Hidden Relationships: Bayesian Estimation with Partial Knowledge

Tomer Michaeli and Yonina C. Eldar, *Senior Member, IEEE*

***Abstract*—We address the problem of Bayesian estimation where the statistical relation between the signal and measurements is only partially known. We propose modeling partial Bayesian knowledge by using an auxiliary random vector called instrument. The statistical relations between the instrument and the signal, and between the instrument and the measurements, are known. However, the joint probability function of the signal and measurements is unknown. Two types of statistical relations are considered, corresponding to second-order moment and complete distribution function knowledge. We propose two approaches for estimation in partial knowledge scenarios. The first is based on replacing the orthogonality principle by an oblique counterpart, and is proven to coincide with the method of instrumental variables from statistics, although developed in a different context. The second is based on a worst-case design strategy and is shown to be advantageous in many aspects. We provide a thorough analysis showing in which situations each of the methods is preferable and propose a non-parametric method for approximating the estimators from a set of examples. Finally, we demonstrate our approach in the context of enhancement of facial images that have undergone unknown degradation and image zooming.**

***Index Terms*—Bayesian estimation, minimax regret, partial knowledge, instrumental variables, nonparametric regression.**

I. Introduction

A common problem in signal processing is that of estimating an unknown random quantity $\boldsymbol{x}$ from a set of noisy measurements $\boldsymbol{y}$. Image denoising and debluring [1], speech enhancement [2], and target tracking [3], are a few examples. The Bayesian framework requires knowledge of the prior distribution of the signal $\boldsymbol{x}$ to be estimated, as well as the conditional probability of the measurements $\boldsymbol{y}$ given $\boldsymbol{x}$ [4]. The former can usually be learned from a set of examples $\{\boldsymbol{x}_i\}$ of "clean" signals. The latter, on the other hand, necessitates either a paired set of examples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ of signals and measurements, or knowledge of the degradation mechanism that yielded the measurements (*e.g.,* additive white Gaussian noise). In many applications, neither assumption is realistic.

In speech enhancement, for example, poor room acoustics and background noise, such as other speakers, are part of the degradation that needs to be overcome [5], [2]. These undesired effects typically vary in time and are very hard to model statistically [6]. Furthermore, no paired examples of clean and degraded signals are available in these scenarios.

Another example is that of enhancement of facial images taken with a low-grade camera (*e.g.,* a web-cam or a cellular-phone camera). The distortion in this case includes blur due to the lens, the nonlinear response of the CCD sensor [7], and non-additive noise [8]. These processes vary with lighting conditions, distance from the camera, etc., and are therefore hard to model. Moreover, obtaining a paired set of examples of clean and degraded images requires a complicated experimental setup consisting of a high-quality camera co-calibrated with the low-grade camera at hand.

A common practice in such scenarios is to resort to simplified model assumptions, such as Gaussian blur and additive white noise in image restoration (see *e.g.,* [1], [9], [10]), and stationary background noise in speech enhancement tasks [2]. These assumptions simplify the treatment but are often far from loyal to the true physical setting. More complicated likelihood models can be treated via approaches such as approximate Bayesian computation [11]. These methods are useful when evaluation of the likelihood is computationally prohibitive. However, they rely on the assumption that data can be simulated from the likelihood, which is not the case if one does not have access to paired examples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ of clean signals and corrupted measurements.

An alternative approach is to make use of many examples of degraded signals $\{\boldsymbol{y}_i\}$, which are typically easy to collect, and only a small number of paired examples $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$, which are hard to obtain. This strategy lies at the heart of the field of semi-supervised learning in general [12] and semi-supervised regression [13] in particular. However, there are situations in which it is highly desired to avoid the need for *any* paired example of signal and measurement.

Bayesian estimation cannot be carried out without knowledge of the joint distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$. Nevertheless, in many applications there is partial knowledge of this statistical relation. Specifically, we may know the joint probability function of $\boldsymbol{x}$ and some auxiliary random vector $\boldsymbol{z}$ as well as that of $\boldsymbol{y}$ and $\boldsymbol{z}$. For instance, to enhance a video sequence $\boldsymbol{y}$ of a speaker without knowing the type of degradation it has undergone, one may use the audio $\boldsymbol{z}$ associated with it. Clearly, we can collect paired examples $\{\boldsymbol{y}_i, \boldsymbol{z}_i\}$ of the noisy video and its associated audio (taken with the given low-quality camcorder), as well as paired examples $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}$ of clean video sequences with their audio (taken from a high-grade video camera). These two sets are unpaired, namely they correspond to video sequences of different scenes. Consequently, they can be used to learn the densities $f_{XZ}(\boldsymbol{x}, \boldsymbol{z})$ and $f_{YZ}(\boldsymbol{y}, \boldsymbol{z})$ but

This work was supported in part by the Israel Science Foundation under Grant no. 170/10 and by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (contract no. 216715).

The authors are with the department of electrical engineering, Technion–Israel Institute of Technology, Haifa, Israel (phone: +972-4-8294682, fax: +972-4-8295757, e-mail: tomermic@tx.technion.ac.il, yonina@ee.technion.ac.il).

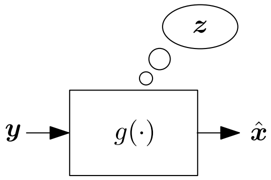


Fig. 1: An estimator $\hat{\boldsymbol{x}} = g(\boldsymbol{y})$ in a partial knowledge setting.

are generally insufficient to determine $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$.

During the last two decades, various approaches have been proposed to enhancing audio or video based on joint audio-visual measurements (see *e.g.,* [14], [15], [16]). There is a fundamental difference, though, from our problem setting. For example, in the scenario described above, the input to the estimator is only the noisy video sequence $\boldsymbol{y}$, *without* the associated audio. The audio data comes into play only in the training sets $\{\boldsymbol{y}_i, \boldsymbol{z}_i\}$ and $\{\boldsymbol{x}_i, \boldsymbol{z}_i\}$ but does not constitute part of the measurements, as schematically shown in Fig. 1. The interesting question that arises, then, is whether audio can aid in enhancing a silent video sequence (or vice versa), namely one that was recorded without sound.

In this paper, we study two partial-knowledge models, which differ in the type of statistical relation between the instrument and the signal/measurements that is assumed to be available. In the first, only the joint second order statistics of $\boldsymbol{x}$ and $\boldsymbol{z}$, as well as of $\boldsymbol{y}$ and $\boldsymbol{z}$ are known. In the second, the entire density functions $f_{XZ}(\boldsymbol{x}, \boldsymbol{z})$ and $f_{YZ}(\boldsymbol{y}, \boldsymbol{z})$ are available. In both scenarios, however, $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$ is unknown.

We propose two strategies for treating Bayesian estimation with partial statistical knowledge. Our first approach is to replace the orthogonality requirement, which characterizes the minimal mean-squared error (MSE) estimator, by an oblique counterpart. Specifically, we seek an estimator $\hat{\boldsymbol{x}} = g(\boldsymbol{y})$ whose error is orthogonal to the instrument $\boldsymbol{z}$ rather than to the measurements $\boldsymbol{y}$. As we show, the resulting estimators coincide with those encountered in instrumental variable regression [17] from the fields of statistics and econometrics, which explains our choice of terminology. The second strategy we consider is based on a worst-case design approach. Here, the estimator is designed to yield the best worst-case performance (over the set of density functions $f_{XYZ}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})$ consistent with the available partial knowledge). We propose explicit ways of approximating both solutions from sets of examples.

We show that each of the proposed methods is optimal in different settings. The performance of the oblique approach, however, can become arbitrarily poor as the statistical dependency between the instrument and measurements weakens. In contrast, the estimation error of the worst-case strategy is guaranteed to be bounded. This property is of great value in practical scenarios, however it comes at the cost of a modest performance at a rather wide variety of settings. Nevertheless, since in typical applications the instrument is often weak, our worst-case design approach is commonly preferable.

We demonstrate the usefulness of our approach in two image processing applications. The first is enhancement of facial images that have undergone unknown degradation. This scenario is highly relevant to face recognition systems working in uncontrolled conditions [18]. There, no paired examples of clean and degraded images can be obtained, thus calling for a partial knowledge treatment. The second application is image zooming. Specifically, many recent works treat this problem by learning the relation between image patches and their down-scaled versions [19]. However, this strategy becomes problematic when the original image is very small, since there are very few training patches left after down-sampling the image. Using our approach, we show how this limitation can be overcome.

The paper is organized as follows. In Section II we provide a concise mathematical formulation of the partial-knowledge Bayesian estimation problem. In Sections III and IV we develop estimators for the second-order moment model which rely on the obliqueness principle and the worst-case design strategy, respectively. We show the relation of these estimators to instrumental variable regression in linear models, and determine in which cases each is preferable. Sections V and VI treat the density-function model via obliqueness and worst-case design respectively. We also discuss the relation of our problem to nonparametric instrumental variable regression in nonlinear models [20] and provide best-case and worst-case analyses for each of the approaches. Section VII is devoted to a quantitative simulation study, which unveils the strengths and weaknesses of the different methods in a wide variety of situations. Finally, in Section VIII, we demonstrate our technique in the context of enhancement of facial images that have undergone unknown distortion, and in Section IX we develop an image zooming algorithm based on our approach.

II. Problem Formulation

We denote random variables (RVs) by capital letters (*e.g.,* X, Y, Z) and the values that they take by bold lower-case letters (*e.g.,* $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$). The pseudo-inverse of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^{\dagger}$. The mean vector and covariance matrix of an RV X are defined as $\boldsymbol{\mu}_X = \mathbb{E}[X]$ and $\boldsymbol{\Gamma}_{XX} = \mathbb{C}\text{ov}[X] = \mathbb{E}[(X - \mu_X)(X - \mu_X)^T]$ respectively. Similarly, the cross-covariance matrix of two RVs X and Y is denoted by $\boldsymbol{\Gamma}_{XY} = \mathbb{C}\text{ov}[X, Y] = \mathbb{E}[(X - \boldsymbol{\mu}_X)(Y - \boldsymbol{\mu}_Y)^T]$. In our setting, X is the quantity to be estimated, also termed “signal”, Y is the measurements, and Z is an auxiliary RV, which we call “instrument”. The RVs X, Y, and Z take values in $\mathbb{R}^M$, $\mathbb{R}^N$, and $\mathbb{R}^Q$, respectively. We denote by $\mathcal{Y}_{\text{L}}$ and $\mathcal{Z}_{\text{L}}$ the sets of all RVs that are affine functions of Y and Z respectively. Specifically, every RV $\hat{X} \in \mathcal{Y}_{\text{L}}$ can be expressed as $\hat{X} = \boldsymbol{A}Y + \boldsymbol{b}$ for some matrix $\boldsymbol{A}$ and vector $\boldsymbol{b}$. Similarly, $\mathcal{Y}$ denotes the set of RVs that are arbitrary (Borel measurable) functions of Y.

We assume that the joint density function $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$ of the signal and measurements is unknown. Nevertheless, we have some knowledge regarding the statistical relation between X and Z and between Y and Z. Specifically, we consider the following two types of partial knowledge models.

M1: Only the first- and second-order moments of $(X^T, Z^T)^T$ and $(Y^T, Z^T)^T$ are known, as depicted in Fig. 2(a). Specifically, we know the mean vectors $\boldsymbol{\mu}_X$, $\boldsymbol{\mu}_Y$, $\boldsymbol{\mu}_Z$, as well as the covariance matrices $\boldsymbol{\Gamma}_{XX}$, $\boldsymbol{\Gamma}_{YY}$, $\boldsymbol{\Gamma}_{ZZ}$, $\boldsymbol{\Gamma}_{XZ}$, $\boldsymbol{\Gamma}_{YZ}$, but we do not know $\boldsymbol{\Gamma}_{XY}$.

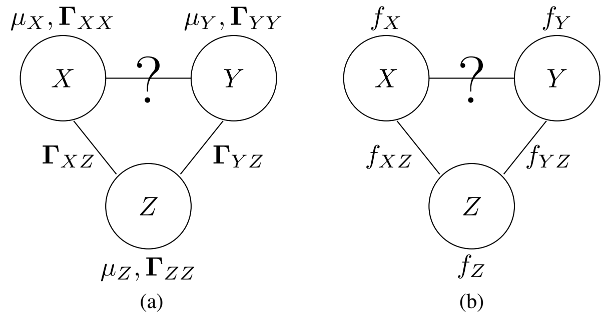


Fig. 2: Two partial knowledge scenarios. (a) Knowledge of moments up to second order. (b) Knowledge of joint density functions.

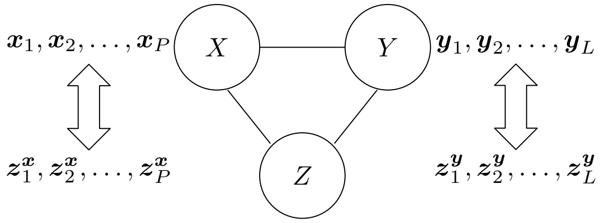

$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_P$
X
Y
$\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_L$
$\boldsymbol{z}_1^{\boldsymbol{x}}, \boldsymbol{z}_2^{\boldsymbol{x}}, \ldots, \boldsymbol{z}_P^{\boldsymbol{x}}$
Z
$\boldsymbol{z}_1^{\boldsymbol{y}}, \boldsymbol{z}_2^{\boldsymbol{y}}, \ldots, \boldsymbol{z}_L^{\boldsymbol{y}}$

Fig. 3: The two unpaired sets of examples $\{\boldsymbol{x}_i, \boldsymbol{z}_i^{\boldsymbol{x}}\}_{i=1}^P$ and $\{\boldsymbol{y}_i, \boldsymbol{z}_i^{\boldsymbol{y}}\}_{i=1}^L$ can be used to learn one of the models in Fig. 2.

M2: The joint density functions $f_{XZ}(\boldsymbol{x}, \boldsymbol{z})$ and $f_{YZ}(\boldsymbol{y}, \boldsymbol{z})$ are known, as schematically shown in Fig. 2(b). This, of course, implies that the marginal densities $f_X(\boldsymbol{x})$, $f_Y(\boldsymbol{y})$ and $f_Z(\boldsymbol{z})$ are known as well.

In practice, both types of information may be unavailable in closed form. Instead, we may only have access to two sets of paired examples $\{\boldsymbol{x}_i, \boldsymbol{z}_i^{\boldsymbol{x}}\}_{i=1}^P$ and $\{\boldsymbol{y}_i, \boldsymbol{z}_i^{\boldsymbol{y}}\}_{i=1}^L$, drawn independently from the densities $f_{XZ}(\boldsymbol{x}, \boldsymbol{z})$ and $f_{YZ}(\boldsymbol{y}, \boldsymbol{z})$ respectively, as shown in Fig. 3. These training sets can be used to estimate the relevant moments and also the entire density functions. The choice of which of the partial-knowledge models to use, then, depends on the cardinalities of the training sets. If the number of training examples is small, then we may only be able to estimate the second-order moments to reasonable accuracy. On the other hand, for large sets, the density functions can be estimated accurately *e.g.,* by nonparametric density estimation methods [21], making the second model relevant.

Since the statistical relation between the signal and the instrument is known, one could theoretically estimate $\boldsymbol{x}$ based on a realization $\boldsymbol{z}$ of Z. However, in our setting we do not observe any realization of the instrument. Thus, the only way Z can be of help is by employing our knowledge of its statistical relation with X and with Y, in order to estimate $\boldsymbol{x}$ from the realization $\boldsymbol{y}$ of Y. In other words, there is a certain symmetry between the instrument Z and the measurements Y, as shown in Table I. The RV Y is measured, but its statistical relation with X is unknown. In contrast, the relation between Z and X is known, but Z is not measured.

TABLE I: Measurement vs. instrument.

	Measured	Known statistical relation with X
Y	✓	×
Z	×	✓

A. Objectives

Ideally, we would like to design an estimator $\hat{\boldsymbol{x}} = g(\boldsymbol{y})$ of the signal $\boldsymbol{x}$ based on the measurements $\boldsymbol{y}$, such that the MSE

$$\text{MSE} = \mathbb{E}\left[\left\|X - \hat{X}\right\|^2\right] \tag{1}$$

is minimized. Unfortunately, the MSE depends on $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$ (since $\hat{X}$ is a function of Y), which is unknown, so that it cannot be computed in our setting.

Had $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$ been known, it would be possible to compute the minimum MSE (MMSE) estimator

$$g(\boldsymbol{y}) = \mathbb{E}[X|Y = \boldsymbol{y}], \tag{2}$$

which depends on $f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) = f_{XY}(\boldsymbol{x}, \boldsymbol{y})/f_Y(\boldsymbol{y})$. Therefore, in the scenario of model M2, our goal is to design an estimator which comes as close as possible to the MMSE method, in some sense.

Under model M1, even if $\mathbf{\Gamma}_{XY}$ was available, we still could not have computed the MMSE estimator (2), as it requires knowledge of the entire density function $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$. Thus, in this setting our goal is to design an estimator that comes as close as possible to the estimator that is optimal among all methods that have access only to the joint first- and second-order moments of X and Y.

A common technique for estimating $\boldsymbol{x}$ from $\boldsymbol{y}$, which relies only on first- and second-order statistics, is the linear MMSE (LMMSE) estimator, given by [4]

$$\hat{X}_{\text{LMMSE}} = \mathbf{\Gamma}_{XY}\mathbf{\Gamma}_{YY}^{\dagger}(Y - \boldsymbol{\mu}_Y) + \boldsymbol{\mu}_X. \tag{3}$$

It is important to note, however, that the fact that the LMMSE estimate happens to be a function of the first- and second order moments, still does not imply that it is optimal in any sense among estimators that solely depend on these quantities. The following theorem shows that the LMMSE estimate is indeed optimal in the sense that its worst-case MSE over all joint distributions $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$ with the given second-order moments, is minimal.

Theorem 1: *The LMMSE estimator* (3) *is the solution to*

$$\min_{\hat{X} \in \mathcal{Y}} \max_{f_{XY} \in \mathcal{A}} \mathbb{E}\left[\left\|X - \hat{X}\right\|^2\right], \tag{4}$$

where $\mathcal{A}$ is the set of densities f_{XY} satisfying $\mathbb{E}[X] = \boldsymbol{\mu}_X$, $\mathbb{E}[Y] = \boldsymbol{\mu}_Y$, $\mathbb{C}\text{ov}[X, Y] = \mathbf{\Gamma}_{XY}$, $\mathbb{C}\text{ov}[X] = \mathbf{\Gamma}_{XX}$ and $\mathbb{C}\text{ov}[Y] = \mathbf{\Gamma}_{YY}$.

Proof: See Appendix A. ■

As a consequence of Theorem 1, in the setting of model M1, our goal is to construct a *linear estimator* whose performance comes close to that of the LMMSE method.

In the next sections, we propose two strategies to estimation in the partial knowledge models M1 and M2, which are based on an obliqueness principle and a worst-case design strategy.

III. Estimation with Moment Knowledge via the Obliqueness Principle

A. Estimation via Obliqueness

We begin by assuming model M1 and rely on an obliqueness principle. To develop our approach, we note that if $\boldsymbol{\Gamma}_{XY}$ were known, then it would have been possible to compute the LMMSE estimator (3). This solution can be interpreted as the orthogonal projection of X onto the set $\mathcal{Y}_{\text{L}}$, which implies that its error $X - \hat{X}_{\text{LMMSE}}$ is uncorrelated with Y. This principle, which is known as the orthogonality criterion, implies that $\hat{X}_{\text{LMMSE}}$ is the (almost surely) unique affine method whose mean and cross-covariance with Y coincide with those of X and Y, namely

$$\boldsymbol{\mu}_{\hat{X}} = \boldsymbol{\mu}_X, \quad \boldsymbol{\Gamma}_{\hat{X}Y} = \boldsymbol{\Gamma}_{XY}. \tag{5}$$

In our setting, we do not know $\boldsymbol{\Gamma}_{XY}$ and thus cannot compute $\hat{X}_{\text{LMMSE}}$. Instead, relying on our knowledge of $\boldsymbol{\mu}_X$ and $\boldsymbol{\Gamma}_{XZ}$, our approach here is to design an affine estimator

$$\hat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}, \tag{6}$$

whose error $X - \hat{X}$ is uncorrelated with Z rather than with Y. In other words, we require that

$$\boldsymbol{\mu}_{\hat{X}} = \boldsymbol{\mu}_X, \quad \boldsymbol{\Gamma}_{\hat{X}Z} = \boldsymbol{\Gamma}_{XZ} \tag{7}$$

in order to determine $\boldsymbol{A}$ and $\boldsymbol{b}$ of (6). We term this requirement the *obliqueness principle* as it results in an estimate $\hat{X}$ that is the oblique projection [22] of X onto $\mathcal{Y}_{\text{L}}$ perpendicular to $\mathcal{Z}_{\text{L}}$. Intuitively, this approach will lead to satisfactory results if Y and Z are "close" in some sense. In Sections III-C and III-D, we quantify this observation in detail.

Taking the expectation of both sides of (6), and equating $\boldsymbol{\mu}_{\hat{X}} = \boldsymbol{\mu}_X$, we find that $\boldsymbol{A}$ and $\boldsymbol{b}$ must satisfy

$$\boldsymbol{\mu}_X = \boldsymbol{A}\boldsymbol{\mu}_Y + \boldsymbol{b}. \tag{8}$$

Similarly, (6) implies that $\boldsymbol{\Gamma}_{\hat{X}Z}$ is given by

$$\boldsymbol{\Gamma}_{\hat{X}Z} = \mathbb{E}\big[(\boldsymbol{A}Y + \boldsymbol{b} - (\boldsymbol{A}\boldsymbol{\mu}_Y + \boldsymbol{b}))(Z - \boldsymbol{\mu}_Z)^T\big] = \boldsymbol{A}\boldsymbol{\Gamma}_{YZ}, \tag{9}$$

where we used (8).

If $Q = N$ and $\boldsymbol{\Gamma}_{YZ}$ is invertible, then (9) implies that $\boldsymbol{A} = \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{YZ}^{-1}$. The vector $\boldsymbol{b}$ can then be computed from (8), resulting in $\boldsymbol{b} = \boldsymbol{\mu}_X - \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{YZ}^{-1}\boldsymbol{\mu}_Y$.

If $Q > N$ then (9) is over-determined. In this case, a solution will typically not exist. To overcome this obstacle, we may seek an affine estimator which comes closest to fulfilling (7). This can be done by minimizing the Frobenius norm $\|\boldsymbol{\Gamma}_{XZ} - \boldsymbol{\Gamma}_{\hat{X}Z}\|_{\text{F}}^2 = \|\boldsymbol{\Gamma}_{XZ} - \boldsymbol{A}\boldsymbol{\Gamma}_{YZ}\|_{\text{F}}^2$. Assuming that $\boldsymbol{\Gamma}_{YZ}$ has full column rank, the solution is given by $\boldsymbol{A} = \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{YZ}^{\dagger}$, resulting in $\boldsymbol{b} = \boldsymbol{\mu}_X - \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{YZ}^{\dagger}\boldsymbol{\mu}_Y$.

When $Q < N$, there are typically infinitely many matrices $\boldsymbol{A}$ satisfying (9). In this case, our knowledge is insufficient for determining a unique oblique linear estimator. One of the solutions is given by $\boldsymbol{A} = \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{YZ}^{\dagger}$. Among all solutions, this matrix has the minimal Frobenius norm. Note, however, that there is no reason to believe that this solution is preferable to others in any sense.

To conclude, assuming that $\boldsymbol{\Gamma}_{YZ}$ has full column rank, the obliqueness requirement leads to the estimate

$$\hat{X}_{\text{OB}}^{\text{M1}} = \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{YZ}^{\dagger}(Y - \boldsymbol{\mu}_Y) + \boldsymbol{\mu}_X. \tag{10}$$

This estimator can be approximated from sets of examples of the type shown in Fig. 3, by replacing $\boldsymbol{\Gamma}_{XZ}$, $\boldsymbol{\Gamma}_{YZ}$, $\boldsymbol{\mu}_Y$ and $\boldsymbol{\mu}_X$ by their associated sample-mean and sample-covariance.

Interestingly, (10) possesses the same structure encountered in the method of linear regression with instrumental variables. This fact can be used to obtain further insight into the obliqueness approach, as we discuss next.

B. Relation to Regression with Instrumental Variables

Assume that $\boldsymbol{x}$ is approximately linearly related to $\boldsymbol{y}$ as

$$\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b} + \boldsymbol{v}, \tag{11}$$

where $\boldsymbol{v}$ is an error term, which is the realization of some zero-mean RV V. To determine $\boldsymbol{A}$ and $\boldsymbol{b}$ based on a set of realizations $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$ drawn independently from the model (11), one can use ordinary least-squares (OLS) regression [17]

$$\hat{\boldsymbol{A}}_{\text{OLS}} = \hat{\boldsymbol{\Gamma}}_{XY}\hat{\boldsymbol{\Gamma}}_{YY}^{-1}, \tag{12}$$

$$\hat{\boldsymbol{b}}_{\text{OLS}} = \hat{\boldsymbol{\mu}}_X - \hat{\boldsymbol{\Gamma}}_{XY}\hat{\boldsymbol{\Gamma}}_{YY}^{-1}\hat{\boldsymbol{\mu}}_Y. \tag{13}$$

Here $\hat{\boldsymbol{\Gamma}}_{XY}$ is the sample cross-covariance of X and Y, $\hat{\boldsymbol{\Gamma}}_{YY}$ is the sample covariance of Y, and $\hat{\boldsymbol{\mu}}_X$ and $\hat{\boldsymbol{\mu}}_Y$ are the sample means of X and Y respectively. If the error V is uncorrelated with Y, and the covariance matrix $\boldsymbol{\Gamma}_{YY}$ is nonsingular, then $\hat{\boldsymbol{A}}_{\text{OLS}}$ and $\hat{\boldsymbol{b}}_{\text{OLS}}$ are known to constitute consistent estimates of $\boldsymbol{A}$ and $\boldsymbol{b}$ respectively[1] [17].

In many situations in statistics, the error V is correlated with Y. In these settings, $\hat{\boldsymbol{A}}_{\text{OLS}}$ and $\hat{\boldsymbol{b}}_{\text{OLS}}$ will not converge to $\boldsymbol{A}$ and $\boldsymbol{b}$. One approach to overcome this difficulty, is to employ an auxiliary RV, Z, referred to as an *instrument*, which is known to be correlated with Y but not with V. Assuming that $Q \geq N$, estimates of $\boldsymbol{A}$ and $\boldsymbol{b}$ can be constructed based on two sets of examples $\{\boldsymbol{x}_n, \boldsymbol{z}_n^{\boldsymbol{x}}\}_{n=1}^P$ and $\{\boldsymbol{y}_n, \boldsymbol{z}_n^{\boldsymbol{y}}\}_{n=1}^L$ drawn from the densities $f_{XZ}(\boldsymbol{x}, \boldsymbol{z})$ and $f_{YZ}(\boldsymbol{y}, \boldsymbol{z})$ respectively. This method is known as instrumental variable regression, and is given by [17]

$$\hat{\boldsymbol{A}}_{\text{IV}} = \hat{\boldsymbol{\Gamma}}_{XZ}\hat{\boldsymbol{\Gamma}}_{YZ}^{\dagger} \tag{14}$$

$$\hat{\boldsymbol{b}}_{\text{IV}} = \hat{\boldsymbol{\mu}}_X - \hat{\boldsymbol{\Gamma}}_{XZ}\hat{\boldsymbol{\Gamma}}_{YZ}^{-1}\hat{\boldsymbol{\mu}}_Y, \tag{15}$$

where $\hat{\boldsymbol{\Gamma}}_{XZ}$, $\hat{\boldsymbol{\Gamma}}_{YZ}$, $\hat{\boldsymbol{\mu}_X}$ and $\hat{\boldsymbol{\mu}}_Y$ are the associated sample covariances and sample means. It can be shown that $\hat{\boldsymbol{A}}_{\text{IV}}$ and $\hat{\boldsymbol{b}}_{\text{IV}}$ tend to $\boldsymbol{A}$ and $\boldsymbol{b}$ in probability as P and L tend to infinity. If $Q < N$ then $\boldsymbol{A}$ and $\boldsymbol{b}$ are unidentifiable [17].

The weak law of large numbers implies that, as the sample sizes increase, $\hat{\boldsymbol{A}}_{\text{IV}}$ and $\hat{\boldsymbol{b}}_{\text{IV}}$ tend to $\boldsymbol{A}$ and $\boldsymbol{b}$ of (10). Therefore, the oblique estimator can also be interpreted as emerging from the assumption that X and Y are related through the linear model (11) with a noise component V uncorrelated with Z. Specifically, once $\boldsymbol{A}$ and $\boldsymbol{b}$ are estimated in this setting, we construct the estimate $\hat{\boldsymbol{x}} = \hat{\boldsymbol{A}}_{\text{IV}}\boldsymbol{y} + \boldsymbol{b}_{\text{IV}}$ by applying the model (11) on $\boldsymbol{y}$, while disregarding $\boldsymbol{v}$. The resulting estimate coincides with our oblique method (10).

[1] Namely, $\hat{\boldsymbol{A}}_{\text{OLS}}$ and $\hat{\boldsymbol{b}}_{\text{OLS}}$ tend to $\boldsymbol{A}$ and $\boldsymbol{b}$ respectively in probability as the sample size N tends to infinity.

C. Best Case Analysis

We now address the question under which situations the oblique method is optimal.

The obliqueness approach relies on the demand that the estimation error be uncorrelated with the instrument Z rather than with the measurements Y. Therefore, in cases where the former implies the latter, this strategy coincides with the LMMSE estimate (3). Comparing (3) and (10), it can be seen that the oblique and LMMSE estimators coincide if

$$\mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{YZ}^{\dagger} = \mathbf{\Gamma}_{XY}\mathbf{\Gamma}_{YY}^{\dagger}. \tag{16}$$

To gain insight into when this occurs, it is instructive to examine the case in which the RVs X, Y and Z are jointly Gaussian. In this situation, the LMMSE method (3) is also optimal among all nonlinear techniques, namely it coincides with the MMSE estimate. For simplicity, we focus on the case in which the dimensions of the measurement and the instrument vectors are equal.

Theorem 2: *Suppose that X, Y and Z are jointly Gaussian RVs that take values in $\mathbb{R}^M$, $\mathbb{R}^N$ and $\mathbb{R}^N$ respectively. Let $A \triangleq (Y^T, Z^T)^T$ and assume that the matrices $\mathbf{\Gamma}_{AA}$, $\mathbf{\Gamma}_{YY}$, $\mathbf{\Gamma}_{YZ}$ and $\mathbf{\Gamma}_{ZZ} - \mathbf{\Gamma}_{ZY}\mathbf{\Gamma}_{YY}^{-1}\mathbf{\Gamma}_{YZ}$ are invertible. Then the oblique estimate* (10) *coincides with the MMSE estimate of X given Y if and only if*

$$f_{X|YZ}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z}) = f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}). \tag{17}$$

Proof: See Appendix B. ■

Theorem 2 states that in the Gaussian setting, the oblique method is optimal if and only if X and Z are independent given Y. To understand this condition, consider the hypothetical scenario in which all cells of Table I are checked. Specifically, assume that we knew the statistical relation between X and Y and we could also measure Z. If in this situation, the MMSE estimate of $\boldsymbol{x}$ given $\boldsymbol{y}$ and $\boldsymbol{z}$ would be only a function of $\boldsymbol{y}$, then the obliqueness approach is optimal.

Unfortunately, in practice, the instrument may carry information on the signal that is not present in the measurements. The larger the amount of this information, the poorer the performance of the oblique estimator will be. For instance, consider the example presented in Section I, where audio constitutes an instrument for enhancing a video sequence from its degraded version. It has been demonstrated by various researchers that joint audio-visual measurements often lead to improved video processing tasks [14], [16]. In other words, in this situation estimation based on Y and Z is preferable to using Y alone. Consequently, the obliqueness approach is expected to be inferior to the MMSE method in this case.

D. Worst Case Analysis

The main disadvantage of the oblique estimator is that its performance becomes arbitrarily poor as the correlation between Y and Z decreases. Indeed, when $Q = N$, direct computation shows that the estimation error is given by

$$\begin{aligned}
&\mathbb{E}\left[\left\|X - \hat{X}_{\mathrm{OB}}^{\mathrm{M1}}\right\|^2\right] = \\
&\quad\mathrm{Tr}\left\{\mathbb{C}\mathrm{ov}\left[X - \hat{X}_{\mathrm{LMMSE}}\right] + \mathbb{C}\mathrm{ov}\left[\hat{X}_{\mathrm{LMMSE}} - \hat{X}_{\mathrm{OB}}^{\mathrm{M1}}\right]\right\} = \\
&\quad\mathrm{Tr}\left\{\mathbf{\Gamma}_{XX} - \mathbf{\Gamma}_{XY}\mathbf{\Gamma}_{YY}^{\dagger}\mathbf{\Gamma}_{YX}\right\} + \mathrm{Tr}\left\{\boldsymbol{D}\mathbf{\Gamma}_{YY}^{\dagger}\boldsymbol{D}^T\right\},
\end{aligned} \tag{18}$$

where we substituted (3) and denoted $\boldsymbol{D} = \mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{YZ}^{-1}\mathbf{\Gamma}_{YY} - \mathbf{\Gamma}_{XY}$. The first term in this expression is the MSE of the LMMSE estimate of X, which could be achieved only by a method that knows $\mathbf{\Gamma}_{XY}$. Since the elements of $\mathbf{\Gamma}_{YZ}^{-1}$ can be arbitrarily large, the second term is unbounded and consequently the MSE can become arbitrarily large.

IV. Estimation with Moment Knowledge via Worst Case Design

As we have seen, one of the major drawbacks of the obliqueness approach is that, if the instrument Z is weakly correlated with the measurements Y, then the estimation error can become arbitrarily large. This phenomenon is rooted in the fact that equating the first and second-order moments of $(\hat{X}^T, Z^T)^T$ and $(X^T, Z^T)^T$ does not necessarily lead to an estimate $\hat{X}$ close to X in an MSE sense. Indeed, as we have seen, this approach is only optimal when all information that Z carries about X, is also present in Y. To overcome this limitation, we now propose an alternative approach, which is based on a worst-case design strategy. As we show, the resulting estimation error is guaranteed to be bounded.

A. Minimax Regret Estimation

Ultimately, we would like to design an estimator $\hat{X} = \boldsymbol{A}Y + \boldsymbol{b}$ that achieves the same MSE as that attained by the LMMSE estimator (3). In practice, though, this is impossible since we do not know the covariance matrix $\mathbf{\Gamma}_{XY} = \mathbb{C}\mathrm{ov}[X, Y]$. The *regret* of an estimator is defined as the difference between the MSE it attains and the MSE of the LMMSE method, which could be achieved if $\mathbf{\Gamma}_{XY}$ was known [23], [24], [25], namely

$$\mathbb{E}\left[\left\|X - \hat{X}\right\|^2\right] - \mathbb{E}\left[\left\|X - \hat{X}_{\mathrm{LMMSE}}\right\|^2\right]. \tag{19}$$

The regret of any estimator is a function of the unknown covariance $\mathbf{\Gamma}_{XY}$. This implies that one estimator can have a lower regret than another for certain choices of $\mathbf{\Gamma}_{XY}$ and a higher regret for others. Our approach here is to design an estimator whose worst-case regret is minimal.

Any RV X can be expressed as $X = \hat{X}_{\mathrm{LMMSE}} + U$, where U is a zero-mean RV uncorrelated with Y. Substituting this expression into (19), the regret becomes

$$\begin{aligned}
&\mathbb{E}\left[\left\|\hat{X}_{\mathrm{LMMSE}} + U - \hat{X}\right\|^2\right] - \mathbb{E}\left[\|U\|^2\right] = \\
&\qquad = \mathbb{E}\left[\left\|\hat{X}_{\mathrm{LMMSE}} - \hat{X}\right\|^2\right] + \mathbb{E}\left[\|U\|^2\right] - \mathbb{E}\left[\|U\|^2\right] \\
&\qquad = \mathbb{E}\left[\left\|\hat{X}_{\mathrm{LMMSE}} - \hat{X}\right\|^2\right],
\end{aligned} \tag{20}$$

where we used the fact that U is uncorrelated with $\hat{X}$ as it is an affine function of Y. In other words, the regret equals the MSE between $\hat{X}$ and the LMMSE solution (3). Substituting (3), the minimax regret problem can be cast as

$$\min_{\hat{X}\in\mathcal{Y}_{\mathrm{L}}} \max_{f_{XYZ}\in\mathcal{A}} \mathbb{E}\left[\left\|\mathbb{C}\mathrm{ov}[X,Y]\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y)+\boldsymbol{\mu}_X-\hat{X}\right\|^2\right], \tag{21}$$

where $\mathcal{A}$ is the set of density functions f_{XYZ} consistent with our moment knowledge, namely for which $\mathbb{E}[X]=\boldsymbol{\mu}_X$, $\mathbb{E}[Y]=\boldsymbol{\mu}_Y$, $\mathbb{E}[Z]=\boldsymbol{\mu}_Z$, $\mathbb{C}\mathrm{ov}[X]=\boldsymbol{\Gamma}_{XX}$, $\mathbb{C}\mathrm{ov}[Y]=\boldsymbol{\Gamma}_{YY}$, $\mathbb{C}\mathrm{ov}[Z]=\boldsymbol{\Gamma}_{ZZ}$, $\mathbb{C}\mathrm{ov}[X,Z]=\boldsymbol{\Gamma}_{XZ}$ and $\mathbb{C}\mathrm{ov}[Y,Z]=\boldsymbol{\Gamma}_{YZ}$.

The optimization problem (21) is challenging because the inner maximization is over a convex function rather than a concave one. This difficulty is typical of minimax regret problems and is encountered in sampling applications [25], [22], deterministic parameter estimation [23], [26] and random parameter estimation [24], [27] to name a few. Nevertheless, as is the case in all these application areas, the minimix-regret problem (21) has a simple closed form solution.

Theorem 3: *The solution to problem* (21) *is given by*

$$\hat{X}_{\mathrm{MX}}^{\mathrm{M1}}=\boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{\dagger}\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y)+\boldsymbol{\mu}_X. \tag{22}$$

Proof: See Appendix C. ■

We note that in contrast with the obliqueness approach, this method does not require inversion of the cross-covariance matrix $\boldsymbol{\Gamma}_{YZ}$ and therefore is especially advantageous over (10) when Y and Z are weakly correlated. In practice, the minimax regret estimator can be approximated from sets of examples, by replacing the means and covariances with their sample counterparts.

B. Equivalence with the Obliqueness Approach

Comparing (22) with (10), we see that the minimax regret and the oblique estimators are equal if

$$\boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{\dagger}\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{\dagger}=\boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{YZ}^{\dagger}. \tag{23}$$

To understand this condition, assume for simplicity that $M=N=Q$ and that the matrices $\boldsymbol{\Gamma}_{XZ}$, $\boldsymbol{\Gamma}_{YZ}$, $\boldsymbol{\Gamma}_{ZZ}$ are invertible. Then condition (23) becomes

$$\boldsymbol{\Gamma}_{ZZ}^{-1}\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{\dagger}=\boldsymbol{\Gamma}_{YZ}^{-1}. \tag{24}$$

Multiplying both sides by $\boldsymbol{\Gamma}_{ZZ}$ from the left and by $\boldsymbol{\Gamma}_{YZ}$ from the right, yields $\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{\dagger}\boldsymbol{\Gamma}_{YZ}=\boldsymbol{\Gamma}_{ZZ}$, or equivalently

$$\boldsymbol{\Gamma}_{ZZ}-\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{\dagger}\boldsymbol{\Gamma}_{YZ}=\mathbf{0}. \tag{25}$$

One may readily recognize this expression as being the covariance of the error of the LMMSE estimate of Z from Y. We thus arrive at the following conclusion.

Corollary 4: *Assume that the covariance matrices $\boldsymbol{\Gamma}_{XZ}$, $\boldsymbol{\Gamma}_{YZ}$ and $\boldsymbol{\Gamma}_{ZZ}$ are invertible. Then the minimax regret estimator* (22) *coincides with the oblique estimator* (10) *if and only if Z can be perfectly linearly estimated from Y.*

The equivalence of the two approaches in the case where Z can be perfectly recovered from Y, is not surprising as the known relation between X and Z can be immediately translated into a relation between X and Y. Thus, this is, in effect, not truly a partial knowledge scenario.

C. Best Case Analysis

The minimax regret estimator (22) was derived from a worst-case perspective. We now take a best-case viewpoint and study in what cases it is optimal. Comparing (22) with (3), we see that the minimax-regret method coincides with the LMMSE estimator if and only if

$$\boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{\dagger}\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{\dagger}=\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{\dagger}. \tag{26}$$

A simple interpretation of this condition can be obtained, as in Section III-C, by examining the case in which the RVs X, Y and Z are jointly Gaussian.

Theorem 5: *Suppose that the RVs X, Y and Z are jointly Gaussian. Let $A\triangleq(Y^T,Z^T)^T$ and assume that the matrices $\boldsymbol{\Gamma}_{AA}$, $\boldsymbol{\Gamma}_{YY}$, $\boldsymbol{\Gamma}_{ZZ}$ and $\boldsymbol{\Gamma}_{YY}-\boldsymbol{\Gamma}_{YZ}\boldsymbol{\Gamma}_{ZZ}^{-1}\boldsymbol{\Gamma}_{ZY}$ are invertible. Then the minimax regret estimate* (22) *coincides with the MMSE estimate of X given Y if and only if*

$$f_{X|YZ}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z})=f_{X|Z}(\boldsymbol{x}|\boldsymbol{z}). \tag{27}$$

Proof: See Appendix D. ■

Theorem 5 implies that the minimax regret solution is optimal if X and Y are independent given Z. In other words, if the MMSE estimate of $\boldsymbol{x}$ given $\boldsymbol{y}$ and $\boldsymbol{z}$ is only a function of $\boldsymbol{z}$, then the minimax regret estimator coincides with the MMSE solution. Thus, as opposed to the obliqueness approach, here we can benefit from an instrument that tells us more about X than Y does. This is particularly true when the information that Y carries about X is contained in the information that Z encompasses about X.

To emphasize the situations in which each of the methods is preferable, consider the following toy example. Suppose we wish to predict whether an individual will become sick with lung cancer based on the subject's smoking habits. In this case, X is a binary variable indicating the illness status, and Y is the average number of cigarettes the subject smokes per day. Now, assume we let Z denote the amount of tobacco accumulated in the subject's lungs. It is reasonable to assume that this instrument tells us about X everything that Y does, and perhaps more (as it is also affected by passive smoking). Thus, in this case the minimax-regret approach is preferable. Suppose, on the other hand, that we use the price of cigarettes as an instrument. In this case Z can affect X only through its effect on Y. Therefore, in this situation the obliqueness approach is preferable.

D. Worst Case Analysis

One of the main advantages of the minimax-regret approach is rooted in its worst-case performance. Specifically, we saw in Section III-D that the MSE of the oblique solution is not bounded from above, unless the correlation between Y and Z is not too weak. As we now show, the MSE of the minimax-regret estimator is guaranteed to be bounded, even if Y and Z are completely uncorrelated.

Theorem 6: *The regret* (19) *of $\hat{X}_{\mathrm{MX}}^{\mathrm{M1}}$ is not larger than $\mathrm{Tr}\{\boldsymbol{\Gamma}_{XX}-\boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{\dagger}\boldsymbol{\Gamma}_{ZX}\}$, the MSE of the LMMSE estimate of X given Z.*

Proof: See Appendix E. ■

Theorem 6 implies that the better X can be linearly recovered from Z, the closer the performance of the minimax regret estimator is to that of the LMMSE method. Therefore, as a rule of thumb, when the instrument is highly correlated with the signal, the minimax-regret method is effective.

V. Estimation with Density Knowledge via Obliqueness

Next, we address the problem of estimating $\boldsymbol{x}$ from $\boldsymbol{y}$ in the partial knowledge model M2. As in Section III, we start with a design strategy which is based on the obliqueness requirement.

Had $f_{XY}(\boldsymbol{x},\boldsymbol{y})$ been known, it would be possible to use the MMSE method $\hat{\boldsymbol{x}} = \mathbb{E}[X|Y=\boldsymbol{y}]$, which is the unique estimator whose error $X-\hat{X}$ is orthogonal to every function of Y. This orthogonality principle implies that the MMSE solution is the unique estimator satisfying

$$\mathbb{E}\left[\hat{X}|Y\right] = \mathbb{E}[X|Y]\,. \tag{28}$$

In our setting, all that is known regarding X is its statistical relation with Z. We therefore propose to replace the orthogonality principle with the demand that the estimation error be orthogonal to every function of Z, leading to the requirement that

$$\mathbb{E}\left[\hat{X}^{\mathrm{M2}}_{\mathrm{OB}}|Z\right] = \mathbb{E}[X|Z]\,. \tag{29}$$

Similarly to Section III, we expect this obliqueness principle to result in satisfactory performance if the instrument Z is close to Y in some sense.

Writing $\hat{X}^{\mathrm{M2}}_{\mathrm{OB}} = g(Y)$ and denoting $\phi(\boldsymbol{z}) = \mathbb{E}[X|Z=\boldsymbol{z}]$, (29) reduces to an integral equation[2] in $g(\boldsymbol{y})$:

$$\int_{\mathbb{R}^N} g(\boldsymbol{y}) f_{Y|Z}(\boldsymbol{y}|\boldsymbol{z}) d\boldsymbol{y} = \phi(\boldsymbol{z}), \quad \forall \boldsymbol{z}\in\mathbb{R}^Q. \tag{30}$$

The functions $f_{X|Z}(\boldsymbol{x}|\boldsymbol{z})$ and $\phi(\boldsymbol{z})$ in this equation are known by the assumptions of model M2. A unique oblique estimator exists if and only if (30) has a unique solution.

Like the oblique method under model M1, the approach taken here is also related to instrumental variable regression. Specifically, an equation very similar to (30) was studied in the context of instrumental variable estimation in nonparametric models [20]. In particular, it was shown that uniqueness of the solution to (30) requires $Q \geq N$ when the distribution of $Y|Z$ belongs to the exponential family. It was conjectured that this necessary condition also holds more generally.

A drawback of the obliqueness approach in the present setting, which did not exist in model M1, is that there is generally no closed form solution to (30). Nevertheless, it is possible to approximate the oblique estimator based on sets of examples of the type shown in Fig. 3. One such nonparametric approach was derived in [20]. Furthermore, despite the lack of a closed form expression, it is possible to draw qualitative conclusions regarding the best- and worst-case scenarios for the oblique estimator, as we discuss next.

Intuitively, if changes in $\boldsymbol{z}$ lead to small changes in $f_{Y|Z}(\boldsymbol{y},\boldsymbol{z})$, then the variance of the solution $\hat{X}^{\mathrm{M2}}_{\mathrm{OB}} = g(Y)$ to (30) is large. In the extreme situation in which Y and Z are independent, $f_{Y|Z}(\boldsymbol{y},\boldsymbol{z})$ is not a function of $\boldsymbol{z}$ at all, and consequently there exists no solution to (30). We thus conclude that as the statistical dependence between Y and Z decreases, the variance of $\hat{X}^{\mathrm{M2}}_{\mathrm{OB}}$ increases without bound. This also implies that the error $\mathbb{E}[\|X-\hat{X}^{\mathrm{M2}}_{\mathrm{OB}}\|^2]$ is unbounded.

A. Best Case Analysis

An interesting question concerning the obliqueness approach, is under what conditions it is optimal. As opposed to the analysis in Sections III-C and IV-C, which focused on the Gaussian case, here we make no assumptions on the structure of the density f_{XYZ}. The next theorem provides a sufficient condition for optimality.

Theorem 7: *Suppose that*

$$f_{X|YZ}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z}) = f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}). \tag{31}$$

Then the MMSE estimate of X given Y is an oblique solution.

Proof: Substituting $g(\boldsymbol{y}) = \mathbb{E}[X|Y=\boldsymbol{y}]$, the left-hand side of (30) becomes

$$\int g(\boldsymbol{y}) f_{Y|Z}(\boldsymbol{y}|\boldsymbol{z}) d\boldsymbol{y} = \iint \boldsymbol{x} f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) f_{Y|Z}(\boldsymbol{y}|\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{y}. \tag{32}$$

Using (31), this expression reduces to

$$\iint \boldsymbol{x} f_{X|YZ}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z}) f_{Y|Z}(\boldsymbol{y}|\boldsymbol{z}) d\boldsymbol{x} d\boldsymbol{y} = \int \boldsymbol{x} f_{X|Z}(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{x} = \mathbb{E}[X|Z=\boldsymbol{z}] \tag{33}$$

so that the MMSE estimator satisfies (30). ■

Note that Theorem 7 does not address the question of uniqueness of the oblique estimator. It merely states that when (31) holds, at least one of the solutions satisfying the obliqueness requirement (30), is the MMSE estimator.

Condition (31) is the same as that of Theorem 2. Therefore, we see that, as in model M1, the obliqueness approach is beneficial if the instrument does not carry any additional information about the signal, beyond that embedded in the measurements. In such a situation, if we knew f_{XYZ} completely, then measuring Z in addition to Y would be superfluous.

VI. Estimation with Density Knowledge via Worst-Case Design

Last, we address estimating $\boldsymbol{x}$ from $\boldsymbol{y}$ in the partial knowledge model M2 via a worst-case design.

As opposed to the obliqueness requirement that $\mathbb{E}[\hat{X}|Z] = \mathbb{E}[X|Z]$, we now seek a solution that minimizes the worst case regret over all RVs (X,Y,Z) with the given conditional expectation $\mathbb{E}[X|Z]$ and the given joint density f_{YZ}. Here we consider the regret

$$\mathbb{E}\left[\left\|X-\hat{X}\right\|^2\right] - \mathbb{E}\left[\|X-E[X|Y]\|^2\right] \tag{34}$$

[2] If Y and Z are discrete RVs, then the equation becomes $\sum_{\boldsymbol{y}} g(\boldsymbol{y}) p_{Y|Z}(\boldsymbol{y}|\boldsymbol{z}) = \phi(\boldsymbol{z})$, $\forall \boldsymbol{z}$.

with respect to the MMSE solution rather than the LMMSE method, as in Section IV. Expressing X as $X = \mathbb{E}[X|Y] + U$, where U is an RV uncorrelated with every function of Y, and using the fact that U is in particular uncorrelated with $\mathbb{E}[X|Y] - \hat{X}$ (as $\hat{X}$ is a function of Y), the regret becomes

$$
\begin{aligned}
E&\left[\left\|\mathbb{E}[X|Y] + U - \hat{X}\right\|^2\right] - \mathbb{E}\left[\|U\|^2\right] = \\
&= \mathbb{E}\left[\left\|\mathbb{E}[X|Y] - \hat{X}\right\|^2\right] + \mathbb{E}\left[\|U\|^2\right] - \mathbb{E}\left[\|U\|^2\right] \\
&= \mathbb{E}\left[\left\|\mathbb{E}[X|Y] - \hat{X}\right\|^2\right]. \qquad (35)
\end{aligned}
$$

Thus, the regret in the setting of model M2 equals the MSE between the estimator $\hat{X}$ and the MMSE solution.

Letting $\phi(Z) = \mathbb{E}[X|Z]$ and $\rho^2 = \mathbb{E}[\|X\|^2]$, which are both known in our setting, our problem is

$$
\min_{\hat{X} \in \mathcal{Y}} \max_{f_{XYZ} \in \mathcal{A}} \mathbb{E}\left[\left\|\mathbb{E}[X|Y] - \hat{X}\right\|^2\right], \qquad (36)
$$

where $\mathcal{A}$ is the set of density functions f_{XYZ} satisfying $\mathbb{E}[X|Z] = \phi(Z)$, $\mathbb{E}[\|X\|^2] = \rho^2$ and $\int_{\mathbb{R}^M} f_{XYZ}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) d\boldsymbol{x} = f_{YZ}(\boldsymbol{y},\boldsymbol{z})$. Note that besides the inner maximization being non-convex, as in model M1, problem (36) is also infinite-dimensional since the outer minimization is now over the set of all functions of Y. Interestingly, though, it has a simple solution, as presented in the next theorem.

Theorem 8: *The solution to problem* (36) *is given by*

$$
\hat{X}_{\mathrm{MX}}^{\mathrm{M2}} = g(Y) = \mathbb{E}[\mathbb{E}[X|Z]|Y]. \qquad (37)
$$

Proof: See Appendix F. ■

We note that (37) can be computed explicitly. This is because the inner and outer expectations are functions of $f_{X|Z}(\boldsymbol{x}|\boldsymbol{z})$ and $f_{Z|Y}(\boldsymbol{z}|\boldsymbol{y})$ respectively, which are both known in our setting.

The partial-knowledge minimax-regret estimator has a simple interpretation. We do not know the statistical relation between X and Y, rendering direct estimation of the signal given the measurements impossible. However, we can calculate the MMSE estimate $\phi(Z) = \mathbb{E}[X|Z]$ of X given Z, as $f_{XZ}(\boldsymbol{x},\boldsymbol{z})$ is available to us. This function cannot be used as an estimator, because we do not observe Z but rather Y. Nevertheless, the statistical relation between $\phi(Z)$ and Y is known, since $f_{YZ}(\boldsymbol{y},\boldsymbol{z})$ is known. Therefore, we can estimate this quantity given the measurements Y in an MMSE sense, leading to $\hat{X} = \mathbb{E}[\phi(Z)|Y] = \mathbb{E}[\mathbb{E}[X|Z]|Y]$.

A. Equivalence with the Obliqueness Approach

We now examine when the minimax regret solution (37) coincides with the oblique method. Although there is no closed form expression for the oblique estimator under model M2, a sufficient condition may easily be obtained from (29) such that the minimax regret estimator (37) is oblique.

Corollary 9: *Assume that $Z = h(Y)$ for some deterministic function $h(\cdot)$. Then the minimax regret estimator* (37) *satisfies the obliqueness principle.*

Proof: Denoting $\phi(Z) = \mathbb{E}[X|Z]$ and substituting $Z = h(Y)$ into (37), we have that

$$
\begin{aligned}
\mathbb{E}\left[\hat{X}_{\mathrm{MX}}^{\mathrm{M2}}|Z\right] &= \mathbb{E}[\mathbb{E}[\phi(Z)|Y]|Z] \\
&= \mathbb{E}[\mathbb{E}[\phi(h(Y))|Y]|Z] \\
&= \mathbb{E}[\phi(h(Y))|Z] \\
&= \mathbb{E}[\phi(Z)|Z] \\
&= \phi(Z), \qquad (38)
\end{aligned}
$$

so that the obliqueness condition (29) is satisfied by $\hat{X}_{\mathrm{MX}}^{\mathrm{M2}}$. ■

Evidently, as in model M1, the minimax-regret and obliqueness approaches result in the same estimator if Z can be perfectly determined from Y. The difference with respect to model M1, is that in Corollary 4 the instrument Z was required to be a linear function of Y, whereas here $h(\cdot)$ is arbitrary.

B. Best-Case Analysis

Next, we analyze which distributions f_{XYZ} are "best" for the minimax regret approach under model M2.

Theorem 10: *Suppose that*

$$
f_{X|YZ}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z}) = f_{X|Z}(\boldsymbol{x}|\boldsymbol{z}). \qquad (39)
$$

Then the minimax regret method (37) *coincides with the MMSE estimate of X given Y.*

Proof: Using condition (39), the estimator (37) becomes

$$
\begin{aligned}
\mathbb{E}[\mathbb{E}[X|Z]|Y] &= \iint \boldsymbol{x} f_{X|Z}(\boldsymbol{x}|\boldsymbol{z}) f_{Z|Y}(\boldsymbol{z}|\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{z} \\
&= \iint \boldsymbol{x} f_{X|YZ}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{z}) f_{Z|Y}(\boldsymbol{z}|\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{z} \\
&= \iint \boldsymbol{x} f_{XZ|Y}(\boldsymbol{x},\boldsymbol{z}|\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{z} \\
&= \int \boldsymbol{x} f_{X|Y}(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} = \mathbb{E}[X|Y], \qquad (40)
\end{aligned}
$$

proving the theorem. ■

Condition (39) is the same as that encountered in Theorem 5 in the context of minimax-regret estimation under model M1. The main difference is that Theorem 10 is relevant for arbitrary distributions, and does not require the normality assumption of Theorem 5.

We see that minimax regret estimation is optimal in situations where the information about X carried by the measurements Y is contained in that carried by the instrument Z. In such scenarios, if f_{XYZ} is completely known, then measuring Y in addition to Z does not help in estimating X. Therefore, the situations in which the oblique and minimax-regret methods are preferable to one another are similar to those of model M1 (see discussion in Section IV-C).

C. Worst-Case Analysis

The minimax-regret solution is especially advantageous over the obliqueness approach because of the fact that its worst-case MSE is finite, as we now show.

Theorem 11: *The regret* (34) *of* $\hat{X}_{\text{MX}}^{\text{M2}}$ *is no larger than* $\mathbb{E}[\|X - \mathbb{E}[X|Z]\|^2]$, *the MSE of the MMSE estimate of* X *given* Z.

Proof: The estimation error $\mathbb{E}[\|X - \hat{X}_{\text{MX}}^{\text{M2}}\|^2]$ is given by

$$\begin{aligned}\mathbb{E}&\Big[\|X - \mathbb{E}[X|Y]\|^2\Big] + \mathbb{E}\Big[\|\mathbb{E}[X|Y] - \mathbb{E}[\mathbb{E}[X|Z]|Y]\|^2\Big] \\ &= \mathbb{E}\Big[\|X - \mathbb{E}[X|Y]\|^2\Big] + \mathbb{E}\Big[\|\mathbb{E}[X - \mathbb{E}[X|Z]|Y]\|^2\Big] \\ &\le \mathbb{E}\Big[\|X - \mathbb{E}[X|Y]\|^2\Big] + \mathbb{E}\Big[\|X - \mathbb{E}[X|Z]\|^2\Big], \qquad (41)\end{aligned}$$

where the first line is a consequence of the fact that $X - \mathbb{E}[X|Y]$ is uncorrelated with every function of Y and, in particular, with $\mathbb{E}[X|Y] - \mathbb{E}[\mathbb{E}[X|Z]|Y]$, completing the proof. ■

As a consequence of Theorem 11, the minimax regret approach yields good results if the signal X could be accurately recovered by observing a realization of Z.

D. Nonparametric Regression

We now propose a nonparametric method for approximating the minimax-regret estimator (37) from two sets of examples $\{\boldsymbol{x}_i, \boldsymbol{z}_i^{\boldsymbol{x}}\}_{i=1}^{P}$ and $\{\boldsymbol{y}_i, \boldsymbol{z}_i^{\boldsymbol{y}}\}_{i=1}^{L}$, drawn independently from the densities $f_{XZ}(\boldsymbol{x}, \boldsymbol{z})$ and $f_{YZ}(\boldsymbol{y}, \boldsymbol{z})$ respectively.

We begin by estimating $\phi(\boldsymbol{z}) = \mathbb{E}[X|Z = \boldsymbol{z}]$ based on $\{\boldsymbol{x}_i, \boldsymbol{z}_i^{\boldsymbol{x}}\}_{i=1}^{P}$. The Nadaraya-Watson nonparametric estimator of $\phi(\boldsymbol{z})$ is given by [28], [29], [30]

$$\hat{\phi}(\boldsymbol{z}) = \frac{\sum_{i=1}^{P} \boldsymbol{x}_i K_Z\big(h_Z^{-1}(\boldsymbol{z} - \boldsymbol{z}_i^{\boldsymbol{x}})\big)}{\sum_{i=1}^{P} K_Z\big(h_Z^{-1}(\boldsymbol{z} - \boldsymbol{z}_i^{\boldsymbol{x}})\big)}, \qquad (42)$$

where $K_Z(\boldsymbol{z})$ is a density function called kernel and h_Z is a positive scalar called bandwidth. Under mild conditions on $K_Z(\boldsymbol{z})$, various converges properties of $\hat{\phi}(\boldsymbol{z})$ to $\phi(\boldsymbol{z})$ are known when $P \to \infty$ and $h \to 0$ at an appropriate rate [28], [29], [30]. The Nadaraya-Watson estimator, which is chosen here merely for concreteness, is a member of the family of local polynomial regression techniques. For these methods, there exist algorithms for automatically selecting the bandwidth parameter h_Z as a function of the sample-size P and possibly also as a function of the data itself [31].

The same nonparametric method could also be used to estimate $g(\boldsymbol{y}) = \mathbb{E}[\phi(Z)|Y = \boldsymbol{y}] = \mathbb{E}[\mathbb{E}[X|Z]|Y = \boldsymbol{y}]$, had we had a set of examples $\{\boldsymbol{y}_i, \phi(\boldsymbol{z}_i^{\boldsymbol{y}})\}$. Such a set is, of course, unavailable since there is no analytic expression for the function $\phi(\boldsymbol{z})$. However, recall that $\hat{\phi}(\boldsymbol{z})$ approximates $\phi(\boldsymbol{z})$ arbitrary well as the sample size P increases. We can thus use the set $\{\boldsymbol{y}_i, \hat{\phi}(\boldsymbol{z}_i^{\boldsymbol{y}})\}_{i=1}^{L}$ to construct a Nadaraya-Watson-like nonparametric estimator of $g(\boldsymbol{y})$, as follows

$$\hat{g}(\boldsymbol{y}) = \frac{\sum_{j=1}^{L} \hat{\phi}(\boldsymbol{z}_j^{\boldsymbol{y}}) K_Y\big(h_Y^{-1}(\boldsymbol{y} - \boldsymbol{y}_j)\big)}{\sum_{j=1}^{L} K_Y\big(h_Y^{-1}(\boldsymbol{y} - \boldsymbol{y}_j)\big)}. \qquad (43)$$

Here $K_Y(\boldsymbol{y})$ and h_Y are the kernel and bandwidth associated with the training set $\{\boldsymbol{y}_i, \hat{\phi}(\boldsymbol{z}_i^{\boldsymbol{y}})\}_{i=1}^{L}$ with $\hat{\phi}(\boldsymbol{z})$ of (42).

VII. Simulations

We now compare the oblique and minimax-regret estimators via simulations.

A. Partial Knowledge of Second-Order Moments

Suppose that X, Y and Z are scalar RVs distributed as

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.85 & \sigma_{XZ} \\ 0.85 & 1.4 & \sigma_{YZ} \\ \sigma_{XZ} & \sigma_{YZ} & 1 \end{pmatrix} \right). \qquad (44)$$

In this case, the LMMSE estimate of X from Y, which is given by $\hat{X}_{\text{LMMSE}} = (\sigma_{XY}/\sigma_{YY})Y \approx 0.61Y$, attains an MSE of $\sigma_{XX} - \sigma_{XY}^2/\sigma_{YY} \approx 0.48$. Assume, however, one does not know the values of the entries $(1, 2)$ and $(2, 1)$ of the covariance matrix of $(X, Y, Z)^T$. Thus, the fact that $\sigma_{XY} = 0.85$ cannot be used to design an estimator. A naive approach in this situation is to use the estimator $\hat{X}_{\text{NAIVE}} = \mu_X = 0$, whose MSE is given by $\sigma_{XX} = 1$. An alternative is to make use of σ_{XZ} and σ_{YZ} via the obliqueness approach of Section III or the minimax-regret method of Section IV.

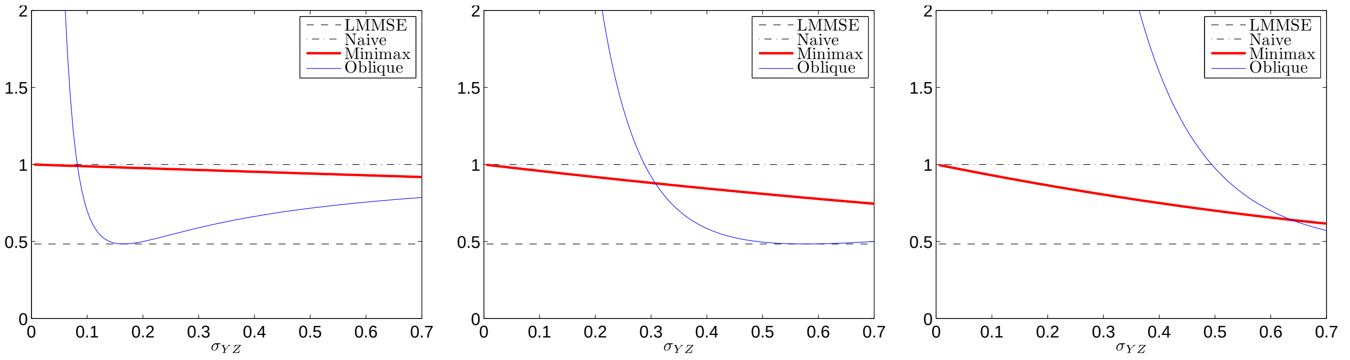
The MSE attained by the oblique estimator (10), which is given by $\hat{X}_{\text{OB}}^{\text{M1}} = (\sigma_{XZ}/\sigma_{YZ})Y$, can be computed explicitly via (18). The minimax-regret estimator (22) is given in our case by $\hat{X}_{\text{MX}}^{\text{M1}} = (\sigma_{XZ}\sigma_{ZY}/(\sigma_{ZZ}\sigma_{YY}))Y = (\sigma_{XZ}\sigma_{ZY}/1.4)Y$ and its MSE can be computed using equation (66) in Appendix E. The performance of both estimators depends on σ_{YZ} and σ_{XZ}. Figure 4 compares MSE_{OB} and MSE_{MX} with $\text{MSE}_{\text{LMMSE}}$ and $\text{MSE}_{\text{NAIVE}}$ as a function of σ_{YZ} for various values of σ_{XZ}. As can be seen, the performance of the oblique method becomes arbitrarily poor as σ_{YZ} is decreased. When σ_{YZ} is significantly smaller than σ_{XZ}, MSE_{OB} is even higher than the MSE of the naive measurement-blind estimator. On the other hand, the MSE of the minimax-regret method, which was designed to yield the best worst-case performance, never exceeds $\text{MSE}_{\text{NAIVE}}$. This behavior is obtained, though, at the expense of a rather modest performance for a wide range of values of σ_{XZ} and σ_{YZ}.

While the MSE of the minimax-regret estimator decreases as σ_{YZ} increases, its performance fails to surpass that of the oblique method at large values of σ_{YZ} when σ_{XZ} is small. This can be seen in Fig. 5, which shows the regions at which each of the estimators is preferable. We conclude that, as a rule of thumb, the oblique estimator should be used when σ_{YZ} is large and σ_{XZ} is small, while the minimax-regret solution is more effective when σ_{YZ} is small and σ_{XZ} is large. This simple test does not depend on σ_{XY} and thus can be performed based on the available partial knowledge.

B. Partial Knowledge of Probablity Functions

Suppose that X, Y and Z are binary RVs distributed as

$$p_{XYZ}(v) = \frac{1}{28}\begin{cases} 10 + \alpha + \beta & v = (0,0,0) \\ 2 - \alpha - \beta & v = (0,0,1) \\ 1 - \alpha & v = (0,1,0) \\ 1 + \alpha & v = (0,1,1) \\ 1 - \beta & v = (1,0,0) \\ 1 + \beta & v = (1,0,1) \\ 2 & v = (1,1,0) \\ 10 & v = (1,1,1), \end{cases} \qquad (45)$$



(a) $\sigma_{XZ} = 0.1$.

(b) $\sigma_{XZ} = 0.35$.

(c) $\sigma_{XZ} = 0.6$.

Fig. 4: MSE as a function of σ_{YZ} of the LMMSE, naive, minimax and oblique methods of model M1 in the setting of (44).

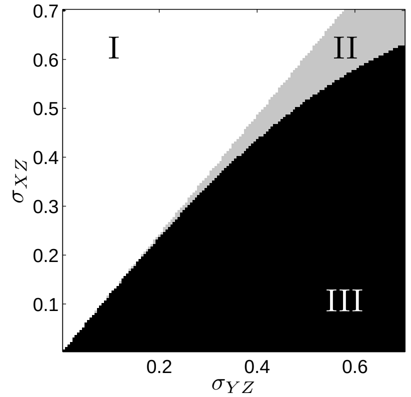


Fig. 5: Comparison between the oblique and minimax methods of model M1 corresponding to (44). Region I: $\text{MSE}_{\text{MX}} < \text{MSE}_{\text{NAIVE}} < \text{MSE}_{\text{OB}}$. Region II: $\text{MSE}_{\text{MX}} < \text{MSE}_{\text{OB}} < \text{MSE}_{\text{NAIVE}}$. Region III: $\text{MSE}_{\text{OB}} < \text{MSE}_{\text{MX}} < \text{MSE}_{\text{NAIVE}}$.

where $v = (x, y, z)$ and α and β are given parameters in the range $[-1, 1]$. In this example the MMSE estimate $\hat{X}_{\text{MMSE}} = \mathbb{E}[X|Y]$ attains an MSE of $6/49 \approx 0.12$ regardless of the values of α and β. Assume, however, that the joint distribution of X and Y is not known so that $\hat{X}_{\text{MMSE}}$ cannot be computed. In this case, we can resort to the naive approach $\hat{X}_{\text{NAIVE}} = \mu_X = 0.5$, which results in an MSE of 0.25. Alternatively, we can rely on our knowledge of $p_{XZ}(x, z)$ and $p_{YZ}(y, z)$ to compute the oblique and minimax-regret solutions of Sections V and VI. The MSE of these two methods, which can be computed explicitly in our case, depends on the parameters α and β. These, in turn, affect the mutual information[3] $I(Y; Z)$ between Y and Z and the mutual information $I(X; Z)$ between X and Z respectively.

Figure 6 depicts the MSE of the oblique and minimax-regret methods as a function of $I(Y; Z)$ for various values of $I(X; Z)$. It can be seen that, as in the linear case, the performance of the oblique estimator deteriorates as $I(Y; Z)$ becomes small, even beyond that of the naive estimator. However, MSE_{OB} is often lower than MSE_{MX} for high values of $I(Y; Z)$, especially when $I(X; Z)$ is small. The regions at which each of the approaches is preferable are shown in Fig. 7. The behavior is very similar to that shown in Fig. 5, leading to similar conclusions.

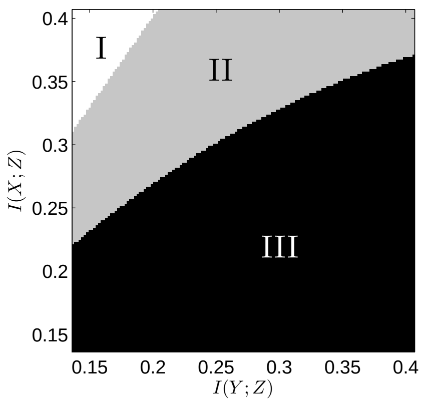


Fig. 7: Comparison between the oblique and minimax solutions of model M2 corresponding to (45). Regions are as in Fig. 5.

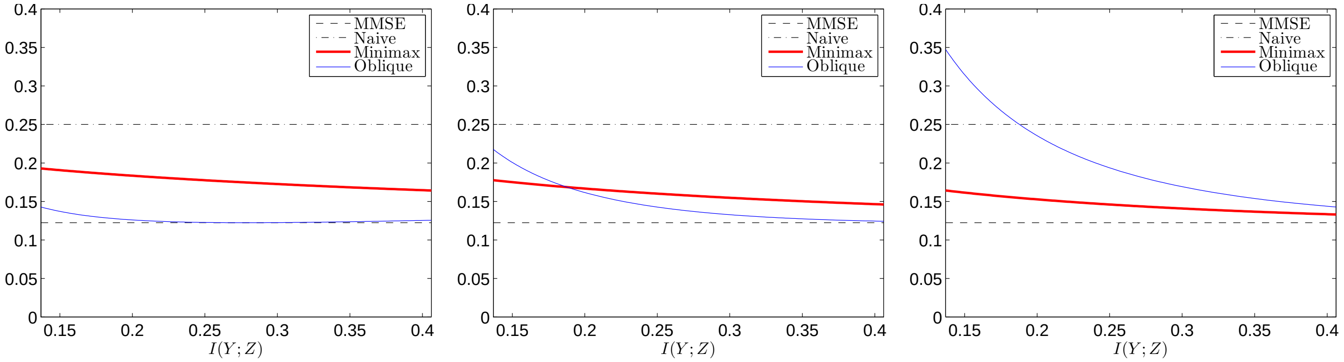
VIII. Application to Facial Feature Recovery

We now demonstrate our approach in the context of facial image enhancement.

Assume we are given an image $\boldsymbol{y}$ of a face taken with a low-grade camera (*e.g.,* a web-cam or a cellular-phone camera) whose degradation model is unknown. Furthermore, a set of paired examples of "clean" and degraded images is unavailable so that standard Bayesian estimation techniques cannot be used since $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$ cannot be learned. More specifically, in such applications we can typically collect many examples of "degraded" images $\{\boldsymbol{y}_i\}$ taken with the low-grade camera as well as many examples of "clean" facial images $\{\boldsymbol{x}_i\}$ taken with some high-quality sensor. However these two separate unpaired sets are not sufficient for learning $f_{XY}(\boldsymbol{x}, \boldsymbol{y})$.

To enhance the degraded image $\boldsymbol{y}$ via our partial-knowledge Bayesian estimation framework, we need to be able to intro-

[3] The mutual information between two RVs A and B is defined by $I(A; B) = \mathbb{E}[\log(p_{AB}(A, B))] - \mathbb{E}[\log(p_A(A)p_B(B))]$. It satisfies $I(A; B) = 0$ if and only if A and B are independent, and becomes larger as the statistical dependence between A and B tightens.

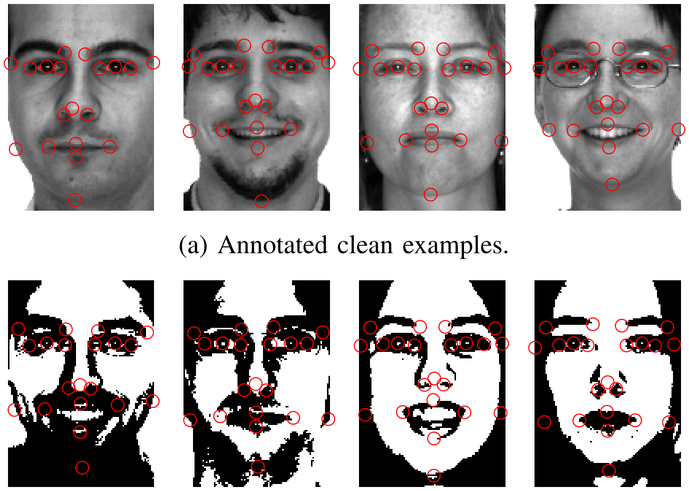


(a) $I(X;Z) = 0.14$.

(b) $I(X;Z) = 0.25$.

(c) $I(X;Z) = 0.41$.

Fig. 6: MSE as a function of $I(Y;Z)$ of the MMSE, naive, minimax and oblique methods of model M2 in the setting of (45).



(a) Annotated clean examples.

(b) Annotated degraded examples.

Fig. 8: Examples from the clean and degraded databases.

duce an instrument $\boldsymbol{z}$ whose relations with $\boldsymbol{x}$ and with $\boldsymbol{y}$ can be learned from examples. This can be done, for instance, by manually marking a set of points in several predefined locations both on the degraded images $\{\boldsymbol{y}_i\}$ and on the clean images $\{\boldsymbol{x}_i\}$. The vector $\boldsymbol{z}$, then, comprises the locations of the annotated points. This enables the construction of the two paired sets of examples $\{\boldsymbol{y}_i, \boldsymbol{z}_i^{\boldsymbol{y}}\}_{i=1}^{L}$, and $\{\boldsymbol{x}_i, \boldsymbol{z}_i^{\boldsymbol{x}}\}_{i=1}^{P}$, as required in our framework.

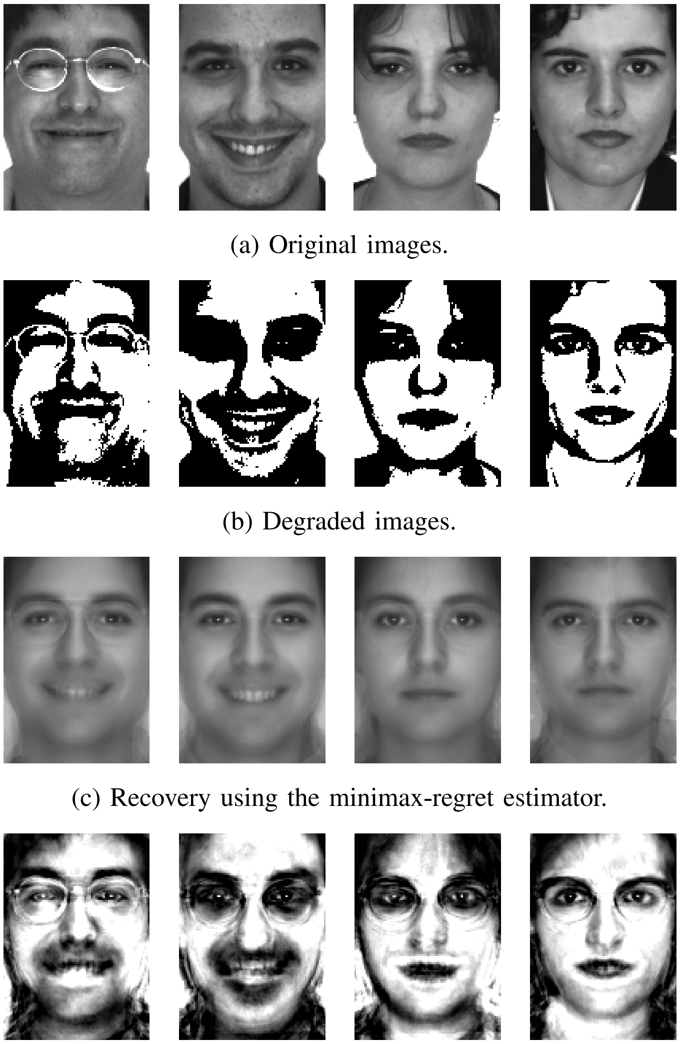
Figure 8 depicts several manually annotated clean and degraded facial images taken from the AR database [32]. The point annotations were taken from [33]. The images were all normalized such that the eyes appear at predefined locations. In practice, this preliminary step can be performed automatically [34], [35]. Here, the degradation (which is unknown to our algorithm) is a threshold operation. Thus, $\boldsymbol{y}$ is a binary image.

It is important to observe that $\boldsymbol{x}$ and $\boldsymbol{y}$ are both images of size 130×92, and thus correspond to vectors in $\mathbb{R}^{11960}$. On the other hand, $\boldsymbol{z}$ comprises 22 points, which means that it corresponds to a vector in $\mathbb{R}^{44}$. This huge difference in dimensionality implies two things. First, since $Q < N$, there are infinitely many estimators satisfying the obliqueness principle so that obliqueness seems an inadequate criterion in this setting. Second, it indicates that the statistical relation between $\boldsymbol{x}$ and $\boldsymbol{y}$ cannot possibly be characterized accurately solely in terms of $f_{XZ}(\boldsymbol{x}, \boldsymbol{z})$ and $f_{YZ}(\boldsymbol{y}, \boldsymbol{z})$. Indeed, $\boldsymbol{z}$ encompasses only geometric information about the face, and completely lacks any gray-level information. Therefore, one cannot expect to loyally recover the original image with this type of instrument, but rather only the expression and dominant facial features.

Figure 9(c) shows the recovery results for several degraded images obtained by our nonparametric approximation (43) to the model-M2 minimax-regret estimator (37). In this experiment, we used $P = 235$ "clean" examples $\{\boldsymbol{x}_i, \boldsymbol{z}_i^{\boldsymbol{x}}\}_{i=1}^{P}$ and $L = 137$ degraded examples $\{\boldsymbol{y}_i, \boldsymbol{z}_i^{\boldsymbol{y}}\}_{i=1}^{L}$ of different subjects. The person whose noisy image $\boldsymbol{y}$ was to be cleaned, was not included in either database. The kernels $K_Y(\boldsymbol{y})$ and $K_Z(\boldsymbol{z})$ were taken to be Gaussians. The same values of h_Y and h_Z were used in all our experiments. In practice, automatic bandwidth selection techniques can be applied [31].

As can be seen, the facial expression, as well as the dominant facial features, were indeed recovered correctly by the minimax estimator. However, the exact gray-level profile, which is among the important cues for distinguishing identity, was not restored accurately.

An alternative approach to treating the facial recovery task is to project the degraded image $\boldsymbol{y}$ onto a low-dimensional subspace learned from the clean examples $\{\boldsymbol{x}_i\}_{i=1}^{P}$ via *e.g.*, PCA [36]. We note that this methodology does not make use of the instrument $\boldsymbol{z}$, neither does it take into account the degraded examples $\{\boldsymbol{y}_i\}_{i=1}^{L}$. Furthermore, it is relevant only for applications where $\boldsymbol{x}$ and $\boldsymbol{y}$ are of the same dimension, whereas our proposed technique is general. Nevertheless, it relies on the observation that facial images approximately lie in a low-dimensional subspace, as experimentally shown in [36]. Therefore, removing from $\boldsymbol{y}$ the component perpendicular to this space, is expected to at least partially compensate for the unknown degradation. Figure 9(d) depicts the results obtained with the PCA approach, where the dimension of the subspace was tuned to account for 95% of the variance in the training set $\{\boldsymbol{x}_i\}_{i=1}^{P}$. As can be seen, the gray-level profile in these images is much closer to the degraded images than to the original



(a) Original images.

(b) Degraded images.

(c) Recovery using the minimax-regret estimator.

(d) Recovery using PCA.

Fig. 9: Recovery of facial images with the minimax regret estimator and with PCA. Each column corresponds to a different subject.

ones. Moreover, this technique produces artifacts which lead to unsatisfactory results. Similar artifacts were observed for different PCA-space dimensions.

IX. Application to Image Zooming

We conclude with an application to image zooming.

Suppose we are given a small image, which we would like to enlarge. Traditional approaches, such as nearest-neighbor, bilinear, bicubic (via Keys' kernel [37]), cubic spline and Lanchoz interpolation, tend to produce overly blurry images, especially at large zooming factors. An illustration of this phenomenon is shown in Fig. 10(b), which depicts the result of enlarging the image of Fig. 10(a) by a factor of 4 using bicubic interpolation.

An attractive alternative to linear interpolation methods, which has gained popularity in recent years, relies on the employment of learning techniques. Specifically, it has been shown that the task of image zooming can greatly benefit from the availability of training sets of high-resolution and down-sampled image patches [38]. Recently, it has also been demonstrated that such a training set can even be constructed from the given image to be enlarged itself [19], thus avoiding the need for a set of training images. Roughly speaking, in this type of methods, to enlarge an image by a factor of F, one first reduces its size by a factor of F and learns the relation between corresponding high-resolution and low-resolution patches. This approach, which relies on the fact that natural images often exhibit self-similarity across different scales, leads to state-of-the-art image zooming results [19]. The major pitfall in this strategy, however, is that it cannot be used on very small images (or with very large zooming factors) since there are simply not enough training patches left after reducing the size of the image. This is demonstrated in Fig. 10(c), which shows the result of enlarging the image of Fig. 10(a) by learning (via first-order local polynomial regression) the relation between each high-resolution pixel x and the corresponding 4×4 low resolution patch $\boldsymbol{y}$. The unsatisfactory result in this experiment can be attributed to the fact that there were only 720 available training patches.

To try and overcome the lack-of-examples barrier in the field of self training for image zooming, we can use our partial knowledge estimation paradigm as follows. To enlarge an image by a factor of F^2, we first reduce its size by a factor of F (rather than F^2). This down-sampled image contains many more training patches than in the standard approach and can be used to learn the statistical relation between an image patch $\boldsymbol{y}$ and its zoomed-by-F version $\boldsymbol{z}$ as well as the relation between a zoomed-by-F patch $\boldsymbol{z}$ and a zoomed-by-F^2 patch $\boldsymbol{x}$. We can thus use our techniques to construct an estimator of $\boldsymbol{x}$ based on $\boldsymbol{y}$ by relying only on the available partial knowledge. The result of applying this method with the M2 minimax-regret estimator is shown in Fig. 10(d). In this experiment, there were 3624 available training patches, and consequently the result is much more satisfactory than that of Fig. 10(c). For comparison, Fig. 10(e) shows the result of zooming by a factor of 3 using the algorithm of[4] [19]. We note that although this image is sharper, it is not necessarily more faithful to the original than Fig. 10(d). For example, the fifth and twelfths letters from the right in the bottom line, which should be "F" and "X" respectively, were recovered correctly in Fig. 10(d) and incorrectly in Fig. 10(c).

X. Conclusions

In this paper we proposed an approach for modeling partial Bayesian knowledge by using an instrumental variable. We considered two types of partial knowledge, which correspond to knowing the joint density functions and the joint second-order moments respectively of the instrument with the signal, and with the measurements. We treated each of these scenarios via two strategies: the obliqueness principle and minimax-regret. We derived closed form expressions for the estimators resulting from each of the design approaches, and analyzed in which situations each is preferable. We showed that the

[4] The image was taken from http://www.wisdom.weizmann.ac.il/~vision/SingleImageSR.html, in which only zooming by a factor of 3 was presented.

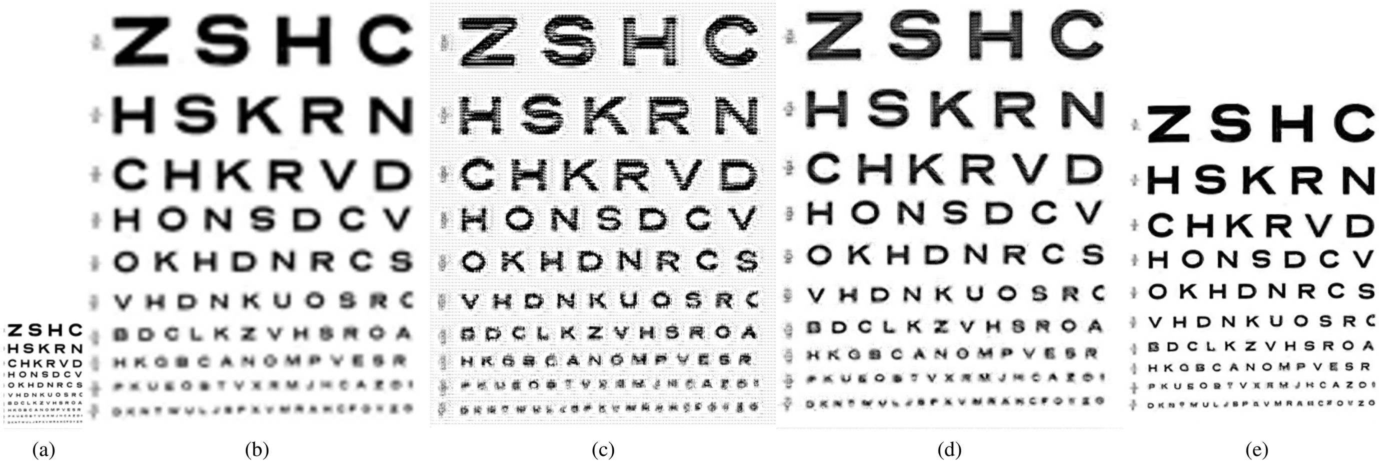


Fig. 10: Several methods for image zooming. (a) Original image. (b) Bicubic interpolation (×4). (c) Direct learning (×4). (d) Minimax learning via two enlargements by 2 (×4). (e) The method of [19] (×3).

oblique estimator coincides with the method of instrumental variable regression. Its main drawback is that its performance becomes arbitrarily poor as the statistical dependency between the instrument and measurements weakens. The performance of the minimax regret method, on the other hand, is guaranteed to be bounded regardless of how weak the instrument is. Nevertheless, this behavior comes at the expense of moderate performance at a wide variety of situations. As an example, we presented experimental results in image zooming and in recovering facial features from images that have undergone unknown degradation.

Appendix A
Proof of Theorem 1

Let $\varepsilon(f_{XY},\hat{X}) = \mathbb{E}_{f_{XY}}[\|\hat{X}-X\|^2]$ denote the MSE incurred by an estimator $\hat{X}\in\mathcal{Y}$ when the joint density of X and Y is $f_{XY}(\boldsymbol{x},\boldsymbol{y})$. It is easily verified that

$$\varepsilon(f_{XY},\hat{X}_{\mathrm{LMMSE}}) = \mathrm{Tr}\{\boldsymbol{\Gamma}_{XX}-\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{\dagger}\boldsymbol{\Gamma}_{YX}\} \quad (46)$$

for all $f_{XY}\in\mathcal{A}$. Consequently (46) is also the worst-case MSE of $\hat{X}_{\mathrm{LMMSE}}$ over $\mathcal{A}$. Now, denoting

$$f_{XY}^{*} = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_X\\ \boldsymbol{\mu}_Y\end{pmatrix},\begin{pmatrix}\boldsymbol{\Gamma}_{XX} & \boldsymbol{\Gamma}_{XY}\\ \boldsymbol{\Gamma}_{YX} & \boldsymbol{\Gamma}_{YY}\end{pmatrix}\right), \quad (47)$$

we note that any estimator $\hat{X}\in\mathcal{Y}$ satisfies

$$\begin{aligned}\max_{f_{XY}\in\mathcal{A}}\varepsilon(f_{XY},\hat{X}) &\geq \varepsilon(f_{XY}^{*},\hat{X})\\ &\geq \min_{\hat{X}}\varepsilon(f_{XY}^{*},\hat{X})\\ &= \mathrm{Tr}\{\boldsymbol{\Gamma}_{XX}-\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{\dagger}\boldsymbol{\Gamma}_{YX}\}, \quad (48)\end{aligned}$$

where the first inequality follows from the fact that $f_{XY}^{*}\in\mathcal{A}$, and the last equality is a result of the fact that the MMSE estimator in the Gaussian setting is linear. We have thus established that the worst-case MSE of any estimator over $\mathcal{A}$ is greater or equal to the worst-case MSE of the LMMSE solution over $\mathcal{A}$, proving that $\hat{X}_{\mathrm{LMMSE}}$ is minimax optimal.

Appendix B
Proof of Theorem 2

We begin by showing that condition (17) implies that the oblique method (10) coincides with the MMSE estimate, which is given by (3) in our setting. Since X and Y are jointly Gaussian, $X|Y$ follows the normal distribution

$$\mathcal{N}\big(\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_Y)+\boldsymbol{\mu}_X,\boldsymbol{\Gamma}_{XX}-\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1}\boldsymbol{\Gamma}_{YX}\big). \quad (49)$$

Similarly, $X|A$ is distributed as

$$\mathcal{N}\big(\boldsymbol{\Gamma}_{XA}\boldsymbol{\Gamma}_{AA}^{-1}(\boldsymbol{a}-\boldsymbol{\mu}_A)+\boldsymbol{\mu}_X,\boldsymbol{\Gamma}_{XX}-\boldsymbol{\Gamma}_{XA}\boldsymbol{\Gamma}_{AA}^{-1}\boldsymbol{\Gamma}_{AX}\big). \quad (50)$$

Equating the covariances of both distributions yields the condition

$$\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1}\boldsymbol{\Gamma}_{YX} = \boldsymbol{\Gamma}_{XA}\boldsymbol{\Gamma}_{AA}^{-1}\boldsymbol{\Gamma}_{AX}. \quad (51)$$

Let $\boldsymbol{C} = (\boldsymbol{\Gamma}_{ZZ}-\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{-1}\boldsymbol{\Gamma}_{YZ})^{-1}$ denote the inverse of the error covariance of the MMSE estimate of Z given Y. Note that by assumption, the inverse exists. Then, using the matrix inversion lemma, the matrix $\boldsymbol{\Gamma}_{AA}^{-1}$ in (51) can be written as

$$\begin{aligned}\boldsymbol{\Gamma}_{AA}^{-1} &= \begin{pmatrix}\boldsymbol{\Gamma}_{YY} & \boldsymbol{\Gamma}_{YZ}\\ \boldsymbol{\Gamma}_{ZY} & \boldsymbol{\Gamma}_{ZZ}\end{pmatrix}^{-1}\\ &= \begin{pmatrix}\boldsymbol{\Gamma}_{YY}^{-1}+\boldsymbol{\Gamma}_{YY}^{-1}\boldsymbol{\Gamma}_{YZ}\boldsymbol{C}\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{-1} & -\boldsymbol{\Gamma}_{YY}^{-1}\boldsymbol{\Gamma}_{YZ}\boldsymbol{C}\\ -\boldsymbol{C}\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{-1} & \boldsymbol{C}\end{pmatrix}.\end{aligned} \quad (52)$$

With this relation, and using the fact that $\boldsymbol{\Gamma}_{XA} = \boldsymbol{\Gamma}_{AX}^{T} = \begin{pmatrix}\boldsymbol{\Gamma}_{XY} & \boldsymbol{\Gamma}_{XZ}\end{pmatrix}$, the right-hand side of (51) becomes

$$\begin{aligned}&\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1}\boldsymbol{\Gamma}_{YX}+\\ &\quad\big(\boldsymbol{\Gamma}_{XZ}-\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1}\boldsymbol{\Gamma}_{YZ}\big)\boldsymbol{C}\big(\boldsymbol{\Gamma}_{XZ}-\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1}\boldsymbol{\Gamma}_{YZ}\big)^{T}.\end{aligned} \quad (53)$$

Therefore, (51) implies that $\boldsymbol{\Gamma}_{XZ} = \boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1}\boldsymbol{\Gamma}_{YZ}$, or equivalently, that

$$\boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{YZ}^{-1} = \boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1}. \quad (54)$$

This, in turn, implies that that the oblique estimate (10) coincides with (3).

Next, we show that if (10) and (3) coincide, then (17) holds. As we have seen, the equivalence of (10) and (3) implies that $\boldsymbol{D} \triangleq \boldsymbol{\Gamma}_{XZ} - \boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1}\boldsymbol{\Gamma}_{YZ} = \boldsymbol{0}$, which in turn implies that the covariances of the distributions of $X|Y$ and $X|A$ are equal. Therefore, all that remains to be shown is that if (10) coincides with (3) then the the means of the distributions (49) and (50) also coincide. Using (52), it is easily verified that

$$\begin{aligned}\boldsymbol{\Gamma}_{XA}\boldsymbol{\Gamma}_{AA}^{-1} &= \begin{pmatrix} \boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1} - \boldsymbol{D}\boldsymbol{C}\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{-1} & \boldsymbol{D}\boldsymbol{C} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1} & \boldsymbol{0} \end{pmatrix}, \qquad (55)\end{aligned}$$

which implies that $\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_Y) = \boldsymbol{\Gamma}_{XA}\boldsymbol{\Gamma}_{AA}^{-1}(\boldsymbol{a}-\boldsymbol{\mu}_A)$ so that the means of (49) and (50) are indeed equal.

Appendix C
Proof of Theorem 3

Every RV X can be uniquely expressed in terms of its LMMSE estimate given Z as $X = \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{\dagger}(Z-\boldsymbol{\mu}_Z) + \boldsymbol{\mu}_X + V$, where V is a zero-mean RV uncorrelated with Z. Direct calculation shows that $\mathbb{C}\mathrm{ov}[X] = \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{\dagger}\boldsymbol{\Gamma}_{ZX} + \mathbb{C}\mathrm{ov}[V]$, so that the constraint $\mathbb{C}\mathrm{ov}[X] = \boldsymbol{\Gamma}_{XX}$ translates into $\mathbb{C}\mathrm{ov}[V] = \boldsymbol{\Gamma}_{XX} - \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{\dagger}\boldsymbol{\Gamma}_{ZX}$. Furthermore, $\mathbb{C}\mathrm{ov}[X,Y] = \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{\dagger}\boldsymbol{\Gamma}_{ZY} + \mathbb{C}\mathrm{ov}[V,Y]$ and therefore the inner maximization in problem (21) is equivalent to

$$\max_{(V,Y,Z)\in\mathcal{B}} \mathbb{E}\left[\left\|\boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{\dagger}\boldsymbol{\Gamma}_{ZY}\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y) + \boldsymbol{\mu}_X \right.\right. \\ \left.\left. + \ \mathbb{C}\mathrm{ov}[V,Y](Y-\boldsymbol{\mu}_Y) - \hat{X}\right\|^2\right], \quad (56)$$

where $\mathcal{B}$ is the set of triplets of RVs (V,Y,Z) satisfying $\mathbb{E}[V] = \boldsymbol{0}$, $\mathbb{E}[Y] = \boldsymbol{\mu}_Y$, $\mathbb{E}[Z] = \boldsymbol{\mu}_Z$, $\mathbb{C}\mathrm{ov}[V] = \boldsymbol{\Gamma}_{XX} - \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{\dagger}\boldsymbol{\Gamma}_{ZX}$, $\mathbb{C}\mathrm{ov}[Y] = \boldsymbol{\Gamma}_{YY}$, $\mathbb{C}\mathrm{ov}[Z] = \boldsymbol{\Gamma}_{ZZ}$, $\mathbb{C}\mathrm{ov}[V,Z] = \boldsymbol{0}$, and $\mathbb{C}\mathrm{ov}[Y,Z] = \boldsymbol{\Gamma}_{YZ}$.

To prove that $\hat{X}_{\mathrm{MX}}^{\mathrm{M1}}$ of (22) is the solution to (21), we establish a lower bound on the minimax regret value and show that $\hat{X}_{\mathrm{MX}}^{\mathrm{M1}}$ achieves this bound. Expanding the norm, (56) becomes

$$\max_{(V,Y,Z)\in\mathcal{B}} \left\{\mathbb{E}\left[\left\|\mathbb{C}\mathrm{ov}[V,Y]\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y)\right\|^2 + \left\|\hat{X}_{\mathrm{MX}}^{\mathrm{M1}} - \hat{X}\right\|^2\right]\right. \\ \left. +2\mathbb{E}\left[\left(\hat{X}_{\mathrm{MX}}^{\mathrm{M1}} - \hat{X}\right)^T \mathbb{C}\mathrm{ov}[V,Y]\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y)\right]\right\}. \quad (57)$$

Our key insight is that for every triplet $(V,Y,Z)\in\mathcal{B}$ we also have $(-V,Y,Z)\in\mathcal{B}$. Furthermore, the first term in (57) is symmetric in V, whereas the second is anti-symmetric in V. This implies that if V maximizes the first term, then either V or $-V$ yields at least the same value for the objective comprising both terms. Consequently,

$$\begin{aligned}&\min_{\hat{X}\in\mathcal{Y}_{\mathrm{L}}} \max_{f_{XYZ}\in\mathcal{A}} \mathbb{E}\left[\left\|\mathbb{C}\mathrm{ov}[X,Y]\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y) + \boldsymbol{\mu}_X - \hat{X}\right\|^2\right] \\ &\geq \min_{\hat{X}\in\mathcal{Y}_{\mathrm{L}}} \max_{(V,Y,Z)\in\mathcal{B}} \left\{\mathbb{E}\left[\left\|\mathbb{C}\mathrm{ov}[V,Y]\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y)\right\|^2\right] + \right. \\ &\qquad\qquad \left. +\mathbb{E}\left[\left\|\hat{X}_{\mathrm{MX}}^{\mathrm{M1}} - \hat{X}\right\|^2\right]\right\} \\ &\geq \max_{(V,Y,Z)\in\mathcal{B}} \min_{\hat{X}\in\mathcal{Y}_{\mathrm{L}}} \left\{\mathbb{E}\left[\left\|\mathbb{C}\mathrm{ov}[V,Y]\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y)\right\|^2\right] + \right. \\ &\qquad\qquad \left. +\mathbb{E}\left[\left\|\hat{X}_{\mathrm{MX}}^{\mathrm{M1}} - \hat{X}\right\|^2\right]\right\} \\ &= \max_{(V,Y,Z)\in\mathcal{B}} \mathbb{E}\left[\left\|\mathbb{C}\mathrm{ov}[V,Y]\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y)\right\|^2\right], \qquad (58)\end{aligned}$$

where the second inequality follows from exchanging the minimum and maximum and the last equality is a result of solving the inner minimization, which is obtained at $\hat{X} = \hat{X}_{\mathrm{MX}}^{\mathrm{M1}}$.

We next show that equality is achieved with $\hat{X} = \hat{X}_{\mathrm{MX}}^{\mathrm{M1}}$. Indeed, (57) implies that for this estimator

$$\begin{aligned}&\min_{\hat{X}\in\mathcal{Y}_{\mathrm{L}}} \max_{f_{XYZ}\in\mathcal{A}} \mathbb{E}\left[\left\|\mathbb{C}\mathrm{ov}[X,Y]\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y) + \boldsymbol{\mu}_X - \hat{X}\right\|^2\right] \\ &= \max_{(V,Y,Z)\in\mathcal{B}} \mathbb{E}\left[\left\|\mathbb{C}\mathrm{ov}[V,Y]\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\boldsymbol{\mu}_Y)\right\|^2\right] \qquad (59)\end{aligned}$$

from which the theorem follows.

Appendix D
Proof of Theorem 5

Assume first that condition (27) holds. We will show that this implies that the minimax regret method (22) coincides with the MMSE estimate, which is given by (3) in our setting. Since X, Y and Z are jointly Gaussian, $X|Z$ follows the normal distribution

$$\mathcal{N}\left(\boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_Z) + \boldsymbol{\mu}_X, \boldsymbol{\Gamma}_{XX} - \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{-1}\boldsymbol{\Gamma}_{ZX}\right), \qquad (60)$$

whereas $X|A$ is distributed according to (50). Equating the covariances of both distributions yields the condition

$$\boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{-1}\boldsymbol{\Gamma}_{ZX} = \boldsymbol{\Gamma}_{XA}\boldsymbol{\Gamma}_{AA}^{-1}\boldsymbol{\Gamma}_{AX}. \qquad (61)$$

Let $\boldsymbol{C} = (\boldsymbol{\Gamma}_{YY} - \boldsymbol{\Gamma}_{YZ}\boldsymbol{\Gamma}_{ZZ}^{-1}\boldsymbol{\Gamma}_{ZY})^{-1}$ denote the inverse of the error covariance of the MMSE estimate of Y given Z, which exists by assumption. Then, using the matrix inversion lemma, the matrix $\boldsymbol{\Gamma}_{AA}^{-1}$ in (51) can be written as

$$\begin{aligned}\boldsymbol{\Gamma}_{AA}^{-1} &= \begin{pmatrix} \boldsymbol{\Gamma}_{YY} & \boldsymbol{\Gamma}_{YZ} \\ \boldsymbol{\Gamma}_{ZY} & \boldsymbol{\Gamma}_{ZZ} \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \boldsymbol{C} & -\boldsymbol{C}\boldsymbol{\Gamma}_{YZ}\boldsymbol{\Gamma}_{ZZ}^{-1} \\ -\boldsymbol{\Gamma}_{ZZ}^{-1}\boldsymbol{\Gamma}_{ZY}\boldsymbol{C} & \boldsymbol{\Gamma}_{ZZ}^{-1}\boldsymbol{\Gamma}_{ZY}\boldsymbol{C}\boldsymbol{\Gamma}_{YZ}\boldsymbol{\Gamma}_{ZZ}^{-1} + \boldsymbol{\Gamma}_{ZZ}^{-1} \end{pmatrix}. \qquad (62)\end{aligned}$$

With this relation, and using the fact that $\boldsymbol{\Gamma}_{XA} = \boldsymbol{\Gamma}_{AX}^T = \begin{pmatrix} \boldsymbol{\Gamma}_{XY} & \boldsymbol{\Gamma}_{XZ} \end{pmatrix}$, the right-hand side of (61) becomes

$$\boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{-1}\boldsymbol{\Gamma}_{ZX} + \\ \left(\boldsymbol{\Gamma}_{XY} - \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{-1}\boldsymbol{\Gamma}_{ZY}\right)\boldsymbol{C}\left(\boldsymbol{\Gamma}_{XY} - \boldsymbol{\Gamma}_{XZ}\boldsymbol{\Gamma}_{ZZ}^{-1}\boldsymbol{\Gamma}_{ZY}\right)^T. \qquad (63)$$

Therefore, (61) implies that $\mathbf{\Gamma}_{XY} = \mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{-1}\mathbf{\Gamma}_{ZY}$, or equivalently, that

$$\mathbf{\Gamma}_{XY}\mathbf{\Gamma}_{YY}^{-1} = \mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{-1}\mathbf{\Gamma}_{ZY}\mathbf{\Gamma}_{YY}^{-1}. \tag{64}$$

This, in turn, implies that that the minimax regret solution (22) coincides with (3).

Next, we show that if (22) and (3) coincide, then (27) holds. As we have seen, the equivalence of (22) and (3) implies that $\boldsymbol{D} \triangleq \mathbf{\Gamma}_{XY} - \mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{-1}\mathbf{\Gamma}_{ZY} = \mathbf{0}$, which in turn implies that the covariances of the distributions of $X|Z$ and $X|A$ are equal. Therefore, all that remains to be shown is that if (22) coincides with (3) then the the means of the distributions (60) and (50) also coincide. Using (62), it is easily verified that

$$\begin{aligned}\mathbf{\Gamma}_{XA}\mathbf{\Gamma}_{AA}^{-1} &= \begin{pmatrix} \boldsymbol{DC} & \mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{-1} - \boldsymbol{DC}\mathbf{\Gamma}_{YZ}\mathbf{\Gamma}_{ZZ}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{0} & \mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{-1} \end{pmatrix},\end{aligned} \tag{65}$$

which implies that $\mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}_Z) = \mathbf{\Gamma}_{XA}\mathbf{\Gamma}_{AA}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}_A)$ so that the means of (60) and (50) are indeed equal.

Appendix E
Proof of Theorem 6

Direct computation of the error $\mathbb{E}[\|X - \hat{X}_{\text{MX}}^{\text{M1}}\|^2]$ yields

$$\begin{aligned}&\text{Tr}\left\{\mathbf{\Gamma}_{XX}\right\} - 2\text{Tr}\left\{\mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{\dagger}\mathbf{\Gamma}_{ZY}\mathbf{\Gamma}_{YY}^{\dagger}\mathbf{\Gamma}_{YX}\right\} \\ &\qquad + \text{Tr}\left\{\mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{\dagger}\mathbf{\Gamma}_{ZY}\mathbf{\Gamma}_{YY}^{\dagger}\mathbf{\Gamma}_{YZ}\mathbf{\Gamma}_{ZZ}^{\dagger}\mathbf{\Gamma}_{ZX}\right\} \\ &= \text{Tr}\left\{\mathbf{\Gamma}_{XX} - \mathbf{\Gamma}_{XY}\mathbf{\Gamma}_{YY}^{\dagger}\mathbf{\Gamma}_{YX}\right\} + \text{Tr}\left\{\mathbf{\Gamma}_{BY}\mathbf{\Gamma}_{YY}^{\dagger}\mathbf{\Gamma}_{YB}\right\},\end{aligned} \tag{66}$$

where $B = \mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{\dagger}(Z - \boldsymbol{\mu}_Z) - (X - \boldsymbol{\mu}_X)$. The first term in this expression is the MSE of the LMMSE estimate of X given Y. Recalling that the MSE of the LMMSE estimate of B from Y is given by $\text{Tr}\{\mathbf{\Gamma}_{BB}\} - \text{Tr}\{\mathbf{\Gamma}_{BY}\mathbf{\Gamma}_{YY}^{\dagger}\mathbf{\Gamma}_{YB}\} \geq 0$, the second term is bounded by $\text{Tr}\{\mathbf{\Gamma}_{BB}\}$.

Letting $A = \mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{\dagger}(Z - \boldsymbol{\mu}_Z)$, so that $B = A - (X - \boldsymbol{\mu}_X)$,

$$\begin{aligned}\text{Tr}\{\mathbf{\Gamma}_{BB}\} &= \text{Tr}\{\mathbf{\Gamma}_{AA}\} - 2\text{Tr}\{\mathbf{\Gamma}_{AX}\} + \text{Tr}\{\mathbf{\Gamma}_{XX}\} \\ &= \text{Tr}\left\{\mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{\dagger}\mathbf{\Gamma}_{ZX} - 2\mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{\dagger}\mathbf{\Gamma}_{ZX} + \mathbf{\Gamma}_{XX}\right\} \\ &= \text{Tr}\left\{\mathbf{\Gamma}_{XX} - \mathbf{\Gamma}_{XZ}\mathbf{\Gamma}_{ZZ}^{\dagger}\mathbf{\Gamma}_{ZX}\right\},\end{aligned} \tag{67}$$

which is the MSE of the LMMSE estimate of X from Z. This completes the proof.

Appendix F
Proof of Theorem 8

We establish a lower bound on the optimal minimax regret value and then show that $\hat{X}_{\text{MX}}^{\text{M2}}$ of (37) achieves this bound, which proves that it is optimal.

The RV X can be uniquely written as

$$X = \mathbb{E}[X|Z] + U = \phi(Z) + U, \tag{68}$$

where U is a zero-mean RV uncorrelated with every function of Z. It follows that $\mathbb{E}[\|X\|^2] = \mathbb{E}[\|\phi(Z)\|^2] + \mathbb{E}[\|U\|^2]$, so that the constraint $\mathbb{E}[\|X\|^2] = \rho^2$ translates into $\mathbb{E}[\|U\|^2] = \rho^2 - \mathbb{E}[\|\phi(Z)\|^2]$. Substituting (68), and noting that $\mathbb{E}[\phi(Z)|Y]$, which equals $\hat{X}_{\text{MX}}^{\text{M2}}$, is fixed over the set $\mathcal{A}$, the inner maximization in (36) becomes

$$\begin{aligned}&\mathbb{E}\left[\left\|\hat{X}_{\text{MX}}^{\text{M2}} - \hat{X}\right\|^2\right] + \\ &\max_{(U,Y,Z)\in\mathcal{B}}\left\{\mathbb{E}\left[\|\mathbb{E}[U|Y]\|^2 + 2\left(\hat{X}_{\text{MX}}^{\text{M2}} - \hat{X}\right)^T\mathbb{E}[U|Y]\right]\right\}\end{aligned} \tag{69}$$

where $\mathcal{B}$ is the set of triplets of RVs (U, Y, Z) such that U is uncorrelated with every function of Z, $\mathbb{E}[U] = \mathbf{0}$, $\mathbb{E}[\|U\|^2] = \rho^2 - \mathbb{E}[\|\phi(Z)\|^2]$ and $\int_{\mathbb{R}^M} f_{UYZ}(\boldsymbol{u}, \boldsymbol{y}, \boldsymbol{z})d\boldsymbol{u} = f_{YZ}(\boldsymbol{y}, \boldsymbol{z})$. The set $\mathcal{B}$ is symmetric in U, namely for every triplet $(U, Y, Z) \in \mathcal{B}$ we also have $(-U, Y, Z) \in \mathcal{B}$. Furthermore, the first term within the maximum in (69) is symmetric in U, whereas the second is anti-symmetric in U. This implies that if U maximizes the first term, then either U or $-U$ yields at least the same value for the objective comprising both terms. Consequently, noting that $\mathbb{E}[\phi(Z)|Y] = \hat{X}_{\text{MX}}^{\text{M2}}$,

$$\begin{aligned}&\min_{\hat{X}\in\mathcal{Y}}\max_{f_{XYZ}\in\mathcal{A}}\mathbb{E}\left[\left\|\mathbb{E}[X|Y] - \hat{X}\right\|^2\right] \geq \\ &\geq \min_{\hat{X}\in\mathcal{Y}}\left\{\mathbb{E}\left[\left\|\hat{X}_{\text{MX}}^{\text{M2}} - \hat{X}\right\|^2\right] + \max_{(U,Y,Z)\in\mathcal{B}}\left\{\mathbb{E}\left[\|\mathbb{E}[U|Y]\|^2\right]\right\}\right\} \\ &= \max_{(U,Y,Z)\in\mathcal{B}}\mathbb{E}\left[\|\mathbb{E}[U|Y]\|^2\right],\end{aligned} \tag{70}$$

where the equality follows from the fact that the solution to the minimization is obtained at $\hat{X} = \hat{X}_{\text{MX}}^{\text{M2}}$.

We now show that the inequality can be achieved with $\hat{X} = \hat{X}_{\text{MX}}^{\text{M2}}$. Indeed, with this choice of $\hat{X}$, (69) implies that

$$\max_{f_{XYZ}\in\mathcal{A}}\mathbb{E}\left[\|\mathbb{E}[X|Y] - \hat{X}\|^2\right] = \max_{(U,Y,Z)\in\mathcal{B}}\mathbb{E}\left[\|\mathbb{E}[U|Y]\|^2\right],$$

from which the theorem follows.

References

[1] M. R. Banham and A. K. Katsaggelos, "Digital image restoration," *IEEE signal processing magazine*, vol. 14, no. 2, pp. 24–41, 1997.

[2] J. Benesty, S. Makino, and J. Chen, *Speech enhancement.* Springer Verlag, 2005.

[3] Y. Bar-Shalom, X. R. Li, and T. Kirubarajan, *Estimation with applications to tracking and navigation.* Wiley-Interscience, 2001.

[4] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, 1993.

[5] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proceedings of the IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992.

[6] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[7] N. M. S. Kawai, M. Morimoto and N. Teranishi, "Photo response analysis in CCD image sensors with a VOD structure," *IEEE Trans. Electron Devices*, vol. 42, no. 4, pp. 652–655, 1995.

[8] D. L. Snyder, C. W. Helstrom, A. D. Lanterman, M. Faisal, and R. L. White, "Compensation for readout noise in CCD images," *Journal of the Optical Society of America A*, vol. 12, no. 2, pp. 272–283, 1995.

[9] A. Buades, B. Coll, and J. M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Modeling and Simulation*, vol. 4, no. 2, pp. 490–530, 2006.

[10] J. Portilla, V. Strela, M. J. Wainwright, and E. P. Simoncelli, "Image denoising using scale mixtures of Gaussians in the wavelet domain," *IEEE Transactions on Image processing*, vol. 12, no. 11, pp. 1338–1351, 2003.

[11] M. A. Beaumont, W. Zhang, and D. J. Balding, "Approximate Bayesian computation in population genetics," *Genetics*, vol. 162, no. 4, pp. 2025–2035, 2002.

[12] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, O. Chapelle, B. Schölkopf, and A. Zien, Eds. The MIT Press, 2010.

[13] J. Lafferty and L. Wasserman, “Statistical analysis of semi-supervised regression,” *Advances in Neural Information Processing Systems*, vol. 20, pp. 801–808, 2007.

[14] T. Chen and R. R. Rao, “Audio-visual integration in multimodal communication,” *Proceedings of the IEEE*, vol. 86, no. 5, p. 837, 1998.

[15] J. W. Fisher, T. Darrell, W. T. Freeman, and P. Viola, “Learning joint statistical models for audio-visual fusion and segregation,” *Advances in Neural Information Processing Systems*, pp. 772–778, 2001.

[16] J. Hershey and J. Movellan, “Audio-vision: Using audio-visual synchrony to locate sounds,” *Advances in Neural Information Processing Systems*, vol. 12, pp. 813–819, 2000.

[17] R. J. Bowden and D. A. Turkington, *Instrumental variables*. Cambridge University Press, 1984.

[18] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, “Overview of the face recognition grand challenge,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005.*, vol. 1, pp. 947–954.

[19] D. Glasner, S. Bagon, and M. Irani, “Super-resolution from a single image,” in *International Conference on Computer Vision*, 2009.

[20] W. K. Newey and J. L. Powell, “Instrumental variable estimation of nonparametric models,” *Econometrica*, vol. 71, no. 5, pp. 1565–1578, 2003.

[21] A. J. Izenman, “Recent developments in nonparametric density estimation,” *Journal of the American Statistical Association*, vol. 86, no. 413, pp. 205–224, 1991.

[22] T. Michaeli and Y. C. Eldar, “Optimization techniques in modern sampling theory,” in *Convex Optimization in Signal Processing and Communications*, Y. C. Eldar and D. Palomar, Eds. Cambridge University Press, 2010, pp. 266–314.

[23] Y. C. Eldar, A. Ben-Tal, and A. Nemirovski, “Linear minimax regret estimation of deterministic parameters with bounded data uncertainties,” *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2177–2188, Aug. 2004.

[24] Y. C. Eldar and N. Merhav, “A competitive minimax approach to robust estimation of random parameters,” *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1931–1946, 2004.

[25] Y. C. Eldar and T. G. Dvorkind, “A minimum squared-error framework for generalized sampling,” *IEEE Transactions on Signal Processing*, vol. 54, no. 6 Part 1, pp. 2155–2167, 2006.

[26] Y. Eldar, “Rethinking biased estimation: Improving maximum likelihood and the Cramér–Rao bound,” *Foundations and Trends in Signal Processing*, vol. 1, no. 4, pp. 305–449.

[27] Y. C. Eldar and N. Merhav, “Minimax MSE-ratio estimation with signal covariance uncertainties,” *IEEE Transactions on Signal Processing*, vol. 53, no. 4, pp. 1335–1347, 2005.

[28] E. A. Nadaraya, “On Estimating Regression,” *Theory of Probability and its Applications*, vol. 9, p. 141, 1964.

[29] G. S. Watson, “Smooth regression analysis,” *Sankhya: The Indian Journal of Statistics, Series A*, vol. 26, no. 4, pp. 359–372, Dec. 1964.

[30] W. Hardle and M. Muller, “Multivariate and semiparametric kernel regression,” Humboldt Universitaet Berlin, Sonderforschungsbereich 373 1997-26, 1997. [Online]. Available: http://ideas.repec.org/p/wop/humbsf/1997-26.html

[31] J. Fan and I. Gijbels, “Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 2, pp. 371–394, 1995.

[32] A. M. Martinez and R. Benavente, “The AR face database,” cVC Technical Report, no. 24, June 1998.

[33] “http://www-prima.inrialpes.fr/FGnet/data/05-ARFace/tarfd_markup.html.”

[34] D. Reisfeld and Y. Yeshurun, “Preprocessing of face images: Detection of features and pose normalization,” *Computer vision and image understanding*, vol. 71, no. 3, pp. 413–430, 1998.

[35] T. Michaeli, “Face normalization for recognition and enrollment,” Patent, 05 2010, EP 1872303 B1. [Online]. Available: http://www.patentlens.net/patentlens/patent/EP_1872303_B1/en/

[36] M. Turk and A. Pentland, “Eigenfaces for recognition,” *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[37] R. Keys, “Cubic convolution interpolation for digital image processing,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 29, no. 6, pp. 1153–1160, 1981.

[38] W. T. Freeman, T. R. Jones, and E. C. Pasztor, “Example-based super-resolution,” *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.