CLASSIFYING LINEAR SYSTEM OUTPUTS BY ROBUST LOCAL BAYESIAN QUADRATIC DISCRIMINANT ANALYSIS ON LINEAR ESTIMATORS

Hyrum S. Anderson and Maya R. Gupta

University of Washington
Department of Electrical Engineering
Seattle, WA 98195

ABSTRACT

We consider the problem of assigning a class label to the noisy output of a linear system, where clean feature examples are available for training. We design a robust classifier that operates on a linear estimate, with uncertainty modeled by a Gaussian distribution with parameters derived from the bias and covariance of a linear estimator. Class-conditional distributions are modeled locally as Gaussians. Since estimation of Gaussian parameters from few training samples can be ill-posed, we extend recent work in Bayesian quadratic discriminant analysis to derive a robust local generative classifier. Experiments show a statistically significant improvement over prior art.

Index Terms— noisy features, robust estimation, MAP classification, pattern classification, supervised learning

1. INTRODUCTION

The problem of classifying a noisy feature vector using clean training features arises often in practice. The cost associated with obtaining quality test data may be prohibitive, or the test environment may not be characterized as well as the training environment. For example, an inexpensive sensor deployed in a sensor network may compare less precise test features to higher quality training features obtained in a laboratory setting. In automatic speech recognition, acoustic conditions during testing are typically much noisier than during training [1, 2]. In remote sensing, training features may be extracted from free-field signals, while test features are extracted from signals that have propagated through an unknown channel [3].

Consider a test feature vector $\mathbf{x} \in \mathbb{R}^d$ and linear observation model $\mathbf{z} = H\mathbf{x}+\mathbf{w}$ for $\mathbf{z} \in \mathbb{R}^n$, known H, $E[\mathbf{w}] = 0$ and $Cov[\mathbf{w}] = \sigma_w^2 I$. We denote vectors in boldface and matrices in uppercase; to preserve this distinction, random vectors will also be boldface, but distinguishable from their realizations by context. Having observed $\mathbf{z}$, it is common to form an estimate $\hat{\mathbf{x}} \in \mathbb{R}^d$ of the clean test feature vector $\mathbf{x}$ and classify $\hat{\mathbf{x}}$ using labeled training examples $\mathcal{X} = \{(\mathbf{x}_i, y_i)\}_{i=1}^M$ for classes $y_i \in \mathcal{G}$; since $\hat{\mathbf{x}}$ is inherently noisy, this is commonly referred to as the noisy features problem [4]. We assume that the training features $\{\mathbf{x}_i\}_{i=1}^M$ are noise-free, or at least contain significantly less noise than the test feature vector $\hat{\mathbf{x}}$ so that they may be treated as noise-free.

Although physical models often include a linear system H, much of the prior work in this area has focused on the case where $H = I$, and $\hat{\mathbf{x}} = \mathbf{x} + \mathbf{w}$ where $\mathbf{w}$ is i.i.d. noise. For example, Papandreou et al. recently demonstrated a Gaussian mixture model and hidden Markov model for environment-robust audio-visual speech recognition [5]. Pawlak and Siu proposed a modified smoothing kernel classifier that achieves provably lower risk than ignoring the noise [6]. Ahmad and Tresp adapted a Gaussian basis function network classifier for noisy features in computer vision [7]. In robust speech recognition, researchers have modeled a convolutional system as additive noise in the cepstral domain; Buera, et al. provide a summary of compensation methods [1].

In fact, for $\mathbf{z} = H\mathbf{x} + \mathbf{w}$ the optimal approach to dealing with noisy observations is to maximize $p(g|\mathbf{z}) \propto p(\mathbf{z}|g)p(g)$ over classes $g \in \mathcal{G}$, which does not rely on an estimate $\hat{\mathbf{x}}$ [4]. For example, we have developed a quadratic discriminant analysis (QDA) classifier that models $p(\mathbf{z}|g)$ as Gaussian and accounts for the disparity between corrupted test and clean training features [3]. A related approach is to train a classifier on artificially corrupted training data, or acquire training data that emulate test conditions. However, classifying $\mathbf{z}$ rather than $\hat{\mathbf{x}}$ may not be viable, e.g., due to high dimensionality of $\mathbf{z}$ or system design issues; this is the premise of this paper.

In Sec. 2, we review a strategy for solving the noisy feature MAP rule when $H = I$ and $\hat{\mathbf{x}} = \mathbf{x}+\mathbf{w}$, where $p(\hat{\mathbf{x}}|\mathbf{x})$ and $p(\mathbf{x}|g)$ are modeled as Gaussian distributions. We investigate generalizing this strategy for a linear estimator $\hat{\mathbf{x}} = G\mathbf{z}$ when $H \neq I$, and consider an alternative approach by defining an expected MAP rule for noisy features. In Sec. 3, we couple the MAP rules for noisy features with a locally Gaussian class-conditional likelihood $p(\mathbf{x}|g)$ to form the robust local Bayesian quadratic discriminant analysis (R-BDA) classifier, which is the chief contribution of this paper. In Sec. 4 and 5, we compare the proposed R-BDA classifier to the robust classifier proposed by Pawlak and Siu in [6] and to several standard non-robust classifiers.

2. NOISY FEATURES AND EXPECTED MAP RULE

The maximum a priori (MAP) rule for noisy features [4, 5, 7] chooses class g that maximizes

$$p(g|\hat{\mathbf{x}}) \quad \propto \quad \int p(\mathbf{x}|g)p(g)p(\hat{\mathbf{x}}|\mathbf{x})\, d\mathbf{x}, \tag{1}$$

where the pdfs in the integrand are assumed to be known. Although the integral in (1) is generally intractable, a closed form solution exists for the case when $p(\hat{\mathbf{x}}|\mathbf{x})$ can be rewritten as a Gaussian with argument $\mathbf{x}$ and $p(\mathbf{x}|g) = \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}_g, \Sigma_g)$. Commonly (e.g., [5]), it is assumed that $\hat{\mathbf{x}} = \mathbf{x} + \mathbf{w}$ with $\mathbf{w} \sim \mathcal{N}(\mathbf{w}; 0, \Sigma_w)$ so that $p(\hat{\mathbf{x}}|\mathbf{x}) = \mathcal{N}(\hat{\mathbf{x}}; \mathbf{x}, \Sigma_w) = \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \Sigma_w)$, and one uses the product of Gaussians identity to rewrite the integrand in (1) as a single Gaussian in $\mathbf{x}$ times a constant:

$$\mathcal{N}(\mathbf{x}; \mathbf{a}, A)\,\mathcal{N}(\mathbf{x}; \mathbf{b}, B) = \mathcal{N}(\mathbf{a}; \mathbf{b}, A + B)\,\mathcal{N}(\mathbf{x}; \mathbf{c}, C) \tag{2}$$

where $\mathbf{c} = C\left(A^{-1}\mathbf{a} + B^{-1}\mathbf{b}\right)$ and $C = \left(A^{-1} + B^{-1}\right)^{-1}$. Since $\mathcal{N}(\mathbf{x}; \mathbf{c}, C)$ integrates to one, the integral in (1) reduces to $p(g|\hat{\mathbf{x}}) \propto \mathcal{N}(\bar{\mathbf{x}}_g; \hat{\mathbf{x}}, \Sigma_g + \Sigma_w)$. Besides mathematical convenience, Gaussians are also motivated by the fact that the Gaussian is the maximum entropy (least assumptive) distribution for fixed mean and covariance.

For $\mathbf{z} = H\mathbf{x} + \mathbf{w}$ and given that $\hat{\mathbf{x}} = G\mathbf{z} = GH\mathbf{x} + G\mathbf{w}$ is some linear estimator with mean $E[\hat{\mathbf{x}}|\mathbf{x}] = GH\mathbf{x}$ and covariance $Cov[\hat{\mathbf{x}}|\mathbf{x}] = \sigma_w^2 GG^T$, we model $p(\hat{\mathbf{x}}|\mathbf{x})$ as $\mathcal{N}(\hat{\mathbf{x}}; GH\mathbf{x}, \sigma_w^2 GG^T)$. As above, we may also obtain a closed-form solution for (1) by re-expressing $p(\hat{\mathbf{x}}|\mathbf{x})$ as a Gaussian with argument $\mathbf{x}$. Rewriting $p(\hat{\mathbf{x}}|\mathbf{x})$ generally requires an approximation when $GH \neq I$, but is exact when $GH = I$, as we will now show.

Proposition. If $GH = I$, then for a linear estimator $\hat{\mathbf{x}} = G\mathbf{z}$ and invertible GG^T, Gaussian $p(\hat{\mathbf{x}}|\mathbf{x})$ can be rewritten as

$$p(\hat{\mathbf{x}}|\mathbf{x}) = \mathcal{N}\left(\mathbf{x}; G\mathbf{z}, \sigma_w^2\left(H^T H\right)^{-1}\right) \triangleq p(\mathbf{x}|\mathbf{z}; G).$$

Proof. Since $E[\hat{\mathbf{x}}|\mathbf{x}] = GH\mathbf{x}$ and $Cov[\hat{\mathbf{x}}|\mathbf{x}] = \sigma_w^2 GG^T$, rewrite the exponent of $\mathcal{N}(\hat{\mathbf{x}}; GH\mathbf{x}, \sigma_w^2 GG^T)$ as

$$-\frac{1}{2}(G\mathbf{z} - GH\mathbf{x})^T(\sigma_w^2 GG^T)^{-1}(G\mathbf{z} - GH\mathbf{x}) \tag{3}$$

$$= -\frac{1}{2}(\mathbf{x} - G\mathbf{z})^T \frac{H^T H}{\sigma_w^2}(\mathbf{x} - G\mathbf{z}), \tag{4}$$

where (4) follows from (3) by the assumption that $GH = I$ and by noting that

$$(\sigma_w^2 GG^T)^{-1} = H^T G^T(\sigma_w^2 GG^T)^{-1}GH = \frac{H^T H}{\sigma_w^2}.$$

To illustrate the proposition, consider the least squares (LS) estimate $\hat{\mathbf{x}} = G\mathbf{z}$ with $G = (H^T H)^{-1}H^T$. Then, since $GH = I$, we have $p(\hat{\mathbf{x}}|\mathbf{x}) = \mathcal{N}(\mathbf{x}; G\mathbf{z}, \sigma_w^2(H^T H)^{-1})$. Commonly $GH \approx I$. For example, for small σ_w^2 and invertible $H\Sigma H^T$ the linear minimum mean squared error (LMMSE) estimator has $G = \Sigma H^T(H\Sigma H^T + \sigma_w^2 I)^{-1}$ where $\Sigma = Cov[\mathbf{x}]$, so that (4) only approximates (3).

Rather than reformulate $p(\hat{\mathbf{x}}|\mathbf{x})$ as $p(\mathbf{x}|\mathbf{z}; G)$ for particular G, we introduce an alternative decision rule in which we model $p(\mathbf{x}|\mathbf{z})$ directly. This may be interpreted as a departure from the traditional MAP paradigm of maximizing $p(\mathbf{z}|g)p(g)$, instead discriminating by the *expected* MAP rule[1]: choose class g that maximizes

$$\int p(\mathbf{x}|g)\, p(g) p(\mathbf{x}|\mathbf{z})\; d\mathbf{x} = E_{\mathbf{x}|\mathbf{z}}\left[p(\mathbf{x}|g)\, p(g)\right]. \tag{5}$$

This rule generalizes the traditional MAP rule in the case of no uncertainty for which $p(\mathbf{x}|\mathbf{z}) = \delta(\mathbf{x} - \mathbf{z})$.

We model $p(\mathbf{x}|\mathbf{z})$ directly by assuming that $\mathbf{x}$ and $\mathbf{z}$ are jointly Gaussian, and thus $E[\mathbf{x}|\mathbf{z}] = F\mathbf{z}$ and $Cov[\mathbf{x}|\mathbf{z}] = \Sigma - FH\Sigma$ with $F = \Sigma H^T(H\Sigma H^T + \sigma_w^2 I)^{-1}$, where $\Sigma = Cov[\mathbf{x}]$. This is not equivalent to having formed the LMMSE estimator $\hat{\mathbf{x}} = F\mathbf{z}$ and rewriting $p(\hat{\mathbf{x}}|\mathbf{x}) \approx p(\mathbf{x}|\mathbf{z}; F)$, since the covariance matrices differ (see Table 1).

3. ROBUST LOCAL BDA

Generally, the likelihood $p(\mathbf{x}|g)$ in (1) or (5) is unknown; we estimate it from the training samples for each class. Here, we propose modeling the gth class-conditional likelihood as *locally* Gaussian. Localizing QDA reduces model bias and also generalizes the local nearest-means classifier [8]. Given a random test sample $\mathbf{x} \sim p(\mathbf{x}|\mathbf{z})$, we fit $p(\mathbf{x}|g)$ to the k nearest neighbor training vectors that belong to class g. Since $\mathbf{x}$ is random, the nearest neighbor to $\mathbf{x}$ is defined in terms of expected distance. Let $\hat{\mathbf{x}} \triangleq E[\mathbf{x}|\mathbf{z}]$ and $\Lambda \triangleq Cov[\mathbf{x}|\mathbf{z}]$, then the nearest neighbor to $\mathbf{x}$ is $\mathbf{x}_{i^*}$, where i^* solves

$$\begin{aligned} & \arg\min_i E_{\mathbf{x}|\mathbf{z}}\left[(\mathbf{x} - \mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i)\right] \\ \equiv\ & \arg\min_i \mathbf{x}_i^T\mathbf{x}_i - 2\mathbf{x}_i^T\hat{\mathbf{x}} + \operatorname{tr}\Lambda + \hat{\mathbf{x}}^T\hat{\mathbf{x}} \\ \equiv\ & \arg\min_i \|\mathbf{x}_i - \hat{\mathbf{x}}\|_2^2. \end{aligned}$$

Thus, the k nearest neighbors to random $\mathbf{x}$ are the k nearest neighbors to $\hat{\mathbf{x}}$. If fewer than k samples are available for class g, then we use all available training samples for the class, so that number of samples used for the gth class is $k_g = \min\left\{\left|\{\mathbf{x}_i \text{ s.t. } y_i = g\}\right|, k\right\}$.

Estimating the mean and covariance of a Gaussian using a small number of feature vectors can be ill-posed. Srivastava et al. addressed the problem of ill-posed Gaussian estimation by using a Bayesian estimate with a data-dependent inverted Wishart prior [9]. The Bayesian estimate is formed by marginalizing the unknown (random) Gaussian pdf over all

[1]The form in (5) is preferred to $E_{\mathbf{x}|\mathbf{z}}[p(g|\mathbf{x})]$ since the latter reduces to maximizing $p(g|\mathbf{z})$—the very problem we are trying to avoid.

Table 1. Estimation methods used to form $N(\mathbf{x}; \hat{\mathbf{x}}, \Lambda)$ and the corresponding $\hat{\mathbf{x}}$ and Λ.

method	$\hat{\mathbf{x}}$	Λ
`LS`: rewrite $p(\hat{\mathbf{x}}\|\mathbf{x}) = \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \Lambda)$	$(H^T H)^{-1} H^T \mathbf{z}$	$\sigma_w^2 (H^T H)^{-1}$
`LMMSE` : rewrite $p(\hat{\mathbf{x}}\|\mathbf{x}) \approx \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \Lambda)$	$\Sigma H^T (H \Sigma H^T + \sigma_w^2 I)^{-1} \mathbf{z}$	$\sigma_w^2 (H^T H)^{-1}$
`joint Gauss`: $p(\mathbf{x}\|\mathbf{z}) \stackrel{\triangle}{=} \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \Lambda)$ assuming $\mathbf{x}$ and $\mathbf{z}$ are jointly Gaussian	$\Sigma H^T (H \Sigma H^T + \sigma_w^2 I)^{-1} \mathbf{z}$	$\left(\Sigma^{-1} + \frac{H^T H}{\sigma_w^2}\right)^{-1}$

Gaussians that could describe the data. The resulting distribution is, in fact, not Gaussian. Rather, the Bayesian quadratic discriminant analysis (BDA) class-conditional likelihood is of the form [9, eq. 15]

$$p(\mathbf{x}|g) = \alpha_g \left(1 + \lambda \left(\mathbf{x} - \bar{\mathbf{x}}_g\right)^T R \left(\mathbf{x} - \bar{\mathbf{x}}_g\right)\right)^{-\gamma}, \quad (6)$$

where $\bar{\mathbf{x}}_g$ is the sample mean of the M_g features in class g, $R = (S_g + B)^{-1}$, $S_g = M_g \bar{\Sigma}_g$ is a scaled version of the maximum likelihood covariance estimate $\bar{\Sigma}_g$, B is a matrix parameter for the inverted Wishart prior with q degrees of freedom, $\gamma = \frac{M_g+q+1}{2}$, $\lambda = \frac{M_g}{M_g+1}$, and α_g is a class-dependent normalizing constant. Srivastava et al. showed good results for a data-dependent scale matrix B that pegs the maximum of the inverted Wishart prior (B/q) at a rough estimate of the class covariance [9]. In that work, Srivastava et al. cross-validate between seven different rough estimates of the class covariance.

We apply the estimator in [9] to the *local* samples for each class so that $p(\mathbf{x}|g)$ is given in (6) with $M_g = k_g$. For simplicity, rather than cross-validating over class covariance estimates, we propose to always use the (local) diagonal pooled sample covariance matrix, regularized slightly by the identity matrix to ensure numerical stability. Let the local pooled sample covariance matrix be denoted $\bar{\Sigma}_{pool}$, then

$$B = q\big(0.95 \, \mathrm{diag} \, \bar{\Sigma}_{pool} + 0.05 \, I\big),$$

where diag discards the off-diagonal elements. We let $q = d + 3$, a choice for which the inverted Wishart prior reduces to the inverted gamma distribution in the scalar case.

Using a Taylor series expansion, we have found a series solution for (1) and (5) with $p(\mathbf{x}|g)$ as in (6), but it exhibits poor convergence properties and is omitted for brevity. Rather, following [9], we approximate (6) by a Gaussian using the fact that $e^{\epsilon} \approx 1 + \epsilon$. For $\epsilon = \lambda \left(\mathbf{x} - \bar{\mathbf{x}}_g\right)^T R \left(\mathbf{x} - \bar{\mathbf{x}}_g\right)$, (6) becomes $p(\mathbf{x}|g) \propto \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}_g, \hat{\Sigma}_g)$, where $\hat{\Sigma}_g = \frac{k_g+1}{k_g+q+1} \left(\frac{S_g+B}{k_g}\right)$.

Given that $p(\mathbf{x}|g)$ and $p(\mathbf{x}|\mathbf{z})$ are Gaussian, by using (2) and the fact that the $\mathcal{N}(\mathbf{x}; \mathbf{c}, C)$ integrates to one, the decision rules in (1) and (5) reduce to

$$\arg\max_{g \in \mathcal{G}} \mathcal{N}\left(\hat{\mathbf{x}}; \bar{\mathbf{x}}_g, \hat{\Sigma}_g + \Lambda\right) p(g). \quad (7)$$

4. EXPERIMENTS

The `optical` benchmark dataset contains 8×8 images of handwritten digits "0" through "9": 3823 training and 1797 test. Prior to experiments, training and test are normalized by the mean and variance of training data. Test data are corrupted by Gaussian blur (std $\sigma = 0.5$ with 4×4 pixel support) then adding zero mean Gaussian white noise with standard deviation σ_w. We compare three different estimation approaches listed in Table 1 to classify each test image $\mathbf{z}$: evaluate (1) with $\hat{\mathbf{x}}$ as the LS estimator (`LS`), evaluate (1) by rewriting $p(\hat{\mathbf{x}}|\mathbf{x})$ with $\hat{\mathbf{x}}$ as the LMMSE estimator (`LMMSE`), and evaluate (5) assuming $\mathbf{x}$ and $\mathbf{z}$ are jointly Gaussian (`joint Gauss`). Note that `LS` differs from `LMMSE` only by choice of $\hat{\mathbf{x}}$, whereas `LMMSE` differs from `joint Gauss` only by Λ. For each of these estimators, we compare the proposed R-BDA classifier in (6) (`r-bda`) to the smoothing kernel introduced by Pawlak and Siu in [6] (`pawlak`):

$$\arg\max_{g \in \mathcal{G}} \sum_{i=1}^{M} I_{y_i = g} W\left(\frac{\hat{\mathbf{x}} - \mathbf{x}_i}{\sqrt{b}}\right) \mathcal{N}(\mathbf{x}_i; \hat{\mathbf{x}}, \Lambda),$$

where W is a Gaussian kernel with bandwidth parameter b. Both `r-bda` and `pawlak` use $\hat{\mathbf{x}}$ and Λ, and $\Sigma = Cov(\mathbf{x})$ is estimated using the pooled sample covariance matrix.We also compare to non-robust classifiers that use only $\hat{\mathbf{x}}$; an SVM with radial basis function (RBF) kernel (`svm`), the k-NN classifier (`knn`) and the BDA classifier (`bda`) in (6) localized to k nearest neighbors.

We vary σ_w to observe the sensitivity of each classifier to uncertainty in the test data. For each choice of σ_w, classifier parameters for each classification scheme are determined via 5-fold cross-validation on a holdout set of training images that are artificially corrupted using the same blur and noise model as the test images. Parameter choices are listed in Table 2. Results shown in Fig. 1 were averaged over 100 runs of i.i.d. test noise per image for each choice of σ_w.

We perform a second experiment by varying the bandwidth of Gaussian blur σ for fixed $\sigma_w = 0.3$.

5. RESULTS AND CONCLUSIONS

Incorporating the test sample uncertainty using the covariance Λ greatly improves performance as seen by comparing the solid and dashed lines, especially for `LS` in both experiments.

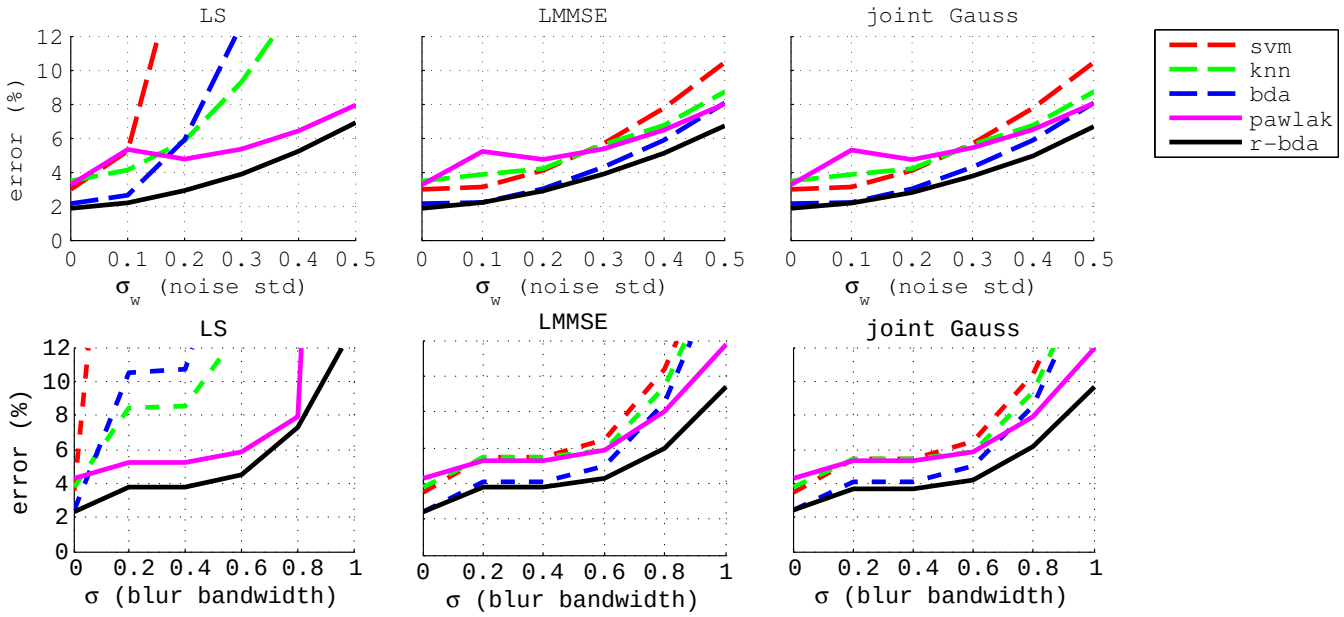

Fig. 1. Classification error as a function of (top row) the noise std σ_w and (bottom row) the blur standard deviation σ. Results for `svm`, `knn` and `bda` are identical for `LMMSE` and `joint Gauss` since $\hat{\mathbf{x}}$ is the same.

Table 2. Cross-validation parameters for each point in Fig. 1.

parameter	cross-validation set
RBF bandwidth (`svm`)	$\{5, 10, 20, 40, 80, 160, 320\}$
k (`knn`, `bda`, `r-bda`)	$\{1, 3, 5, 9, 17, 33, 65\}$
bandwidth b (`pawlak`)	$\{1, 2, 5, 10, 20, 50, 100\}$

Although `pawlak` also incorporates Λ, it generally performs poorly when σ_w or σ are small, but increases in relative performance with increasing uncertainty. Among the non-robust classifiers, `bda` is the best performer for `LMMSE` (equivalently, `joint Gauss`), and all classifiers perform objectionably for `LS`. For high noise and high blur, the `r-bda` classifier performs better with `LMMSE` and `joint Gauss` than with `LS`.

For the first experiment at $\sigma_w = 0$, `bda` and `r-bda` both chose the same neighborhood size $k = 17$ in cross-validation, thus, any discrepancy should be ascribed to the Gaussian approximation used in `r-bda`, but according to a one-sided Wilcoxon rank test with 95% confidence, the methods are statistically tied.

6. REFERENCES

[1] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz, "Cepstral vector normalization based on stereo data for robust speech recognition," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 15, no. 3, pp. 109–1113, March 2007.

[2] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, no. 3, pp. 267–285, June 2001.

[3] H. S. Anderson and M. R. Gupta, "Joint deconvolution and classification with applications to passive acoustic underwater multipath," *J. Acous. Soc. Am.*, vol. 124, no. 5, pp. 2973–2983, November 2008.

[4] R. Duda, E. Hart, and D. Stork, *Pattern Classification*, Wiley-Interscience, 2nd edition, 2001.

[5] G. Papandreou, A. Katsamanis, V. Pitsikalis, and P. Maragos, "Adaptive multimodal fusion by uncertainty compensation with application to audio-visual speech recognition," in *Multimodal Processing and Interaction: Audio, Video, Text*. Springer, 2008.

[6] M. Pawlak and D. Siu, "Pattern classification with noisy features," *Lec. Notes in Comp. Sci.*, vol. 1451, pp. 845–852, 1998.

[7] S. Ahmad and V. Tresp, "Some solutions to the missing feature problem in vision," *NIPS*, 1993.

[8] Y. Mitani and Y. Hamamoto, "A local mean-based nonparametric classifier," *Patt. Rec. Letters*, vol. 27, pp. 1151–1159, 2006.

[9] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis," *J. Mach. Learn. Res.*, vol. 8, pp. 1277–1305, 2007.