

A MINIMAX APPROACH TO BAYESIAN ESTIMATION WITH PARTIAL KNOWLEDGE OF THE OBSERVATION MODEL

Tomer Michaeli and Yonina C. Eldar

Department of Electrical Engineering
Technion–Israel Institute of Technology, Haifa, Israel
{tomermic@tx, yonina@ee}.technion.ac.il

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 joint probability distributions of the instrument and the signal, and of the instrument and the measurements, are known. However, the joint probability function of the signal and measurements is unknown. Our model generalizes that underlying the method of instrumental variables from statistics, in that the instrument does not have to satisfy any requirements and no parametric form for the optimal regressor needs to be available.

We begin by deriving an estimator for this scenario, via a worst-case design strategy. We then propose a non-parametric method for learning this estimator from a set of examples. Finally, we demonstrate our approach in the context of enhancement of facial images that have undergone an unknown degradation.

Index Terms— Bayesian estimation, minimax regret, partial knowledge, learning.

1. INTRODUCTION

A common problem in signal processing is that of estimating an unknown random quantity \mathbf{x} from a set of noisy measurements \mathbf{y} . Image denoising and deblurring, speech enhancement, and target tracking, are a few examples of problems that are frequently addressed via a Bayesian strategy. The Bayesian framework requires knowledge of the prior distribution of the signal to be estimated, as well as the conditional probability of the measurements given the signal. The former can usually be learned from a set of examples $\{\mathbf{x}_i\}$ of “clean” signals. The latter, on the other hand, necessitates either a paired set of examples $\{\mathbf{x}_i, \mathbf{y}_i\}$ of signals and measurements, or knowledge of the degradation mechanism that yielded the measurements (e.g., blurring, additive Gaussian noise, etc.). In many applications, neither of these assumptions 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. These undesired effects typically vary in time and are very hard to model statistically. Furthermore, no set of paired examples of clean and degraded signals are available in these scenarios.

This work was supported in part by the Israel Science Foundation under Grant no. 1081/07 and by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMMunications NEWCOM++ (contract no. 216715).

As another example, consider the problem of enhancement of facial images taken with a low-grade camera (e.g., a web-cam or a cellular-phone camera). The degradation in this case includes the blur due to the lens, the nonlinear response of the CCD sensor, and non-additive noise. 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 is to resort to simplified degradation model assumptions, such as Gaussian blur kernel and additive white noise in image restoration applications, and stationary background noise in speech enhancement tasks. These models simplify the treatment but are often far from being loyal to the true physical setting.

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

During the last decade, various approaches have been proposed to enhancing audio or video based on joint audio-visual measurements (see e.g., [1]). There is a fundamental difference, though, from our approach. For example, in the scenario described above, the input to the estimator is only the noisy video sequence \mathbf{y} , *without* the associated audio. Thus, it does not fall into the category of sensor fusion but rather it belongs to the field of video enhancement. The audio data comes into play only in the training sets $\{\mathbf{y}_i, \mathbf{z}_i\}$ and $\{\mathbf{x}_i, \mathbf{z}_i\}$ but does not constitute part of the measurements. 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 address the problem of Bayesian estimation with partial statistical knowledge via a worst-case design framework. We also provide a nonparametric algorithm for approximating the resulting estimator from two sets of paired examples $\{\mathbf{y}_i, \mathbf{z}_i\}$ and $\{\mathbf{x}_i, \mathbf{z}_i\}$. Finally, we demonstrate our approach in the context of enhancement of facial images that have undergone unknown distortion.

In statistics, a similar setting is encountered in the context of

nonlinear regression with instrumental variables (IV) [2]. Our modeling, however, is more general. Specifically, in the traditional IV formulation, the instrument Z is required to be independent of the error $X - E[X|Y]$, its dimension cannot be smaller than that of X , and a parametric form for the optimal regressor $E[X|Y]$ needs to be available. In our formulation, on the other hand, none of these restrictions is needed.

2. MINIMAX REGRET ESTIMATION

Random variables (RVs) are denoted by capital letters (e.g., X, Y, Z) and bold lower-case letters denote the values that they take (e.g., $\mathbf{x}, \mathbf{y}, \mathbf{z}$). The RV X is the quantity to be estimated, also termed ‘‘signal’’, Y is the measurements RV, 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 assume that the joint density functions $f_{X,Z}(\mathbf{x}, \mathbf{z})$ and $f_{Y,Z}(\mathbf{y}, \mathbf{z})$ are known, whereas $f_{X,Y}(\mathbf{x}, \mathbf{y})$ is unknown.

Our goal is to produce an estimate $\hat{X} = g(Y)$ of the signal X based on the measurements Y , such that the mean-squared error (MSE)

$$\text{MSE} = E[\|X - \hat{X}\|^2] \quad (1)$$

is minimized. Careful investigation reveals that MSE minimization is not a well posed problem in our setting. This follows from the fact that the MSE depends on $f_{X,Y}(\mathbf{x}, \mathbf{y})$ (since \hat{X} is a function of Y), which is unknown. More specifically, it is well known that the minimum MSE (MMSE) estimator is given by the conditional expectation $E[X|Y]$. This quantity cannot be computed without knowledge of $f_{X|Y}(\mathbf{x}|\mathbf{y}) = f_{X,Y}(\mathbf{x}, \mathbf{y})/f_Y(\mathbf{y})$.

To tackle this difficulty, we use the fact that $f_{X,Y}(\mathbf{x}, \mathbf{y})$ is not completely arbitrary, as it has to be consistent with our knowledge of $f_{X,Z}(\mathbf{x}, \mathbf{z})$ and $f_{Y,Z}(\mathbf{y}, \mathbf{z})$. Specifically, X can be written as

$$X = E[X|Z] + U, \quad (2)$$

where U is a zero-mean RV independent of Z . The first term depends on $f_{X|Z}(\mathbf{x}|\mathbf{z})$, which is known. The second term, however, is arbitrary. Namely, any zero-mean U that is independent of Z and has an arbitrary statistical relation with Y , is consistent with our prior knowledge.

The above discussion motivates a worst-case design strategy where we minimize the *regret*, for the worst-case RV X that is consistent with our knowledge of $E[X|Z]$. The regret is defined as the difference between the squared-norm error and the smallest possible error that can be achieved if $f_{X,Y}(\mathbf{x}, \mathbf{y})$ was known [3]:

$$E[\|X - \hat{X}\|^2] - E[\|X - E[X|Y]\|^2]. \quad (3)$$

Any RV X can be expressed as $X = E[X|Y] + V$, where V is zero-mean and independent of Y . Substituting this expression into (3) and using the fact that V is also independent of $E[X|Y] - \hat{X}$ (as \hat{X} is a function of Y), the regret becomes

$$\begin{aligned} & E[\|E[X|Y] + V - \hat{X}\|^2] - E[\|V\|^2] = \\ & = E[\|E[X|Y] - \hat{X}\|^2] + E[\|V\|^2] - E[\|V\|^2] \\ & = E[\|E[X|Y] - \hat{X}\|^2]. \end{aligned} \quad (4)$$

Therefore, letting $\phi(Z)$ denote the known conditional expectation $E[X|Z]$, we wish to solve

$$\min_{\hat{X} \in \mathcal{Y}} \max_{f_{X,Y,Z} \in \mathcal{A}} E[\|E[X|Y] - \hat{X}\|^2], \quad (5)$$

where

$$\mathcal{A} = \{f_{X,Y,Z} : E[X|Z] = \phi(Z), E[\|X\|^2] \leq \rho^2\} \quad (6)$$

is the set of feasible probability density functions $f_{X,Y,Z}(\mathbf{x}, \mathbf{y}, \mathbf{z})$, and \mathcal{Y} denotes the set of RVs who are (Borel measurable) functions of Y . The bound ρ^2 on the variance of X is imposed to guarantee that the error cannot grow indefinitely. However, as we will see, the value of ρ does not effect the solution. The following theorem provides the solution to the minimax-regret problem.

Theorem 1 *The solution to problem (5) is given by*

$$\hat{X} = g(Y) = E[E[X|Z]|Y]. \quad (7)$$

Before proving the theorem, we note that (7) can be computed explicitly. This is because the inner and outer expectations are functions of $f_{X|Z}(\mathbf{x}|\mathbf{z})$ and $f_{Z|Y}(\mathbf{z}|\mathbf{y})$ respectively, which are both known in our setting.

Proof. We will first establish a lower bound on the optimal minimax regret value and then show that \hat{X} of (7) achieves this bound, which proves that it is optimal.

Substituting (2), the inner maximization in (5) becomes

$$\begin{aligned} & E[\|\hat{X} - E[\phi(Z)|Y]\|^2] + \\ & \max_{U \in \mathcal{B}} \left\{ E[\|E[U|Y]\|^2] - 2E[(\hat{X} - E[\phi(Z)|Y])^T E[U|Y]] \right\}, \end{aligned} \quad (8)$$

where $\mathcal{B} = \{U : E[\|U\|^2] \leq \rho^2 - E[\|\phi(Z)\|^2]\}$ is the set of feasible RVs U in the decomposition (2). Clearly, at the maximum value of U we have that $E[(\hat{X} - E[\phi(Z)|Y])^T E[U|Y]] \leq 0$ since we can change the sign of U without effecting the constraint. Therefore,

$$\begin{aligned} & \max_{U \in \mathcal{B}} \left\{ E[\|E[U|Y]\|^2] - 2E[(\hat{X} - E[\phi(Z)|Y])^T E[U|Y]] \right\} \\ & \geq \max_{U \in \mathcal{B}} \{E[\|E[U|Y]\|^2]\}. \end{aligned} \quad (9)$$

Combining (9) and (8),

$$\begin{aligned} & \min_{\hat{X} \in \mathcal{Y}} \max_{f_{X,Y,Z} \in \mathcal{A}} E[\|E[X|Y] - \hat{X}\|^2] \geq \\ & \geq \min_{\hat{X} \in \mathcal{Y}} \left\{ E[\|\hat{X} - E[\phi(Z)|Y]\|^2] + \max_{U \in \mathcal{B}} \{E[\|E[U|Y]\|^2]\} \right\} \\ & = \max_{U \in \mathcal{B}} \{E[\|E[U|Y]\|^2]\}, \end{aligned} \quad (10)$$

where the equality is a result of solving the outer minimization, obtained at $\hat{X} = E[\phi(Z)|Y] = E[E[X|Z]|Y]$.

We now show that the inequality can be achieved with $\hat{X} = E[\phi(Z)|Y] = E[E[X|Z]|Y]$. Indeed, with this choice of \hat{X} , (8) implies that

$$\max_{f_{X,Y,Z} \in \mathcal{A}} E[\|E[X|Y] - \hat{X}\|^2] = \max_{U \in \mathcal{B}} \{E[\|E[U|Y]\|^2]\}, \quad (11)$$

from which the theorem follows. ■

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 $\phi(Z) = E[X|Z]$ of X given Z , as $f_{X,Z}(\mathbf{x}, \mathbf{z})$ 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 $f_{Y,Z}(\mathbf{y}, \mathbf{z})$ is known. Therefore, we can estimate this quantity given the measurements Y in an MMSE sense, leading to $\hat{X} = E[\phi(Z)|Y] = E[E[X|Z]|Y]$.

2.1. Best-Case Analysis

The above estimator was derived from a worst-case perspective. We now take a best-case viewpoint. Namely, we analyze which distributions $f_{X,Y,Z}$ are the “best” for our approach.

Perhaps the simplest “optimal” situation arises when the signal is a deterministic function of the instrument, $X = \psi(Z)$. In this case, (7) coincides with the MMSE estimate:

$$\hat{X} = E[E[\psi(Z)|Z]|Y] = E[\psi(Z)|Y] = E[X|Y]. \quad (12)$$

A more complicated scenario happens when $\{X, Z, Y\}$ form a Markov triplet in the sense that $f_{X|Y,Z}(\mathbf{x}|\mathbf{y}, \mathbf{z}) = f_{X|Z}(\mathbf{x}, \mathbf{z})$. In this case (7) reduces to

$$\begin{aligned} E[E[X|Z]|Y] &= \int \int \mathbf{x} f_{X|Z}(\mathbf{x}|\mathbf{z}) f_{Z|Y}(\mathbf{z}|\mathbf{y}) d\mathbf{x} d\mathbf{z} \\ &= \int \int \mathbf{x} f_{X|Y,Z}(\mathbf{x}|\mathbf{y}, \mathbf{z}) f_{Z|Y}(\mathbf{z}|\mathbf{y}) d\mathbf{x} d\mathbf{z} \\ &= \int \int \mathbf{x} f_{X,Z|Y}(\mathbf{x}, \mathbf{z}|\mathbf{y}) d\mathbf{x} d\mathbf{z} \\ &= \int \mathbf{x} f_{X|Y}(\mathbf{x}|\mathbf{y}) d\mathbf{x} = E[X|Y]. \end{aligned} \quad (13)$$

3. NONPARAMETRIC REGRESSION

In many practical scenarios no explicit expressions for $f_{X,Z}(\mathbf{x}, \mathbf{z})$ and $f_{Y,Z}(\mathbf{y}, \mathbf{z})$ are available, but rather only sets of examples drawn from these densities. We now propose a nonparametric method for approximating the minimax-regret estimator using such sets of examples.

Suppose that one has access to two sets of paired examples $\{\mathbf{x}_i, \mathbf{z}_i^x\}_{i=1}^P$ and $\{\mathbf{y}_i, \mathbf{z}_i^y\}_{i=1}^L$, drawn independently from the densities $f_{X,Z}(\mathbf{x}, \mathbf{z})$ and $f_{Y,Z}(\mathbf{y}, \mathbf{z})$ respectively. To obtain a nonparametric approximation of the minimax-regret estimator, we begin by estimating $\phi(\mathbf{z}) = E[X|Z = \mathbf{z}]$ based on $\{\mathbf{x}_i, \mathbf{z}_i^x\}_{i=1}^P$.

The Nadaraya-Watson nonparametric estimator of $\phi(\mathbf{z})$ is given by [4, 5, 6]

$$\hat{\phi}(\mathbf{z}) = \frac{\sum_{i=1}^P \mathbf{x}_i K_Z(h_Z^{-1}(\mathbf{z} - \mathbf{z}_i^x))}{\sum_{i=1}^P K_Z(h_Z^{-1}(\mathbf{z} - \mathbf{z}_i^x))}, \quad (14)$$

where $K_Z(\mathbf{z})$ is a density function called kernel and h_Z is a positive scalar called bandwidth. Under mild conditions on $K_Z(\mathbf{z})$, various convergence properties of $\hat{\phi}(\mathbf{z})$ to $\phi(\mathbf{z})$ are known when $P \rightarrow \infty$ and $h \rightarrow 0$ at an appropriate rate [4, 5, 6].

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

$$\hat{g}(\mathbf{y}) = \frac{\sum_{j=1}^L \left(\frac{\sum_{i=1}^P \mathbf{x}_i K_Z(h_Z^{-1}(\mathbf{z}_j^y - \mathbf{z}_i^x))}{\sum_{i=1}^P K_Z(h_Z^{-1}(\mathbf{z}_j^y - \mathbf{z}_i^x))} \right) K_Y(h_Y^{-1}(\mathbf{y} - \mathbf{y}_j))}{\sum_{j=1}^L K_Y(h_Y^{-1}(\mathbf{y} - \mathbf{y}_j))}, \quad (15)$$

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

The above nonparametric regression works as follows. For each training example $(\mathbf{y}_j, \mathbf{z}_j^y)$, we approximate the conditional expectation $\hat{\phi}(\mathbf{z}_j^y) = E[X|Z = \mathbf{z}_j^y]$ based on the set $\{\mathbf{x}_i, \mathbf{z}_i^x\}_{i=1}^P$. The resulting set of estimates $\{\hat{\phi}(\mathbf{z}_j^y)\}_{j=1}^L$, is then linearly combined with weights that depend on the resemblance of the measurements \mathbf{y} to each of the examples $\{\mathbf{y}_j\}_{j=1}^L$, to produce the final estimate $\hat{g}(\mathbf{y})$.

4. APPLICATION TO FACIAL FEATURE RECOVERY

We now demonstrate the partial-knowledge Bayesian estimation framework introduced above, in the context of facial image enhancement.

Assume we are given an image \mathbf{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. In such situations, one cannot use standard Bayesian estimation techniques to enhance the image, since $f_{X,Y}(\mathbf{x}, \mathbf{y})$ is unknown and cannot be learned from examples. More specifically, in applications of this sort we can usually collect many examples of “degraded” images $\{\mathbf{y}_i\}$ taken with the low-grade camera as well as many examples of “clean” facial images $\{\mathbf{x}_i\}$ taken with some high-quality sensor. However these two separate un-paired sets are not sufficient for learning the joint density $f_{X,Y}(\mathbf{x}, \mathbf{y})$.

To enhance the degraded image \mathbf{y} via our partial-knowledge Bayesian estimation framework, we need to be able to introduce an instrument \mathbf{z} whose relations with \mathbf{x} and with \mathbf{y} can be learned from examples. This can be done, for example, by manually marking a set of points in several predefined locations both on the degraded images $\{\mathbf{y}_i\}$ and on the clean images $\{\mathbf{x}_i\}$. The vector \mathbf{z} , then, comprises the locations of the annotated points. This enables the construction of the two paired sets of examples $\{\mathbf{y}_i, \mathbf{z}_i^y\}_{i=1}^L$, and $\{\mathbf{x}_i, \mathbf{z}_i^x\}_{i=1}^P$, as required in our framework.

Figure 1 depicts several manually annotated clean and degraded facial images taken from the AR database [7]. The point annotations were taken from [8]. The images were all normalized such that the eyes appear at predefined locations. In practice, this preliminary step can be performed automatically [9]. Here, the degradation (which is unknown to our algorithm) is a threshold operation. Thus, \mathbf{y} is a binary image.

It is important to observe that \mathbf{x} and \mathbf{y} are both images of size 130×92 , and thus correspond to vectors in \mathbb{R}^{11960} . On the other hand, \mathbf{z} comprises 22 points, which means that it corresponds to a vector in \mathbb{R}^{44} . This huge difference in dimensionality indicates that the statistical relation between \mathbf{x} and \mathbf{y} cannot possibly be characterized accurately solely in terms of $f_{X,Z}(\mathbf{x}, \mathbf{z})$ and $f_{Y,Z}(\mathbf{y}, \mathbf{z})$. Indeed, \mathbf{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 2(c) shows the recovery results for several degraded images. In this experiment, we used $P = 235$ “clean” examples $\{\mathbf{x}_i, \mathbf{z}_i^x\}_{i=1}^P$ and $L = 137$ degraded examples $\{\mathbf{y}_i, \mathbf{z}_i^y\}_{i=1}^L$ of different subjects. The person whose noisy image \mathbf{y} was to be cleaned, was not included in either database. The kernels $K_Y(\mathbf{y})$ and $K_Z(\mathbf{z})$ 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.

As can be seen, the facial expression, as well as the dominant

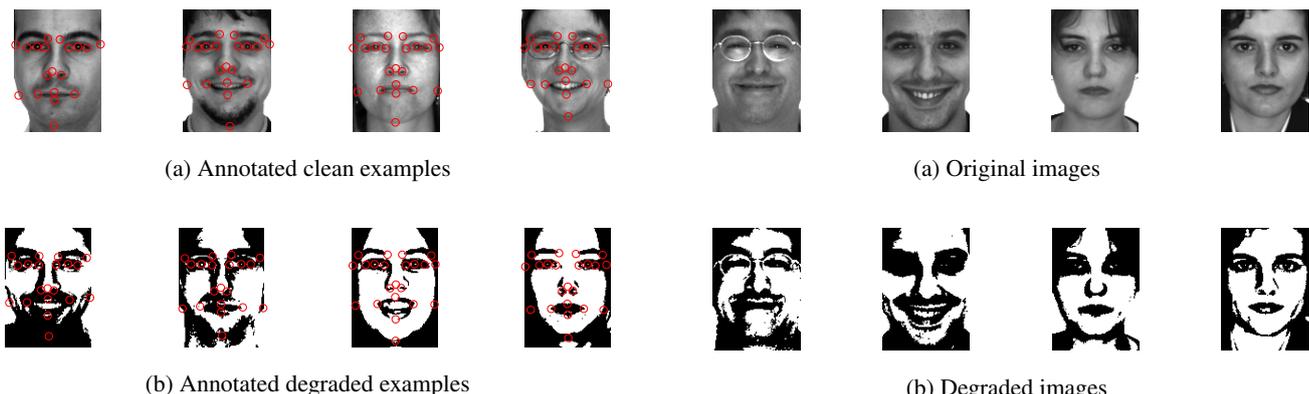


Fig. 1. Examples from the clean and degraded databases.

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 tackling the facial recovery task is to project the degraded image \mathbf{y} onto a low-dimensional subspace learned from the clean examples $\{\mathbf{x}_i\}_{i=1}^P$ via *e.g.*, PCA [10]. We note that this methodology does not make use of the instrument \mathbf{z} , neither does it take into account the degraded examples $\{\mathbf{y}_i\}_{i=1}^L$. Furthermore, it is relevant only for applications where \mathbf{x} and \mathbf{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 [10]. Therefore, removing from \mathbf{y} the component perpendicular to this space, is expected to at least partially compensate for the unknown degradation. Figure 2(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 $\{\mathbf{x}_i\}_{i=1}^P$. As can be seen, the gray-level profile in these images is much closer to the degraded images than to the original ones. Moreover, this technique produces artifacts which lead to unsatisfactory results. Similar artifacts were observed for different PCA-space dimensions.

5. CONCLUSIONS

In this paper we proposed a novel approach for modeling partial Bayesian knowledge by using an instrumental variable. Our approach generalizes the instrumental variables methodology used in statistics in that no restrictions on the instrument are imposed. We derived an estimator suited for our model, via a worst-case design strategy. Finally, we presented experimental results on recovering facial features from images that have undergone unknown degradation. As the technique is general and does not impose restrictions on the instrument, it can be applied to a wide variety of applications in signal processing.

6. REFERENCES

[1] 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.

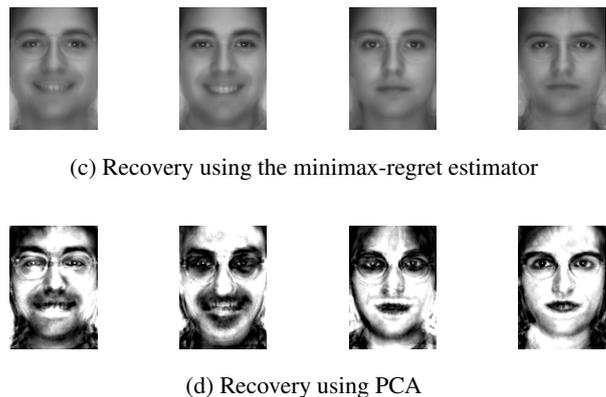


Fig. 2. Recovery of facial images with the minimax regret estimator and with PCA. Each column corresponds to a different subject.

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

[3] 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.

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

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

[6] W. Hardle and M. Muller, “Multivariate and semiparametric kernel regression,” Sonderforschungsbereich 373 1997-26, Humboldt Universitaet Berlin, 1997.

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

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

[9] 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.

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