QUARTERLY OF APPLIED MATHEMATICS
VOLUME LXVII, NUMBER 1
MARCH 2009, PAGES 137–161
S 0033-569X(09)01120-7
Article electronically published on January 8, 2009

MINIMAX ENTROPY SOLUTIONS OF ILL-POSED PROBLEMS

By

FRED GREENSITE

Department of Radiological Sciences, University of California, Orange, California 92868

Abstract. Convergent methodology for ill-posed problems is typically equivalent to application of an operator dependent on a single parameter derived from the noise level and the data (a regularization parameter or terminal iteration number). In the context of a given problem discretized for purposes of numerical analysis, these methods can be viewed as resulting from imposed prior constraints bearing the same amount of information content. We identify a new convergent method for the treatment of certain multivariate ill-posed problems, which imposes constraints of a much lower information content (i.e., having much lower bias), based on the operator's dependence on many data-derived parameters. The associated marked performance improvements that are possible are illustrated with solution estimates for a Lyapunov equation structured by an ill-conditioned matrix. The methodology can be understood in terms of a Minimax Entropy Principle, which emerges from the Maximum Entropy Principle in some multivariate settings.

1. Introduction. For particular Hilbert spaces $\mathcal{X}, \mathcal{Y}$, consider the vector space $\mathcal{F}$ consisting of operators $F : \mathcal{X} \to \mathcal{Y}$, where F is either a compact or Fredholm operator. The members of $\mathcal{F}$ pertain to the classical first kind and second kind integral equations, these generally representing Hadamard ill-posed versus well-posed problems, as the operator applied to the unknown is compact versus Fredholm. If $\mathcal{X}$ is the space of Lebesgue square-integrable functions on a compact subset of $\mathbb{R}$, then in the generalization to multivariate problems it is natural to consider the operator space given by the tensor product $\mathcal{F} \otimes \mathcal{F}$. This space contains compact operators and Fredholm operators, but it also contains a third class of operator that is neither compact nor Fredholm, and which is associated with ill-posed problems of a different character than those of the univariate setting. In discretized form, these include the problems of solving Lyapunov and Sylvester equations [4] when the latter involve ill-conditioned matrices. It will be shown that that such problems are amenable to an unusual regularization procedure (based on use of a derived nonlinear "recombinant" operator) capable of much higher accuracy than

Received August 7, 2007.
2000 *Mathematics Subject Classification.* Primary 47A52, 45Q05; Secondary 65J20, 34A55.
Key words and phrases. Inverse problems, ill-posed problems, Sylvester equation.
E-mail address: fred.greensite@uci.edu

that produced by an application of the standard regularization methodologies. The underlying mechanism can be understood in the context of a "Minimax Entropy Principle" which emerges from the Maximum Entropy Principle when tensor products of operator spaces pertain.

1.1. *Entropy and ill-posed problems.* A problem is considered ill-posed if its solution nominally requires inversion of an operator lacking a continuous inverse. Though classical examples such as differentiation have been successfully treated for hundreds of years, ill-posed problems become highly problematic if uncertainty or noise pertain to the data, since the inverse operator (if it exists) is unbounded, implying grossly unstable solutions. Systematic approaches to such problems have been developed only since the mid-twentieth century. The deterministic theory recognizes methodologies that are "convergent". These supply solution estimates via application of a noise level-dependent operator to the data such that the resulting estimates converge to the solution of the noiseless setting as the noise level becomes small [2]. Though comparative convergence rates between competing methods are sometimes available, in general this theory does not distinguish between convergent methods in the sense that there are usually no available criteria for choosing which one of the competing solution estimates derived from different convergent methods should be preferred. In the discretized settings that typically arise in the numerical analysis of ill-posed problems, the stochastic theory can address this ambiguity through the Maximum Entropy Principle [8]. This approach is able to utilize available prior constraints to provide such selection criteria through a Bayesian formulation and identification of a prior. The concepts of random functions and covariance operators allow generalization of this discretized approach to more general function space settings [10], and the notion of convergence continues to apply here as well. In fact, varieties of a standard deterministic theory method such as Tikhonov Regularization can be presented in a stochastic theory guise, each being equivalent to the imposition of a particular prior covariance operator. If (as is usually the case) constraints inherent in the problem formulation are insufficient to uniquely supply a full prior (or prior covariance) operator, then the Maximum Entropy Principle can be invoked to identify needed remaining constraints.

The Maximum Entropy Principle grew out of the Bernoulli-Laplace Principle of Insufficient Reason and Poincare's introduction of group theory into probabilistic arguments [8]. A constraint not inherent in the problem formulation is favored if its replacement by a different constraint necessarily leads to a prior that has lower information theoretic entropy than that associated with the original constraint. This principle has proved quite useful in various applications, and is a well-established concept. On the other hand, its use in inverse problems in the "minimum information" setting is less than compelling, perhaps because in general there are available constraints that are either ignored or difficult to encode (we will touch on this issue again in Section 1.2). Nevertheless, zero-order Tikhonov regularization is useful and widely used. In discretized settings, its penalty operator corresponds to the zero mean Gaussian prior with covariance matrix proportional to the identity (this is derived in [9] as the maximum entropy Gaussian prior in minimum information settings).

As we shall see, the Maximum Entropy Principle can be looked at as a special case of a more general Minimax Entropy Principle pertaining to the situation where there is more than one set of maximum entropy constraints that can be used to provide a solution estimate. Unlike the Maximum Entropy Principle, the Minimax Entropy Principle *is* compelling in the treatment of the relevant ill-posed problems in the minimum information setting. Thus, we have the following loosely-stated principles related to the treatment of ill-posed problems:

(1) *Convergence Principle.* Any convergent method can be used. Solution estimate preferences are based on more rapid convergence, if such can be established.
(2) *Maximum Entropy Principle.* A prior is constructed such that in the more general function space setting the corresponding method remains convergent, and such that replacement of a constraint used to construct this prior (that was not part of the original problem formulation) by a different constraint (not part of the original problem formulation) lowers the entropy of the implied prior probability density.
(3) *Minimax Entropy Principle.* Given a set of convergent methods satisfying the Maximum Entropy Principle in the sense of (2) above, the favored solution derives from the method whose implied prior probability density has minimum entropy.

The Minimax Entropy Principle is distinct from the Maximum Entropy Principle (i.e., nontrivial) only when more than one convergent maximum entropy constraint set can be identified. Indeed, its novelty lies first in the demonstration that there are sometimes more than one such set of constraints. Intuitively, its attractiveness derives from its insistence on sufficient constraints having least unjustified bias based on their relative amount of information content. As with the Principle of Insufficient Reason and the Maximum Entropy Principle, the Minimax Entropy Principle is compelling on the bases of need, simplicity, and performance.

1.2. *The interplay of deterministic and stochastic viewpoints.* The deterministic formulation of ill-posed problems [2] can be applied to premises such as

ASSUMPTION D. We are given a bounded operator $F : \mathcal{L}^2[\mathcal{S}] \to \mathcal{L}^2[\mathcal{T}]$, with $\mathcal{S}$ and $\mathcal{T}$ compact subsets of $\mathbb{R}^n$ and with $\mathcal{L}^2[\cdot]$ indicating the space of Lebesgue square-integrable functions. For a particular $x \in \mathcal{L}^2[\mathcal{S}]$ and a particular $y \in \mathcal{L}^2[\mathcal{T}]$,

$$y = F[x], \tag{1.1}$$

though x and y are unknown, and we are instead given $y_\delta \in \mathcal{L}^2[\mathcal{T}]$ such that $\|y - y_\delta\| \leq \delta$. It is desired to compute $x_\delta \in \mathcal{L}^2[\mathcal{S}]$ depending continuously on y_δ, and such that $x_\delta \to x^\dagger$ as $\delta \to 0$, where $x^\dagger \in \mathcal{L}^2[\mathcal{S}]$ is such that $y = F[x^\dagger]$.

The task of estimating x in (1.1) given y_δ is ill-posed under rather general conditions [2]. Tikhonov regularization combined with the Discrepancy Principle [2] is a well-known technique for providing a solution estimate satisfying the criteria at the end of Assumption D. In a typical incarnation of this method, the estimate is chosen as

$$x_\delta = \arg \min_{z \in D(R)} \left(\|F[z] - y_\delta\|^2 + \lambda_\delta \|R[z]\|^2 \right), \tag{1.2}$$

where "penalty operator" $R : D(R) \to \mathcal{L}^2[\mathcal{S}]$ has domain $D(R) \subset \mathcal{L}^2[\mathcal{S}]$ and where $\lambda_\delta \in \mathbb{R}^+$ is a "regularization parameter" selected such that

$$\|F[x_\delta] - y_\delta\| = \delta \tag{1.3}$$

(the left-hand side above is the "discrepancy"). Thus, from the standpoint of Lagrange multipliers, x_δ minimizes the penalty functional $\|R[z]\|$ subject to the discrepancy equation (1.3). Though useful solution estimates often follow from this program (and related methods), there are ambiguities stemming from the selection of one or another penalty operator R (or one of the competing deterministic iterative methods). In principle, this ambiguity can be exploited by allowing incorporation of "prior knowledge" in the selection of the penalty operator. According to the usual interpretation [6, 2], if possible, R should be selected to reflect the belief that $R[x]$ is square-integrable (indeed, any solution x_δ to (1.2) will have the feature that $R[x_\delta]$ is a member of $\mathcal{L}^2[\mathcal{S}]$). Thus, if some information regarding the smoothness of the solution is available, such as it is in the domain of a particular differential operator, then (presumably) the latter could be a good choice of penalty operator.

The alternative stochastic approach is based on a formulation of the inverse problem in terms of random processes and probability theory (e.g., [3, 10]). In contrast to Assumption D, it can proceed from premises such as

Assumption S. We are given

$$y = F[x] + e, \tag{1.4}$$

where F is as in Assumption D, but x is a zero mean Gaussian random function, e is a zero mean Gaussian white noise process with variance δ^2 at each point of the process, and we are given a realization y_δ of random function y. It is desired to compute x_δ as the mean of the posterior random function (and further, to assess the reliability of x_δ as a solution estimate using the computed posterior covariance operator).

This approach can be understood by considering the analogous program in the discretized settings that usually arise in the numerical analysis of ill-posed problems. In this case, one conceives of y, x, e as random vectors, associated with probability densities $P(y)$, $P(x)$ (the prior), $P(e)$ (the noise model), $P(y|x)$ (the likelihood), and $P(x|y)$ (the posterior density), where $P(\cdot|\cdot)$ indicates conditional density. Bayes Theorem states that

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}. \tag{1.5}$$

The objective is to compute the posterior density (or its most important features) subject to data y_δ (i.e., compute $P(x|y_\delta)$), as this describes the relative likelihood of any relevant choice of solution estimate, given the supplied data realization and noise model. One might then favor the maximizer of this density as a useful solution estimate (which corresponds to the mean in the case of Gaussian densities) and use the covariance of the posterior density to gauge the utility of the estimate.

In the general Hilbert space setting, mean and covariance of a prior Gaussian random function play a similar role to the mean and covariance of a prior density in the discretized setting. For linear and many nonlinear F, mean and covariance of a posterior random function can be calculated and have similar significance to the mean and covariance of the

posterior density in the discretized setting (despite the fact that probability distributions do not exist). Specification of the prior covariance operator associated with x is in some ways analogous to selection of the penalty operator R in the deterministic approach.

However, the following expresses a general discordancy between the deterministic and stochastic viewpoints.

THEOREM 1.1. In the context of Assumption D, suppose F is linear and Tikhonov regularization, with regularization parameter λ_δ and an invertible linear penalty operator R, is applied to produce the solution estimate x_δ. Then with respect to Assumption S, this x_δ corresponds to a prior specification that x is a random function such that realizations of $R[x]$ are expected not to be members of $\mathcal{L}^2[\mathcal{S}]$.

Proof. As is well known [2], a simple variational argument applied to (1.2) leads to the solution estimate

$$(F'F + \lambda_\delta R'R)^{-1}F'y_\delta, \tag{1.6}$$

where F', R' denote the transposes. As derived in [3], Assumption S with linear F leads to the solution estimate

$$C_xF'(FC_xF' + C_e)^{-1}y_\delta, \tag{1.7}$$

where C_x and C_e are the covariance operators associated with x and e. According to Assumption S, $C_e = \delta^2 I$, where I is the identity operator. Setting

$$C_x = \frac{\delta^2}{\lambda_\delta}(R'R)^{-1}, \tag{1.8}$$

a simple manipulation then shows that (1.6) and (1.7) coincide. In that case, the covariance of $R[x]$ is $RC_xR' = (\delta^2/\lambda_\delta)I$. Evidently, the latter is not a trace class operator. But the expectation of the square of the $\mathcal{L}^2$-norm of Rx is

$$\mathcal{E}[\|Rx\|^2] = \mathcal{E}[\mathrm{trace}(Rxx'R')] = \mathrm{trace}(\mathcal{E}[Rxx'R']) = \mathrm{trace}(RC_xR'),$$

where $\mathcal{E}[\cdot]$ denotes expectation. Thus, the theorem assertion follows. □

Under the theorem hypotheses, the deterministic approach requires a solution of (1.2), which implies $R[x_\delta] \in \mathcal{L}^2[\mathcal{S}]$ (i.e., $x_\delta \in D(R)$), with x_δ given by (1.6)). Thus, from the deterministic viewpoint, it is considered advantageous to select the penalty functional to reflect the prior supposition that $R[x]$ is square-integrable, since in fact the variational format of Tihonov regularization restricts the class of admissible solution estimates to such functions. However, while the stochastic approach above also requires that the solution estimate be given by (1.6) (via (1.7)), it does *not* consider $R[x^\dagger]$ to be a member of $\mathcal{L}^2[\mathcal{S}]$. It proceeds from the opposing viewpoint, namely that R should be chosen such that $(R'R)^{-1}$ is proportional to the prior covariance C_x. But in that case, it is expected that realizations of $R[x]$ are *not* square-integrable.

This issue is most sharply drawn when R is taken to be the identity operator (zero-order Tikhonov regularization). From the stochastic viewpoint, this choice actually corresponds to the supposition that prior information regarding the solution is described by a random function whose realizations are expected not to be square-integrable. We could view this choice as implying prior information that it is quite possible (in fact, "likely") that (1.1) is insoluble, which contradicts the usual supposition that the equation *does*

have a square-integrable solution. Indeed, the standard regularization algorithms supply well-behaved solution estimates to first-kind Fredholm integral equations which in the noiseless setting fail the Picard Condition, and thereby have no meaningful square-integrable solutions. To avoid such nonsensical well-behaved "solutions", as a practical matter one commits to being only interested in considering problems for which one has some prior belief that a well-behaved solution exists for the underlying "noiseless" case (evidence for this belief can be accessed with the so-called "Discrete Picard Condition" [12, 6]). It would be natural to consider such a belief to be "priori" information not reflected in the choice of the identity as the penalty operator. As we remarked earlier, this could be partly responsible for the fact that zero-order Tihkhonov regularization usually does not perform better than alternative regularization methods.

So, first, from the stochastic viewpoint it would appear that the most desirable penalty operator R for the deterministic approach would have the feature that the given data suggests that $R[x]$ is not a member of $\mathcal{L}^2[\mathcal{S}]$, since this would indicate that (in this respect) the covariance implied by the penalty operator (via (1.8)) is consistent with the "true" solution x (which actually generates the data) being a typical realization of the imposed prior random function. Secondly, as discussed in the preceding paragraph, from the deterministic viewpoint it would appear that the most desirable prior for the stochastic approach would have the feature that realizations of the associated random function are expected to be square-integrable, violated by the prior covariance given by the identity operator, which is the maximum entropy choice under minimum information conditions [9]. Unfortunately, it is unclear how to uniquely satisfy either of these desires in the context of ordinary regularization procedures. However, both conditions can be met for a particular class of ill-posed problems using a novel regularization method. In addition, use of the resulting structures leads to a reduction in the amount of bias necessary to introduce obtaining a stable solution estimate. That is, both conditions can be met when the Minimax Entropy Principle can be nontrivially applied.

NOTATION. In the finite-dimensional setting, where $\mathcal{L}^2[\mathcal{S}], \mathcal{L}^2[\mathcal{T}]$ in Assumption D are replaced by $\mathbb{R}^m, \mathbb{R}^p$, the derivation of our program is very compactly expressed with tensor indices notation and the Einstein summation convention. This notation continues to serve quite well in the functional analytic setting arising here when linear F is compact or Fredholm, or is a linear combination of tensor products of such operators. Thus, we will formally write real square-integrable functions, random functions, and the latter operators, using this array notation (when the context is clear, we sometimes omit the indices of an operator we have previously written with indices). These "arrays" will be notated as uppercase Roman letters that have Greek subscripts or superscripts. Roman letter subscripts are used to distinguish different arrays, rather than indexing their entries. In all instances, it can first be verified that the expressions make sense in the discretized case, after which it can be seen that the obvious notational generalization pertains. Thus, X^μ can be either a real vector or a real function on a compact subset of $\mathbb{R}$, or their random counterparts. $X^{\mu\nu}$ can be a real second-rank tensor or a real function on a compact subset of $\mathbb{R}^2$, or their random counterparts. $F^\alpha{}_\mu$ (with transpose $F_\alpha{}^\mu$) can be a matrix, a compact operator, or a Fredholm operator (its higher rank analogues, such as $F^{\alpha\beta}{}_{\mu\nu}$, are linear combinations of tensor products of these). $\mathcal{E}[\cdot]$ denotes expectation.

We use the formal integral expression $\mathcal{E}[X^\mu X_\xi]$ to denote the covariance of random function X^μ (recognizing, as usual, that this kernel of the covariance operator might exist only in the sense of a distribution). $F^\alpha{}_\mu X^\mu$ and $X^\mu X_\mu$ are integral generalizations of the Einstein convention - the former gives the linear map of X^μ to Y^α, and the latter is the squared-norm of real X^μ ($\mathcal{E}[X^\mu X_\mu]$ is expected square-norm of random function X^μ). Raised and lowered indices are unimportant in the context of the $\mathcal{L}^2$-metric, and are used only as a part of the traditional summation convention - i.e., outside of that context, X^μ and X_μ are essentially the same thing (e.g., for random function X^μ the expression X_μ is not a member of the dual space). In analogy with nomenclature in finite-dimensional settings, we will refer to operations involving repeated indices as "contractions". δ^μ_ν indicates either the identity matrix or the identity operator. We emphasize that the indices notation is strictly formal and could at any time be replaced by more cumbersome notation involving integrals. We will denote the Frechet derivative of operator F as $\mathbf{d}F$. Finally, we will write Fx and Rx to indicate the expressions previously written as $F[x]$ and $R[x]$.

2. A recombinant operator for computation of the prior covariance. Proceeding under Assumption S, suppose $\mathcal{S}$ and $\mathcal{T}$ are compact subsets of $\mathbb{R}$ (in the univariate setting), with vector space $\mathcal{F}_{s,t}$ being the space of operators from $\mathcal{L}^2[\mathcal{S}]$ to $\mathcal{L}^2[\mathcal{T}]$, where each operator is of the form $aI + bK$, with K compact and $a, b \in \mathbb{R}$. All Fredholm operators between the above function spaces can be mapped to a member of $\mathcal{F}_{s,t}$ by the application of an isomorphism.

In our indices notation, (1.4) takes the form

$$Y^\alpha = F^\alpha{}_\mu X^\mu + N^\alpha. \tag{2.1}$$

A zero mean Gaussian random function X^μ is fully determined by its covariance $\mathcal{E}[X^\mu X_\xi]$, and so determines a zero mean Gaussian prior and therefore (given the data and noise model) a possible posterior random function [10]. Let $\mathcal{V}_\mathcal{S}$ be the set of zero mean Gaussian random functions over domain $\mathcal{S}$. $\mathcal{V}_\mathcal{S}$ is a vector space, and the members of $\mathcal{V}_\mathcal{S}$ each determine a posterior random function given a realization of Y^α and a noise model. $\mathcal{V}_\mathcal{S}$ is associated with a dual space, $\mathcal{V}^*_\mathcal{S}$. An element $J^\mu \in \mathcal{V}_\mathcal{S}$ generates an element $J^*_\mu \in \mathcal{V}^*_\mathcal{S}$ such that for any $M^\mu \in \mathcal{V}_\mathcal{S}$,

$$J^*_\mu M^\mu = \mathcal{E}[J_\mu M^\mu].$$

Now we consider the multivariate setting. Suppose $\mathcal{S} = \mathcal{W} \times \mathcal{Z}$, and $\mathcal{T} = \mathcal{P} \times \mathcal{Q}$, with $\mathcal{W}, \mathcal{Z}, \mathcal{P}, \mathcal{Q}$ compact subsets of $\mathbb{R}$. Then (2.1) generalizes to

$$Y^{\alpha\beta} = F^{\alpha\beta}{}_{\mu\nu} X^{\mu\nu} + N^{\alpha\beta}, \tag{2.2}$$

where we now assume $F^{\alpha\beta}{}_{\mu\nu} \in \mathcal{F}_{w,p} \otimes \mathcal{F}_{z,q}$ (the two factors in the tensor product are a vector space of operators from $\mathcal{L}^2[\mathcal{W}]$ to $\mathcal{L}^2[\mathcal{P}]$ and a vector space of operators from $\mathcal{L}^2[\mathcal{Z}]$ to $\mathcal{L}^2[\mathcal{Q}]$, respectively). According to Assumption S, random function $X^{\mu\nu}$ has zero mean. Hence, it is a member of the tensor product of vector spaces $\mathcal{V}_\mathcal{W} \otimes \mathcal{V}_\mathcal{Z}$, where it is understood that all members of the latter space have zero mean. Each member of

this space is a linear combination of tensor products of members of $\mathcal{V}_{\mathcal{W}}$ with members of $\mathcal{V}_{\mathcal{Z}}$. Thus, we can write

$$X^{\mu\nu} = \sum_i (W_i)^\mu (Z_i)^\nu, \tag{2.3}$$

with $(W_i)^\mu \in \mathcal{V}_{\mathcal{W}}$ and $(Z_i)^\nu \in \mathcal{V}_{\mathcal{Z}}$.

To solve the problem defined by (2.1) (i.e., where the solution is understood to be a posterior random function), the statistics of random function X^μ must be antecedently supplied virtually in their entirety. The minimum information prior (in discretized settings) has zero mean and covariance operator proportional to the identity [9] corresponding to the penalty operator used in zero-order Tikhonov regularization. As regards the choice of the prior random function $X^{\mu\nu}$ in (2.2), if we are given no prior constraints we can proceed in several ways. For example, we have no knowledge of how correlations related to the first variable are influenced by the second variable. Thus, as a first step we are motivated to restrict our attention to "second rank" (two variable) random functions, each having the property that all contractions with members of $\mathcal{V}^*_{\mathcal{Z}}$ (i.e., with respect to the second index) lead to first rank random functions that are proportional to each other. That is, manipulations with respect to the second variable cannot influence statistics related to the first variable (since we assume there are no prior constraints indicating how such distinctions could be assigned).

Definition 2.1. $X^{\mu\nu} \in \mathcal{V}_{\mathcal{W}} \otimes \mathcal{V}_{\mathcal{Z}}$ is *admissible* if there is a $W^\mu \in \mathcal{V}_{\mathcal{W}}$ such that for any $H^*_\nu \in \mathcal{V}^*_{\mathcal{Z}}$,

$$H^*_\nu X^{\mu\nu} = aW^\mu, \tag{2.4}$$

where a is a scalar dependent on H^*_ν.

In discretized settings, admissibility is a maximum entropy type of condition in that it requires that normalized contractions with random tensor $X^{\mu\nu}$ be assigned the same probability distribution. Admissibility is a proper subset of the full set of maximum entropy conditions on $X^{\mu\nu}$, which *in toto* would specify that the covariance of $X^{\mu\nu}$ is proportional to the identity. But note that admissibility as an expression of this particular subset of such constraints is not uniquely favored, and other such subsets could be imposed giving an alternative definition of admissibility, with equal justification.

Theorem 2.2. If $X^{\mu\nu}$ is admissible, then the covariance of $X^{\mu\nu}$ is

$$\mathcal{E}[X^{\mu\nu} X_{\xi\eta}] = \frac{\mathcal{E}[X^{\mu\sigma} X_{\xi\sigma}] \mathcal{E}[X^{\gamma\nu} X_{\gamma\eta}]}{\mathcal{E}[X^{\gamma\sigma} X_{\gamma\sigma}]}. \tag{2.5}$$

Proof. For $W^\mu \in \mathcal{V}_{\mathcal{W}}$ and $Z^\nu \in \mathcal{V}_{\mathcal{Z}}$, we have $\mathcal{E}[W^\mu Z^\nu] = 0$, since the members of $\mathcal{V}_{\mathcal{W}} \otimes \mathcal{V}_{\mathcal{Z}}$ have zero mean. Since W^μ and Z^ν are Gaussian, it then follows that they are independent (i.e., the members of $\mathcal{V}_{\mathcal{W}}$ are independent of those of $\mathcal{V}_{\mathcal{Z}}$). Using (2.3), equation (2.4) can be written as

$$H^*_\nu X^{\mu\nu} = \sum_i H^*_\nu [(W_i)^\mu (Z_i)^\nu] = \sum_i (W_i)^\mu H^*_\nu (Z_i)^\nu = aW^\mu. \tag{2.6}$$

The first equation follows from the equality of the integral of the limit of a sequence of measurable functions with the limit of the integral of the sequence. The second equality follows from independence of members of $\mathcal{V}_{\mathcal{W}}$ from those of $\mathcal{V}_{\mathcal{Z}}$. Without loss, we can

assume that the members of the set $\{(Z_i)^\nu\}$ in (2.3) are independent of each other. For a particular j and a particular k, define $H_\nu^{*(1)}, H_\nu^{*(2)} \in \mathcal{V}_\mathcal{Z}^*$ such that $H_\nu^{*(1)}(Z_i)^\nu = \delta_{ji}$, and $H_\nu^{*(2)}(Z_i)^\nu = \delta_{ki}$, where δ_{mn} is the Kronecker delta. Then $H_\nu^{*(1)}(X^{\mu\nu}) = (W_j)^\mu$ and $H_\nu^{*(2)}(X^{\mu\nu}) = (W_k)^\mu$. But by (2.4), this implies that $(W_j)^\mu \propto (W_k)^\mu$. Since k, j are arbitrary in the definition of $H_\nu^{*(1)}$ and $H_\nu^{*(2)}$, this implies that all members of $\{(W_i)^\mu\}$ are proportional to each other. Thus, for admissible $X^{\mu\nu}$ there exists $W^\mu \in \mathcal{V}_\mathcal{W}$ and $Z^\nu \in \mathcal{V}_\mathcal{Z}$ such that

$$X^{\mu\nu} = W^\mu Z^\nu. \tag{2.7}$$

Equation (2.7) implies

$$\begin{aligned}
\mathcal{E}[X^{\gamma\sigma}X_{\gamma\sigma}]\mathcal{E}[X^{\mu\nu}X_{\xi\eta}] &= \mathcal{E}[W^\gamma Z^\sigma W_\gamma Z_\sigma]\mathcal{E}[W^\mu Z^\nu W_\xi Z_\eta] \\
&= (\mathcal{E}[W^\gamma W_\gamma]\mathcal{E}[Z^\sigma Z_\sigma])(\mathcal{E}[W^\mu W_\xi]\mathcal{E}[Z^\nu Z_\eta]) \\
&= (\mathcal{E}[W^\gamma W_\gamma]\mathcal{E}[Z^\nu Z_\eta])(\mathcal{E}[W^\mu W_\xi]\mathcal{E}[Z^\sigma Z_\sigma]) \\
&= \mathcal{E}[W^\gamma Z^\nu W_\gamma Z_\eta]\mathcal{E}[W^\mu Z^\sigma W_\xi Z_\sigma] \\
&= \mathcal{E}[X^{\gamma\nu}X_{\gamma\eta}]\mathcal{E}[X^{\mu\sigma}X_{\xi\sigma}],
\end{aligned} \tag{2.8}$$

where the second and fourth equalities above follow from the previously noted independence of members of $\mathcal{V}_\mathcal{W}$ with respect to members of $\mathcal{V}_\mathcal{Z}$. Equation (2.8) implies (2.5). □

We now turn to the ill-posed problem defined by (2.2). The new treatment mechanism will depend on

Definition 2.3. Let $F \in \mathcal{F}_{w,p} \otimes \mathcal{F}_{z,q}$ be written as

$$F^{\alpha\beta}{}_{\mu\nu} = \sum_i (A_i)^\alpha{}_\mu (B_i)_\nu{}^\beta, \tag{2.9}$$

where $(A_i)^\alpha{}_\mu \in \mathcal{F}_{w,p}$, $(B_i)_\nu{}^\beta \in \mathcal{F}_{z,q}$, the sum is convergent in the operator norm, each member of $\{(A_i)^\alpha{}_\mu\}$ is linearly independent of the other members, and each member of $\{(B_i)_\nu{}^\beta\}$ is linearly independent of the other members. The *recombinant operator* $\mathcal{R}_F : \mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}] \times \mathcal{L}^2[\mathcal{W}\times\mathcal{W}] \to \mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}] \times \mathcal{L}^2[\mathcal{W}\times\mathcal{W}]$ is given by

$$\mathcal{R}_{F^{\alpha\beta}{}_{\mu\nu}} \begin{bmatrix} U^\nu{}_\eta \\ V^\mu{}_\xi \end{bmatrix} = \begin{bmatrix} \sum_{i,j} (A_i)_\alpha{}^\xi (A_j)^\alpha{}_\mu V^\mu{}_\xi \; (B_i)^\eta{}_{\beta'} (B_j)_\nu{}^\beta U^\nu{}_\eta \\ \sum_{i,j} (A_i)^\alpha{}_\mu (A_j)_{\alpha'}{}^\xi V^\mu{}_\xi \; (B_i)^\beta{}_\nu (B_j)_\beta{}^\eta U^\nu{}_\eta \end{bmatrix}. \tag{2.10}$$

$\mathcal{R}_{F^{\alpha\beta}{}_{\mu\nu}}$ is independent of the particular decomposition of $F^{\alpha\beta}{}_{\mu\nu}$ in (2.9), as will now be seen.

Corollary 2.4. Given the hypotheses of Assumption S, with linear $F \in \mathcal{F}_{w,p} \otimes \mathcal{F}_{z,q}$ and (1.4) written as (2.2), if $X^{\mu\nu}$ is admissible, then

$$\begin{bmatrix} \mathcal{E}[Y^{\alpha\beta}Y_{\alpha\beta'}] - \mathcal{E}[N^{\alpha\beta}N_{\alpha\beta'}] \\ \mathcal{E}[Y^{\alpha\beta}Y_{\alpha'\beta}] - \mathcal{E}[N^{\alpha\beta}N_{\alpha'\beta}] \end{bmatrix} = \mathcal{R}_{F^{\alpha\beta}{}_{\mu\nu}} \begin{bmatrix} \mathcal{E}[X^{\gamma\nu}X_{\gamma\eta}]/\mathcal{E}[X^{\gamma\sigma}X_{\gamma\sigma}]^{1/2} \\ \mathcal{E}[X^{\mu\sigma}X_{\xi\sigma}]/\mathcal{E}[X^{\gamma\sigma}X_{\gamma\sigma}]^{1/2} \end{bmatrix}. \tag{2.11}$$

Proof. Since $F \in \mathcal{F}_{w,p} \otimes \mathcal{F}_{z,q}$, the operator can be expanded as in (2.9). Thus, (2.2) becomes

$$Y^{\alpha\beta} = F^{\alpha\beta}{}_{\mu\nu} X^{\mu\nu} + N^{\alpha\beta} = \sum_i (A_i)^\alpha{}_\mu (B_i)_\nu{}^\beta X^{\mu\nu} + N^{\alpha\beta}. \tag{2.12}$$

We then have

$$\mathcal{E}[Y^{\alpha\beta}Y_{\alpha\beta'}] = \sum_{i,j}(A_i)_{\alpha}{}^{\xi}(A_j)^{\alpha}{}_{\mu}(B_i)^{\eta}{}_{\beta'}(B_j)_{\nu}{}^{\beta}\mathcal{E}[X^{\mu\nu}X_{\xi\eta}] + \mathcal{E}[N^{\alpha\beta}N_{\alpha\beta'}]. \tag{2.13}$$

Substitution of (2.5) immediately gives

$$\begin{aligned}&(\mathcal{E}[Y^{\alpha\beta}Y_{\alpha\beta'}] - \mathcal{E}[N^{\alpha\beta}N_{\alpha\beta'}])\\ &\quad= \left\{\sum_{i,j}\left((A_i)_{\alpha}{}^{\xi}(A_j)^{\alpha}{}_{\mu}\frac{\mathcal{E}[X^{\mu\sigma}X_{\xi\sigma}]}{\mathcal{E}[X^{\gamma\sigma}X_{\gamma\sigma}]^{1/2}}\right)(B_i)^{\eta}{}_{\beta'}(B_j)_{\nu}{}^{\beta}\right\}\frac{\mathcal{E}[X^{\gamma\nu}X_{\gamma\eta}]}{\mathcal{E}[X^{\gamma\sigma}X_{\gamma\sigma}]^{1/2}}.\end{aligned} \tag{2.14}$$

A similar observation related to $\mathcal{E}[Y^{\alpha\beta}Y_{\alpha'\beta}]$ leads to

$$\begin{aligned}&(\mathcal{E}[Y^{\alpha\beta}Y_{\alpha'\beta}] - \mathcal{E}[N^{\alpha\beta}N_{\alpha'\beta}])\\ &\quad= \left\{\sum_{i,j}\left((B_i)^{\beta}{}_{\nu}(B_j)_{\beta}{}^{\eta}\frac{\mathcal{E}[X^{\gamma\nu}X_{\gamma\eta}]}{\mathcal{E}[X^{\gamma\sigma}X_{\gamma\sigma}]^{1/2}}\right)(A_i)^{\alpha}{}_{\mu}(A_j)_{\alpha'}{}^{\xi}\right\}\frac{\mathcal{E}[X^{\mu\sigma}X_{\xi\sigma}]}{\mathcal{E}[X^{\gamma\sigma}X_{\gamma\sigma}]^{1/2}}.\end{aligned} \tag{2.15}$$

Equations (2.14), (2.15), and (2.10), imply (2.11). □

Corollary 2.5. $\mathcal{R}_F$ is independent of the particular decomposition of F given by the right-hand side of (2.9).

Proof. Equation (2.11) must hold for any decomposition of F having the form of the right-hand side of (2.9). But the expression operated upon by $\mathcal{R}_F$ on the right-hand side of (2.11) is an arbitrary member of $\mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}]\times\mathcal{L}^2[\mathcal{W}\times\mathcal{W}]$ that is independent of the decomposition of F. The left-hand side of (2.11) is also independent of that decomposition. Therefore, $\mathcal{R}_F$ must be independent of the decomposition as well. □

According to Theorem 2.2, the two components of the expression operated upon by $\mathcal{R}_F$ on the right-hand side of (2.11) yield the prior covariance as their tensor product. A noisy version of the left-hand side of (2.11) is available to us via manipulation of the data y_δ (data realizations differ only by the noise realization). This motivates us to begin examining under what conditions (2.11) can be treated to obtain a useful estimate of the prior covariance from available data.

We denote the Frechet derivative of $\mathcal{R}_F$ at (U,V) as $\mathbf{d}\mathcal{R}_F|_{(U,V)}$. From (2.10), it is evident that this derivative exists. In terms of the decomposition in (2.9),

$$\mathbf{d}\mathcal{R}_{F^{\alpha\beta}{}_{\mu\nu}}|_{(U^{\nu}{}_{\eta},V^{\mu}{}_{\xi})} = \begin{pmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{D}\end{pmatrix}, \tag{2.16}$$

where

$$\mathbf{A} = \sum_{i,j} \left((A_i)_\alpha{}^\xi (A_j)^\alpha{}_\mu V^\mu{}_\xi \right) (B_i)^\eta{}_{\beta'} (B_j)_\nu{}^\beta, \tag{2.17}$$

$$\mathbf{B} = \sum_{i,j} (A_i)_\alpha{}^\xi (A_j)^\alpha{}_\mu (B_i)^\eta{}_{\beta'} (B_j)_\nu{}^\beta U^\nu{}_\eta, \tag{2.18}$$

$$\mathbf{C} = \sum_{i,j} (A_i)^\alpha{}_\mu (A_j)_\alpha{}^\xi V^\mu{}_\xi (B_i)^\beta{}_\nu (B_j)_\beta{}^\eta, \tag{2.19}$$

$$\mathbf{D} = \sum_{i,j} (A_i)^\alpha{}_\mu (A_j)_\alpha{}^\xi \left((B_i)^\beta{}_\nu (B_j)_\beta{}^\eta U^\nu{}_\eta \right). \tag{2.20}$$

Prior to application of Theorem 2.2, it is (in practical terms) out of the question to consider estimation of the covariance of $X^{\mu\nu}$ using a data realization of $Y^{\alpha\beta}$ (on dimensionality grounds). As Corollary 2.4 shows, such a determination at least becomes feasible in the context of Theorem 2.2 (i.e., via the admissibility assumption). Although the utility of a *regularized* solution to the equations of Corollary 2.4 is still questionable (i.e., solution of the nonlinear simultaneous equations (2.14), (2.15)), the recombinant operator often defines a *well-posed problem*.

LEMMA 2.6. For $F^{\alpha\beta}{}_{\mu\nu} \in \mathcal{F}_{w,p} \otimes \mathcal{F}_{z,q}$, there exist $a \in \mathbb{R}$ and compact $A^\alpha{}_\mu$, $B_\nu{}^\beta$, $K^{\alpha\beta}{}_{\mu\nu}$, such that

$$F^{\alpha\beta}{}_{\mu\nu} = a\, \delta^\alpha_\mu \delta^\beta_\nu + A^\alpha{}_\mu \delta^\beta_\nu + \delta^\alpha_\mu B_\nu{}^\beta + K^{\alpha\beta}{}_{\mu\nu}. \tag{2.21}$$

Proof. Each term in the sum on the right-hand side of (2.9) can be expanded to be in the form of the right-hand side of (2.21). The sum of any number of such terms can evidently still be written in the form of the right-hand side of (2.21) (because the sum of compact operators remains compact). The lemma assertion then follows since the sum on the right-hand side of (2.9) converges in the operator norm. □

If $a \neq 0$ in (2.21), then (2.2) defines a well-posed problem apart from some exceptional values of a (related to the eigenvalues of the sum of the last three terms on the right-hand side of (2.21)). If $a = 0$, $A^\alpha{}_\mu = 0$ and $B_\nu{}^\beta = 0$, then $F^{\alpha\beta}{}_{\mu\nu}$ is compact, and (2.2) defines an "ordinary" ill-posed problem. However, the intermediate cases define a different class of problem.

THEOREM 2.7. Suppose $F^{\alpha\beta}{}_{\mu\nu} \in \mathcal{F}_{w,p} \otimes \mathcal{F}_{z,q}$ is such that in (2.21), $a = 0$, $A^\alpha{}_\mu \neq 0$ and $B_\nu{}^\beta \neq 0$. Then $\mathbf{d}\mathcal{R}_{F^{\alpha\beta}{}_{\mu\nu}}|_{(U^\nu{}_\eta, V^\mu{}_\xi)}$ has a continuous inverse for $(U^\nu{}_\eta, V^\mu{}_\xi)$ in a dense open subset of $\mathcal{L}^2[\mathcal{Z} \times \mathcal{Z}] \times \mathcal{L}^2[\mathcal{W} \times \mathcal{W}]$, so that $\mathcal{R}_{F^{\alpha\beta}{}_{\mu\nu}}$ is a local homeomorphism around each point of this subset.

Proof. From the theorem hypotheses and Lemma 2.6, we can rewrite (2.9) by setting $(A_1)^\alpha{}_\mu = A^\alpha{}_\mu$, $(A_2)^\alpha{}_\mu = \delta^\alpha_\mu$, $(B_1)_\nu{}^\beta = \delta^\beta_\nu$, $(B_2)_\nu{}^\beta = B_\nu{}^\beta$, where $A^\alpha{}_\mu$ and $B_\nu{}^\beta$ are compact, as are all remaining terms of the sum on the right-hand side of (2.9). Substituting into (2.17), we can then write

$$\mathbf{A} = a_1 \delta^\eta_{\beta'} \delta^\beta_\nu + \delta^\eta_{\beta'} (K_1)^\beta{}_\nu + (K_2)^\eta{}_{\beta'} \delta^\beta_\nu + (K_3)^{\eta\beta}{}_{\beta'\nu} \equiv a_1 \delta^\eta_{\beta'} \delta^\beta_\nu - H^{\eta\beta}{}_{\beta'\nu}, \tag{2.22}$$

where $a_1 = A_\alpha{}^\xi A^\alpha{}_\mu V^\mu{}_\xi$, while K_1, K_2, K_3 are compact operators continuously dependent on $V^\mu{}_\xi$. For a fixed $V^\mu{}_\xi$, the spectrum of $\mathbf{A}$ is evidently discrete. If $\mathbf{A}$ fails to have an

inverse for some $(V_1)^{\mu}{}_{\xi}$, then $H^{\eta\beta}{}_{\beta'\nu}$ has an eigenvalue equal to a_1. But in that case, there are $V^{\mu}{}_{\xi}$ arbitrarily close to $(V_1)^{\mu}{}_{\xi}$ such that $\mathbf{A}$ *does* have a continuous inverse, since otherwise $H^{\eta\beta}{}_{\beta'\nu}$ would necessarily have an eigenvalue varying exactly with $a_1 = A_{\alpha}{}^{\xi}A^{\alpha}{}_{\mu}V^{\mu}{}_{\xi}$ (this is impossible since the $(A_i)^{\alpha}{}_{\mu}$ are linearly independent). Thus, the points $V^{\mu}{}_{\xi}$ for which $\mathbf{A}$ has a continuous inverse are dense in $\mathcal{L}^2[\mathcal{S}]$. Now suppose $\mathbf{A}$ has a continuous inverse for some $(V_2)^{\mu}{}_{\xi}$. In that case, $H^{\eta\beta}{}_{\beta'\nu}$ does not have an eigenvalue equal to a_1. For sufficiently small perturbations of $(V_2)^{\mu}{}_{\xi}$, the terms a_1 and $H^{\eta\beta}{}_{\beta'\nu}$ (which vary continuously with $V^{\mu}{}_{\xi}$) will not change sufficiently for this fact to be altered. Thus, if $\mathbf{A}$ has an inverse at some point, then it will have an inverse at all points of an open set around that point. Thus, the points where $\mathbf{A}$ has an inverse are open and dense in $\mathcal{L}^2[\mathcal{W}\times\mathcal{W}]$. The same argument holds for $\mathbf{D}$, since (as seen by substitution in (2.20)) it has precisely the same form as $\mathbf{A}$,

$$\mathbf{D} = a_2\delta^{\alpha}_{\mu}\delta^{\xi}_{\alpha'} + \delta^{\alpha}_{\mu}(K_4)^{\xi}{}_{\alpha'} + (K_5)^{\alpha}{}_{\mu}\delta^{\xi}_{\alpha'} + (K_6)^{\alpha\xi}{}_{\mu\alpha'},$$

where K_4, K_5, K_6 are compact. That is, $\mathbf{D}$ has an inverse at each $U^{\nu}{}_{\eta}$ in a dense open subset of $\mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}]$.

Using (2.21), it is also easy to show that (2.18) becomes

$$\mathbf{B} = \delta^{\xi}_{\mu}(K_7)^{\beta}{}_{\beta'} + (K_8)^{\xi\beta}{}_{\mu\beta'}$$

for K_7, K_8 compact. Since $\mathbf{C}$ has the same form as $\mathbf{B}$, we can write (2.19) as

$$\mathbf{C} = (K_9)^{\alpha}{}_{\alpha'}\delta^{\eta}_{\nu} + (K_{10})^{\alpha\eta}{}_{\alpha'\nu}$$

for K_9, K_{10} compact.

Now, the formal inverse of the right-hand side of (2.16) is

$$\begin{pmatrix} (\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & (\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{pmatrix}. \tag{2.23}$$

As we have noted, $\mathbf{D}^{-1}$ exists on a dense open subset of $\mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}]$. On this subset,

$$\mathbf{B}\mathbf{D}^{-1}\mathbf{C} = \big(\delta^{\xi}_{\mu}(K_7)^{\beta}{}_{\beta'} + (K_8)^{\xi\beta}{}_{\mu\beta'}\big)(\mathbf{D}^{-1})^{\mu\alpha'}{}_{\alpha\xi}\big((K_9)^{\alpha}{}_{\alpha'}\delta^{\eta}_{\nu} + (K_{10})^{\alpha\eta}{}_{\alpha'\nu}\big),$$

where on the right-hand side above we have written the indices associated with $\mathbf{D}^{-1}$. On the dense open subset where continuous $\mathbf{D}^{-1}$ exists, it is then evident that

$$\mathbf{B}\mathbf{D}^{-1}\mathbf{C} = (K_{11})^{\beta}{}_{\beta'}\delta^{\eta}_{\nu} + (K_{12})^{\beta\eta}{}_{\beta'\nu} \tag{2.24}$$

for K_{11}, K_{12} compact. Now, suppose there is $\big((U_3)^{\nu}{}_{\eta}, (V_3)^{\mu}{}_{\xi}\big) \in \mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}]\times\mathcal{L}^2[\mathcal{W}\times\mathcal{W}]$ for which $\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ fails to have an inverse. In view of (2.22) and (2.24), we see that $\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ has the same form as $\mathbf{A}$. Thus, we can apply to $\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}$ the same argument relating to the existence of the inverse of $\mathbf{A}$. That is, if the former lacks a continuous inverse at $\big((U_3)^{\nu}{}_{\eta}, (V_3)^{\mu}{}_{\xi}\big)$, then there are arbitrarily small variations in $(V_3)^{\mu}{}_{\xi}$ that will produce a point at which a continuous inverse exists. Again, all points in a sufficiently small neighborhood of a point where the inverse exists must have continuous inverses as well, by the prior reasoning. The same argument holds for $(\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B})$ as regards perturbing $U^{\nu}{}_{\eta} \in \mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}]$. Thus, (2.23) exists for a dense open set of points in $\mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}]\times\mathcal{L}^2[\mathcal{W}\times\mathcal{W}]$. □

Since the recombinant operator is a local homeomorphism around any point of a dense open set in $\mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}]\times\mathcal{L}^2[\mathcal{W}\times\mathcal{W}]$, for a sufficiently good seed one can anticipate a stable iterative solution mechanism for the equations of Corollary 2.4 (i.e., (2.14) and (2.15)) when the hypotheses of Theorem 2.7 are satisfied.

We can apply this result to various multivariate ill-posed problems. For example, suppose

$$F^{\alpha\beta}{}_{\mu\nu} = A^{\alpha}{}_{\mu}\delta^{\beta}_{\nu} + \delta^{\alpha}_{\mu}B_{\nu}{}^{\beta}, \tag{2.25}$$

with $A^{\alpha}{}_{\mu}$ and $B_{\nu}{}^{\beta}$ compact. The discretized noiseless form of (2.12) is then a Sylvester equation $Y = AX + XB'$ (where Y, A, X, B are matrices). This can be alternatively written as

$$Y_{:} = (I\otimes A + B\otimes I)X_{:},$$

where $Y_{:}$ is the vector formed by sequentially appending the columns of Y (and similarly for $X_{:}$) and $\otimes$ is the Kronecker product. Since A and B are discretizations of compact operators, the inverse of $(I\otimes A + B\otimes I)$ has arbitrarily large operator norm as the discretization becomes finer and finer (unless $A^{\alpha}{}_{\mu}$ and $B_{\nu}{}^{\beta}$ are degenerate). Zero-order Tikhonov regularization can be cast as a maximum entropy method here. But, we can also derive a solution from the admissibility criterion of Definition 2.1, which uses a proper subset of the zero-order Tikhonov constraints. Formally, this must imply lesser bias, as we note in Corollary 2.10 below.

As a second example, consider (2.25) where $B_{\nu}{}^{\beta} = 0$, so that

$$F^{\alpha\beta}{}_{\mu\nu} = A^{\alpha}{}_{\mu}\delta^{\beta}_{\nu}. \tag{2.26}$$

Again, this operator lacks a continuous inverse. This operator arises in so-called nonstationary inverse problems,

$$y(t,u) = \int_{\mathcal{S}} f(s,t)x(s,u)ds, \tag{2.27}$$

where $f(s,t)$ is square-integrable on $\mathcal{S}$, and it is required to estimate $x(s,u)$ given $y(t,u)$. Identifying $A^{\alpha}{}_{\mu}\delta^{\beta}_{\nu}$ with $\int_{\mathcal{S}} f(s,t)(\cdot)ds$, $X^{\mu\nu}$ with $x(s,u)$, and $Y^{\alpha\beta}$ with $y(t,u)$, this is of the form (2.12). If a candidate for $\mathcal{E}[X^{\mu\sigma}X_{\xi\sigma}]$ is supplied *a priori*, it is possible to treat (2.14) by itself to obtain $\mathcal{E}[X^{\gamma\nu}X_{\gamma\eta}]$, thus supplying the prior covariance operator. However, in all discretized settings, this program has the same amount of constraint information content as the usual approach, since the latter also only requires a choice of $\mathcal{E}[X^{\mu\sigma}X_{\xi\sigma}]$ (from which it can treat (2.27) for any fixed value of u individually).

Thus, neither approach is necessarily favored over the other when the operator is as in (2.26). On the other hand, for an operator

$$F^{\alpha\beta}{}_{\mu\nu} = A^{\alpha}{}_{\mu}\delta^{\beta}_{\nu} + K^{\alpha\beta}{}_{\mu\nu},$$

with K compact, the new approach would be favored according to the Minimax Entropy Principle (assuming K is nondegenerate).

REMARK 2.8. As with standard regularization approaches, generalization to nonlinear F is more difficult. We only observe that the above mechanism relevant to linear equations can evidently be incorporated into every iteration used in a standard Levenberg-Marquardt approach for nonlinear equations, assuming F has a Frechet derivative of the above requisite form (i.e., in line with the hypotheses of Theorem 2.7).

REMARK 2.9. Under Assumption D, if $\mathbf{d}\mathcal{R}_{\mathbf{d}F|x^\dagger}$ has a continuous inverse, then a nontrivial Minimax Entropy approach (MME) is possible.

COROLLARY 2.10. Suppose we are given the hypotheses of Assumption S and Theorem 2.7, with linear F, admissible $x = X^{\mu\nu}$, and we are also given $(\widehat{Y}_i)^{\alpha\beta} \in \mathcal{L}^2[\mathcal{T}] = \mathcal{L}^2[\mathcal{P}\times\mathcal{Q}]$ as a realization of y when $\delta = \delta_i$, for $i = 1, 2, \ldots$, with $\lim_{i\to\infty}\delta_i = 0$, and $(\widehat{Y}_0)^{\alpha\beta} = \lim_{i\to\infty}(\widehat{Y}_i)^{\alpha\beta} \in \mathcal{L}^2[\mathcal{T}]$. Suppose the factors comprising covariance function C_x as given by Theorem 2.2 reside in the dense open subset of $\mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}]\times\mathcal{L}^2[\mathcal{W}\times\mathcal{W}]$ for which $\mathbf{d}\mathcal{R}_F|_{(U,V)}$ has a continuous inverse. Let C_{x,δ_i} be the tensor product of the solutions of the simultaneous equations of Corollary 2.4 (equations (2.14) and (2.15)) when $(\widehat{Y}_i)^{\alpha\beta}(\widehat{Y}_i)_{\alpha\beta'}$ and $(\widehat{Y}_i)^{\alpha\beta}(\widehat{Y}_i)_{\alpha'\beta}$ replace the left-hand sides of the respective equations. Then,

(1) The covariance function of x is such that $C_x = \lim_{i\to\infty} C_{x,\delta_i}$. If R is such that $(R'R)^{-1}$ is the covariance operator with covariance function C_x, then realizations of Rx do not have expected finite $\mathcal{L}^2$-norm.
(2) Realizations of x have expected finite $\mathcal{L}^2$-norm.
(3) Consider a discretization of x, i.e., x is replaced by a zero mean Gaussian random vector having finitely many components. Then the information content embodied by the admissibility condition is less than the information content of an imposed prior covariance, and this difference tends to infinity as the number of components in the discretization of x tends to infinity.

Proof. (1) By hypothesis, the two factors forming C_x (i.e., the two terms outside the braces on the right-hand sides of (2.14) and (2.15)) comprise a point in the dense open subset of $\mathcal{L}^2[\mathcal{Z}\times\mathcal{Z}]\times\mathcal{L}^2[\mathcal{W}\times\mathcal{W}]$ where $\mathcal{R}_F$ is a local homeomorphism. Thus, as $(\widehat{Y}_i)^{\alpha\beta} \to (\widehat{Y}_0)^{\alpha\beta}$, we have $C_{x,\delta_i} \to C_x$. By an identical argument as presented at the end of the proof of Theorem 1.1, the assertion follows since Rx is a random function whose covariance operator is the identity.

(2) By hypothesis, $\|(\widehat{Y}_0)^{\alpha\beta}(\widehat{Y}_0)_{\alpha\beta'} - \big(\mathcal{E}[Y^{\alpha\beta}Y_{\alpha\beta'}] - \mathcal{E}[N^{\alpha\beta}N_{\alpha\beta'}]\big)\|^2 = 0$ and $\|(\widehat{Y}_0)^{\alpha\beta}(\widehat{Y}_0)_{\alpha'\beta} - \big(\mathcal{E}[Y^{\alpha\beta}Y_{\alpha'\beta}] - \mathcal{E}[N^{\alpha\beta}N_{\alpha'\beta}]\big)\|^2 = 0$, with $(\widehat{Y}_0)^{\alpha\beta}(\widehat{Y}_0)_{\alpha\beta'} \in \mathcal{L}^2[\mathcal{Q}\times\mathcal{Q}]$ and $(\widehat{Y}_0)^{\alpha\beta}(\widehat{Y}_0)_{\alpha'\beta} \in \mathcal{L}^2[\mathcal{P}\times\mathcal{P}]$. Also by hypothesis, $\mathcal{R}_F$ in (2.11) is a local homeomorphism. Thus, $\frac{\mathcal{E}[X^{\gamma\nu}X_{\gamma\eta}]}{\mathcal{E}[X^{\gamma\sigma}X_{\gamma\sigma}]^{1/2}}$ and $\frac{\mathcal{E}[X^{\mu\sigma}X_{\xi\sigma}]}{\mathcal{E}[X^{\gamma\sigma}X_{\gamma\sigma}]^{1/2}}$ are each square-integrable. Hence, C_x is square-integrable, which implies that the realizations of x have expected finite $\mathcal{L}^2$-norm.

(3) In the context of discretized x, the information content of a specified prior covariance must be the sum of the information content of the admissibility constraint plus the information content of prior specification of the two factors in the numerator on the right-hand-side of (2.5) (since the admissibility constraint does not itself influence the particular values of the entries of the latter two matrices, which can be thought of as distinct "spatial" and "temporal" covariances). Thus, the prior covariance information content is strictly greater than the information content of the admissibility condition. Also, as the number of components of the spatial and temporal covariance matrices tend to infinity, their information content must tend to infinity. □

REMARK 2.11. Properties (1) and (2) of Corollary 2.10 indicate that the nontrivial MME treatment of applicable ill-posed problems via a solution of the equations of Corollary 2.4 ((2.14) and (2.15)) resolves the deterministic versus stochastic discordances

described in Section 1.2. Specifically, given the prior assumption of admissibility (Definition 2.1), a penalty operator R is supplied such that Rx is expected not to have a finite norm, this being a feature consistent with a good choice for the associated imposed covariance operator. Also, the selected prior covariance implies an expected square-integrable realization of the prior random function, unlike the (default Maximum Entropy) zero-order Tikhonov prior (covariance proportional to the identity operator). Finally, property (3) of Corollary 2.10 indicates that the bias introduced by the admissibility constraint (Minimax Entropy) in this setting is less than that introduced by prior imposition of a probability density (such as with the usual Maximum Entropy default).

3. Computational issues.

3.1. *Preliminaries.* In the setting of Tikhonov regularization, a solution estimate for a linear ill-posed problem is taken to be as in (1.2). Assuming F and R are linear and R has an inverse, this estimate corresponds to the mean of the posterior random function for a Gaussian white noise model, where $(\lambda R'R)^{-1}$ is the covariance of x multiplied by the inverse of the noise variance. This can be verified from the proof of Theorem 1.1, where explicit computation of the minimizer in (1.2) leads to (1.6) with the above specifications. Thus, for a white noise model, both the MME treatment and Tikhonov regularization are formally described by (1.2), the difference being that (if the hypotheses of Theorem 2.7 are satisfied) MME supplies $(R'R)^{-1}$ from the data and admissibility criterion (Definition 2.1).

In analyzing the numerical issues in the discretized setting, the generalized singular value decomposition of a matrix pair (GSVD) [11], the L-curve associated with regularization parameter selection [7], and a discretized counterpart to the Picard Condition [6, 12], are of interest.

3.1.1. *The L-curve and GSVD.* When the noise variance is unknown (precluding use of the Discrepancy Principle), a number of methods are employed for suggesting a favored value of λ in (1.2). One useful method is that of Generalized Cross-Validation, particularly since it has a desirable convergence property in discretized settings [14]. Another popular method is that of the L-curve [7]. This is the plot $(\log \|Fx_\lambda - y\|^2, \log \|Rx_\lambda\|^2)$ parametrized by the regularization parameter λ. The L-curve often has a pronounced corner corresponding to a heuristically reasonable value for λ. As it happens, for finer and finer discretizations, the curvature of the L-curve corner either tends to zero or the point of greatest curvature does not actually correspond to the optimal regularization parameter [13]. Also, for nondiscretized problems it has been shown that as $\delta \to 0$ the solution estimate chosen by the maximum curvature point on the L-curve corner does not converge to the noiseless solution [5]. Nevertheless, the L-curve is a revealing depiction of the relative behaviors of the discrepancy (the abscissa) and penalty functional (the ordinate). For each value of the regularization parameter, the latter are the two co-minimized terms in (1.2).

The GSVD is helpful in analysis of the L-curve. Assuming that R is a nonsingular matrix, we can write the GSVD of (F, R) in a simplified form:

$$F = U\Sigma W^{-1}, \tag{3.1}$$

$$R = VMW^{-1} \tag{3.2}$$

with $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$, $\sigma_i \geq 0$, $M = \text{diag}(\mu_1, \ldots, \mu_n)$, $\mu_i > 0$, $\{\sigma_i\}$ monotonic increasing, $\{\mu_i\}$ monotonic decreasing, and $\sigma_i^2 + \mu_i^2 = 1$. The monotonically increasing set of elements $\{\gamma_1, \ldots, \gamma_n\}$, $\gamma_i = \sigma_i/\mu_i$, is the set of generalized singular values. In this case (since R is nonsingular) the GSVD follows trivially from the singular value decomposition of FR^{-1}. Although it is the usual convention to order the general singular values such that $\{\gamma_i\}$ is an increasing sequence, from now on we will assume that these are ordered so that $\{\gamma_i\}$ is a decreasing sequence.

We can use the GSVD to write the regularized solution estimate (1.6) as

$$x_\lambda = \sum_i \frac{\gamma_i^2}{\gamma_i^2 + \lambda} \frac{u_i' y}{\sigma_i} w_i, \tag{3.3}$$

where u_i is the i-th column of U (so that $u_i' y$ is the inner product of u_i and y) and w_i is the i-th column of W. From (1.7), with C_e proportional to the identity matrix and $C_x^{-1} C_e = \lambda R' R$, equation (3.3) follows from direct substitution of the GSVD expressions. Note that U, V, Σ, M are stably obtained from the SVD of FR^{-1}, and $W = R^{-1} VM$. Thus, there is no inversion of ill-conditioned matrices given the stable computation of $(R'R)^{-1}$ via the solution of the equations of Corollary 2.4 and use of Theorem 2.2.

We can now observe that the L-curve is monotonically decreasing [7]. That is, using (3.3) and the GSVD of (F, R), we have

$$\|Rx_\lambda\|^2 = \sum_i \left(\frac{\gamma_i^2}{\gamma_i^2 + \lambda} \frac{u_i' y}{\gamma_i} \right)^2 \tag{3.4}$$

and

$$\|Fx_\lambda - y\|^2 = \sum_i \left(\frac{\lambda}{\gamma_i^2 + \lambda} u_i' y \right)^2 + \epsilon^2, \tag{3.5}$$

where ϵ is the norm of the component of y outside the range of F. The monotonicity then immediately follows, since (3.4) is monotonic decreasing as λ increases and (3.5) is monotonic increasing as λ increases. As noted, the L-curve often has a pronounced corner corresponding to solution estimates "near" the mean of the posterior density, as discussed in [7].

3.1.2. *The Discrete Picard Condition.* We now return to the general Hilbert space setting. The singular value expansion of a square-integrable function $f(s,t)$ can be written as $f(s,t) = \sum_i \alpha_i u_i(s) v_i(t)$, where $\{\alpha_i, u_i(s)\}$ is the set of eigensolutions of the symmetric kernel $\int_S f(s,t) f(s',t) dt$ and $\{\alpha_i, v_i(t)\}$ is the set of eigensolutions of the symmetric kernel $\int_S f(s,t) f(s,t') ds$. The Picard condition states that the first-kind Fredholm equation,

$$y(t) = \int_S f(s,t) x(s) ds,$$

has a square-integrable solution if and only if

$$\|x(s)\|^2 = \sum_i \frac{1}{\alpha_i^2} \left(\int_S y(t) u_i(t) dt \right)^2 < \infty.$$

If the Picard Condition is satisfied, then

$$x(s) = \sum_i \frac{1}{\alpha_i} \left(\int_{\mathcal{S}} y(t) u_i(t) dt \right) v_i(s). \tag{3.6}$$

Evidently, the Picard Condition requires that the Fourier coefficients $\int_S y(t)u_i(t)dt$ tend to zero faster than the singular values α_i. In the noiseless discretized setting for $\lambda = 0$, with the solution estimate given by (1.6), equation (3.4) implies

$$\|Rx\|^2 = \sum_i \left(\frac{u_i' y}{\gamma_i} \right)^2. \tag{3.7}$$

In fact, for arbitrary R, there is no reason to suppose that $\|Rx]\|^2$ is small. Nevertheless, if R is selected as a penalty matrix, Tikhonov regularization calls for the co-minimization of $\|Rx\|^2$ (i.e., (1.2)). Based on (3.7), it has been argued that in order for x_λ in the noisy setting to be a reasonable mimic of x in the noiseless setting (taking into account the prior significance R has for the true solution x), $|u_i' y|$ should decrease faster than γ_i "on average" for those terms of the sum that are not dominated by noise (again, we assume reversal of the usual ordering of the generalized singular values, so that the sequence is monotonic decreasing). This is the Discrete Picard Condition (DPC). This argument is expanded in [6, 12], where evidence is presented that DPC is satisfied for penalty operators typically selected for Tikhonov regularization. Yet, from Theorem 1.1, we know that a feature of a good penalty operator R is that $\|Rx\|^2$ should not be finite; i.e., a good choice or R should be such that DPC is *violated* for the given data y and F.

3.2. *A numerical illustration.* According to the three parts of Corollary 2.10, analysis of the results of MME treatments should show that

(1) MME utilizes a prior covariance operator suggesting that Rx is not square-integrable (where x is the unknown true solution generating the data),
(2) the prior covariance selected by the MME treatment is consistent with the presumption of square-integrable x,
(3) the accuracy of the MME treatment exceeds that of zero-order Tikhonov regularization (the usual Maximum Entropy treatment) in a particular signal-to-noise interval between essentially well-posed conditions (where all treatments work extremely well) and grossly ill-posed conditions (where all treatments are virtually worthless). This expected result follows from the presumption that stable methods imposing less "unjustified" bias will tend to be more accurate in general.

We will illustrate such numerical results by treatment of an ill-posed problem described by the Sylvester equation.

The (noisy) Sylvester equation is of the form (2.12) with the specification given by (2.25). Accordingly, the terms in (2.25) can be substituted into (2.14) and (2.15) to obtain equations for computation of the regularization tensor. There is no further need to continue with the indices notation, and we will employ matrix notation for the Sylvester

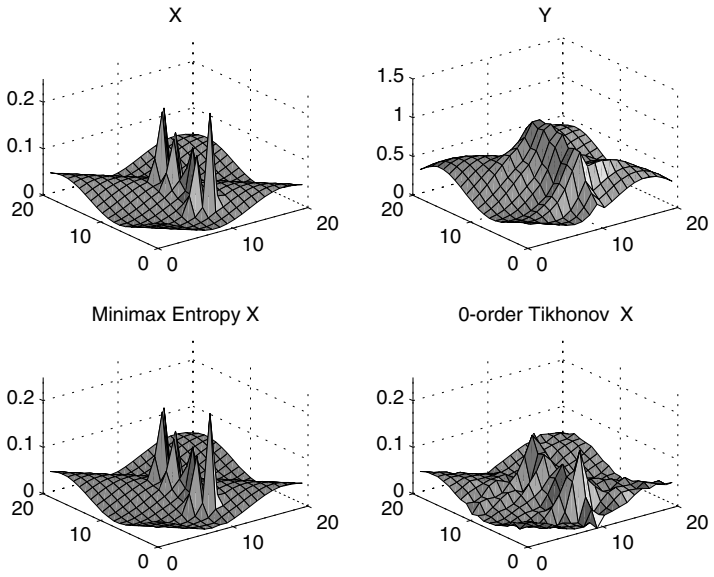

FIG. 1. Plots of the true solution (X), data (Y), the MME estimate, and the zero-order Tikhonov estimate, for the numerical example in Section 3.2. The MME estimate has much greater accuracy than the estimate resulting from zero-order Tikhonov regularization.

equation as

$$Y = AX + XB' + N \tag{3.8}$$

to represent the resulting form of (2.12).

We choose the "unknown" X to consist of a combination of various peaks (constructed from Gaussians) on a wave background (Figure 1). For $m = 1, \ldots, 20$ and $n = 1, \ldots, 20$, the entries of X are explicitly given by

$$\begin{aligned} X(m,n) &= 0.05\cos^2((m+n)/6) \\ &+ 0.1e^{-5(m-20/2)^2}e^{-3(n-20/2)^2} \\ &+ 0.2e^{-5(m-20/2.3)^2}e^{-3(n-20/2.3)^2} \\ &+ 0.2e^{-5(m-20/1.8)^2}e^{-3(n-20/2.3)^2} \\ &+ 0.2e^{-5(m-20/2.7)^2}e^{-3(n-20/2)^2} \\ &+ 0.2e^{-5(m-20/3.8)^2}e^{-3(n-20/2.3)^2} \\ &+ 0.2e^{-5(m-20/1.7)^2}e^{-3(n-20/2)^2} \\ &+ 0.2e^{-5(m-20/5)^2}e^{-3(n-20/2)^2}. \end{aligned}$$

Since our example actually illustrates treatment of a Lyapunov equation, we have $A = B$. To preserve reference to the more general Sylvester equation, we will continue to refer to A and B below. We take A to be a Gaussian convolution matrix (leading to Y as in Figure 1). For $m = 1, \ldots, 20$ and $n = 1, \ldots, 20$, the entries of A are explicity given by

$$A(m,n) = e^{-(m-n)^2/16}.$$

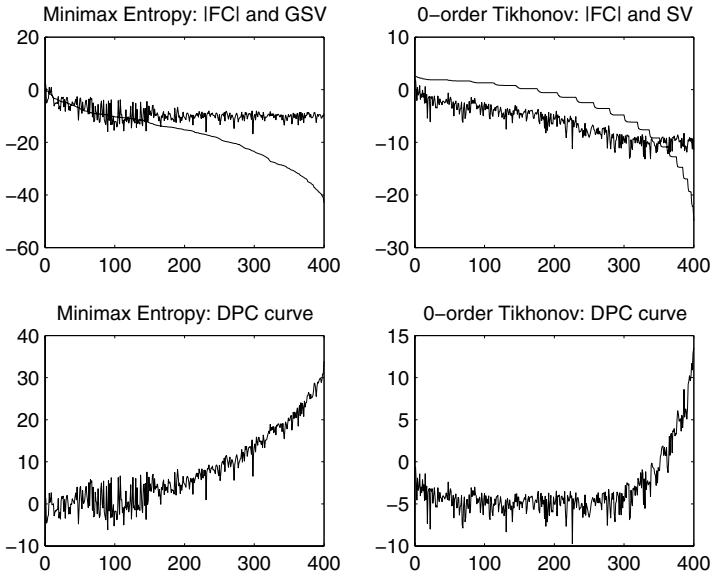

FIG. 2. *Left, upper row:* MME regularization treatment. Plot of the sequence of (natural) logarithms of the Fourier coefficient absolute values and corresponding plot (superimposed) of the logarithms of the generalized singular values (proceeding from largest to smallest). The "noisy" appearing plot is the sequence of Fourier coefficient absolute values. *Right, upper row:* Zero-order Tikhonov regularization. Plot of the sequence of logarithms of the Fourier coefficient absolute values and corresponding plot (superimposed) of the logarithms of the singular values. *Left, lower row:* MME regularization. Plot of the logarithms of the ratios of the Fourier coefficient absolute values to the corresponding generalized singular values. *Right, lower row:* Zero-order Tikhonov regularization. Plot of the logarithms of the ratios of the Fourier coefficient absolute values to the corresponding singular values. The Discrete Picard Condition is apparently violated by the MME regularization (consistent with the stochastic theory specifications for a "desirable" prior covariance matrix), but the Discrete Picard Condition is satisfied in the zero-order Tikhonov regularization.

N is a realization of zero mean white Gaussian (pseudo)noise. In estimating X, we take the inverse of $R'R$ in (1.6) to be the Kronecker product of a particular linear treatment of the nonlinearly coupled equations (2.14) and (2.15) as follows:

Note that the terms enclosed by the large parentheses inside the sums in (2.14) and (2.15) represent a set of scalars that nonlinearly couple those simultaneous equations. Suppose we bring the term $\mathcal{E}[X^{\gamma\sigma}X_{\gamma\sigma}]^{1/2}$ that is outside the braces of those equations, inside the braces so that the denominators no longer have the exponent (1/2). Using matrix rather than tensor notation, the scalars in the large parentheses are now the Frobenius product of $A_i'A_j$ (or $B_i'B_j$) with the unit trace matrix $\mathcal{E}[XX']/\mathcal{E}[\|X\|^2]$ (or $\mathcal{E}[X'X]/\mathcal{E}[\|X\|^2]$). Rather than utilizing a scheme to directly address the nonlinear

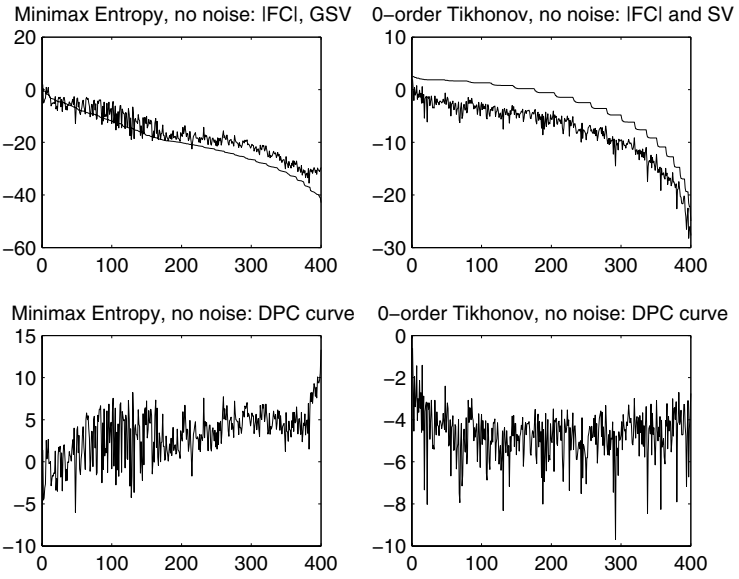

FIG. 3. The plots above are analogous to those of Figure 2, except that they were obtained in the noiseless state $N = 0$. Note that the ordinate axes in the lower row are scaled differently than in Figure 2. Again, the findings are consistent with unbounded $\|RX_{:}\|^2$ for the MME selected prior covariance R and the actual $X_{:}$ that generates data Y (no portions of the lower left plot have a decreasing trend).

problem, we instead choose the scalars as resulting from the Frobenius products of $A_i'A_j$ (and $B_i'B_j$) with unit trace versions of the identity matrix (rather than unit trace versions of $\mathcal{E}[XX']$ and $\mathcal{E}[X'X]$). Thus, these scalars are simply taken to be the traces of $A_i'A_j$ (and $B_i'B_j$). Hence, the two equations are now uncoupled well-posed linear equations that we can solve individually to obtain estimates of $\mathcal{E}[XX']$ and $\mathcal{E}[X'X]$. Formally, the inverse of $\lambda(R'R)$ in (1.6) is taken to be proportional to the resulting estimate of $\mathcal{E}[X'X]\otimes\mathcal{E}[XX']$. The actual computed estimates are obtained using a GSVD as in (3.3). The regularization parameter λ is varied over thirty orders of magnitude. To apply the GSVD, the algorithm addresses (3.8) in the standard linear form:

$$Y_{:} = (I \otimes A + B \otimes I)X_{:} + N_{:}.$$

The optimal regularization parameter value occurs near the L-curve corner. This approach is compared with optimal zero-order Tikhonov regularization ($R = I$), where the computed estimates are obtained analogously using the SVD. All computations were performed using MatLab 7.3 (The MathWorks, Inc., Natick, MA, USA).

In line with the three parts of Corollary 2.10, the following results are obtained.

3.2.1. *The MME choice for the prior covariance of x is such that the Discrete Picard Condition (DPC) is violated, consistent with (1) in Corollary* 2.10. That is, for the penalty operator computed under MME, the results in our example are consistent with

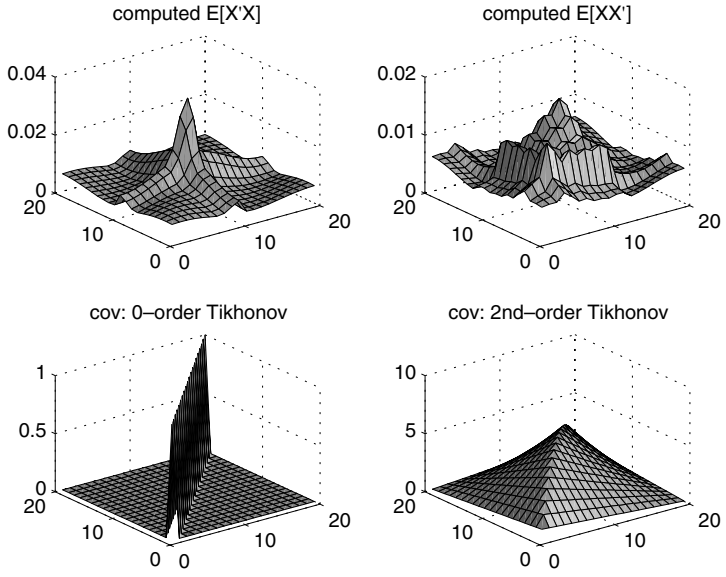

FIG. 4. The MME prior covariance matrix for the example in Section 3.2 is proportional to the Kronecker product of the matrices depicted in the upper row, resulting from a numerical solution of the equations of Corollary 2.4 ((2.14) and (2.15)). Shapes of these matrices are consistent with square-integrablility of the underlying function to be estimated. The identity matrix shape (lower left) is not consistent with such an expectation (the latter is used as the penalty matrix for zero-order Tikhonov regularization). The inverse of the penalty matrix used in second-order Tikhonov regularization is shown for comparison in the lower right.

the presumption that the data y is generated by a true x such that Rx does not have a finite $\mathcal{L}^2$-norm.

According to the usual deterministic interpretation, penalty matrix R will be useful only if the (reversed) generalized singular value sequence (i.e., proceeding from the largest to smallest generalized singular values) decreases "on average" less steeply than the corresponding sequence of absolute values of the Fourier components of the data, for the portion of the sequence that is not noise-dominated [6, 2]. In [6], evidence is presented that this condition seems to be satisfied for usual Tikhonov regularization matrices. The right-sided columns of Figure 2 and Figure 3 indicate that this is the case for our example utilizing zero-order Tikhonov regularization, consistent with the notion that (for the corresponding "nondiscretized" problem) Rx is square-integrable (with R being the identity matrix for zero-order Tikhonov regularization). However, the Discrete Picard Condition appears not to be satisfied by the MME treatment, as is clear from the left-sided columns of Figure 2 and Figure 3. This result is compatible with the notion that (for the corresponding "non-discretized" problem) Rx is not square-integrable for $(R'R)^{-1}$ computed as proportional to the prior covariance in the admissibility constraint approach (where x is the true x generating the data y, rather than the regularized estimate of x). Although this is in conflict with usual deterministic notions regarding useful penalty

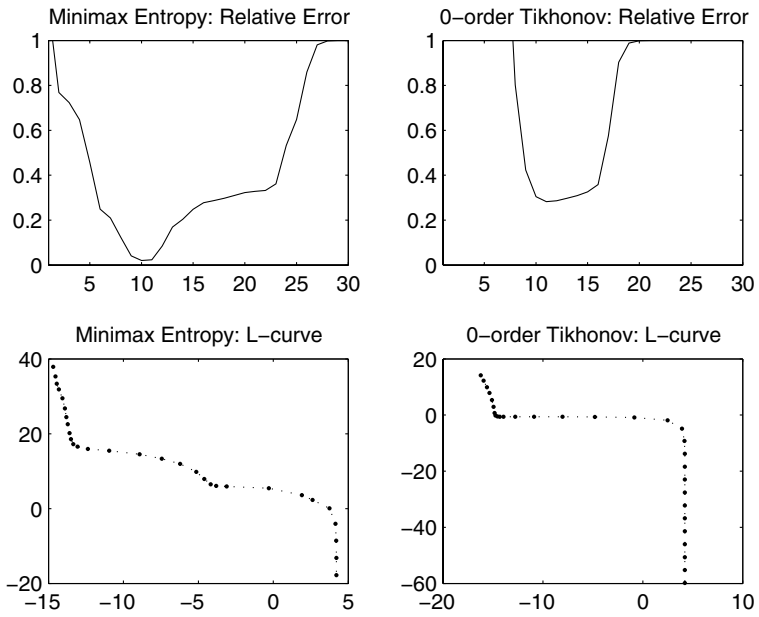

FIG. 5. Respective relative errors (versus regularization parameter choices over 30 orders of magnitude) and L-curves for MME and zero-order Tikhonov regularization. Markedly superior accuracy of the MME treatment compared to zero-order Tikhonov is evident in this example.

operators, this latter finding is, in fact, in accord with the stochastic viewpoint (i.e., the meaning of a covariance operator). Thus, in line with Corollary 2.10, we see that the penalty operator R derived by the MME method is such that Rx is expected *not* to be square-integrable, and, indeed, the terms of the sum corresponding to the right-hand side of (3.7) never appear to have a decreasing trend, according to Figure 2 and Figure 3. As we noted in Section 1.2, this is a feature that an optimal penalty operator should have - a feature not fulfilled by zero-order Tikhonov regularization. Thus, the MME treatment is capable of supplying a deterministic penalty operator consistent with the stochastic viewpoint.

3.2.2. *The prior covariance computed by the MME treatment is compatible with x being square-integrable, consistent with (2) in Corollary* 2.10. The prior covariance used in the admissibility constraint approach is proportional to an estimate of $\mathcal{E}[X'X] \otimes E[XX']$. The two factors in this Kronecker product (computed under the MME treatment) are shown (up to a scalar magnitude) in the upper row of Figure 4, these being of quite unusual shape compared to typical regularization matrices (shown in the lower row of the figure). However, their shapes are consistent with smoothing kernels, such that (for the nondiscretized problem) realizations of x compatible with this covariance would be expected to be square-integrable. This contrasts with the prior covariance associated

with zero-order Tikhonov regularization, which implies that a realization is not expected to be square-integrable.

3.2.3. *The MME treatment for applicable inverse problems has greater accuracy than the 1-parameter Maximum Entropy approach (zero-order Tikhonov regularization) in a suitable signal-to-noise range, consistent with (3) in Corollary* 2.10. In Figure 5, Relative Error is defined as $\|X - X_{\text{est}}\|_{\text{Frob}}/\|X\|_{\text{Frob}}$, where X_{est} is the estimate of X using the MME treatment or zero-order Tikhonov regularization, and $\|\cdot\|_{\text{Frob}}$ is the Frobenius norm. Figure 1 and Figure 5 show that the MME treatment results in much higher resolution and accuracy than zero-order Tikhonov regularization in the example. This advantage decreases as high resolution features of the source are reduced in prominence, and with increases in either noise power or variance of the convolving Gaussians (where all methods lead to poor results), and with decreases in noise power or variance of the convolving Gaussians (where all methods tend to perfect results).

4. Discussion. The stochastic theory applied to ill-posed problems implies that a desirable feature of an imposed penalty operator to be used in the deterministic theory is that the given data are inconsistent with having been generated by a (true) solution that is in the domain of this penalty operator, despite the necessity that the calculated solution estimate must ultimately be in the domain of this operator (Theorem 1.1). As discussed in Section 3, evidence for the true solution's inclusion or exclusion from the domain of a proposed penalty operator can be accessed by application of the Dicrete Picard Condition. From the standpoint of deterministic regularization, the desirability of the true solution's exclusion from the domain of the selected penalty operator must be considered surprising, even if it is not immediately clear how such a penalty operator is to be chosen in general. However, for a particular class of ill-posed problems, it is indeed possible to compute such an operator from the data and forward operator, as shown in Section 2. This raises the question of exploring how to do this in more general circumstances. Such a program would have to find some way to choose between members of the set of covariance operators whose inverses generate (with the given data and forward operator) DPC curves that on average do not decrease. After discretization of the problem, one approach to such a question would ask for the associated prior probability distribution of maximum entropy that satisfies such a constraint. Such a program, if proved practical and useful, might have much to say about the regularization of general inverse problems, rather than only the restricted class considered in this paper.

In deriving the requisite covariance operator for the subset of ill-posed problems considered in this paper, a Minmax Entropy conception arises in a natural way as a means to select between different sets of imposed constraints, based on the desire for imposing least bias. That is, the constraint set should be such that

- replacing an imposed component constraint with an alternative constraint should not increase the entropy - one wants *maximum* entropy in this sense, but
- using the fewest constraints to complete the identification of a prior is also desirable - the constraint set should have *minimum* entropy (information content) in this sense.

Although we are unaware of prior use of a Minimax Entropy principle in the context of the solution of ill-posed problems, a somewhat analogous motivation has been applied in building statistical models for signals, particularly for the case of images. This entails identification of a favored probability distribution associated with a class of images of interest (e.g., images having a particular 'texture'). Thus, one uses a maximum entropy principle to bind a necessarily incomplete set of observed statistics into a complete probability distribution, but subsequently selects the distribution of minimum entropy among the maximum entropy distributions that are individually constructed from different plausible subsets of observed feature statistics [15]. Though it cannot address the setting of a transformed image with no given information concerning statistics of the underlying image (inverse problems), a notable feature that this work has in common with the present work is the use of a single image to construct a plausible probability density (as we use the given data set to fashion the prior from a smaller than usual set of imposed constraints). In contrast, classical regularization procedures use the given data of the problem to identify only a single regularization *parameter*, while the penalty *operator* is imposed 'arbitrarily' and without reference to the data. The use of one rather than many computed regularization parameters appears to be another 'arbitrary' choice of current regularization methods. The method of this paper, of course, uses the data to fashion both the regularization parameter and the penalty operator (for the subclass of applicable ill-posed problems).

Since Sylvester equations appear prominently in Sections 2 and 3, it may be useful to observe that these have arisen elsewhere in regularization theory. For example, Bouhamidi and Jbilou [1] have derived a 'Sylvester Tikhonov' regularization methodology in the context of image restoration. This addresses the case of ill-posed problems where the forward operator can be expressed as a Kronecker product (i.e., is separable). The authors show that a Tikhonov regularization treatment can then be expressed in terms of a generalized Sylvester equation which (given a selected regularization parameter) in this case defines a well-posed problem. However, the problem addressed there is very different from that treated in the present paper.

References

[1] A. Bouhamidi and K. Jbilou, *Sylvester Tikhonov-regularization methods in image restoration*, J. Comput. Appl. Math. **206** (2007), 86-98. MR2333837

[2] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996. MR1408680 (97k:65145)

[3] J. N. Franklin, *Well-posed stochatic extensions of ill-posed problems*, J. Math. Anal. Appl. **31** (1970), 682–716. MR0267654 (42:2556)

[4] Z. Gajic and M. T. J. Quereshi, *Lyapunov Matrix Equation in System Stability and Control*, Academic Press, San Diego, 1995. MR1343974 (96g:93001)

[5] M. Hanke, *Limitations of the L-curve method in ill-posed problems*, BIT **36** (1996), 287-301. MR1432249 (97j:65098)

[6] P. C. Hansen, *The discrete Picard condition for discrete ill-posed problems*, BIT **30** (1990), 658–672. MR1082808 (91m:65119)

[7] P. C. Hansen, *Analysis of discrete ill-posed problems by means of the L-curve*, SIAM Review **34** (1992), 561-580. MR1193012 (93k:65035)

[8] E. T. Jaynes, *Probability Theory*, Cambridge University Press, Cambridge, United Kingdom, 2003. MR1992316 (2004g:62006)

[9] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1984. MR836136 (87d:60001)
[10] A. Tarantola, *Inverse Problem Theory*, SIAM, Philadelphia, 2005. MR2130010 (2007b:62011)
[11] C. F. van Loan, *Generalizing the singular value decomposition*, SIAM J. Numer. Anal. **11** (1976), 76-83. MR0411152 (53:14891)
[12] J. M. Varah, *Pitfalls in the numerical solution of linear ill-posed problems*, SIAM J. Sci. Stat. Comput. **4** (1983), 164-176. MR697171 (84g:65052)
[13] C. Vogel, *Non-convergence of the L-curve regularization parameter selection method*, Inverse Problems **12**, (1996) 535-547. MR1402108 (97k:65149)
[14] G. Wahba, *Practical approximate solutions to linear operator equations when the data are noisy*, SIAM J. Numer. Anal. **14** (1977), 651-667. MR0471299 (57:11036)
[15] S. C. Zhu, Y. N. Wu, and D. Mumford, *Minimax entropy principle and its application to texture modeling*, Neural Computation **9** (1997), 1627-1660.