol{A}^T\boldsymbol{y}\\ &= \left\{\left[\boldsymbol{V}^r,\left(\boldsymbol{V}^r\right)^\perp\right]\left(\left[\begin{array}{c|c}\boldsymbol{\Sigma}_r^2 & 0\\ \hline 0 & \boldsymbol{D}\end{array}\right]+\alpha\left[\begin{array}{c|c}\boldsymbol{I} & 0\\ \hline 0 & \boldsymbol{I}\end{array}\right]-\left[\begin{array}{c|c}\alpha\boldsymbol{I} & 0\\ \hline 0 & 0\end{array}\right]\right)\left[\boldsymbol{V}^r,\left(\boldsymbol{V}^r\right)^\perp\right]^T\right\}^{-1}\boldsymbol{A}^T\boldsymbol{y}.\end{aligned}$$

The last equality suggests that the DI approach is equivalent to *first applying the same*[1] *(finite) regularization for all parameter directions and then removing regularizations in the data-informed directiions.* A few observations are in order: 1) When

[1] Note that α need not be the same for all directions.

$r = 0$, DI becomes the standard Tikhonov regularization; 2) When $\alpha \to \infty$ DI approaches the truncated SVD; and 3) when $\alpha \ll \sigma_i$ for $i \leq r$, i.e. under-regularization, Tikhonov is close to DI as the contribution of the regularization to data-informed modes is small. These observations are clearly demonstrated in Figure 3 for a 1D deconvolution with $\lambda = 0.05$ with various combination of regularization parameter α and the number of retained data-informed modes r. An important feature of the DI technique that can be seen from this result is that for each r the DI solution is robust with the regularization parameter, that is, the solution does not alter significantly, especially for moderate-to-large regularization, while Tikhonov solution is damped out as the regularization parameter increases. The last column of Figure 3 shows that for $r = 20$ the DI solution retains high frequency modes which are not regularized and thus oscillatory.

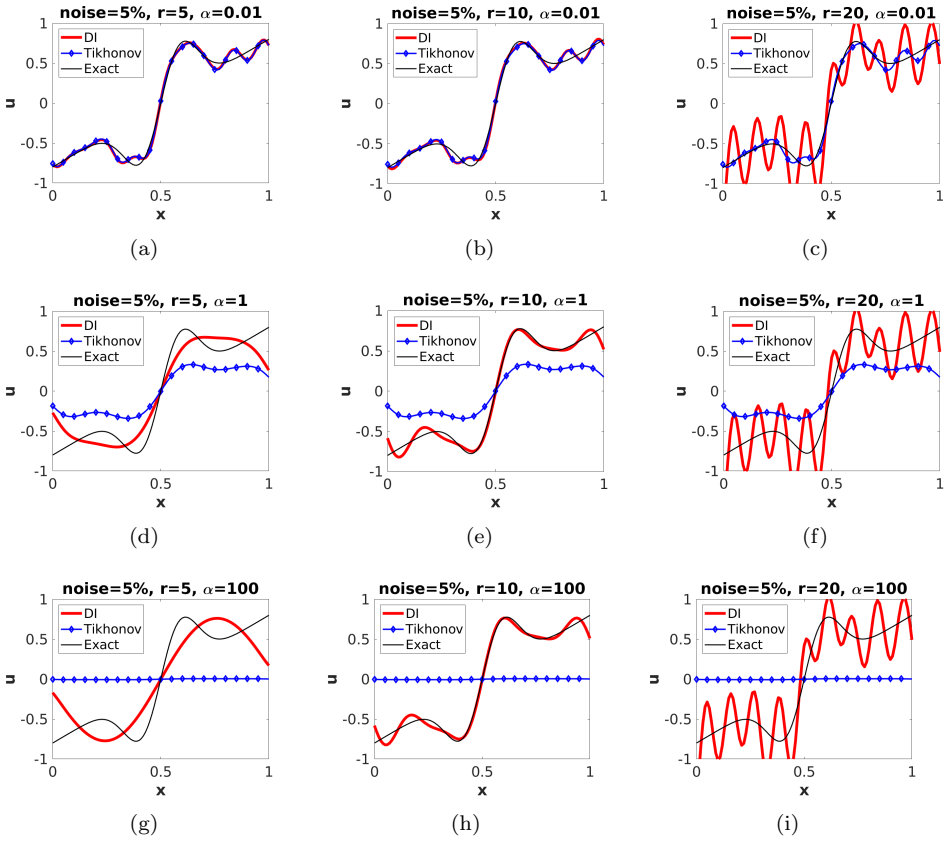


FIG. 3. *Deconvolution with noise level $\lambda = 5\%$ using DI and Tikhonov regularization for various values of regularization parameter and r.*

In order to gain more insights into the behavior of DI regularization, we compute the relative error between the solutions using the DI approach and the truth for a wide range of regularization parameters and a few values of r. The result is shown in Figure 4. As can be seen, when $r = 1$, DI is essentially Tikhonov, which is not surprising as all modes in the DI solution are regularized exactly the same as Tikhonov except for the first one (lowest frequency). For $r = \{5, 10\}$, the DI solution behaves the same as Tikhonov for the under-regularization regime ($\alpha < 0.01$) as expected, and

it outperforms Tikhonov for $\alpha > 0.01$ as the retained data-informed modes, which determine the quality of the deconvolution solution, are left untouched. For $r = 20$, the retained modes now also include high frequency modes and hence the DI approach is not as accurate as Tikhonov for $\alpha < 1$. For all cases with significant number of modes retained, i.e. $r > 5$, the DI solution quality is insensitive to a large range of the regularization parameter.

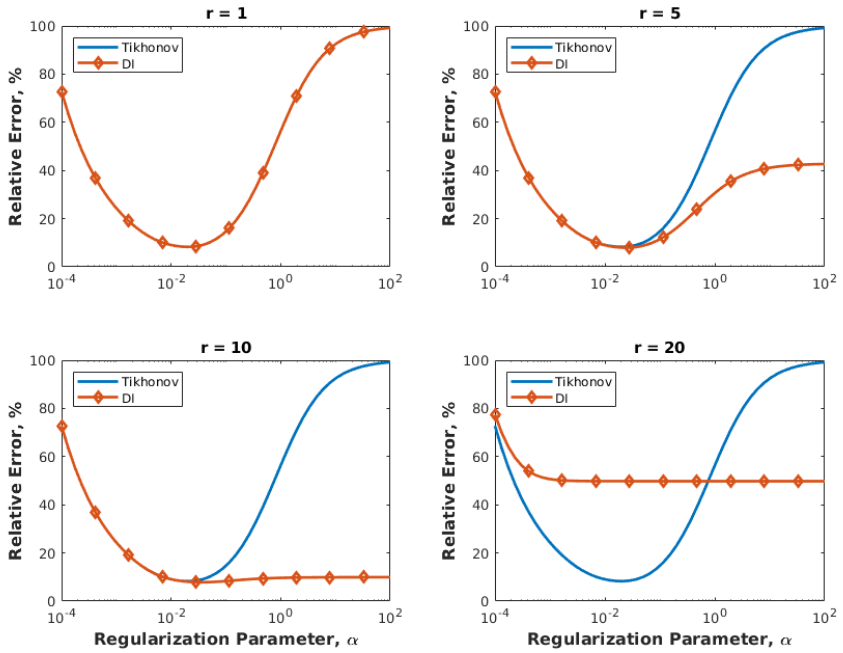


FIG. 4. *Deconvolution with noise level* $\lambda = 5\%$ *using DI and Tikhonov regularizations for* $\alpha = \left[10^{-4}, 10^{2}\right]$ *and* $r = \{1, 5, 10, 20\}$.

It may appear that we have to first construct the rank-r data-informed subspace $\boldsymbol{V}^r$ before applying regularization in its orthogonal complement. However, since $\boldsymbol{V}^r \left(\boldsymbol{V}^r\right)^T$ is the orthogonal projection into the row space of $\boldsymbol{A}$ when $\sigma_i = 0$ for $i > r$, i.e. $\boldsymbol{V}^r \left(\boldsymbol{V}^r\right)^T = \boldsymbol{A}^T \left(\boldsymbol{A}\boldsymbol{A}^T\right)^\dagger \boldsymbol{A}$, we can rewrite the inverse (optimization) problem (2.5) as

$$\min_{\boldsymbol{x}} J := \frac{1}{2} \left\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\right\|^2 + \frac{1}{2} \left\|L\left(\boldsymbol{x} - \boldsymbol{x}_0\right)\right\|^2, \tag{2.6}$$

where

$$L^T L := \alpha \left(\boldsymbol{I} - \boldsymbol{A}^T \left(\boldsymbol{A}\boldsymbol{A}^T\right)^\dagger \boldsymbol{A}\right).$$

In this form, the DI regularization approach (2.6) not only avoids using $\boldsymbol{V}^r$ explicitly but also brings us to a statistical data-informed inverse framework in the next section.

2.2. A statistical data-informed (DI) inverse framework. The cost function in (2.6) can be rewritten as

$$\exp\left(-J\right) = \frac{\exp\left(-\frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\right\|^2\right) \times \exp\left(-\frac{\alpha}{2}\left\|\boldsymbol{x}-\boldsymbol{x}_0\right\|^2\right)}{\exp\left(-\frac{\alpha}{2}\left(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}\boldsymbol{x}_0\right)^T\left(\boldsymbol{A}\boldsymbol{A}^T\right)^\dagger\left(\boldsymbol{A}\boldsymbol{x}-\boldsymbol{A}\boldsymbol{x}_0\right)\right)}.$$

From a Bayesian inverse perspective [6, 13, 3, 8, 7, 12, 10], the numerator is the product of the likelihood

$$\pi_{\text{like}}\left(\boldsymbol{y}|\boldsymbol{x}\right) \propto \exp\left(-\frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\right\|^2\right)$$

from the observational model (2.1) with the noise $\boldsymbol{e} \sim \mathcal{N}\left(\boldsymbol{O},\boldsymbol{I}\right)$, and the Gaussian prior

$$\pi_{\text{prior}}\left(\boldsymbol{x}\right) \propto \exp\left(-\frac{\alpha}{2}\left\|\boldsymbol{x}-\boldsymbol{x}_0\right\|^2\right) \tag{2.7}$$

with mean $\boldsymbol{x}_0$ and $\boldsymbol{I}/\alpha$ covariance matrix. In other words, the numerator is a Bayesian posterior with the aforementioned likelihood and Gaussian prior. *The key difference compared to the Bayesian approach is the denominator.*

We now show that the denominator is nothing more than the pushforward of the prior (2.7) via the forward map $\boldsymbol{A}$. Indeed, let $\tilde{\boldsymbol{y}} := \boldsymbol{A}\boldsymbol{x}$ be a random variable induced by the forward map $\boldsymbol{A}$. With $\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{x}_0, \boldsymbol{I}/\alpha\right)$, $\tilde{\boldsymbol{y}}$ is also a Gaussian with mean $\tilde{\boldsymbol{y}}_0$ and covariance matrix $\mathbb{C}$ where

$$\begin{aligned}
\tilde{\boldsymbol{y}}_0 &:= \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{A}\boldsymbol{x}\right] = \boldsymbol{A}\boldsymbol{x}_0 \\
\mathbb{C} &:= \mathbb{E}_{\boldsymbol{x}}\left[\left(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{y}}_0\right)\left(\tilde{\boldsymbol{y}}-\tilde{\boldsymbol{y}}_0\right)^T\right] = \mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^T\boldsymbol{A}^T\right] = \frac{1}{\alpha}\boldsymbol{A}\boldsymbol{A}^T.
\end{aligned}$$

Note that it is necessary to use the pseudo-inverse for the inverse of the covariance $\mathbb{C}$, i.e., $\mathbb{C}^{-1} := \alpha\left(\boldsymbol{A}\boldsymbol{A}^T\right)^\dagger$ since $\boldsymbol{A}$ may not have full row rank and thus the pushforward distribution can be a degenerate Gaussian.

Remark 2.1. The push-forward of the prior through the parameter-to-observable map $\boldsymbol{A}$ depends on $\boldsymbol{x}$. *It is through this push-forward term that the data-informed (DI) approach learns the data-informed parameter directions.* Indeed, this new approach, through the push-forward term, changes the original prior

$$\exp\left(-\frac{\alpha}{2}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^T\boldsymbol{I}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)\right)$$

to the new one

$$\exp\left(-\frac{\alpha}{2}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)^T\left[\boldsymbol{I}-\boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{A}^T\right)^\dagger\boldsymbol{A}\right]\left(\boldsymbol{x}-\boldsymbol{x}_0\right)\right)$$

in such a way that the new prior leaves the data-informed directions, i.e. the row space of $\boldsymbol{A}$, untouched, and hence only regularizes data-uninformed directions. The data-informed approach accomplishes this by the pushforward of the prior via the parameter-to-observable map $\boldsymbol{A}$.

We can now define the DI posterior as

$$\pi_{\mathrm{DI}}(\boldsymbol{x}|\boldsymbol{y}) = \frac{\pi_{\mathrm{like}}(\boldsymbol{y}|\boldsymbol{x}) \times \pi_{\mathrm{prior}}(\boldsymbol{x})}{\boldsymbol{A}_{\#}\pi_{\mathrm{prior}}(\boldsymbol{x})}, \tag{2.8}$$

where $\boldsymbol{A}_{\#}\pi_{\mathrm{prior}}(\boldsymbol{x})$ denotes the pushforward of $\pi_{\mathrm{prior}}(\boldsymbol{x})$ via the parameter-to-observable map $\boldsymbol{A}$.

We have constructively derived the DI approach by modifying the truncated SVD method and Gaussian prior with scaled-identity covariance matrix. In practice, the prior can be more informative about the correlations among components of $\boldsymbol{x}$ and in that case the covariance matrix is no longer an identity matrix. Let us denote by $\pi_{\mathrm{prior}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}_0, \boldsymbol{\Gamma}/\alpha)$ the Gaussian prior with covariance matrix $\boldsymbol{\Gamma}/\alpha$. Let us also consider a more general data distribution where, for a given parameter $\boldsymbol{x}$, the data is distributed by the Gaussian $\mathcal{N}(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{\Lambda})$. In order to use most of the above results, let us whiten both the parameter and observations. In particular, $\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{y}$ is the whitened observations (inducing $\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}$ as the new parameter-to-observable forward map) and $\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x}$ is the prior-whitened parameter. The pushforward of the prior via $\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}$ now reads[2]

$$\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}_{\#}\pi_{\mathrm{prior}}(\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{x}_0, \frac{1}{\alpha}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\Gamma\boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right), \tag{2.9}$$

The DI posterior (2.8) with whitened parameter, whitened observations, and induced parameter-to-observable map now becomes

$$\begin{aligned}\pi_{\mathrm{DI}}(\boldsymbol{x}|\boldsymbol{y}) &= \frac{\pi_{\mathrm{like}}(\boldsymbol{y}|\boldsymbol{x}) \times \pi_{\mathrm{prior}}(\boldsymbol{x})}{\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}_{\#}\pi_{\mathrm{prior}}(\boldsymbol{x})} \\ &\propto \frac{\exp\left(-\frac{1}{2}\left\|\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{y}\right\|^2\right) \times \exp\left(-\frac{\alpha}{2}\left\|\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x} - \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x}_0\right\|^2\right)}{\exp\left(-\frac{\alpha}{2}\left\|\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{x}_0\right\|^2_{\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\Gamma\boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)^\dagger}\right)},\end{aligned} \tag{2.10}$$

which, after writing the push-forward measure in terms of the whitened parameter, reads

$$\pi_{\mathrm{DI}}(\boldsymbol{x}|\boldsymbol{y}) \propto \frac{\exp\left(-\frac{1}{2}\left\|\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{y}\right\|^2\right) \times \exp\left(-\frac{\alpha}{2}\left\|\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x} - \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x}_0\right\|^2\right)}{\exp\left(-\frac{\alpha}{2}\left\|\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x} - \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x}_0\right\|^2_{\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{A}^T\boldsymbol{\Lambda}^{\frac{1}{2}}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\Gamma\boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)^\dagger\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}^{\frac{1}{2}}}\right)},$$

or equivalently

$$-\log\left(\pi_{\mathrm{DI}}(\boldsymbol{x}|\boldsymbol{y})\right) \propto \frac{1}{2}\left\|\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{y}\right\|^2 + \frac{1}{2}\left\|L\left(\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x} - \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x}_0\right)\right\|^2, \tag{2.11}$$

[2]Note that using the modified forward map $\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}$ is not necessary as using the original map $\boldsymbol{A}$ yields the same result.

where

$$
\begin{aligned}
L^T L &= \alpha \left(\boldsymbol{I} - \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{A}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \left(\boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Gamma} \boldsymbol{A}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \right)^{\dagger} \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Gamma}^{\frac{1}{2}} \right) \\
&= \alpha \left(\boldsymbol{I} - \boldsymbol{V}^n \boldsymbol{V}^{nT} \right) = \alpha \left(\boldsymbol{V}^n \right)^{\perp} \left(\left(\boldsymbol{V}^n \right)^{\perp} \right)^T \\
&= \left[\boldsymbol{V}^n, \left(\boldsymbol{V}^n \right)^{\perp} \right] \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & \alpha \boldsymbol{I} \end{array} \right] \left[\boldsymbol{V}^n, \left(\boldsymbol{V}^n \right)^{\perp} \right]^T, \qquad (2.12)
\end{aligned}
$$

where $\boldsymbol{V}^n$ contains the first n right singular vectors of the following SVD

$$
\boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Gamma}^{\frac{1}{2}} := \boldsymbol{U} \boldsymbol{\Sigma} \boldsymbol{V}^T. \qquad (2.13)
$$

As can be seen, the push-forward measure seeks to find the first n columns of $\boldsymbol{V}$ associated with the n non-zero singular values. The DI method then avoids regularizing these "data-informed directions" $\boldsymbol{V}^n$. In other words, in the whitened parameter, the induced regularization by the prior is identity and the DI approach removes regularization in the parameter subspace spanned by $\boldsymbol{V}^n$. From (2.13) it is clear that $\boldsymbol{V}^n$ now depends on both the prior covariance $\boldsymbol{\Gamma}$ and the observational covariance $\boldsymbol{\Lambda}$ in addition to $\boldsymbol{A}$. *So how do we understand the parameter subspace spanned by $\boldsymbol{V}^n$, and hence the DI approach?* To that end, let us define $\boldsymbol{\Sigma}^{\frac{1}{2}}$ to be the same as $\boldsymbol{\Sigma}$ except on the main the diagonal where $\boldsymbol{\Sigma}^{\frac{1}{2}}(i,i) = \sqrt{\boldsymbol{\Sigma}(i,i)} = \sqrt{\sigma_i}$ (note that $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is nothing more than the square-root of $\boldsymbol{\Sigma}$ when $\boldsymbol{\Sigma}$ is a square matrix). Let $\boldsymbol{\Psi}$ be the first n rows of $\boldsymbol{\Sigma}^{\frac{1}{2}}$ and $\boldsymbol{\Phi}$ be the first n columns of $\boldsymbol{\Sigma}^{\frac{1}{2}}$. Clearly, by definition $\boldsymbol{\Psi}(i,i) = \boldsymbol{\Phi}(i,i) = \sqrt{\sigma_i}$ for $i \leq n$.

Let us define the following maps

$$
\boldsymbol{z} := \boldsymbol{T}\boldsymbol{x}, \quad \text{where} \quad \boldsymbol{T} := \boldsymbol{\Psi} \boldsymbol{V}^T \boldsymbol{\Gamma}^{-\frac{1}{2}}, \qquad (2.14)
$$

$$
\boldsymbol{w} := \boldsymbol{S}\boldsymbol{y}, \quad \text{where} \quad \boldsymbol{S} := \boldsymbol{\Phi}^T \boldsymbol{U}^T \boldsymbol{\Lambda}^{-\frac{1}{2}}. \qquad (2.15)
$$

where $\boldsymbol{z}$ are the first n coordinates of $\boldsymbol{x}$ in $\boldsymbol{V}$, after whitening via $\boldsymbol{\Gamma}^{-\frac{1}{2}}$ and then being scaled by $\boldsymbol{\Psi}$. Similarly, $\boldsymbol{w}$ are the first n coordinates of $\boldsymbol{y}$ in $\boldsymbol{U}$, after whitening via $\boldsymbol{\Lambda}^{-\frac{1}{2}}$ and then being scaled by $\boldsymbol{\Phi}$. The map $\boldsymbol{T}$ pushes forward the prior in $\boldsymbol{x}$ to the prior in $\boldsymbol{z}$ as

$$
\pi_{\text{prior}}(\boldsymbol{z}) \sim \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \sigma_i^{-1} \left(\boldsymbol{z}_i - \overline{\boldsymbol{z}}_i \right)^2 \right), \qquad (2.16)
$$

where $\overline{\boldsymbol{z}} = \boldsymbol{T}\boldsymbol{x}_0$. Similarly, given $\boldsymbol{x}$ (and hence $\boldsymbol{z}$), the induced likelihood in terms of $\boldsymbol{w}$ is given by

$$
\pi_{\text{like}}(\boldsymbol{w}|\boldsymbol{z}) \sim \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \sigma_i^{-1} \left(\boldsymbol{w}_i - \sigma_i \boldsymbol{z}_i \right)^2 \right). \qquad (2.17)
$$

As can be seen from (2.16) and (2.17), the maps $\boldsymbol{T}$ and $\boldsymbol{S}$ transform the original parameter $\boldsymbol{x}$ and original data $\boldsymbol{y}$ to new parameter $\boldsymbol{z}$ and new data $\boldsymbol{w}$. Two observations are in order: 1) though in general the original parameter and data dimensions are different, the new parameter and data have the same dimension; and 2) the new data $\boldsymbol{w}$ and new parameter $\boldsymbol{z}$, up to the difference in the mean, have the

same distribution. In particular, both $\boldsymbol{z}$ and $\boldsymbol{w}$ are $\mathbb{R}^n$-vector of independent Gaussian distributions with diagonal covariance matrix $\boldsymbol{\Theta} \in \mathbb{R}^{n\times n}$ with $\boldsymbol{\Theta}_{ii} = \sigma_i$. Both $\boldsymbol{z}_i$ and $\boldsymbol{w}_i$, up to the difference in the mean, are the same Gaussian distribution with variance σ_i. Since $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n > 0$, the independent random variable $\boldsymbol{z}_i$ (and hence $\boldsymbol{w}_i$) are ranked from the one with most variance to the one with least variance.

Let us call the ith column of $\boldsymbol{U}$, namely $\boldsymbol{U}_i$, the ith important direction in the data space, and the ith column of $\boldsymbol{V}$, namely $\boldsymbol{V}_i$, the ith important direction in the parameter space. Let us also rank the degree of importance of $\boldsymbol{U}_i$ and $\boldsymbol{V}_i$ by the magnitude of σ_i. It follows that the transformations $\boldsymbol{T}$ and $\boldsymbol{S}$ map the original parameter $\boldsymbol{x}$ and data $\boldsymbol{y}$ into new parameter $\boldsymbol{z}$ and data $\boldsymbol{w}$ in which the corresponding parameter $\boldsymbol{z}_i$ and data $\boldsymbol{w}_i$ are equally important. This is similar to the concept of balanced transformation in control theory (see, e.g., [4, 1] and the references therein). The new parameter $\boldsymbol{z}$ is thus equally data-informed and prior-informed. In particular $\boldsymbol{z}_i$ is equally less data-informed and prior-informed relatively to $\boldsymbol{z}_j$ for $j < i$.

The DI method thus regularizes only the (equally) data-uninformed and prior-uninformed parameters/directions.

2.3. Properties of the DI regularization approach.

2.3.1. Deterministic properties. It is easy to see the optimality condition of the optimization problem $\max_{\boldsymbol{x}} \log\left(\pi_{\mathrm{DI}}\left(\boldsymbol{x}|\boldsymbol{y}\right)\right)$ is given by

$$\boldsymbol{H}\boldsymbol{x}_{\mathrm{DI}} = \boldsymbol{b}, \tag{2.18}$$

where

$$\begin{aligned}
\boldsymbol{H} &:= \left\{\boldsymbol{A}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{A} + \alpha\left[\boldsymbol{\Gamma}^{-1} - \boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)^{\dagger}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\right]\right\},\\
\boldsymbol{b} &:= \boldsymbol{A}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{y} + \alpha\left[\boldsymbol{\Gamma}^{-1} - \boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)^{\dagger}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\right]\boldsymbol{x}_0
\end{aligned}$$

In order to solve the optimality condition (2.18) in practice, we can use the rank-r approximation

$$\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}^{\frac{1}{2}} = \boldsymbol{U}^n\boldsymbol{\Sigma}^n\left(\boldsymbol{V}^n\right)^T \approx \boldsymbol{U}^r\boldsymbol{\Sigma}^r\left(\boldsymbol{V}^r\right)^T \tag{2.19}$$

for the push-forward matrix $\boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)^{\dagger}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}$, where n is largest index for which $\sigma_n > 0$. Thus rank-r approximations for $\boldsymbol{H}$ and $\boldsymbol{y}$ are given by

$$\begin{aligned}
\boldsymbol{H}^r &:= \boldsymbol{A}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{A} + \alpha\left(\boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{V}^r\left(\boldsymbol{V}^r\right)^T\boldsymbol{\Gamma}^{-\frac{1}{2}}\right),\\
\boldsymbol{b}^r &:= \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{V}^n\boldsymbol{\Sigma}^n\left(\boldsymbol{U}^n\right)^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{y} + \alpha\left[\boldsymbol{\Gamma}^{-1} - \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{V}^r\left(\boldsymbol{V}^r\right)^T\boldsymbol{\Gamma}^{-\frac{1}{2}}\right]\boldsymbol{x}_0.
\end{aligned}$$

Note that we don't perform low rank approximation for the term $\boldsymbol{A}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{y}$ in $\boldsymbol{y}$ as it requires only a matrix-vector product. We also leave the first term in $\boldsymbol{H}^r$ as is, since we invert $\boldsymbol{H}^r$ using the conjugate gradient (CG) method which requires only matrix-vector products. In the numerical results section, section 3, we present a nested optimization method that avoids the low rank approximation altogether. The analysis of such method is, however, more technical and thus left for future work. The rank-r approximation to the solution of the optimality condition (2.18) is defined as

$$\boldsymbol{H}^r\boldsymbol{x}^r_{\mathrm{DI}} = \boldsymbol{b}^r, \tag{2.20}$$

which is solved approximately using the conjugate gradient (CG) method (see Section 3). For the following analysis, however, we compute the inverse of $\boldsymbol{H}^r$ analytically. To that end, we rewrite $\boldsymbol{H}^r$ in terms of n singular vectors corresponding to the n nonzero singular values as

$$\boldsymbol{H}^r = \alpha \boldsymbol{\Gamma}^{-\frac{1}{2}} \left[\boldsymbol{I} + \boldsymbol{V}^n \boldsymbol{D}^n \left(\boldsymbol{V}^n\right)^T \right] \boldsymbol{\Gamma}^{-\frac{1}{2}},$$

where $\boldsymbol{D}^n$ is an $n \times n$ diagonal matrix with $\boldsymbol{D}^n(i,i) = \left(\sigma_i^2 - \alpha\right)/\alpha$ for $i \leq r$ and $\boldsymbol{D}^n(i,i) = \sigma_i^2/\alpha$ for $r < i \leq n$.

LEMMA 2.2. *The DI solution with r data-informed modes reads*

$$\boldsymbol{x}_{DI}^r := \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{V}^n \boldsymbol{\Theta}^n \left(\boldsymbol{U}^n\right)^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{y} + \left[\boldsymbol{I} - \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{V}^n \overline{\boldsymbol{I}}^n \left(\boldsymbol{V}^n\right)^T \boldsymbol{\Gamma}^{-\frac{1}{2}} \right] \boldsymbol{x}_0, \tag{2.21}$$

where $\boldsymbol{\Theta}^n$ is an $n \times n$ diagonal matrix with $\boldsymbol{\Theta}^n(i,i) = \sigma_i^{-1}$ for $i \leq r$ and $\boldsymbol{\Theta}^n(i,i) = \sigma_i/\left(\sigma_i^2 + \alpha\right)$ for $r < i \leq n$. Here, $\overline{\boldsymbol{I}}^n$ is an $n \times n$ diagonal matrix with $\overline{\boldsymbol{I}}^n(i,i) = 1$ for $i \leq r$ and $\overline{\boldsymbol{I}}^n(i,i) = \sigma_i^2/\left(\sigma_i^2 + \alpha\right)$ for $r < i \leq n$. Furthermore,

$$\boldsymbol{A}\boldsymbol{x}_{DI}^n = \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{U}^n \left(\boldsymbol{U}^n\right)^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{y}. \tag{2.22}$$

Proof. Using a Woodbury formula we have

$$\left(H^r\right)^{-1} = \frac{1}{\alpha} \boldsymbol{\Gamma}^{\frac{1}{2}} \left[\boldsymbol{I} - \boldsymbol{V}^n \overline{\boldsymbol{D}}_{\mathrm{DI}}^{n,r} \left(\boldsymbol{V}^n\right)^T \right] \boldsymbol{\Gamma}^{\frac{1}{2}}, \tag{2.23}$$

where $\overline{\boldsymbol{D}}^{n,r}$ is an $n \times n$ diagonal matrix with $\overline{\boldsymbol{D}}_{\mathrm{DI}}^{n,r}(i,i) = \left(\sigma_i^2 - \alpha\right)/\sigma_i^2$ for $i \leq r$ and $\overline{\boldsymbol{D}}_{\mathrm{DI}}^{n,r}(i,i) = \sigma_i^2/\left(\sigma_i^2 + \alpha\right)$ for $r < i \leq n$. The computation of the product $\left(H^r\right)^{-1} \boldsymbol{y}^r$ to arrive at the assertion is straightforward algebra manipulation and hence omitted. ∎

The result (2.22) shows that the image of the DI solution $\boldsymbol{x}_{\mathrm{DI}}$ through the parameter-to-observable map is exactly the data if $\boldsymbol{U}^n \left(\boldsymbol{U}^n\right)^T = \boldsymbol{I}$ or $\boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{y}$ resides in the column space of $\boldsymbol{U}^n$. This happens, for example, when $\boldsymbol{A}$ has full row-rank and the number of data is not more than the dimension of the parameter, i.e., $d \leq p$. In this case, retaining all modes corresponding to non-zero singular values in the DI solution makes the data misfit vanish, that is, the DI solution in this case would match the noise, which is undesirable. As discussed in Section 2.1, r should be smaller than n for the solution to be meaningful. Let us define

$$r_\varepsilon := \max\left\{ i : 1 \leq i \leq n \text{ and } \sigma_i \geq \varepsilon \right\},$$

for some $\varepsilon > 0$, and

$$\mathcal{R}_\varepsilon := \left(\boldsymbol{H}^{r_\varepsilon}\right)^{-1} \boldsymbol{A}^T \boldsymbol{\Lambda}^{-\frac{1}{2}}.$$

THEOREM 2.3. *For any $\varepsilon > 0$ and $\alpha > 0$, consider the inverse problem*

$$\min_{\boldsymbol{x}} \mathcal{J} = \frac{1}{2} \left\| \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{y} \right\|^2 + \frac{1}{2} \left\| L \boldsymbol{\Gamma}^{-\frac{1}{2}} \left(\boldsymbol{x} - \boldsymbol{x}_0\right) \right\|^2, \tag{2.24}$$

using the DI technique with rank-r_ε approximation, where

$$L^T L = \alpha \left(\boldsymbol{I} - \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{A}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \left(\boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Gamma} \boldsymbol{A}^T \boldsymbol{\Lambda}^{-\frac{1}{2}} \right)^\dagger \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Gamma}^{\frac{1}{2}} \right).$$

The following hold:

i) The inverse problem with rank r_ε DI technique is well-posed.

ii) Suppose that the nullspace of $\boldsymbol{A}$ is trivial, i.e. $\mathcal{N}(\boldsymbol{A}) = \{0\}$, then the DI technique is a regularization strategy in the following sense

$$\lim_{\varepsilon \to 0} \mathcal{R}_\varepsilon \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}.$$

iii) If $\alpha = \mathcal{O}(\varepsilon)$ and $\mathcal{N}(\boldsymbol{A}) = \{0\}$, then the rank-$r_\varepsilon$ DI technique is an admissible regularization method.

Proof. From Lemma 2.2 we see that the DI solution $\boldsymbol{x}_{\mathrm{DI}}^{r_\varepsilon}$ is unique and furthermore

$$\|\boldsymbol{x}_{\mathrm{DI}}^{r_\varepsilon}\| \leq \beta(\varepsilon, \alpha) \left\|\boldsymbol{\Gamma}^{\frac{1}{2}}\right\| \left\|\boldsymbol{\Lambda}^{-\frac{1}{2}}\right\| \|\boldsymbol{y}\| + \left(1 + \sqrt{\kappa(\boldsymbol{\Gamma})}\right) \|\boldsymbol{x}_0\|,$$

where $\kappa(\boldsymbol{\Gamma})$ denotes the condition number of $\boldsymbol{\Gamma}$, $\beta(\varepsilon, \alpha)$ is a constant defined as

$$\beta(\varepsilon, \alpha) := \frac{1}{\min_{r_\varepsilon < i \leq n} \{r_\varepsilon, \sigma_i + \alpha/\sigma_i\}},$$

which shows the the DI solution is stable, which in turn proves *i)*. To see assertion ii), we use the definition of $\mathcal{R}_\varepsilon$ and the SVD of $\boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Gamma}^{\frac{1}{2}}$ to arrive at

$$\mathcal{R}_\varepsilon \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A} = \frac{1}{\alpha} \boldsymbol{\Gamma}^{\frac{1}{2}} \left[\boldsymbol{I} - \boldsymbol{V}^n \overline{\boldsymbol{D}}^{n, r_\varepsilon} (\boldsymbol{V}^n)^T\right] \boldsymbol{V}^n (\boldsymbol{\Sigma}^n)^2 (\boldsymbol{V}^n)^T \boldsymbol{\Gamma}^{-\frac{1}{2}} =$$
$$\boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{V}^n \left[\begin{array}{c|c} \boldsymbol{I} & 0 \\ \hline 0 & \operatorname{diag}\left(\frac{\sigma_i^2}{\sigma_i^2 + \alpha}\right)_{r_\varepsilon < i \leq n} \end{array}\right] (\boldsymbol{V}^n)^T \boldsymbol{\Gamma}^{-\frac{1}{2}},$$

which implies

$$\lim_{\varepsilon \to 0} \mathcal{R}_\varepsilon \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A}\boldsymbol{x} = \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{V}^n \boldsymbol{I} (\boldsymbol{V}^n)^T \boldsymbol{\Gamma}^{-\frac{1}{2}} \boldsymbol{x} = \boldsymbol{x},$$

where we have used the fact that $r_\varepsilon \to n$ as $\varepsilon \to 0$, and that $\boldsymbol{V}^n (\boldsymbol{V}^n)^T = \boldsymbol{I}$ since $\mathcal{N}(\boldsymbol{A}) = \{0\}$.

For assertion *iii)*, it is sufficient to show that

$$\sup_{\boldsymbol{y}} \left\{ \left\| \mathcal{R}_\varepsilon \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{y} - \boldsymbol{x} \right\| : \left\| \boldsymbol{\Lambda}^{-\frac{1}{2}} (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) \right\| \leq \varepsilon \right\} \to 0 \text{ as } \varepsilon \to 0,$$

for any $\boldsymbol{x}$. We have

$$\left\| \mathcal{R}_\varepsilon \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{y} - \boldsymbol{x} \right\| \leq \left\| \mathcal{R}_\varepsilon \boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A}\boldsymbol{x} - \boldsymbol{x} \right\| + \left\| \mathcal{R}_\varepsilon \boldsymbol{\Lambda}^{-\frac{1}{2}} (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}) \right\|$$
$$\leq \left\| \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{V}^n \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & \operatorname{diag}\left(\frac{-\alpha}{\sigma_i^2 + \alpha}\right)_{r_\varepsilon < i \leq n} \end{array}\right] (\boldsymbol{V}^n)^T \boldsymbol{\Gamma}^{-\frac{1}{2}} \right\| \|\boldsymbol{x}\| + \|\mathcal{R}_\varepsilon\| \varepsilon$$
$$\leq \frac{\alpha}{\sigma_n^2 + \alpha} \sqrt{\kappa(\boldsymbol{\Gamma})} \|x\| + \varepsilon \left\|\boldsymbol{\Gamma}^{\frac{1}{2}}\right\| \left\| \left[\begin{array}{c|c} \operatorname{diag}\left(\frac{1}{\sigma_i}\right)_{i \leq r_\varepsilon} & 0 \\ \hline 0 & \operatorname{diag}\left(\frac{\sigma_i}{\sigma_i^2 + \alpha}\right)_{r_\varepsilon < i \leq n} \end{array}\right] \right\|$$
$$\leq \frac{\alpha}{\sigma_n^2 + \alpha} \sqrt{\kappa(\boldsymbol{\Gamma})} \|x\| + \varepsilon \sigma_n^{-1} \left\|\boldsymbol{\Gamma}^{\frac{1}{2}}\right\|,$$

where we have used the result from *ii)*, definition of $\mathcal{R}_\varepsilon$, and the orthonormality of $\boldsymbol{V}$ and $\boldsymbol{U}$. Using the assumption $\alpha = \mathcal{O}(\varepsilon)$ concludes the proof. ∎

Remark 2.4. Note that most of the above arguments are still valid for infinite dimensional setting, i.e. $p = \infty$, assuming that $\boldsymbol{\Gamma}$ is a trace class. Indeed, $\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}^{\frac{1}{2}}$ is then a compact operator and we can invoke the infinite dimensional singular value decomposition [2] for $\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}^{\frac{1}{2}}$. Note that all the matrices are now interpreted as operators, transpose operator (superscript T) as adjoint operator, and $\boldsymbol{\Gamma}^{-\frac{1}{2}}$ as pseudo-inverse if $\mathcal{N}(\boldsymbol{\Gamma}) \neq \{0\}$. We leave out the details for the sake of brevity.

2.3.2. Statistical properties. Now we discuss some probabilistic aspects of the DI prior and the DI posterior. Since the regularization parameter α plays no role in the following discussion, we absorb it into $\boldsymbol{\Gamma}$. We define the DI prior as

$$\pi_{\text{DI-prior}}(\boldsymbol{x}) \sim \exp\left\{-\frac{1}{2}\left\|L\boldsymbol{\Gamma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{x}_0)\right\|^2\right\}. \tag{2.25}$$

From (2.12), the DI prior (pseudo-) inverse covariance is given by

$$\begin{aligned}(\mathbb{C}^n)^\dagger &:= \boldsymbol{\Gamma}^{-\frac{1}{2}}\left[\boldsymbol{I}-\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)^\dagger\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}^{\frac{1}{2}}\right]\boldsymbol{\Gamma}^{-\frac{1}{2}}\\ &= \boldsymbol{\Gamma}^{-\frac{1}{2}}\left[\boldsymbol{I}-\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^T\right)^\dagger\boldsymbol{A}\boldsymbol{\Gamma}^{\frac{1}{2}}\right]\boldsymbol{\Gamma}^{-\frac{1}{2}} = \boldsymbol{\Gamma}^{-\frac{1}{2}}(\boldsymbol{V}^n)^\perp\left((\boldsymbol{V}^n)^\perp\right)^T\boldsymbol{\Gamma}^{-\frac{1}{2}},\end{aligned}$$

where we have used the fact that $\boldsymbol{\Lambda}$ is invertible in the second equality. Thus $\boldsymbol{\Lambda}$ actually contribute to neither the DI prior nor its rank-r version

$$(\mathbb{C}^r)^\dagger := \boldsymbol{\Gamma}^{-\frac{1}{2}}(\boldsymbol{V}^r)^\perp\left((\boldsymbol{V}^r)^\perp\right)^T\boldsymbol{\Gamma}^{-\frac{1}{2}}.$$

The rank-r DI covariance thus reads

$$\mathbb{C}^r := \boldsymbol{\Gamma}^{\frac{1}{2}}(\boldsymbol{V}^r)^\perp\left[(\boldsymbol{V}^r)^\perp\right]^T\boldsymbol{\Gamma}^{\frac{1}{2}} = \boldsymbol{\Gamma}^{\frac{1}{2}}\left(\boldsymbol{I}-\boldsymbol{V}^r(\boldsymbol{V}^r)^T\right)\boldsymbol{\Gamma}^{\frac{1}{2}} \tag{2.26}$$

which is clearly symmetric positive semidefinite in $\mathbb{R}^p$, though degenerate.[3] The DI-prior (2.25) is not a well defined density in $\mathbb{R}^p$, that is, it is not absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}^p$. This is not surprising as we argue above the the DI-prior is the prior on the less data-informed directions. Let us define

$$\boldsymbol{z}^\perp := \boldsymbol{T}^\perp\boldsymbol{x}, \quad \text{where} \quad \boldsymbol{T}^\perp := \left((\boldsymbol{V}^r)^\perp\right)^T\boldsymbol{\Gamma}^{-\frac{1}{2}}.$$

THEOREM 2.5. *The following hold true:*

i) $\boldsymbol{z}$ and $\boldsymbol{z}^\perp$ are distributed by the push-forward density of the prior through $\boldsymbol{T}$ and $\boldsymbol{T}^\perp$, respectively. In particular, $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{T}\boldsymbol{x}_0, \boldsymbol{I})$ and $\boldsymbol{z}^\perp \sim \mathcal{N}\left(\boldsymbol{T}^\perp\boldsymbol{x}_0, \boldsymbol{I}\right)$.

ii) the DI-prior density is the density of $\boldsymbol{z}^\perp$, and hence is well-defined.

iii) The DI-prior density is the conditional density of $\boldsymbol{x}$ given $\boldsymbol{z}$.

Proof. Assertion *i)* is straightforward. To see the the second assertion, we note that the density of $\boldsymbol{z}^\perp$ can be written as

$$\begin{aligned}\exp\left\{-\frac{1}{2}\left\|\boldsymbol{z}^\perp-\boldsymbol{T}^\perp\boldsymbol{x}_0\right\|^2\right\} &= \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)\left(\boldsymbol{T}^\perp\right)^T\boldsymbol{T}^\perp(\boldsymbol{x}-\boldsymbol{x}_0)\right\}\\ &= \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)\boldsymbol{\Gamma}^{-\frac{1}{2}}(\boldsymbol{V}^r)^\perp\left((\boldsymbol{V}^r)^\perp\right)^T\boldsymbol{\Gamma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{x}_0)\right\},\end{aligned}$$

[3] The nullspace of $\mathbb{C}^r$: $\mathcal{N}(\mathbb{C}^r) := \left\{\boldsymbol{x} : \boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{x} \in \mathcal{R}(\boldsymbol{V}^r)\right\}$, where $\mathcal{R}(\cdot)$ denotes the range space.

which is exactly the DI prior (2.25). In other words, we have shown that the DI prior is a well defined density on $\boldsymbol{z}^{\perp}$. To see assertion *iii*) we observe that

$$\pi_{\text{prior}}(\boldsymbol{x}) = \pi_{\text{prior}}\left(\boldsymbol{V}^r\boldsymbol{z} + (\boldsymbol{V}^r)^{\perp}\boldsymbol{z}^{\perp}\right),$$

and thus

$$\pi_{\text{prior}}(\boldsymbol{x}|\boldsymbol{z}) = \frac{\pi_{\text{prior}}(\boldsymbol{x})}{\pi(\boldsymbol{z})} = \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)\boldsymbol{\Gamma}^{-\frac{1}{2}}(\boldsymbol{V}^r)^{\perp}\left((\boldsymbol{V}^r)^{\perp}\right)^T\boldsymbol{\Gamma}^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{x}_0)\right\},$$

which is exactly the DI prior since $\pi(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{T}\boldsymbol{x}_0, \boldsymbol{I})$ is exactly the push-forward density of $\pi_{\text{prior}}(\boldsymbol{x})$ via the map $\boldsymbol{T}$. ∎

Remark 2.6. Note that the above decomposition of $\boldsymbol{x}$ into $\boldsymbol{z}$ and $\boldsymbol{z}^{\perp}$, through the maps $\boldsymbol{T}$ and $\boldsymbol{T}^{\perp}$, is still valid for infinite dimensional settings. However, $\boldsymbol{z}^{\perp}$ would be distributed by an infinite dimensional Gaussian measure with identity covariance operator, which is not a valid Gaussian measure. A more general understanding of the DI prior is through disintegration. Indeed, under mild conditions on the map $\boldsymbol{T}$ and its push-forward measure of the prior measure, the DI prior is nothing more than a disintegration of the prior measure via the map $\boldsymbol{T}$ and this view is also valid for infinite dimensional settings.

To quantify the uncertainty in the DI inverse solution (2.21), we can use the covariance matrix of the DI posterior (2.10). For linear inverse problems with Gaussian prior and Gaussian noise—the problems considered in this chapter—the covariance matrix is exactly the inverse of the Hessian. For rank-r DI approach, the DI posterior covariance matrix $\mathcal{C}_{\text{DI}}^{\text{post}}$ is given in (2.23), i.e.,

$$\mathcal{C}_{\text{DI}}^{\text{post}} = \frac{1}{\alpha}\boldsymbol{\Gamma} - \frac{1}{\alpha}\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{V}^n\overline{\boldsymbol{D}}_{\text{DI}}^{n,r}(\boldsymbol{V}^n)^T\boldsymbol{\Gamma}^{\frac{1}{2}} \tag{2.27}$$

It is easy to see that the covariance matrix corresponding to the Tikhonov regularization is given by

$$\mathcal{C}_{\text{Tik}}^{\text{post}} = \frac{1}{\alpha}\boldsymbol{\Gamma} - \frac{1}{\alpha}\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{V}^n\overline{\boldsymbol{D}}_{\text{Tik}}^{n,r}(\boldsymbol{V}^n)^T\boldsymbol{\Gamma}^{\frac{1}{2}}, \tag{2.28}$$

where both $\overline{\boldsymbol{D}}_{\text{DI}}^{n,r}$ and $\overline{\boldsymbol{D}}_{\text{Tik}}^{n,r}$ are diagonal matrices given in Table 1. Note that we have used α as the magnitude of the regularization to study the robustness and accuracy of all methods. If not needed, α can be straightforwardly absorbed into $\boldsymbol{\Gamma}$, and hence σ_i^2; in that case α is simply replaced by 1 everywhere (including those in Table 1) it appears. As can be seen, $\overline{\boldsymbol{D}}_{\text{Tik}}^{n,r}(i,i)$ is always non-negative for all i,

	$i \le r$	$r < i \le n$
DI posterior	$\overline{\boldsymbol{D}}_{\text{DI}}^{n,r}(i,i) = \dfrac{\sigma_i^2 - \alpha}{\sigma_i^2}$	$\overline{\boldsymbol{D}}_{\text{DI}}^{n,r}(i,i) = \dfrac{\sigma_i^2}{\sigma_i^2 + \alpha}$
Tikhonov posterior	$\overline{\boldsymbol{D}}_{\text{Tik}}^{n,r}(i,i) = \dfrac{\sigma_i^2}{\sigma_i^2 + \alpha}$	$\overline{\boldsymbol{D}}_{\text{Tik}}^{n,r}(i,i) = \dfrac{\sigma_i^2}{\sigma_i^2 + \alpha}$

TABLE 1
The difference between the DI and the Tikhonov Covariance matrices.

while $\overline{\boldsymbol{D}}_{\text{DI}}^{n,r}(i,i)$ is negative when $\sigma_i^2 < \alpha$ for $i \le r$. That is, while the Tikhonov

posterior uncertainty, $\mathcal{C}_{\text{Tik}}^{\text{post}}$ (Bayesian posterior with standard Gaussian prior), is always smaller than the prior uncertainty $\boldsymbol{\Gamma}$ no matter how much informed the data is, the DI posterior uncertainty could be higher than the prior counterpart if the data supports this. In other words, standard (or typical) Gaussian prior does not allow the data to increase the uncertainty, and hence prone to producing over-confident results (see Section 3). The DI prior, on the other hand, takes the parameter-to-observable map (the proxy to the data) into account, and thus along parameter directions that are more data-informed, i.e. $\sigma_i^2 \geq \alpha$, the posterior uncertainty is reduced relative to the prior uncertainty. Along parameter directions that is less data-informed, i.e. $\sigma_i^2 < \alpha$, the posterior uncertainty increases relative to the prior uncertainty.

3. Applications to Imaging Problems.

3.1. Image Deblurring. One typical inverse problem in imaging is image deblurring. Given some blurry image, we want to recover the true, sharp image. To understand the deblurring process, we must first understand how an image becomes blurred in the first place. A simple and effective mathematical model of the blurring process is convolution of a sharp image with a blurring kernel. This blurring kernel is often described mathematically as a point spread function (PSF). The PSF describes how energy from a point source (ie. a single pixel) is *smeared* out among neighboring pixels, resulting in a blur.

Since convolution is a linear operation, it can be expressed mathematically as

$$\mathcal{A}\boldsymbol{X_{true}} = \boldsymbol{B} \tag{3.1}$$

where $\mathcal{A}$ is the blurring (convolution) operator acting on the true image $\boldsymbol{X_{true}} \in \mathbb{R}^{m_1 \times m_2}$ resulting in the blurred image $\boldsymbol{B} \in \mathbb{R}^{m_1 \times m_2}$. By stacking (or *vectorizing*) the columns of $\boldsymbol{X_{true}}$, we can write (3.1) as a linear algebraic equation. Let us denote by $\boldsymbol{x_{true}}$ the vectorized true image and by $\boldsymbol{y}$ the vectorized blurred image, i.e.,

$$\boldsymbol{x_{true}} = \text{vec}(\boldsymbol{X_{true}}) \ \in \mathbb{R}^{m_1 m_2}, \quad \boldsymbol{y} = \text{vec}(\boldsymbol{B}) \ \in \mathbb{R}^{m_1 m_2}$$

Also, since $\mathcal{A}$ is a linear operator acting on a vector, it has a matrix representation denoted by $\boldsymbol{A} \in \mathbb{R}^{m_1 m_2 \times m_1 m_2}$. Finally, (3.1) becomes

$$\boldsymbol{A}\boldsymbol{x_{true}} = \boldsymbol{y} \tag{3.2}$$

Note that while this notation is convenient for manipulating mathematically, it is not efficient to construct the two-dimensional convolution matrix. $\boldsymbol{A}$ is a large sparse matrix, which, for large problems, cannot be stored in memory. Even on problems small enough to fit in memory, it is computationally expensive to explicitly construct this matrix. Fortunately, there are efficient methods for computing spectral decompositions of the matrices arising from convolution operators using the fast fourier transform and discrete cosine transform. While interesting in their own right, these implementation details are not necessary for the following discussion. For a detailed treatment of image deblurring problems and algorithms, the interested reader is encouraged to consult [5].

For all examples considered in this chapter

$$\boldsymbol{\Lambda} = \lambda^2 \boldsymbol{I}, \quad \text{and } \boldsymbol{\Gamma} = \boldsymbol{I},$$

where λ is the noise level (the standard deviation).

Since truncated SVD (TSVD) and Tikhonov are spectral filtering methods, the regularized solution using these methods can be written using the following common form

$$\boldsymbol{x}_{filt} = \sum_{i=1}^{p} \phi_i \frac{\boldsymbol{U}_i^T \boldsymbol{y}}{\sigma_i} \boldsymbol{V}_i, \tag{3.3}$$

where ϕ_i is usually called the *filter factor* as it has the effect of filtering (damping) when ϕ_i is close to 0. It can be shown that the filter factor for rank-r TSVD is given by

$$\phi_i = \begin{cases} 1, & i \leq r \\ 0, & \text{otherwise.} \end{cases}$$

Likewise, the filter factor for Tikhonov regularization is given by

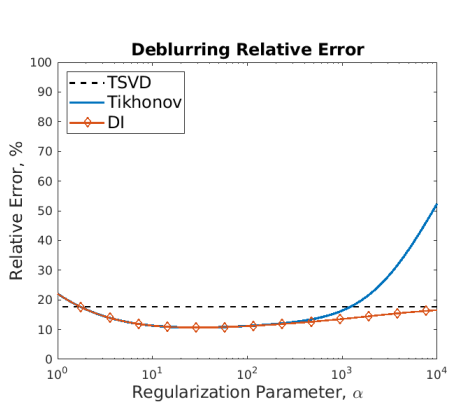
$$\phi_i = \frac{\sigma_i^2}{\sigma_i^2 + \alpha}$$

As discussed in Section 2, the DI method with rank-r approximation removes regularization on the first r directions $\boldsymbol{V}_i$, $1 \leq i \leq r$, while being the same as Tikhonov on the other directions. For $\boldsymbol{\Gamma} = I$ and $\boldsymbol{x}_0 = 0$ the DI solution (see Lemma 2.2) can be written in the filtered form as

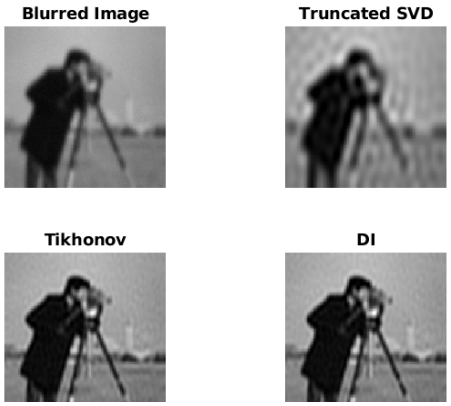
$$\phi_i = \begin{cases} 1, & i \leq r \\ \dfrac{\sigma_i^2}{\sigma_i^2 + \alpha}, & \text{otherwise.} \end{cases}$$

Remark 3.1. It should be emphasized that the DI method also shares the same spectral decomposition form in this case because $\boldsymbol{\Gamma} = \boldsymbol{I}$ and $\boldsymbol{x}_0 = 0$. When $\boldsymbol{\Gamma} \neq \boldsymbol{I}$, singular vectors of $\boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A}$ do not necessarily diagonalize both $\boldsymbol{A}$ and $\boldsymbol{\Gamma}$ simultaneously. In other words, the filtered form (3.3) is not valid for the DI approach unless $\boldsymbol{U}$ and $\boldsymbol{V}$ are singular vectors of $\boldsymbol{\Lambda}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{\Gamma}^{\frac{1}{2}}$ and $\boldsymbol{x}_0 = 0$. When $\boldsymbol{x}_0 \neq 0$, there is an additional term contributed from $\boldsymbol{x}_0$ as shown in the DI solution given in Lemma 2.2.

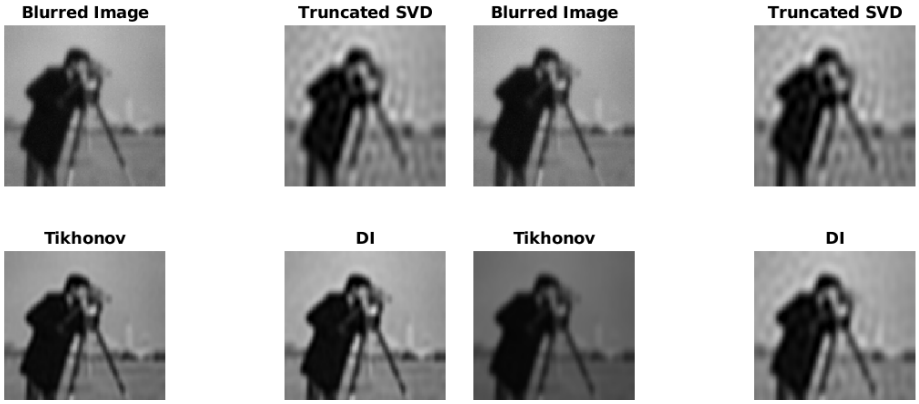
We can see here again that 1) when $r \to 0$, DI approaches Tikhonov; 2) when $\alpha \ll \sigma_i$ for $i \leq r$, Tikhonov is close to DI; and 3) when $\alpha \to \infty$, DI converges to TSVD. This can be clearly seen in Figure 5(a) for a deblurring problem in which we plot the relative error between the deblurred images and the original ones for $m_1 = m_2 = 128$, $\lambda = 0.01$, $r = 400$, and a wide range of α. For the under-regularization regime, i.e. $\alpha < 1$, which should be avoided, the regularization is not sufficient to suppress the oscillations due to the high frequency modes for both Tikhonov and DI methods, resulting in inaccurate reconstructions. For reasonable-to-over regularization regimes, i.e. $\alpha > 1$, DI is the best compared to both Tikhonov and TSVD method as it combines the advantages from both sides. That is: 1) DI behaves similar to Tikhonov for reasonable (but small) regularization and outperforms Tikhonov in reasonable-to-over regularization regimes; and 2) Compared to TSVD, DI is more accurate for reasonable regularization parameters as it maintains the benefits of keeping useful information from all parameter directions while avoiding potential errors caused by over-regularization. Consequently, the DI error is the smallest of the three methods discussed for all $\alpha > 10^3$ and DI is robust with respect to the regularization parameter.



(a) Relative error versus α



(b) Deblurring results with $\alpha = 100$



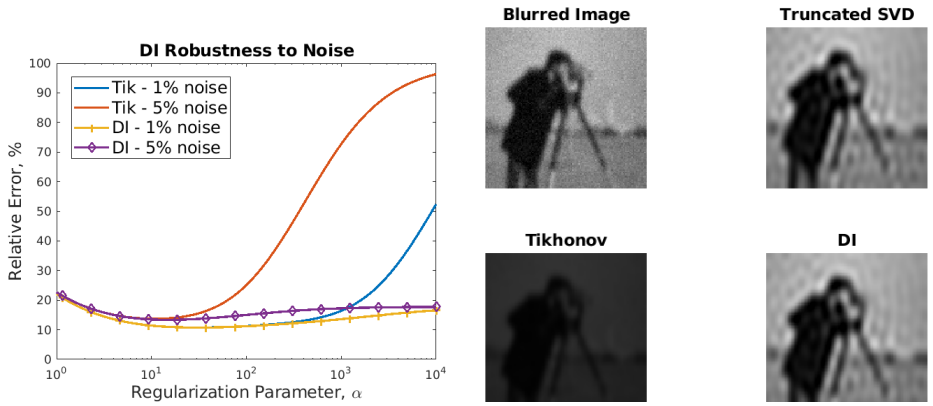
(c) Deblurring results with $\alpha = 1000$ (d) Deblurring results with $\alpha = 5000$

FIG. 5. *Deblurring results for* $m_1 = m_2 = 128$, $\lambda = 0.01$, $r = 400$. *Top Left: relative error between deblurred images and the truth for a range of regularization parameter* $\alpha \in [1, 10^4]$. *Top Right: the DI deblurred image with* $\alpha = 100$. *Bottom Left: the DI deblurred image with* $\alpha = 1000$. *Bottom Right: the DI deblurred image with* $\alpha = 5000$.

In Figure 5(b) are the deblurred images for $\alpha = 100$ corresponding to the smallest deblurring error for both DI and Tikhonov. As can be seen, Tikhonov result is similar to the DI one, while truncated SVD result is blurry as it removes (putting infinite regularization on) useful information in directions $\boldsymbol{V}_i$ for $i > r$. Figures 5(c) and (d) show the deblurred images for $\alpha = 1000$ and $\alpha = 5000$, respectively, corresponding to cases where DI outperforms both Tikhonov and TSVD (see Figure 5(a)). Indeed, the DI deblurred image has higher quality.

In order to see if the DI method is sensitive to noise, we now consider the case with $\lambda = 5\%$ noise. Deblurring accuracy for this case (purple) is shown in Figure 6(a) together with the accuracy for the case of 1% noise (yellow). As can be seen, the solution quality of the DI method does not degrade significantly due to the presence of noise. Compare this to the difference seen in the Tikhonov method (red and blue curves) with the increase in noise level, we can see that the solution quality of the Tikhonov method degrades rapidly in the presence of noise. It can also be seen that Tikhonov regularization becomes more sensitive to the choice of α as the

noise increases. Since the DI method regularizes only the data-uninformed directions, which also contain much of the noise, increasing the noise level has little effect on the solution quality.



(a) Effect of noise on DI and Tikhonov solutions (b) Deblurring results with $\alpha = 1000$

FIG. 6. *Deblurring results for* $m_1 = m_2 = 128$, $\lambda = 0.05$, $r = 400$. *Left: relative error of DI and Tikhonov solutions with respect to true solution for noise levels of* 1% *and* 5% *and* $\alpha \in [1, 10^4]$. *Right: the DI, Tikhonov, and TSVD deblurred images with* $\alpha = 1000$.

For the rest of this section, we consider the more challenging cases with $\lambda = 5\%$ noise. To make the problem even more challenging, we consider images with missing pixels to simulate more interesting cases when images are damaged or incomplete. Figure 7 show the deblurring results using DI, TSVD, and Tikhonov (Tik) regularizations for damaged images with $m_1 = m_2 = 128$, $r = 400$. First column contains four scenarios with 10% random data, 25% random data, 50% random data, and 100% data, all with noise. Note that we plot the damaged images by filling the missing data with 0. The second column are the corresponding TSVD deblurring results. The last four columns are the results from DI and Tikhonov with $\alpha = 10$ and 20. As can be observed, all methods are able to deblur and at the same time recover the true image quite well even with only 10% data. Both DI and Tikhonov yield clearer images compared to TSVD. The Tikhonov results are "darker", especially with $\alpha = 20$, indicating over-regularization, while the DI images are insensitive to regularization parameter as the data-informed modes are left untouched. Indeed, Figure 8 clearly demonstrates these expected results for larger regularization parameters ($\alpha = 50$ and $\alpha = 100$).

Recall the goal of sections 2.2 and 2.3.2 is to gain insights into statistical properties of the DI prior. For linear parameter-to-observable maps—which are the cases for this chapter—with Gaussian observational noise, the posterior is also a Gaussian. As a result, the result at the end of section 2.3.2 also allows us to use the posterior covariances (2.27) and (2.28) to estimate the uncertainty in the corresponding inverse solutions. Since the posterior using either Tikhonov or DI prior is Gaussian, its diagonal contains the marginal pixel-wise variances, which can be used as a measure of uncertainty for each pixel. We now study the uncertainty estimation in the solution of deblurring problems.

To begin, it is important to distinguish the following two cases:

- *Case I*: using only rank-r DI regularization in which rank-r approximation for

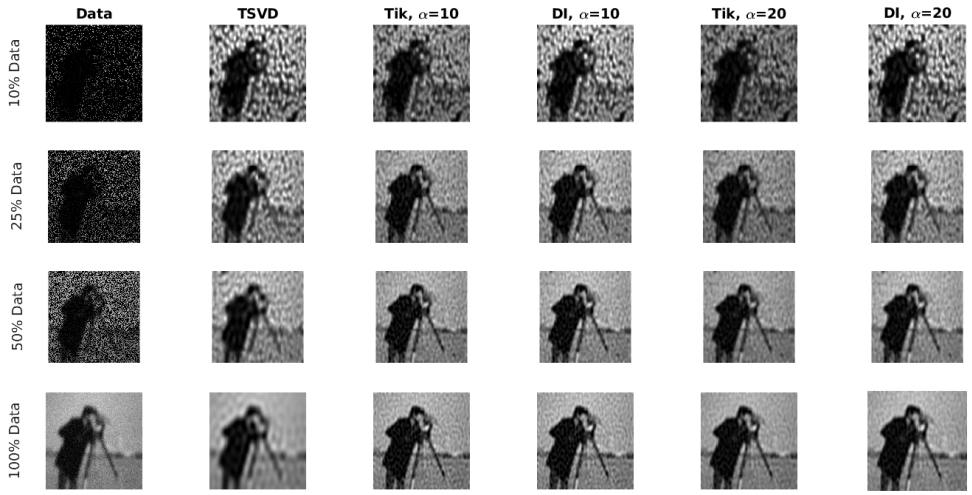


FIG. 7. *Deblurring results using DI, TSVD, and Tikhonov (Tik) regularizations for damaged images with* $m_1 = m_2 = 128$, $\lambda = 0.05$, $r = 400$. *First column consists of four scenarios with 10%, 25%, 50%, and 100% data. The second column are the corresponding TSVD deblurring results. The last four columns are the results from DI and Tikhonov with* $\alpha = 10$ *and* 20.

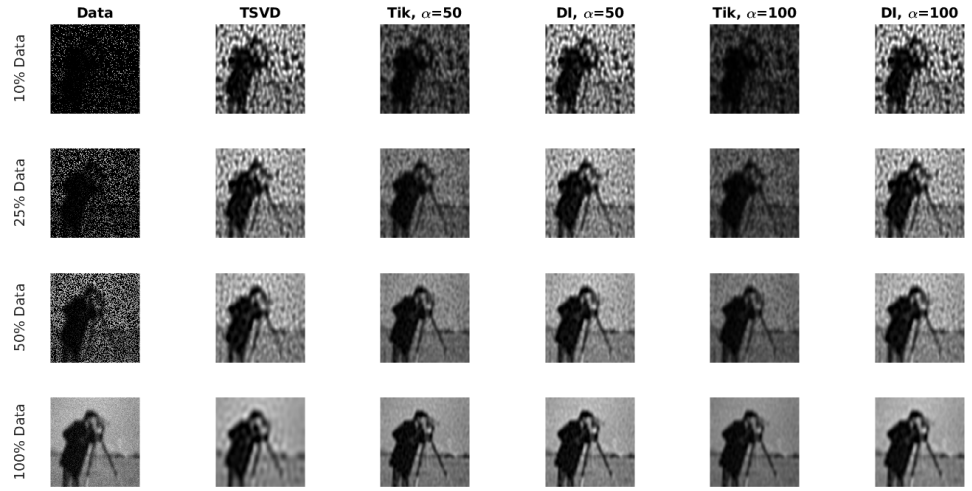


FIG. 8. *Deblurring results using DI, TSVD, and Tikhonov (Tik) regularizations for damaged images with* $m_1 = m_2 = 128$, $\lambda = 0.05$, $r = 400$. *First column consists of four scenarios with 10%, 25%, 50%, and 100% data. The second column are the corresponding TSVD deblurring results. The last four columns are the results from DI and Tikhonov with* $\alpha = 50$ *and* 100.

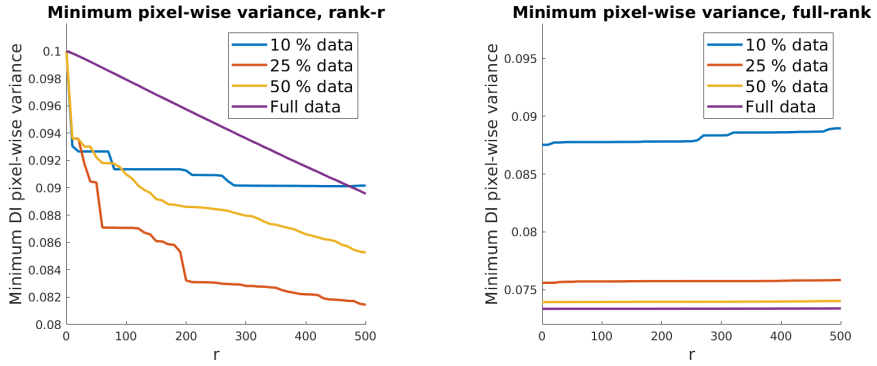
the pseudo-inverse $\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^T\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)^\dagger$ is done as we have presented. The DI posterior covariance (2.27) thus involves second and third columns in Table 1 and a rank-n SVD (2.13) is needed.

- *Case II*: performing rank-r low-rank approximation of the posterior covariance in addition to rank-r DI regularization. This amounts to using only the

second column of Table 1 for the DI posterior covariance in (2.27). This case is typically more practical for large-scale problems as only rank-r SVD (2.19) is needed.

In Figure 9(a) are the minimum pixel-wise variances for four scenarios with 10% random data, 25% random data, 50% random data, and 100% random data for *Case II*. As can be seen, the uncertainty corresponding to the case of missing data is lower than the uncertainty for full data case! We expect the opposite, that is, more available (supposedly) informative data is expected to lead to lower uncertainty in the inverse solution. The observation is twofold: first, care needs to be taken for *Case II* results as rank-r approximation may not provide accurate uncertainty; second, for 10% data case, when $r > 500$ the uncertainty is larger compared to the full data case. This suggests that r needs to be sufficiently large for an accurate uncertainty estimation, and this will be confirmed in the discussion below for *Case I* in which we use the full rank (rank-n) decomposition (2.13). The criteria for estimating such a value of r is a subject for future research.[4]

We next discuss the results for *Case I*. Again, this requires a rank-n SVD decomposition (2.13), where n is the rank of $\boldsymbol{A}$, to compute (2.27) using Table 1. Figure 9(b) shows that the minimum uncertainty for any missing data case is higher than the full data case regardless of any value of r in rank-r DI regularization. As also expected, the uncertainty scales inversely with the amount of available data, i.e., the more informative data we have the smaller the uncertainty in the inverse solution. Note that the result and the conclusion for the largest pixel-wise variances are similar and hence omitted here.



(a) *Case II* (b) *Case I*

FIG. 9. *Rank-r DI posterior pixel-wise uncertainty using rank-n SVD decomposition (*Case I *with both second and third columns of Table 1) and using rank-r SVD decomposition (*Case II *with only the second column of Table 1).*

We now compare the DI and Tikhonov posterior uncertainty estimations. Since *Case I*, though more expensive, provides more accurate uncertainty estimation, it is used for computing DI posterior pixel-wise variances. To be fair, we also use the full decomposition for Tikhonov regularization. In other words, the following comparison is based on (2.27) and (2.28) and Table 1. As discussed above in Figure 6(a) and Figure 7, $\alpha = 10$ corresponds to a case in the region where DI and Tikhonov give

[4] At the moment of writing this chapter we have not yet found such a criteria.

nearly the same reconstructions (in fact Tikhonov slightly over-regularizes), so let us start with this case first. Figure 10 shows that the DI posterior has higher pixel-wise variance than the Tikhonov posterior, this is consistent with the result and the discussion of Table 1 and Figure 7, that is, the Tikhonov posterior is not only over-regularing but also overconfident. For both methods, regions of higher uncertainty are visually discernible where data is missing. In the case of 100% data, the result is the same, namely, Tikhonov uncertainty estimation subjectively is less than the DI uncertainty estimation. In this case, the uncertainty estimate is not very interesting: both DI and Tikhonov have approximately uniform uncertainty everywhere as we have data everywhere. We next consider the case with $\alpha = 1000$ where Tikhonov significantly over-regularizes (see Figure 6(b)). Figure 11, shows that while Tikhonov is uniformly (very) overconfident, i.e., having small posterior uncertainty everywhere, DI gives informative UQ results. The latter can be clearly seen for the case with 10% data in which the uncertainty is higher for missing pixels. This implies that the DI priors could provide more useful UQ results than the Tikhonov (standard Gaussian) ones.

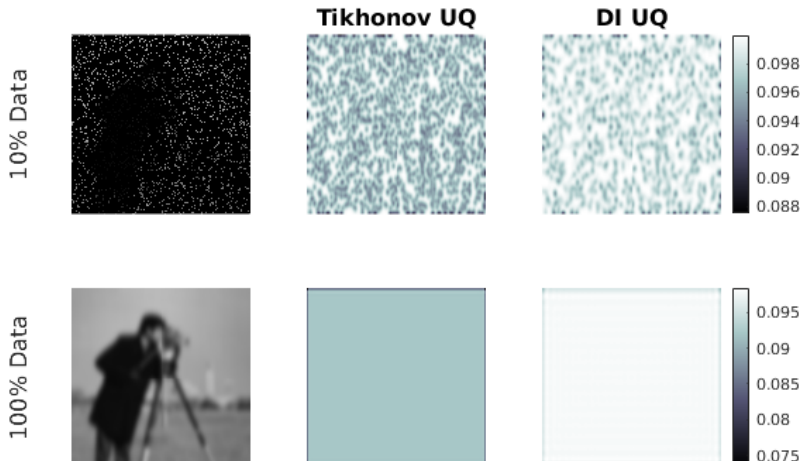


FIG. 10. *Visualization of pixel-wise variance estimates for the deblurring problem with* $\lambda = 0.05$, $r = 400$, *and* $\alpha = 10$. *In the left column are the noisy images with* 10% *data and* 100% *data. In the second column are the Tikhonov uncertainty estimates for 10% data (top) and 100% data (bottom). Likewise, the third column contains the DI uncertainty estimates for 10% data (top) and 100% data (bottom).*

3.2. Image Denoising. We can extend the idea of data-informed (DI) regularization to the image denoising problem. Since noise typically resides in the high frequency portion of the image, denoising can be performed by applying spectral filtering techniques directly to the noisy image. These noisy high-frequency modes are also the *less informative* modes in the DI setting. Taking the SVD of the noisy image, $\boldsymbol{X}_{noisy}$, we have

$$\boldsymbol{X}_{noisy} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \sum_i \sigma_i \boldsymbol{U}_i \boldsymbol{V}_i^T,$$

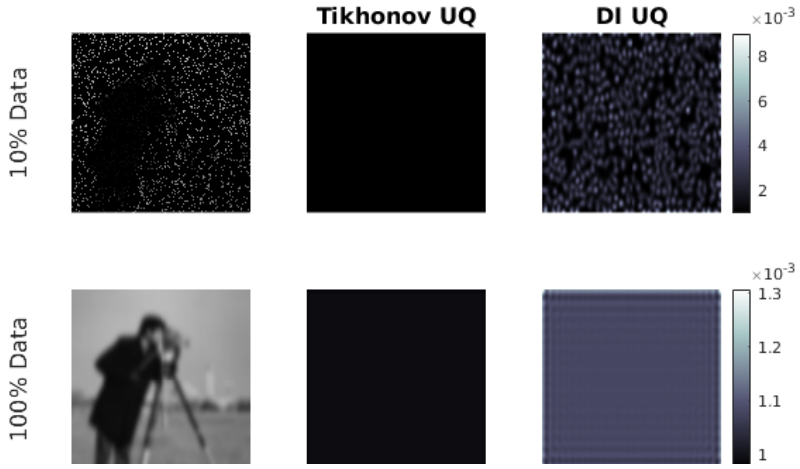


FIG. 11. *Visualization of pixel-wise variances for the deblurring problem with* $\lambda = 0.05$, $r = 400$, *and* $\alpha = 1000$. *In the left column are the noisy images with* 10% *data and* 100% *data. In the second column are the Tikhonov uncertainty estimates for 10% data (top) and 100% data (bottom). The third column contains the DI uncertainty estimates for 10% data (top) and 100% data (bottom).*

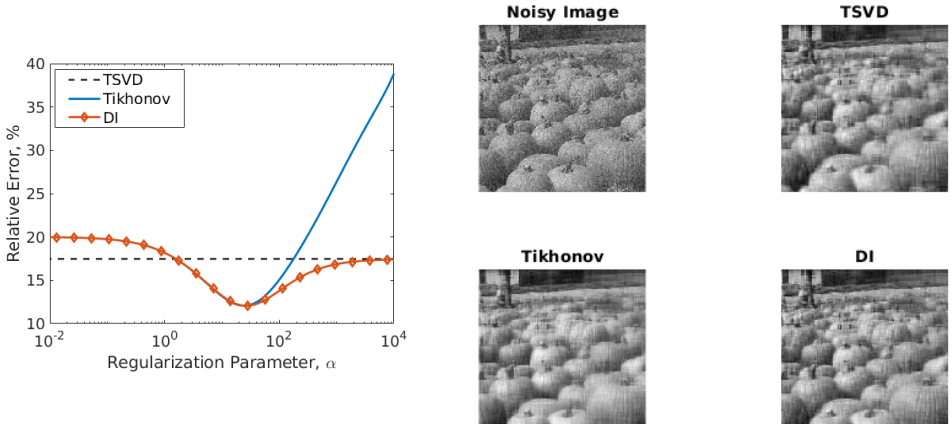
The denoised image can be obtained by "filtering" the noise as

$$\boldsymbol{X}_{filt} = \boldsymbol{U}\boldsymbol{\Sigma}^{filt}\boldsymbol{V}^T = \sum_i \phi_i \sigma_i \boldsymbol{U}_i \boldsymbol{V}_i^T,$$

where Σ^{filt} is the diagonal matrix with $\Sigma_{ii}^{filt} = \phi_i \sigma_i$. The filter factors ϕ_i are the same as those defined for the deblurring case. For a numerical demonstration, we pick a noisy image [5] with 5% noise (see the top left sub-figure of Figure 12(a)). Shown in Figure 12(a) are denoised results using DI with $r = 20$ and $\alpha = 100$, TSVD with $r = 20$, and Tikhonov with $\alpha = 100$. Though the difference in the results is not clearly visible, the DI has smaller error compared the other two methods. This can be verified in Figure 12(b) where the relative error between the denoised image and the true one for a wide range of "regularization parameter" $\alpha \in \left[10^{-2}, 10^4\right]$ is presented. Clearly, we would not choose $\alpha < 1$ as these correspond to under-regularization. For $\alpha > 1$, DI is the best compared to both Tikhonov and TSVD method as it combines the advantages from both methods. Indeed, the DI error is smallest for all $\alpha > 1$ and DI is robust with regularization parameter.

3.3. X-ray Tomography. In the previous two examples, we have been able to implement spectral filtering methods directly by introducing filter factors which effectively modified the singular values to minimize the impact of noise on the inversion process.[5] Each method relied on computing a full factorization of $\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}$ and then applying filters. While this is an effective and straightforward method to solve small-to-moderate inverse problems that helps provide insight into each approach, it can be cumbersome or even computationally infeasible to compute full factorizations for large-scale problems. It is not uncommon that inverse problems arising in imaging applications can lead to very large matrix operators. Indeed, we have seen even in

[5] Recall that the DI method also shares the same spectral decomposition form in this case because $\boldsymbol{\Gamma} = \boldsymbol{I}$ and $\boldsymbol{x}_0 = 0$.



(a) Relative denoised error versus α (b) Denoised Results

FIG. 12. *Denoising with DI, Tikhonov, and TSVD methods. Left: The relative error between the denoised image and the true one for a wide range of "regularization parameter"* $\alpha \in [10^{-2}, 10^{4}]$. *The DI error is smallest for all* $\alpha > 1$ *(corresponding reasonable to over-regularization regimes). Right: Denoised results using DI with* $r = 20$ *and* $\alpha = 100$*, TSVD with* $r = 20$*, and Tikhonov with* $\alpha = 100$.

the toy image deblurring problem in Section 3.1 that matrix size of 16384×16384 is significantly large and we have employed more sophisticated methods to compute the factorization of the convolution operator. For many problems, however, such efficient factorizations may not exist or it is computationally prohibitive to compute a full factorization.

One way to overcome the challenge of factorizing large matrices is to solve the optimality condition (2.18) iteratively. Since $\boldsymbol{H}$ is symmetric positive definite, we choose the conjugate gradient (CG) method (see, e.g, [11] and the references therein) which requires only matrix-vector products, which in turn avoids forming any matrices (including $\boldsymbol{A}$ or $\boldsymbol{H}$) completely. We consider two variants: a) Using CG to solve for (2.20), that is, we still require rank-r approximation of the DI regularization; and b) Using CG to solve for (2.18), that is, a rank-r approximation of the DI regularization is not required and in this case we use a least squares optimization method to compute the pseudo-inverse $\left(\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{A}^{T}\boldsymbol{\Lambda}^{-\frac{1}{2}}\right)^{\dagger}$ acting on a vector for each CG iteration.

The detailed computational procedure for the a)-variant is given in Algorithm 3.1 To demonstrate the effectiveness of this approach for the DI method, we choose to solve the inverse problem of reconstructing an image from X-ray measurements. The forward model of generating X-ray measurements, $\boldsymbol{A}$, is given by the Radon transform and $\boldsymbol{A}^{T}$ is given by the inverse Radon transform. A more detailed description of the X-ray tomography inverse problem is given in [9] (and the references therein). We use the MATLAB Image Processing Toolbox to compute the product of the Radon transform $\boldsymbol{A}$ and and its inverse $\boldsymbol{A}^{T}$ with a vector. Results using Algorithm 3.1 for a popular 256×256 phantom image are shown in Figure 13 for various values of the regularization parameter α and the rank r. Each row contains the results for each regularization parameter with different values of r. The corresponding values for α and r can also be found in the rows and columns of Table 2. Note that below each figure is the relative error of the corresponding reconstruction and the actual phantom

Algorithm 3.1 Data-informed Inversion Using Randomized Eigensolver and CG

Input: Data $\boldsymbol{y}$, number of eigenvectors r, prior $\boldsymbol{x}_0$, prior covariance matrix $\boldsymbol{\Gamma}$, noise covariance matrix $\boldsymbol{\Lambda}$, regularization parameter α

1: Define $\boldsymbol{F} := \boldsymbol{\Lambda}^{-\frac{1}{2}} A \boldsymbol{\Gamma}^{\frac{1}{2}}$.
2: Create functions to compute matrix-vector products $\boldsymbol{F}\boldsymbol{x}$ and $\boldsymbol{F}^T\boldsymbol{x}$
3: Compute the first r eigenvectors of $\boldsymbol{F}^T\boldsymbol{F}$ using a randomized eigensolver
4: Solve linear equation (2.20), i.e.,

$$\boldsymbol{\Gamma}^{-\frac{1}{2}}\Big[\boldsymbol{F}^T\boldsymbol{F} + \alpha\big(\boldsymbol{I} - \boldsymbol{V_r}\boldsymbol{V_r}^T\big)\Big]\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x} = \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{F}^T\boldsymbol{y} + \alpha\boldsymbol{\Gamma}^{-\frac{1}{2}}\big(\boldsymbol{I} - \boldsymbol{V_r}\boldsymbol{V_r}^T\big)\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x}_0$$

using the conjugate gradient method

image. These relative errors are collected in Table 2 for clarity. Note that for the last two images on the last row of Figure 13, CG does not converge and this issue is still under investigation. Other than that the observations are similar to the previous section. That is, compared to Tikhonov, DI is robust to the regularization parameter and it is at least as good as Tikhonov regardless the values of regularization parameter α and rank r.

α	Relative Error, %					
	$r = 0$ (Tik)	$r = 10$	$r = 50$	$r = 100$	$r = 200$	$r = 400$
1	33.52	33.52	33.52	33.52	33.52	33.52
10	31.73	31.73	31.73	31.73	31.73	31.73
100	24.44	24.45	24.45	24.45	24.45	24.45
1000	29.81	29.80	29.72	29.66	29.51	29.09
10^4	58.76	58.52	56.93	55.92	54.03	50.43
10^5	81.77	77.10	70.33	67.78	63.84	57.84
10^6	96.09	81.29	72.44	69.50	81.80	299.73

TABLE 2

Comparison of the relative errors of the DI solution estimate for various regularization parameters α and various values for r. The noise level here is $\lambda = 1\%$.

Next we present the detailed computational procedure for the b)-variant in Algorithm 3.2. In order to compare variant b) with variant a), we compute the relative error of the reconstruction and the true image for various values of regularization parameter α. From the results in Figure 13, we choose $r = 200$ to balance the accuracy and the cost of the eigensolver. The result is in Figure 14, which shows that the b)-variant (red curve) is at least as good as the a)-variant (blue curve) while not requiring low rank approximations. Indeed, to demonstrate this, we pick $\alpha = 100$ for which Figure 14 shows that both variants give similar reconstruction quality, and the reconstruction from both variants are shown in Figure 15. As can be seen, the result from the b)-variant looks much clearer, which is expected in this case, as $r = 200$ is not sufficient to capture all the data-informed modes for the a)-variant.

By using the pseudoinverse formulation, we can still get excellent results while avoiding the computation of a large factorization.

4. Conclusions. We have presented a new regularization technique called data-informed (DI) regularization that combines advantages of the classical truncated SVD

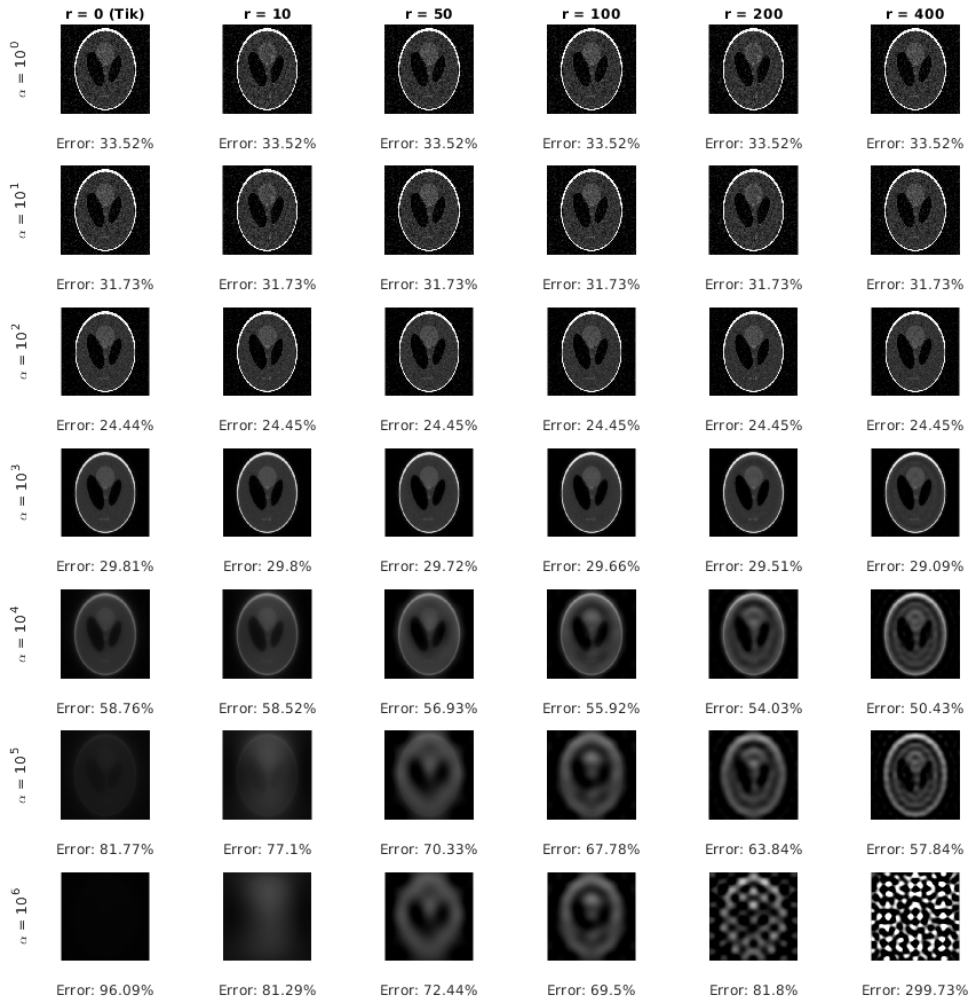


FIG. 13. *DI reconstructions for various values of the regularization parameter α and the rank r. Each row contains the results for each regularization parameter with different values of r. The corresponding values for α and r can be found in the rows and columns of Table 2. Below each figure is the relative error of the corresponding reconstruction and the actual phantom image.*

and Tikhonov regularization. In particular, the DI approach does not pollute the data-informed modes, and regularizes only less data-informed ones. As a direct consequence, the DI approach is at least as good as the Tikhonov method for any value of the regularization parameter and it is more accurate than the TSVD (for reasonable regularization parameter). Due to the blending of these two classical methods, DI is expected to be robust with regularization parameter and this is verified numerically. We have shown that our DI approach has an interesting statistical interpretation, that is, it transforms both the data distribution (i.e. the likelihood) and prior distribution (induced by Tikhonov regularization) to the same Gaussian distribution whose covariance matrix is diagonal and the diagonal elements are exactly the singular values of a composition of the prior covariance matrix, the forward map, and the noise covari-

Algorithm 3.2 Data-informed Inversion Using Nested CG

Input: Data $\boldsymbol{y}$, number of eigenvectors r, prior $\boldsymbol{x}_0$, prior covariance matrix $\boldsymbol{\Gamma}$, noise covariance matrix $\boldsymbol{\Lambda}$, regularization parameter α

1: Define $\boldsymbol{F} = \boldsymbol{\Lambda}^{-\frac{1}{2}} A \boldsymbol{\Gamma}^{\frac{1}{2}}$
2: Create functions to compute matrix-vector products $\boldsymbol{F}\boldsymbol{x}$ and $\boldsymbol{F}^T\boldsymbol{x}$
3: Solve linear equation (2.18), i.e.,

$$\boldsymbol{\Gamma}^{-\frac{1}{2}}\Big[\boldsymbol{F}^T\boldsymbol{F}+\alpha\big(\boldsymbol{I}-\boldsymbol{F}^T(\boldsymbol{F}\boldsymbol{F}^T)^\dagger\boldsymbol{F}\big)\Big]\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x} = \boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{F}^T\boldsymbol{y}+\alpha\boldsymbol{\Gamma}^{-\frac{1}{2}}\big(\boldsymbol{I}-\boldsymbol{F}^T(\boldsymbol{F}\boldsymbol{F}^T)^\dagger\boldsymbol{F}\big)\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x}_0$$

using the conjugate gradient method. For each CG iterations, compute the product of $\boldsymbol{F}^T(\boldsymbol{F}\boldsymbol{F}^T)^\dagger\boldsymbol{F}\boldsymbol{\Gamma}^{-\frac{1}{2}}$ with any vector $\boldsymbol{x}$ using matrix-free Algorithm 3.3.

Algorithm 3.3 Compute the product of $\boldsymbol{F}^T(\boldsymbol{F}\boldsymbol{F}^T)^\dagger\boldsymbol{F}\boldsymbol{\Gamma}^{-\frac{1}{2}}$ with any vector using optimization

Input: functions to compute $\boldsymbol{F}\boldsymbol{x}$ and $\boldsymbol{F}^T\boldsymbol{x}$, current estimate of $\boldsymbol{x}$, prior covariance matrix $\boldsymbol{\Gamma}$

1: Compute $\boldsymbol{b} = \boldsymbol{F}\boldsymbol{\Gamma}^{-\frac{1}{2}}\boldsymbol{x}$
2: Using conjugate gradient method, solve linear equation

$$\boldsymbol{F}\boldsymbol{F}^T\boldsymbol{z} = \boldsymbol{b}.$$

3: Return $\boldsymbol{F}^T\boldsymbol{z}$

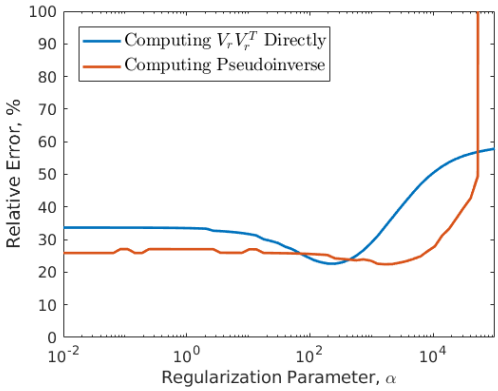
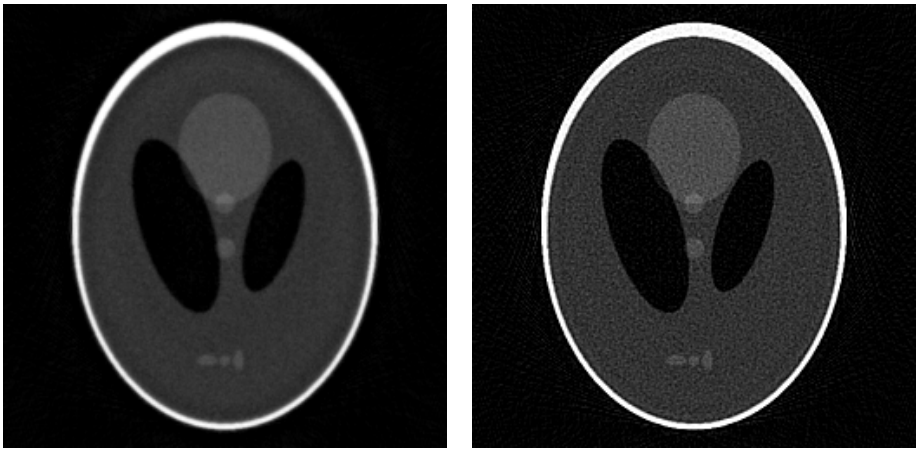


FIG. 14. *A comparison between variant b) (red curve) and variant a) with $r = 200$ (blue curve). Here, we compute the relative error of the reconstruction and the truth image for various values of regularization parameter α.*

ance matrix. In other words, DI finds the modes that are most equally data-informed and prior-informed and leaves these modes untouched so that the inverse solution receive the best possible (balanced) information from both prior and the data. We



(a) X-ray tomography using a)-variant with $r = 200$

(b) X-ray tomography using b)-variant

FIG. 15. *X-Ray tomography reconstruction with* 1% *noise and* $\alpha = 100$*: (a) the result from the a)-variant with* $r = 200$*, and (b) the result from the b)-variant.*

have shown that DI is a regularization strategy. To demonstrate and to support our findings, we have presented various results for popular computer vision and imaging problems including deblurring, denosing, and X-ray tomography.

Acknowledgments. This work is inspired by an ongoing collaborative work on data-consistent inverse framework with Tim Wildey and Brad Marvin, and the authors would like to thank them for many fruitful discussions.

REFERENCES

[1] A. C. ANTOULAS, *Approximation of Large-Scale Systems*, SIAM, Philadelphia, 2005.
[2] D. COLTON AND R. KRESS, *Integral equation methods in scattering theory*, John Wiley & Sons, 1983.
[3] J. N. FRANKLIN, *Well-posed stochastic extensions of ill–posed linear problems*, Journal of Mathematical Analysis and Applications, 31 (1970), pp. 682–716.
[4] S. GUGERCIN AND A. C. ANTOULAS, *A survey of model reduction by balanced truncation and some new results*, International Journal of Control, 77 (2004), pp. 748–766.
[5] N. J. G. HANSEN, P. C. AND D. P. O'LEARY, *Deblurring Images: Matrices, Spectra, and Filtering*, SIAM, Philadelphia, 2006.
[6] J. KAIPIO AND E. SOMERSALO, *Statistical and Computational Inverse Problems*, vol. 160 of Applied Mathematical Sciences, Springer-Verlag, New York, 2005.
[7] S. LASANEN, *Discretizations of generalized random variables with applications to inverse problems*, PhD thesis, University of Oulu, 2002.
[8] M. S. LEHTINEN, L. PÄIVÄRINTA, AND E. SOMERSALO, *Linear inverse problems for generalized random variables*, Inverse Problems, 5 (1989), pp. 599–612.
[9] S. S. MUELLER, J. L., *Linear and Nonlinear Inverse Problems with Practical Applications*, SIAM, Philadelphia, 2012.
[10] P. PIIROINEN, *Statistical measurements, experiments, and applications*, PhD thesis, Department of Mathematics and Statistics, University of Helsinki, 2005.
[11] J. R. SHEWCHUK, *An introduction of the conjugate gradient method without the agonizing pain*, (1994), https://www.cs.cmu.edu/~quake-papers/painless-conjugate-gradient.pdf.
[12] A. M. STUART, *Inverse problems: A Bayesian perspective*, Acta Numerica, 19 (2010), pp. 451–559, https://doi.org/doi:10.1017/S0962492910000061.
[13] A. TARANTOLA, *Inverse Problem Theory and Methods for Model Parameter Estimation*, SIAM, Philadelphia, PA, 2005.