REFERENCES

[1] H. Jaeger, "The 'echo state' approach to analysing and training recurrent neural networks," German Nat. Res. Cntr. Inf. Technol., Sankt Augustin, Germany, 2001, GMD Report 148.

[2] H. Jaeger, "A tutorial on training recurrent neural networks, covering BPPT, RTRL, EKF and the 'echo state network' approach," German Nat. Res. Cntr. Inf. Technol., Sankt Augustin, Germany, 2002, GMD Report 159.

[3] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *Science*, vol. 304, pp. 78–80, 2004.

[4] G.-B. Huang, L. Chen, and C.-K. Siew, "Universal approximation using incremental constructive feedforward networks with random hidden nodes," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 879–892, Jul. 2006.

[5] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Real-time learning capability of neural networks," *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 863–878, Jul. 2006.

[6] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, pp. 489–501, 2006.

[7] G.-B. Huang and L. Chen, "Convex incremental extreme learning machine," *Neurocomputing*, vol. 70, pp. 3056–3062, 2007.

[8] D. Erdogmus, O. Fontenla-Romero, J. C. Principe, A. Alonso-Betanzos, and E. Castillo, "Linear-least-squares initialization of multilayer perceptrons through backpropagation of the desired response," *IEEE Trans. Neural Netw.*, vol. 16, no. 2, pp. 325–337, Mar. 2005.

[9] N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," *IEEE Trans. Neural Netw.*, vol. 17, no. 6, pp. 1411–1423, Nov. 2006.

[10] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, Budapest, Hungary, Jul. 25–29, 2004, vol. 2, pp. 985–990.

[11] G.-B. Huang, N.-Y. Liang, H.-J. Rong, P. Saratchandran, and N. Sundararajan, "On-line sequential extreme learning machine," in *Proc. IASTED Int. Conf. Comput. Intell.*, Calgary, AB, Canada, Jul. 4–6, 2005, pp. 232–237.

[12] T. Kim and T. Adali, "Approximation by fully complex multilayer perceptrons," *Neural Comput.*, vol. 15, pp. 1641–1666, 2003.

[13] M.-B. Li, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "Fully complex extreme learning machine," *Neurocomputing*, vol. 68, pp. 306–314, 2005.

[14] L. Breiman, "Hinging hyperplanes for regression, classification, and function approximation," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 999–1013, May 1993.

[15] C. L. Giles and T. Maxwell, "Learning, invariance, and generalization in high-order neural networks," *Appl. Opt.*, vol. 26, no. 23, pp. 4972–4978, 1987.

[16] Y. Shin and J. Ghosh, "Approximation of multivariate functions using ridge polynomial networks," in *Proc. Int. Joint Conf. Neural Netw.*, Baltimore, MD, Jun. 2002, pp. 380–385.

[17] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 41, pp. 909–996, 1988.

[18] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 961–1005, Sep. 1990.

[19] F. Han and D.-S. Huang, "Improved extreme learning machine for function approximation by encoding a priori information," *Neurocomputing*, vol. 69, pp. 2369–2373, 2006.

[20] G.-B. Huang, "Learning capability and storage capacity of two-hidden-layer feedforward networks," *IEEE Trans. Neural Netw.*, vol. 14, no. 2, pp. 274–281, Mar. 2003.

[21] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and its Applications*. New York: Wiley, 1971.

[22] S. Boyd, L. E. Ghaoul, E. Feron, and V. Balakrishnan, *Linear Matrix Inequalities in System and Control Theory*. Philadelphia, PA: SIAM, 1994.

[23] J. Platt, "A resource-allocating network for function interpolation," *Neural Comput.*, vol. 3, pp. 213–225, 1991.

[24] L. Yingwei, N. Sundararajan, and P. Saratchandran, "A sequential learning scheme for function approximation using minimal radial basis function (RBF) neural networks," *Neural Comput.*, vol. 9, pp. 461–478, 1997.

[25] C. Blake and C. Merz, UCI Repository of Machine Learning Databases, Dept. Inf. Comput. Sci., Univ. California, Irvine, CA, 1998 [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

Model Selection Criteria for Image Restoration

Abd-Krim Seghouane

***Abstract*—In this brief, the image restoration problem is approached as a learning system problem, in which a model is to be selected and parameters are estimated. Although the parameters which correspond to the restored image can easily be obtained, their quality depend heavily on a proper choice of the regularization parameter that controls the tradeoff between fidelity to the blurred noisy observed image and the smoothness of the restored image. By analogy between the model selection philosophy that constitutes a fundamental task in systems learning and the choice of the regularization parameter, two criteria are proposed in this brief for selecting the regularization parameter. These criteria are based on Bayesian arguments and the Kullback–Leibler divergence and they can be considered as extensions of the Bayesian information criterion (BIC) and the Akaike information criterion (AIC) for the image restoration problem.**

***Index Terms*—Akaike information criterion (AIC), Bayesian information criterion (BIC), image restoration, model selection, regularization.**

I. INTRODUCTION

Image restoration is an important problem of image processing which has been extensively studied [1]–[5]. The aim is to construct a good estimate of the original image from noisy, degraded observations. This problem is, for example, encountered in astronomical imaging, ultrasound imaging, and radar imaging. In most cases, the standard statistical model used to relate the observations to the unknown underlying image is given by

$$g = Hf + w \tag{1}$$

where the $n \times 1$ vectors f, g, and w represent, respectively, the original image, the observed image, and the noise with independent identically distributed (i.i.d.) Gaussian elements of mean zero and variance σ_w^2. The matrix $H \in R^{n\times n}$ represents the known distortion blurring matrix, and it has as elements samples of the point spread function (PSF) of the image system. H can have a special structure depending on the properties of the PSF. For the general convolutional case, H is a block diagonal matrix. The images are assumed to be of support $p \times m$ pixels where p and m, respectively, represent the length and the width of the image. The support of the above vectors is then $n = p \times m$.

The simplest way to obtain an estimate $\hat{f}$ from (1) is to select $\hat{f}$ as a minimizer of the least squares error criterion

$$J(f, g) = \|g - Hf\|^2 \tag{2}$$

Manuscript received July 11, 2007; revised October 23, 2008, January 30, 2009, April 21, 2009, and April 28, 2009; accepted May 21, 2009. First published July 10, 2009; current version published August 05, 2009. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

The author is with the Canberra Research Laboratory, National ICT Australia (NICTA), Canberra, A.C.T. 2601, Australia and also with the College of Engineering and Computer Science, The Australian National University, Canberra, A.C.T. 0200, Australia (e-mail: Abd-krim.seghouane@nicta.com.au).

Digital Object Identifier 10.1109/TNN.2009.2024146

1045-9227/$26.00 © 2009 IEEE

which results in the pseudoinverse estimate

$$\hat{f} = (H^T H)^{-1} H^T g. \tag{3}$$

However, as is well known, this is an ill-posed problem, which means that this estimate does not lead to a suitable restoration [1].

A classical framework in which to solve such an image estimation problem is regularized estimation. The estimator is given by

$$\hat{f}_\lambda = \arg\min_f \|g - Hf\|^2 + \lambda \|Qf\|^2 \tag{4}$$

which yields the estimate

$$\hat{f}_\lambda = (H^T H + \lambda Q^T Q)^{-1} H^T g. \tag{5}$$

Assume that

$$p\left(g|f, \sigma_w^2\right) = \frac{1}{(2\pi\sigma_w^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_w^2}\|g - Hf\|^2\right\} \tag{6}$$

represents the probability density of g given f and that

$$p\left(f|\sigma_f^2\right) \propto \sigma_f^{-n} \exp\left\{-\frac{1}{2\sigma_f^2} S(f)\right\} \tag{7}$$

where $S(f)$ is a nonnegative quadratic form, denotes the prior probability density of f dependent on a parameter σ_f^2, which in the Gaussian case represents the variance of the prior. A conventional choice for $S(f)$, which is widely used in image processing, is the quadratic smoothness penalty $S(f) = \|Qf\|^2$ [1] (implicitly assuming a Gaussian prior on f), where Q is a differentiating operator [6], [7]. A smoothness requirement on the solution can be imposed by requiring Q being a high-pass filter, for example, Q is the 2-D Laplacian operator [8]. This choice of penalty yields the estimate (5) which penalizes large local variations in the image. For the Gaussian model (1), the estimate $\hat{f}_\lambda$ is linear in the data g.

The Bayesian interpretation of (4) is given by

$$\begin{aligned}\hat{f}_\lambda &= \arg\max_f p\left(f|g, \sigma_w^2, \sigma_f^2\right) \\ &= \arg\max_f p\left(g|f, \sigma_w^2\right) p\left(f|\sigma_f^2\right) \\ &= \arg\max_f p(g|f) p(f|\lambda).\end{aligned} \tag{8}$$

It is commonly assumed in image restoration that $\hat{f}_\lambda$ is a much more preferable solution than $\hat{f}$. However, the quality of the restored image depends heavily on a proper choice of the parameter $\lambda = \sigma_w^2/\sigma_f^2$ called the regularization parameter. Therefore, it is important to choose the regularization parameter judiciously because it controls the tradeoff between fidelity to the observations (the log-likelihood term) and the smoothness of the estimated image (the regularization term). While it is sometimes the case that a range of values of the regularization provides an acceptable estimated image, it is important to be in the correct window of values.

Various approaches have been introduced in the literature for estimating λ. Among them, one can cite mean square error (MSE)-based approaches, the Chi-squared method (CHI), and the equivalent degrees of freedom method (EDF) [9]. MSE-based approaches require the *a priori* knowledge of f and σ_w^2, whereas CHI and EDF methods require only the knowledge of σ_w^2. However, in many practical situations, the noise variance σ_w^2 is not known; therefore, these methods cannot be applied directly. Maximum likelihood (ML) and cross validation (CV) are the only methods that allow the selection of the regularization parameter when σ_w^2 is unknown [10]. These methods are totally data based, requiring no *a priori* knowledge whatsoever. However, the ML method has the tendency to produce oversmoothed solutions [11] whereas CV can fail in some circumstances producing either no positive smoothing parameter or a grossly underestimated regularization parameter [12].

Learning from an experimental data set consists of two tasks, the choice of an appropriate model structure and estimating its parameters. The task of parameter estimation is generally achieved by ML or least squares procedures given the structure or dimension of the model. In image restoration problems, the unknown original image represents the parameter vector and its estimate is given by (5). The choice of the dimension of a model is often facilitated by the use of a model selection criterion where one only has to evaluate two simple terms. The underlying idea of model selection criteria is the parsimonious principle which says that there should be a tradeoff between data fitting and complexity. In image restoration problems, the choice of the model structure corresponds to the choice of the regularization parameter that controls the tradeoff between fidelity to the blurred noisy observed image and the smoothness of the restored image.

Model selection criteria are powerful tools that have not yet been applied to image restoration problems. Based on different approaches, different model selection criteria have been proposed in the literature. The first criterion which has gained widespread acceptance was the Akaike information criterion (AIC), which is based on information theoretic arguments [13]. In [14], based on Bayesian arguments and maximum *a posteriori* probability, the Bayesian information criterion (BIC) was introduced.

In this brief, the application of the model selection approach to the problem of image restoration is proposed. By extending the AIC and BIC derivation ideas, improved variants of AIC and BIC are developed for selecting the regularization parameters to control the amount of smoothness of the restored image $\hat{f}_\lambda$. These proposed criteria can be considered as simple totally data-based alternatives to the ML and CV methods for choosing the regularization parameter without knowledge of σ_w^2.

II. The Kullback–Leibler Divergence Approach

A. Review of the AIC

Suppose a collection of observed data $\mathbf{y} = (y_1, \ldots, y_n)$ has been sampled from an unknown distribution $G(\mathbf{y})$ having density function $g(\mathbf{y})$. Estimation of $g(\mathbf{y})$ is done within a set of candidate models $M_1, \ldots, M_K$ characterized by probability densities $f(\mathbf{y}|\theta_k)$, $k = 1, \ldots, K$, where $\theta_k \in \Theta_k \subset R^k$. It is also assumed that the generating density model g is a member of the approximating family of models; $\exists\theta_0 \in \Theta_K$ such that $g(\cdot) = f(\cdot, \theta_0)$ [13]. Let $\hat{\theta}(\mathbf{y})$ denote the vector of estimated parameters obtained by maximizing the likelihood $f(\mathbf{y}|\theta_k)$ over Θ_k and let $f(\mathbf{y}|\hat{\theta}_k)$ denote the corresponding fitted model.

To determine which candidate density model best approximates the true unknown model $g(\mathbf{y})$, we require a measure which provide a suitable reflection of the separation between $g(\mathbf{y})$ and an approximating model $f(\mathbf{y}|\hat{\theta}_k)$. The Kullback–Leibler divergence also known as the cross entropy or discrepancy is one of such measure.

For the two probability densities $g(\mathbf{y})$ and $f(\mathbf{y}|\hat{\theta}_k)$, the Kullback–Leibler divergence between $g(\mathbf{y})$ and $f(\mathbf{y}|\hat{\theta}_k)$ with respect to $g(\mathbf{y})$ is defined as

$$\begin{aligned}I_n(g(\cdot), f(\cdot|\hat{\theta}_k)) &= E_g\left\{2\ln\frac{g(\mathbf{y})}{f(\mathbf{y}|\hat{\theta}_k)}\right\} \\ &= E_g\left\{-2\ln f(\mathbf{y}|\hat{\theta}_k)\right\} - E_g\left\{-2\ln g(\mathbf{y})\right\} \\ &= d_n(g, f_k) - d_n(g, g)\end{aligned}$$

where

$$d_n(g, f_k) = E_g\{-2\ln f(\mathbf{y}|\hat{\theta}_k)\} \tag{9}$$

and $E_g\{\cdot\}$ represents the expectation with respect to $g(\mathbf{y})$. By splitting the observed data into several subsamples $\mathbf{y}^{(1)} = (y_1, \ldots, y_r)$,$\mathbf{y}^{(2)} = (y_{r+1}, \ldots, y_{2r})$,$\ldots, \mathbf{y}^{(k)} = (y_{(k-1)r+1}, \ldots, y_n)$, the empirical computation of $E_g\{.\}$ and (9) is obtained by

$$\hat{d}_n(g, f_k) = \frac{1}{k}\sum_{i=1}^{k} -2\ln f\left(\mathbf{y}^{(i)}|\hat{\theta}_k(\mathbf{y}^{(i)})\right).$$

Since $d_n(g,g)$ does not depend on θ_k, any ranking of the candidate models according to $I_n(g(\cdot), f(\cdot|\hat{\theta}_k))$ would be identical to ranking them according to $d_n(g, f_k)$.

The above discussion suggests that

$$\begin{aligned} d_n(g, f_k) &= E_g\left\{-2\ln f(\mathbf{y}|\hat{\theta}_k)\right\} \\ &= -2\ln f(\mathbf{y}|\hat{\theta}_k) \\ &\quad + E_g\left\{-2\ln f(\mathbf{y}|\hat{\theta}_k)\right\} - \{-2\ln f(\mathbf{y}|\hat{\theta}_k)\} \end{aligned} \quad (10)$$

would provide a suitable estimate of the Kullback–Leibler divergence up to the order of a constant between the generating model $g(\mathbf{y})$ and the candidate model $f(\mathbf{y}|\hat{\theta}_k)$. Yet evaluating $d_n(g, f_k)$ is not possible, since doing so requires the knowledge of $g(\mathbf{y})$.

However, as noted in [13], $-2\ln f(\mathbf{y}|\hat{\theta}_k)$ serves as a biased estimator of (9) and that, under proper regularity conditions [15], the bias adjustment

$$E_g\left\{E_g\left\{-2\ln f(\mathbf{y}|\hat{\theta}_k)\right\}\right\} - E_g\left\{-2\ln f(\mathbf{y}|\hat{\theta}_k)\right\} \quad (11)$$

can often be asymptotically estimated by $2k$. Based on such observation, the use of

$$\text{AIC} = -2\ln f(\mathbf{y}|\hat{\theta}_k) + 2k \quad (12)$$

was proposed in [13] as a criterion for model selection

$$\hat{k} = \arg\min_{k\in\{1,\ldots,K\}} \text{AIC}.$$

If we denote

$$\Delta_n(k, g) = E_g\{d_n(g, f_k)\}$$

then the following approximation holds [13]:

$$\Delta_n(k, g) = E_g\{\text{AIC}\} + o(1).$$

The penalty term in AIC is a simple minded bias correction to the log maximum likelihood and there is no assurance that such a bias correction yields a good estimate of the Kullback–Leibler divergence. Indeed, in [16], it was shown that in parametric linear regression and autoregressive time-series contexts the bias corrected AIC takes the form

$$\text{AIC}c = -2\ln f(\mathbf{y}, \hat{\theta}_k(\mathbf{y})) + 2\frac{(k+1)n}{n-k-2}. \quad (13)$$

B. Improved AIC for Regularized Parameter Selection

In this section, an information approach to the regularization parameter selection for the image restoration problem is introduced.

Given a family of probability densities $p(g|f, \lambda)$ characterized by a parameter vector f, the regularization parameter value corresponding to the density function from the specified family that matches the unknown density $p(g)$ most closely is chosen.

When the unknown image f is estimated by (4), each particular choice of the regularization parameter λ yields some approximating density $p(g|\hat{f}_\lambda, \lambda)$. The closeness of this approximating density $p(g|\hat{f}_\lambda, \lambda)$ to the unknown density $p(g)$ can be evaluated by the Kullback–Leibler divergence between these densities

$$I(p(g), p(g|\hat{f}_\lambda, \lambda)) = E_{p(g)}\{\log p(g)\} - E_{p(g)}\{\log p(g|\hat{f}_\lambda, \lambda)\} \quad (14)$$

where $E_{p(g)}\{\cdot\}$ represents the expectation with respect to $p(g)$.

The regularization parameter can then be estimated to be the minimizer of the Kullback–Leibler divergence

$$\hat{\lambda} = \arg\min_{\lambda} I(p(g), p(g|\hat{f}_\lambda, \lambda)).$$

As before, the minimization of $I(p(g), p(g|\hat{f}_\lambda, \lambda))$ is equivalent to the maximization of the expected log likelihood $E_{p(g)}\{\log p(g|\hat{f}_\lambda, \lambda)\}$. Since $p(g)$ is unknown, the empirical distribution is used instead, and the expected log likelihood is estimated by the average log likelihood $\log p(g|\hat{f}_\lambda, \lambda)/n$. However, it is well known that this estimation introduces a bias in estimating

$$\text{bias} = E_{p(g)}\left\{\frac{1}{n}\log p(g|\hat{f}_\lambda, \lambda) - E_{p(g)}\{\log p(g|\hat{f}_\lambda, \lambda)\}\right\} \quad (15)$$

that should be corrected.

Theorem: The asymptotic bias of the average log likelihood $\log p(g|\hat{f}_\lambda, \lambda)/n$ fitted by (4) in estimating the expected log likelihood $E_{p(g)}\{\log p(g|\hat{f}_\lambda, \lambda)\}$ is given by

$$\text{bias} = \frac{1}{n}\text{tr}[H(H^T H + \lambda Q^T Q)^{-1} H^T] + o(1). \quad (16)$$

Proof: See the Appendix.

Therefore

$$\text{AIC}_\lambda = -\frac{2}{n}\log p(g|\hat{f}_\lambda, \lambda) + \frac{2}{n}\text{tr}[H(H^T H + \lambda Q^T Q)^{-1} H^T] \quad (17)$$

is an asymptotically unbiased estimator of $E_{p(g)}\{\log p(g|\hat{f}_\lambda, \lambda)\}$. Minimizing (17) provides the smoothing parameter estimate $\hat{\lambda}_{\text{AIC}}$

$$\hat{\lambda}_{\text{AIC}} = \arg\min_{\lambda} \text{AIC}_\lambda.$$

A corrected version of AIC for smoothing parameter selection in image restoration problem can be obtained by replacing k in (13) by the trace of the matrix $H(H^T H + \lambda Q^T Q)^{-1} H^T$.

III. The Bayesian Approach

A. Review of the BIC

The motivation behind BIC [17] can be seen through a Bayesian development of the model selection problem. Given the observed data $\mathbf{y}$ and the set of family of candidate models $M_1, \ldots, M_K$, the approximating model M_k which is *a posteriori* most probable is chosen. By Bayes theorem, the posterior probability of the kth family of candidate models is defined by

$$p(M_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k)p(M_k)}{p(\mathbf{y})} \quad (18)$$

where $p(\mathbf{y}|M_k)$ represents the marginal density of the data given they are generated by the model M_k, $p(M_k)$ represents the prior probability of the model M_k, and

$$p(\mathbf{y}) = \sum_{k=1}^{K} p(\mathbf{y}|M_k)p(M_k)$$

represents the marginal density of the data.

To find the best family of candidate models, we evaluate $p(M_k|\mathbf{y})$ for $k = 1, \ldots, K$ and select the model that maximizes $p(M_k|\mathbf{y})$

$$\hat{k} = \arg \max_{k \in \{1,\ldots,K\}} p(M_k|\mathbf{y}).$$

Since $p(\mathbf{y})$ is not a function of M_k, it is a common factor for all models and then does not affect the model selection procedure. If the prior probabilities $p(M_k)$ are assumed equal for all the families of candidate models, the best model corresponds then to the one that maximizes the probability $p(\mathbf{y}|M_k)$. This density can be evaluated from

$$p(\mathbf{y}|M_k) = \int p(\mathbf{y}|\theta_k, M_k) p(\theta_k|M_k) d\theta_k \tag{19}$$

where $p(\mathbf{y}|\theta_k, M_k)$ represents the likelihood for $\mathbf{y}$ based on M_k and $p(\theta_k|M_k)$ denote the prior on the parameter vector θ_k given the model M_k. The evaluation of the marginal density $p(\mathbf{y}|M_k)$ requires, in general, multidimensional integration. One way of evaluating it is using Laplace approximation [18]. In the case of flat prior $p(\theta_k|M_k) = 1$ and under certain regularity conditions [19], the Laplace approximation to (19) is

$$p(\mathbf{y}|M_k) = \frac{(2\pi)^{k/2}}{n^{k/2}|Q(\hat{\theta}_k)|^{1/2}} \exp\{p(\mathbf{y}|\hat{\theta}_k, M_k)\}\left(1 + O_p(n^{-1})\right) \tag{20}$$

where $\hat{\theta}_k$, as defined in Section II-A, is the ML estimate and

$$Q(\hat{\theta}_k) = -\frac{1}{n} \frac{\partial^2 \ln p(\mathbf{y}|\theta_k, M_k)}{\partial\theta\partial\theta^T}\bigg|_{\theta_k = \hat{\theta}_k}.$$

The BIC is obtained by taking $-2\log$ of (20) and ignoring the terms of order $O(1)$ and higher

$$\text{BIC} = -2\ln p(\mathbf{y}|\hat{\theta}_k, M_k) + k\ln(n). \tag{21}$$

The expression (21), however, cannot be used directly for the choice of the regularization parameter. In what follows, the above method used to derive BIC is used to approach the regularization parameter estimation problem in a model selection framework.

B. Improved BIC for Regularized Parameter Selection

In this section, BIC is extended so that it can be applied to the selection of the regularization parameter in image restoration problems.

The restored image $\hat{f}_\lambda$ is estimated by maximizing the penalized log-likelihood function, which is defined by (8)

$$l(f_\lambda) \propto 2\ln p(g|f_\lambda) - \lambda\|Qf\|^2.$$

The penalty term corresponds to a multivariate normal prior density $p(f|\lambda)$

$$p(f|\lambda) = (2\pi)^{-n/2}\lambda^{n/2}|Q|\exp\left(-\frac{\lambda}{2}f^TQ^2f\right)$$

where Q is an $n \times n$ matrix. The quantity of interest $p(g|\lambda)$ is obtained by integrating over $f \in R^n$

$$\begin{aligned} p(g|\lambda) &= \int \exp\{\ln(p(f|g,\lambda))\}df \\ &= \int \exp\{\ln p(g|f) + \ln p(f|\lambda)\}df. \end{aligned} \tag{22}$$

The Laplace approximation of (22) is given by

$$p(g|\lambda) = \frac{(2\pi)^{n/2}}{n^{n/2}|Q(\hat{f}_\lambda)|^{1/2}} \exp\{p(\hat{f}_\lambda|g,\lambda)\} \tag{23}$$

where

$$Q(\hat{f}_\lambda) = -\frac{1}{n}\frac{\partial^2 \ln p(f|g,\lambda)}{\partial f \partial f^T}\bigg|_{f=\hat{f}_\lambda} = \frac{1}{n}(H^TH + \lambda Q^TQ).$$

It follows from (8) and (23) that $-2\ln p(g|\lambda)$ can be approximated by

$$\begin{aligned} -2\ln p(g|\lambda) \simeq -2\ln p(g|\hat{f}_\lambda) + \lambda\hat{f}_\lambda^TQ^TQ\hat{f}_\lambda - n\ln(\lambda) \\ + \ln(|H^TH + \lambda Q^TQ|) - 2\ln(|Q|). \end{aligned}$$

Removing the terms that do not dependent on λ provides

$$\begin{aligned} \text{BIC}_\lambda = -2\ln p(g|\hat{f}_\lambda) + \lambda\hat{f}_\lambda^TQ^TQ\hat{f}_\lambda - n\ln(\lambda) \\ + \ln(|H^TH + \lambda Q^TQ|) \end{aligned} \tag{24}$$

where $\hat{f}_\lambda$ is a solution to the equation

$$2\frac{\partial \ln p(f|g,\lambda)}{\partial f} = \frac{\partial}{\partial f}\left\{2\ln p(g|f) - \lambda f^TQ^2f\right\} = 0$$

which for model (1) corresponds to $\hat{f}_\lambda = (H^TH + \lambda Q^TQ)^{-1}H^Tg$.

Minimizing (24) provides the smoothing parameter estimate $\hat{\lambda}_{\text{BIC}}$

$$\hat{\lambda}_{\text{BIC}} = \arg\min_\lambda \text{BIC}_\lambda.$$

IV. Simulation Examples

The performances of the proposed techniques for choosing the regularization parameter λ are illustrated using the three 512 $times$ 512 images. The PSF, used to blur these original images, is the Gaussian-shaped expressed as

$$h(i,j) = c \cdot \exp\left\{-\frac{i^2 + j^2}{2\sigma^2}\right\}, \qquad i,j = 0,1,\ldots,n-1$$

where c is a constant, which assures the response system to be lossless. Due to the high-pass character of Laplace operator filter, this operator is selected as the regularization operator.

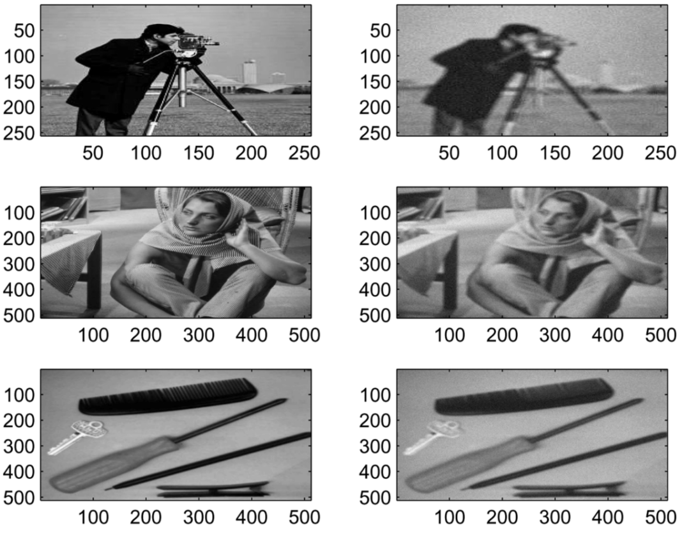
Of particular interest is the comparative performance of the two proposed criteria for choosing λ and the two other totally data-based methods cited in the introduction. As objective measurements of the performance of these different methods, the improvement in signal-to-noise ratio (ISNR) defined by

$$\text{ISNR} = 20\log_{10}\frac{\|f - g\|_2}{\|f - \bar{f}\|_2}$$

and the peak signal-to-noise ratio (PSNR) defined by

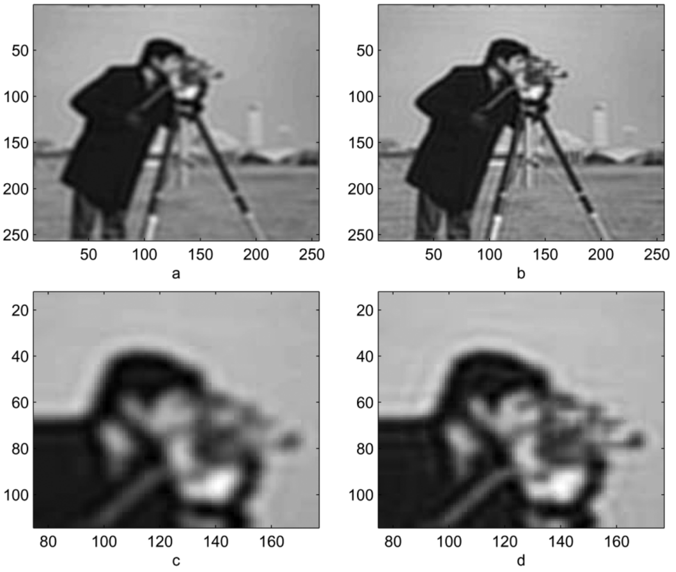
$$\text{PSNR} = 10\log_{10}\frac{nV^2}{\|f - \bar{f}\|^2}$$

where f, g, and $\bar{f}$ are, respectively, the original, degraded, and restored images, are used. Since the images tested are gray level images, the value of $V = 255$. The Gaussian-shaped PSF used to blur



(a) (b)

Fig. 1. (a) Original images and (b) noisy blurred image versions $\sigma_w^2 = 100$.



a b c d

Fig. 2. Restored cameraman image obtained form the noisy blurred cameraman image $\sigma_w^2 = 150$ with $\hat{\lambda}_{\rm AIC}$ for (a), (c-zoom) and $\hat{\lambda}_{\rm BIC}$ for (b) (d-zoom).

the original image had variance $\sigma^2 = 3$. Each blurred image was further degraded by adding independent Gaussian noise having variance $\sigma_w^2 = 10$, $\sigma_w^2 = 100$, and $\sigma_w^2 = 150$ and uniform noise. The original and noisy blurred images with $\sigma_w^2 = 100$ are displayed in Fig. 1.

An example of restored images obtained by the two choices of the regularization parameter $\lambda_{\rm BIC}$ and $\lambda_{\rm AIC}$ are displayed in Fig. 2. This figure illustrates the difference between the restored images obtained form the noisy blurred cameraman image with $\sigma_w^2 = 150$; a zoom on the image is included to highlight the difference.

In the tabulated results of Tables I and II, the three blurred images are denoted I1, I2, and I3, while the four different noises are denoted G1 (for Gaussian noise with variance $\sigma_w^2 = 10$), G2 (for $\sigma_w^2 = 100$), G3 (for $\sigma_w^2 = 150$), and U (for uniform noise). The values of the regularization parameter were found in all the above cases using the methods described in Sections II and III and [10] (CV and ML); and they are tabulated in Table I.

TABLE I
λ VALUES OF THE DIFFERENT METHODS FOR VARIOUS NOISES

Images	λ_{BIC}	λ_{AIC}	λ_{ML}	λ_{CV}
I1G1	0.09	0.28	0.33	0.12
I1G2	0.13	0.53	0.73	0.17
I1G3	0.18	0.65	0.77	0.22
I1U	0.08	0.18	0.22	0.04
I2G1	0.07	0.23	0.29	0.08
I2G2	0.30	0.57	0.77	0.18
I2G3	0.35	0.72	0.83	0.27
I2U	0.05	0.17	0.27	0.06
I3G1	0.22	0.35	0.67	0.16
I3G2	0.62	0.74	0.81	0.57
I3G3	0.71	0.82	0.88	0.68
I3U	0.16	0.34	0.43	0.14

TABLE II
ISNR AND PSNR VALUES CORRESPONDING TO THE DIFFERENT CHOICES OF λ FOR THE DIFFERENT IMAGES

Images	λ_{BIC}	λ_{AIC}	λ_{ML}	λ_{CV}
I1G1	1.94 (13.81)	1.44 (12.66)	1.36 (12.53)	1.83 (13.01)
I1G2	2.52 (12.96)	2.11 (12.50)	2.00 (12.38)	2.36 (12.78)
I1G3	2.79 (12.30)	2.32 (11.73)	2.24 (11.57)	2.66 (12.14)
I1U	0.72 (13.52)	0.56 (12.83)	0.55 (12.74)	0.65 (13.29)
I2G1	0.92 (15.82)	0.67 (15.59)	0.60 (15.52)	0.78 (15.63)
I2G2	1.74 (15.42)	1.49 (15.13)	1.47 (15.04)	1.58 (15.27)
I2G3	2.53 (14.95)	2.25 (14.47)	2.18 (14.36)	2.47 (14.83)
I2U	0.81 (15.74)	0.56 (15.46)	0.43 (15.32)	0.73 (15.59)
I3G1	4.68 (20.88)	4.32 (20.55)	3.78 (20.30)	4.57 (20.73)
I3G2	7.46 (20.27)	7.23 (20.09)	6.92 (19.88)	7.32 (20.12)
I3G3	7.83 (19.49)	7.50 (19.27)	7.36 (19.18)	7.73(19.38)
I3U	4.13 (20.43)	3.30 (20.15)	3.05 (20.03)	4.01 (20.31)

From Table I, it is clear that the ML method always yields the largest value of $\hat{\lambda}$ and thus an over regularized estimate which oversmooths the restored image. Further, the lower the SNR is, the larger the values of $\hat{\lambda}$ are.

The results of Table I indicate that $\hat{\lambda}_{\rm BIC} < \hat{\lambda}_{\rm AIC}$. As a result, it appears that $\hat{\lambda}_{\rm AIC}$ is likely to oversmooth. This is in accordance with the AIC criterion which tends to overfit. The tendency of AIC to oversmooth in comparison to BIC is a result of AIC only providing an asymptotically unbiased estimator of the Kullback–Leibler divergence whereas BIC is based on maximum *a posteriori* probability. Therefore, $\hat{\lambda}_{\rm AIC}$ will only provide an asymptotically unbiased estimator of the true ML as detailed in (10) and (11). Nevertheless, this estimate is better

TABLE III
COMPARISON OF REGULARIZATION WITH λ_{BIC} AND BILATERAL FILTERING IN TERM OF PSNR FOR THE CAMERAMAN IMAGE

Image	σ_w^2	regularization with λ_{BIC}	Bilateral filtering
I1	10	13.86	13.98
I1	20	13.44	13.15
I1	30	13.32	12.67

than the one provided by the ML method as it provides the bias correction estimate (15). A reason why λ_{CV} provides better results than λ_{AIC} is that λ_{AIC} only provides an asymptotically unbiased estimator of the true ML, whereas λ_{CV} is asymptotically equal to λ_{PMSE}, where PMSE stands for predicted mean square error [20]. As for AIC$_c$ in model selection, a refinement of the bias correction estimate (15) will lead to better image restoration with AIC. If $\hat{\lambda}_{\mathrm{BIC}}$ provides better results than $\hat{\lambda}_{\mathrm{AIC}}$, it is also because BIC has a much smaller probability of overfitting than AIC [21] in the case of linear regression models. Therefore, $\hat{\lambda}_{\mathrm{BIC}}$ is less prone to oversmoothing than $\hat{\lambda}_{\mathrm{AIC}}$.

In Table II, the results of the proposed methods for estimating the regularization parameter in image restoration are compared in terms of ISNR and PSNR. According to Table II, it can be seen that the restoration obtained with λ_{BIC} generate the highest ISNR and PSNR in comparison to the other totally data-based methods used in this example. λ_{BIC} should, therefore, be preferred in applications.

In Table III, the regularization method with a regularization parameter estimated with the proposed BIC variant BIC is compared to the method of bilateral filtering [22] in terms of PSNR. The cameraman image degraded by adding independent Gaussian noise having variance $\sigma_w^2 = 10$, $\sigma_w^2 = 20$, and $\sigma_w^2 = 30$ was used for this comparison. The bilateral filter was applied with $\sigma_d = 3$ and $\sigma_s = 30$ and a window size $= 11 \times 11$.

V. CONCLUSION

Regularization parameter estimation has traditionally been recognized as a critical task in image restoration problems. In this brief, estimation of the regularization parameter value for image restoration problems is cast in a model selection framework. Based on the Kullback–Leibler divergence and maximum *a posteriori* probability two totally data-based criteria are proposed for selecting the regularization parameter. These criteria can be seen as improved variants of AIC and BIC for image restoration and can be considered as alternatives to the ML and CV methods. As for model selection, it can be seen that the BIC variant provides better results than the AIC variant. Indeed, the AIC variant still presents the same inconvenience as AIC because it tends to oversmooth. This is in accordance with AIC which presents a higher asymptotic probability of overfitting than BIC for the linear regression models. A way to avoid oversmoothing and improve the performance of the AIC variant is as for AIC$_c$ to derive a more precise approximation of the bias adjustment (15) than (16). Future work will aim to refine approximation to obtain a variant of AIC$_c$ for estimating the regularization parameter in image restoration problems.

APPENDIX

As defined in [23], the bias is given by

$$E_{p(g)}\left\{\frac{1}{n}\log p(g|\hat{f}_\lambda,\lambda) - E_{p(g)}\{\log p(g|\hat{f}_\lambda,\lambda)\}\right\} = \frac{1}{n}b + o(1)$$

with

$$b = \mathrm{tr}\left\{\int T^{(1)}(g|f;P)\left.\frac{\partial p(g|f,\lambda)}{\partial f^T}\right|_{\hat{f}_\lambda} dP\right\} \tag{25}$$

where P is the unknown distribution and $T^{(1)}(g|f;P)$ is the influence function.

The influence function of the estimator $\hat{f}_\lambda$ for the model $\log p(g|\hat{f}_\lambda,\lambda)$ is determined by the n-dimensional statistical functional $T(\cdot)$ defined by

$$\int \frac{\partial}{\partial f}\left\{\|g - Hf\|^2 + \lambda\|Qf\|^2\right\}\bigg|_{T(P)} dP = 0.$$

From [23], it can be deduced that the influence function of the maximum penalized likelihood estimator $\hat{f}_\lambda$ is of the form

$$T^{(1)}(g|f;P) = \left[\frac{\partial(\|g - Hf\|^2 + \lambda\|Qf\|^2)}{\partial f \partial f^T}\right]^{-1} \times \left.\frac{\partial(\|g - Hf\|^2 + \lambda\|Qf\|^2)}{\partial f}\right|_{\hat{f}_\lambda}. \tag{26}$$

Inserting (26) in (25) and evaluating the expectation gives

$$b = \mathrm{tr}[H(H^T H + \lambda Q^T Q)^{-1} H^T].$$

REFERENCES

[1] M. R. Banham and A. K. Katsaggelos, “Digital image restoration,” *IEEE Signal Process. Mag.*, vol. 14, no. 2, pp. 24–41, Mar. 1997.

[2] L. Guan, A. Anderson, and J. Sutton, “A network of networks processing model for image regularization,” *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 169–174, Jan. 1997.

[3] H. S. Wong and L. Guan, “A neural learning approach for adaptive image restoration using a fuzzy model-based network architecture,” *IEEE Trans. Neural Netw.*, vol. 12, no. 3, pp. 516–531, May 2001.

[4] R. Molina, A. K. Katsaggelos, and J. Mateos, “Bayesian and regularization methods for hyperparameter estimation in image restoration,” *IEEE Trans. Image Process.*, vol. 8, no. 2, pp. 231–246, Feb. 1999.

[5] S. W. Perry and L. Guan, “Weight assignment for adaptive image restoration by neural networks,” *IEEE Trans. Neural Netw.*, vol. 11, no. 1, pp. 156–170, Jan. 2000.

[6] G. Archer and D. M. Titterington, “On some Bayesian/regularization methods for image restoration,” *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 989–995, Jul. 1995.

[7] M. G. Kang and A. K. Katsaggelos, “General choice of the regularization functional in regularized image restoration,” *IEEE Trans. Image Process.*, vol. 4, no. 5, pp. 594–602, May 1995.

[8] B. R. Hunt, “The application of constrained least-squares estimation to image restoration by digital computer,” *IEEE Trans. Comput.*, vol. 22, no. 9, pp. 805–812, Sep. 1973.

[9] A. M. Thompson, J. C. Brown, J. W. Kay, and D. M. Titterington, “A study of methods of choosing the smoothing parameter in image restoration by regularization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 4, pp. 326–339, Apr. 1991.

[10] N. P. Galatsanos and A. K. Katsaggelos, “Methods for choosing the regularization parameter and estimating the noise variance in image restoration and their relation,” *IEEE Trans. Image Process.*, vol. 1, no. 3, pp. 322–336, Jul. 1992.

[11] N. Fortier, G. Demoment, and Y. Goussard, “GCV and ml methods of determining parameters in image restoration by regularization: Fast computation in the spatial domain and experimental comparison,” *J. Vis. Commun. Image Represent.*, vol. 4, pp. 157–170, 1993.

[12] A. M. Thompson, J. W. Kay, and D. M. Titterington, “A cautionary note about crossvalidatory choice,” *J. Statist. Comput. Simulat.*, vol. 33, pp. 199–216, 1989.

[13] H. Akaike, “A new look at the statistical model identification,” *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.

[14] G. Schwarz, “Estimating the dimension of a model,” *Ann. Statist.*, vol. 6, pp. 461–464, 1978.

[15] R. Shibata, "Statistical aspect of model selection," in *From Data to Model*, J. C. Willems, Ed. New York: Springer-Verlag, 1989, pp. 215–240.
[16] C. M. Hurvich and C. L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, pp. 297–307, 1989.
[17] H. Akaike*, Time Series Analysis and Control Through Parametric Models*, ser. Applied Time Series Analysis, D. F. Findley, Ed. New York: Academic, 1978, pp. 1–23.
[18] L. Tierney, R. E. Kass, and J. B. Kadane, "Fully exponential Laplace approximations to expectations and variances of nonpositive functions," *J. Amer. Statist. Assoc.*, vol. 84, pp. 710–716, 1989.
[19] A. A. Neath and J. E. Cavanaugh, "Regression and time series model selection using variants of the Schwartz information criterion," *Commun. Statist.*, vol. 26, pp. 559–580, 1997.
[20] G. H. Golub, M. Heath, and G. Wahba, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, vol. 21, pp. 215–223, 1979.
[21] A. D. R. McQuarie and C. L. Tsai*, Regression and Time Series Model Selection*. Singapore: World Scientific, 1998.
[22] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Washington, DC, 1998, pp. 839–846.
[23] S. Konishi and G. Kitagawa, "Generalized information criteria in model selection," *Biometrika*, vol. 83, pp. 875–890, 1996.

"Vague-to-Crisp" Neural Mechanism of Perception

Leonid I. Perlovsky

***Abstract*—This brief describes neural modeling fields (NMFs) for object perception, a bio-inspired paradigm. We discuss previous difficulties in object perception algorithms encountered since the 1950s, and describe how NMF overcomes these difficulties. NMF mechanisms are compared to recent experimental neuroimaging observations, which have demonstrated that initial top-down signals are vague and during perception they evolve into crisp representations matching the bottom-up signals from observed objects. Neural and mathematical mechanisms are described and future research directions outlined.**

***Index Terms*—Brain imaging, brain mechanisms, cognition, cognitive engineering, dynamic logic, neural mechanisms, neural networks, object recognition, perception, "vague-to-crisp".**

I. Neural Networks and Mechanisms of the Mind

The field of neural networks has achieved significant success in engineering applications [1]–[4] and modeling mechanisms of the brain-mind [5]–[10]. Still, most neural paradigms have not addressed a recently discovered property of the visual perception mechanisms, a vague-fuzzy nature of internal representations [11], which gives rise to top-down priming signals. In this brief, we argue that this property is essential for understanding the workings of the mind at lower levels such as object perception, as well as at higher levels associated with abstract concepts, higher emotions including the beautiful and sublime, and their roles in cognition, imagination, and intuition. At lower levels, a process "from vague-to-crisp" is essential for fast operation of perception. Mathematically, it reduces the complexity of computation from (often) combinatorial to linear [1], [12], [13]. At higher levels, in addition to reducing complexity, it is essential for understanding mechanisms that were not previously understood and seemed mysterious [13]–[15].

This brief also touches on a role of logic in the mind mechanisms. For the first time, we describe how logical states emerge from vague-fuzzy states in the continuous process "from vague-to-crisp." Whereas fuzzy logic [16] is usually perceived in opposition to Aristotelian logic [17], we note that Aristotle did not consider logic a fundamental mechanism of the mind. His views on the mind operations incorporated vague-fuzzy states of the mind and were closer to the process "from vague-to-crisp" considered here [18].

The next three sections summarize neural modeling fields (NMFs) and dynamic logic (DL) forming the mathematical foundation for the paper. Section II summarizes difficulties of the past algorithms, which NMF-DL overcomes; these difficulties are related to complexity and logic. Section III describes the neural architecture of NMF-DL and its operations. Section IV presents an example illustrating application of NMF-DL to object detection in clutter, which significantly exceeds the performance of previously known algorithms and neural networks. Experimental validation of DL in neuroimaging experiments is discussed in Section V. Section VI discusses further directions.

II. Past Difficulties, Complexity and Logic

Biological object perception involves signals from sensory organs and the internal mind's representations (memories) of objects. During perception, the mind associates subsets of signals corresponding to objects with representations of object [5], [6], [11], [19], the so-called matching of bottom-up and top-down signals. This matching produces object recognition.

Mathematical descriptions of the very first *recognition* step in this seemingly simple association–recognition–understanding process have met a number of difficulties during the past 50 years. These difficulties were summarized under the notion of combinatorial complexity (CC) [12]. CC refers to multiple combinations of various elements in a complex system. For example, recognition of a scene often requires concurrent recognition of its multiple elements that could be encountered in various combinations. CC is prohibitive because the number of combinations is very large: for example, consider 100 elements (not too large a number); the number of combinations of 100 elements is 100^{100}, exceeding the number of all elementary particle events in life of the Universe. No computer would ever be able to compute that many combinations.

Algorithmic complexity was first identified in pattern recognition and classification research in the 1960s and was named "the curse of dimensionality" [20]. It seemed that adaptive self-learning algorithms and neural networks (see [21]) could learn solutions to any problem "on their own," if provided with a sufficient number of training examples. The following 30 years of developing adaptive statistical pattern recognition [22] and neural network algorithms led to the conclusion that the required number of training examples often was combinatorially large. Not only individual objects had to be presented for training, but also combinations of objects. Thus, self-learning approaches encountered *CC of learning requirements* [1], [12], [13].

Rule-based systems were proposed in the 1970s to solve the problem of learning complexity [23], [24]. An initial idea was that rules would capture the required knowledge and eliminate the need for learning.

Manuscript received November 25, 2008; revised May 18, 2009; accepted June 09, 2009. First published July 21, 2009; current version published August 05, 2009. This work was supported by the Air Force Office of Scientific Research (AFOSR) under the Laboratory Task 05SN02COR (PM Dr. Jon Sjogren).

The author is with Harvard University, Cambridge, MA 02138 USA and also with the U.S. Air Force Research Laboratory, Sensors Directorate, Hanscom, MA 01731 USA (e-mail: leonid@seas.harvard.edu).

Color versions of one or more of the figures in this brief are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNN.2009.2025501