Tomographic Considerations in Ensemble Bias/Variance Decomposition

David Windridge

CVSSP, University of Surrey, Guildford, UK

Abstract. Classifier decision fusion has been shown to act in a manner analogous to the back-projection of Radon transformations when individual classifier feature sets are non or partially overlapping. It is possible, via this analogy, to demonstrate that standard linear classifier fusion introduces a morphological bias into the decision space due to the implicit angular undersampling of the feature selection process. In standard image-based (eg medical) tomography, removal of this bias involves a filtration process, and an analogous n-dimensional processes can be shown to exist for decision fusion using Högbom deconvolution.
Countering the biasing process implicit in linear fusion, however, is the fact that back projection of Radon transformation (being additive) should act to reduce variance within the composite decision space. In principle, this additive variance-reduction should still apply to tomographically-filtered back-projection, unless the filtration process contravenes.
We therefore argue that when feature selection is carried-out independently for each classifier (as in e.g. multi-modal problems) unfiltered decision fusion, while in general being variance-decreasing, is typically also bias-increasing. By employing a shot noise model, we seek to quantify how far filtration acts to rectify this problem, such that feature selection can be made *both* bias and variance reducing within an ensemble fusion context.

1 Introduction

A central result of both the MCS and regression ensemble fields is that of the bias-variance-covariance decomposition of the mean squared error (MSE) [8, 1]. Whereas in individual classifiers we are concerned only with a bias-variance trade-off (i.e. assessing flexibility verses structural risk), Ueda and Nakano [4] demonstrated that ensembles must also consider correlations between estimators either implicitly or explicitly. This can be related to the Tumer and Gosh [7] framework for describing fused classifier error in terms of the effect on the margin. Thus (adopting the nomenclature of Brown et al. [1]), we have that $f(X, y, params)$ defines an estimator of some true underlying function $t(X, y)$ defined over the feature space X w.r.t. to the class y. We denote the combination of M estimators as:

$$\bar{f} = \frac{1}{M} \Sigma_{i=1}^{M} f_i(X, y, params) \qquad (1)$$

(Henceforth, we drop the parameter-denotation from f and t)

Thus, we have that for an ensemble of M estimators, the estimated mean square error can be decomposed via eqn. 1 as follows:

$$\begin{aligned} E\{MSE\} = E\{(\bar{f}-t)^2\} &= (E\{\bar{f}\}-t)^2 + E\{(\bar{f}-E\{\bar{f}\})^2\} && (2)\\ &= bias(\bar{f})^2 + variance(\bar{f}). && (3)\\ &= \bar{bias}^2 + \frac{1}{M}\bar{var} + \left(1-\frac{1}{M}\right)co\bar{v}ar && (4) \end{aligned}$$

where the bar dictates an ensemble average quantity, i.e.:

$$\bar{bias} = \frac{1}{M}\Sigma_i(E\{(f_i - t)\})\ ,\ \bar{var} = \frac{1}{M}\Sigma_i E\{(f_i - E\{f_i\})^2\} \tag{5}$$

$$co\bar{v}ar = \frac{1}{M}\Sigma_i\Sigma_{j\neq i}E\{(f_i - E\{f_i\})(f_j - E\{f_j\})\} \tag{6}$$

$E\{.\}$ is the expectation over all samples and all X. Error minimization hence requires that we seek to reduce the bias, variance and covariance of the constituent classifiers as far as possible. It is thus clear that various advantages accrue from using ensembles: we allow for the *possibility* for uncorrelated biases to cancel each other out and for the relative suppression of absolute deviations via the additivity of variance. It is also possible that, unlike the other terms, covariance can be negative, permitting a further avenue for error minimization in the ensemble.

In the current scenario, we consider only estimators f_i that include an (explicit or implicit[1]) feature selection stage for each classifier, such that the rejected features are treated by the omission of ordinates from the data vectors, i.e.:

$\forall X^n, y, params\ f_i(X^n, y, params) \propto f_i(x^{n_i}, y, params)$

with $n_i \subset n$ (specifically, x^{n_i} is a projective subset of X^n, a convention we shall adopt throughout). Classifiers are hence taken to model marginal distributions of $t(X^n, y)$ (though we will still consider the classifier to be defined over all X^n, as this will become important later). This consideration potentially complicates all 3 aspects (bias, variance and covariance) of the standard analysis, and reveals other strategies for reducing the overall MSE.

For example, ensemble covariance should tend towards zero when it is legitimate to make a naive Bayes assumption about the data irrespective of the underlying classifiers; that is, classifier diversity may be brought about by the feature selection process rather than the intrinsic nature of the classifiers in an ensemble. (Arguably the strongest motivation for feature selection in an MCS context is projection of largely independent data into independently-classified marginal distributions in order to maximize sampling (and thereby minimize

[1] In multi-modal decision fusion, we can consider the individual modalities as being a feature-selected subset of some composite space.

structural risk) at no cost to the feature-space coverage; however, we here consider the more general case in which features may be associated to classifiers for purely instrumental reasons). In rejecting the naive Bayes assumption as generally unrepresentative, it is evident that classifier variance can only be reduced by a factor related to the increase in bias within a feature-selection context (classification within marginal projections implies fewer parameters to represent the data). Feature-selected classier ensembles can thus not take advantage of any intrinsic decorrelation in bias in the same way as non-feature-selected ensembles.

Feature selection thus, in general, acts to reduce variance via the reduction in dimensionality, reduce *co*-variance via the introduction of the possibility of classifiers relating to potentially independent subspaces, but increases bias by eliminating information-bearing feature compositions (cf [3]).

The argument of this paper is that this issue is not necessarily as clear-cut as is usually considered: that we can, in fact, exploit feature selection to reduce variance and yet offset some the increase in bias by appropriate treatment (eg deconvolution) of the morphological sampling artifacts induced by the feature selection process. To do this we need to consider the tomographic nature of the feature-selection/combination process. Section 2 will therefore outline this analogy and its practical application. Section 3 will extend this work by quantifying a theoretical application of this approach in bias/variance terms and section 4 will apply an experimental test of the idea.

2 The Tomographic Analogy for classifier fusion

Full details of the tomographic approach to removing the morphological bias from decision fusion are given in [9]. Put briefly, the tomographic analogy assumes that the projective nature of feature selection process with respect to the original feature space, followed by the subsequent representation of marginals within generative (and to a lesser extent discriminative) classifiers can be modelled as n-dimensional Radon transformation (Radon transformation being the 'line-of-sight' integration carried out by eg X-ray detectors in medical imaging). Thus we assume:

$$f(\boldsymbol{x_i}, y)dx_i \approx \int_{x}^{\boldsymbol{x_i}+dx_i} t_i(x_i, y)dx_i \equiv \int_{x}^{\boldsymbol{x_i}+dx_i} \int_{all\ x'} t(X^n, y)dx'\ dx_i \qquad (7)$$

where $\boldsymbol{x_i} \in \boldsymbol{X}^{n_i}$ & $\boldsymbol{x'_i} \in \boldsymbol{X'}^{n_i}$ with $X^{n_i} \subset X$ and $X'^{n_i} = X^{n\perp}$ (i.e. X'^{n_i} is the orthogonal complement of X_i^n; $t_i(x_i, y, params)$ is thus the true marginal distribution.

If multiple classifiers are derived in this fashion (i.e. so that it is not necessarily the case that feature sets within the individual classifiers are fully coincident) then it can be shown that linear classifier combination (eg Sum Rule, Product Rule) is either equivalent to, or bounded by, back-projection, the inverse operation to Radon projection; $p_b(\boldsymbol{X^n}) = \frac{1}{M}\Sigma_{i=1}^{M} f_i(\boldsymbol{x_i}, y)$. However, this introduces an axially aligned artefact, $A(\boldsymbol{X^n}) = \Sigma_i dx_i. \int_{all} dX'^{n_i}$, that is a consequence of the fact that the Radon projections induced by feature selection represent only a small fraction of the total angular sample-space required for lossless reconstruction of the function $t(\boldsymbol{X^n}, y)$. What *is* recovered by back-projection (i.e. linear classifier fusion) is an estimation of the function $t(\boldsymbol{X^n}, y) \star A(\boldsymbol{X^n})$ (ie the true function t convolved with the artefact A). What we actually require is an estimate of $t(\boldsymbol{X^n}, y)$. While, in general, it is impossible to recover all of the 'lost' information brought about by feature selection by deconvolving A from the back-projected (ie fused classifier) output, performing this deconvolution does give rise to a *morphologically unbiased* estimate of t.

Rather than explicitly perform this deconvolution, pre-filtration of the Radon integrals is generally used prior to back-projection in medical imaging. However, since this approach can gives rise to negative values unrepresentative of stochastic estimates, the present paper considers a post-filtration approach, via iterative Högbom deconvolution of the biasing artefact[2].

Högbom deconvolution consists in iterative removal of the biasing artefact and replacing it by a Dirac delta function (or a coarse approximation to it) in the composite decision space. This process can be shown [10] to be equivalent to seeking correlated morphology in the classifiers and progressively reconstructing the morphology giving rise to this correlation in the composite decision space. It thus outperforms linear combination methods by using the correlated morphologies of classifiers in the ensemble to give more information about the sampled point than would otherwise be available. (Note that this approach works even for discriminative classifiers, though is optimal for generative classifiers). As a pseudo-code, the Högbom methodology is as set out in Appendix 1.

3 Theoretical study: Morphologically Induced Bias Following Variance-motivated Feature-Selection

For the present study, we assume a generative model of classification, in which classifiers (even if feature-selected) estimate the overall class distribution. If we were further to assume a unimodal model in which classes are represented by an arbitrary single-peaked distribution then the Högbom algorithm is provably optimal (ie can potentially recover the entire composite distribution $t(\boldsymbol{X^n}, y)$) from the marginal classifiers, provided the unimodality is of known cross-sectional form. This covers a wide range of possibilities included Gaussians with arbitrary covariance matrices. Under more realistic conditions (ie with an unknown cross-section), the Högbom algorithm is generally sub-optimal, but well-behaved, making only conservative (ie non-biasing) assumptions about the ambiguities arising from deconvolution, such that unimodal distributions of $t(\boldsymbol{X^n}, y)$ will give rise to identically unimodal estimates $t_{est}(\boldsymbol{X^n}, y)$ for all possible Radon projections and

[2] Note that Högbom deconvolution is not necessarily an approach that would be economic in practise; we here consider it because of its guaranteed positivity preserving characteristics. Note that efficient implementations of post-filtration *are* possible by appropriate kernelisation of the method (though beyond the scope of the current paper to set-out).

back-projections of X^n. The same is not generally true of linear fusion methods; an arbitrary change of basis of $t_{est}(\boldsymbol{X^n}, y)$ can potentially introduce differing numbers of modes within transformed marginal distributions.

In order to estimate the effect that this axial bias has on a standard feature-selection/classifier-fusion approach, we consider instead a simplified shot noise distribution model within X^n, such that n individual classifiers consist of non-intersecting unidimensional marginal distribution estimates (i.e. one feature is allocated for each of the n classifiers). The shot noise model considered consists of a random placement of K_{tot} distribution centroids such that, within each feature i, the marginal projection of each one of the K individual distributions has a well-defined width of ω_i (such that the marginal density projection has a value of exactly zero elsewhere). This occurs within a bounded width Δ_i attributable to the feature as a whole.

The marginal distribution estimate is the integral over the remaining $n-1$ components (assuming 1 selected feature per classifier) of the K distributions, which thus have density distributions D in X^n and marginal distributions D_i in X^n_i, i.e.:

$$P_i(x_i) = \int_{\forall \boldsymbol{x_j}:\boldsymbol{j}\neq \boldsymbol{i}} \Sigma_{K=1}^{K_{tot}} D(\boldsymbol{x} - \boldsymbol{c^K})\, d\boldsymbol{x} \equiv \Sigma_{K_i=1}^{K_{tot}} D_i(x_i - c_i^K) \tag{8}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots x_n)$ and $\boldsymbol{c^K} = (c_1^K, c_2^K, \ldots c_n^K)$ is the Kth cluster center.

For the sake of the current analysis, we initially assume that marginal distributions are sufficiently densely-sampled for there to be no significant variance issues when classifiers C_i are assigned to each feature, with undersampling only evident in the composite space $\boldsymbol{x}$ (perhaps motivating the feature selection in the first place). That is, we wish to isolate the tomographic influence on bias at this stage.

Since $D_i(x_i - c_i^K)$ is only non-zero for $x_i = c_i^K \pm \omega_i/2$, we can write:

$$P_i(x_i) = \Sigma_{k(x_i)} D_i(x_i - c_i^k) \tag{9}$$

where $k(x_i) \leq K_{tot}$ indexes the set of cluster centers for which $c_i^K = x_i \pm \omega_i/2$ (ie the clusters that become 'merged' under marginal projection).

The Högbom algorithm iteratively identifies and removes either whole or partial $D(\boldsymbol{x} - \boldsymbol{c^K})$ components from the back-projected (sum rule) composite space by recursively selecting the peaks in the density functions of classifiers defined over each marginal distribution. In general, the peaks of each marginal distribution estimate will be defined by the peaks of $k(x_i)$ which is, in turn, determined solely by the distribution of shot noise in the model (this is always true if $D_i(x_i - c_i^k) = const$ for $x_i = (c_i^K \pm \omega_i/2)$. This means that $P_i(x_i) \approx P(k(x_i)) = {}^{K_{tot}}C_k p^k (1-p)^{K_{tot}-k}$ (ie k is Binomially distributed, with $p = \omega_i/\Delta_i$.)

In the recursive Högbom deconvolution, all marginal distribution components of D at 'density level' $k/(K_{tot}\omega_i) < h < (k+1)/(K_{tot}\omega_i)$ are removed at the k-th iteration (these components are the level sets parameterized by h, i.e. the closed topological sets created by the truncation of the marginal density at value

h). The original shot noise components to which these level sets refer cannot be disambiguated by the procedure if the cardinality of closed topologies is greater than 1 for more than one of the features, and the reconstituted space must consist in all possible compositions of components ie $\{\{c_1^{k_h}\} \times \{c_2^{k_h}\} \times \ldots\}$, with k_h the number of marginal components for which $h \leq k.const$. The cluster centers generated in the composite space are thus the set of *all* ordered n-tuples: $\{(a_1, a_2, \ldots a_I) | a_1 \in \{c_1^{k_h}\}, a_2 \in \{c_2^{k_h}\} \ldots a_I \in \{c_2^{k_h}\}\}$) .

From the Binomial distribution of marginal components, we have that the mean density level of the marginal distributions is $K_{tot} . \frac{\omega_i}{\Delta_i} . \frac{1}{K_{tot}}$, meaning that there will be $\approx \left(\frac{\Delta_i}{\omega_i}\right)^n$ cluster centers in the reconstituted space following Högbom deconvolution if K_{tot} is large. Under sparse conditions, however, this figure will be $(K_{tot})^n$. All but K_{tot} of these reconstituted cluster centers are excess with respect to the true distribution of t in X^n; however we are guaranteed that these K_{tot} cluster centers are accurately represented if the marginal density estimate is accurate. These excess cluster centers (of cardinality $\approx ((K_{tot})^n - K_{tot})$) constitute the main source of remaining bias in the tomographic fusion model when applied to the shot noise model, representing the irrecoverable information loss implicit in feature selection.

Without Högbom deconvolution (ie using standard linear decision fusion), cluster centers are not explicitly identified in the composite space X^n. However, each point of the back-projected composite space *does* contain a contribution from the correct center. In fact, the unfiltered back-projected space consists of Dirac delta functions located at the correct centers (ie $\delta(c_1^1, c_2^1), \delta(c_1^2, c_2^2) \ldots \delta(c_1^K, c_2^K)$) that are convolved with the axially-aligned-artefact $A(X^n)$, such that the reconstituted classifier density generated by standard linear classifier combination is $\Sigma_K D(\boldsymbol{x} - \boldsymbol{c^K})) \star A(X^n)$. However, the interstices of the convolved artefacts themselves produce further Dirac delta functions (eg $\delta(c_1^1, c_2^2), \delta(c_1^2, c_2^1), etc$), that are equivalent to the novel cluster centers produced by the Högbom algorithm. Thus, filtered and unfiltered decision fusion are identical in terms of the generation of spurious cluster centers within the decision space under a shot noise model. This represents the unavoidable bias in decision fusion. However, in the absence of Högbom filtration, there is also the additional ambiguity created by the convolution artefacts. This represents the excess bias created by standard linear combiners.

Thus, to quantify these biases, we have that the excess bias generated by linear combination followed by Högbom filtration is:

$$\mathrm{Bias}_{Tom} \approx \int_{\forall \boldsymbol{x_j}:\boldsymbol{j}\neq \boldsymbol{i}} \Sigma_{K=1}^{K_{tot}} D(\boldsymbol{x} - \boldsymbol{c^K})\ d\boldsymbol{x}\ .\ \frac{1}{K_{tot}} \left((K_{tot})^n - K_{tot}\right)^2 \tag{10}$$

$$\approx \left[\frac{1}{K_{tot}} \left((K_{tot})^n - K_{tot}\right)^2\right] \tag{11}$$

The corresponding quantity for the sum rule decision scheme (representing linear fusion), which includes the axially-aligned artefacts generated by the convolution of reconstructed cluster centers with A, is the following:

$$\text{Bias}_{Sum} = \sum_{i=1}^{n} \frac{1}{n} \int_{\forall \boldsymbol{x_{j:j \neq i}}} d\boldsymbol{x'_i}. \int_{\boldsymbol{x_i}} \sum_{K=1}^{K_{tot}} D_i(x_i - c_i^K)^2 \, d\boldsymbol{x_i} - \int_{\forall X^n} D(\boldsymbol{x} - \boldsymbol{c^K})^2 \, d\boldsymbol{x} \tag{12}$$

$$\approx \Sigma_{i=1}^{n} \left(\frac{1}{nK_{tot}} \left[K_{tot} \left(\Pi_{i=1}^{n} \frac{\Delta_i}{\omega_i} \right) - K_{tot} \right] \right)^2 \tag{13}$$

(weighted sums, while not directly considered, should behave similarly)

Since the condition of sparsity is that $\frac{\Delta_i}{\omega_i} >> K_{tot}$ this implies that linear decision schemes will *always* have higher bias than the tomographically-filtered equivalent for sparse distributions for the case of single features per classifier with distribution centers as per the shot noise model. The problem worsens with both increasing dimensionality and increasing sparsity.

3.1 Variance Estimation

Variance is imported into this simplified model by considering sampling variation with respect to the marginal histograms standing in for classifiers. Variance within single-feature classifiers is hence reduced by a factor relating to the marginal integration; we denote this post-feature-selected marginal variance: v_i. However, in addition to this variance reduction mechanism, the back projection implicit in feature selection has the potential to reduce this variance further via summation. Specifically, in the composite decision space, we expect this further reduction at the points of convergence of the non-zero marginal histograms due to the combination. For the (normalized) sum rule, as for the Högbom-filtered decision space, this implies a variance of $\frac{1}{n^2}\Sigma_{i=1}^{n} v_i$ at the cluster centers (on the assumption of decorrelation). However, elsewhere the non-zero component caused by the Radon artefacts in the sum rule will not experience this reduction; variance for the axial component will be as for the marginal distributions (i.e. v_i). In general, these will dominate for a sparse distribution, giving a total variance of:

$\text{Variance}_{Sum} \approx \Sigma_i(\frac{\Delta_i}{\omega_i} - K_{tot})\omega_i.v_i + K_{tot}^n \frac{1}{n^2} \Sigma^n v_i \omega_i$

The corresponding variance for the filtered combination is simply:

$\text{Variance}_{Tom} \approx K_{tot}^n \frac{1}{n^2} \Sigma^n v_i \omega_i$

However, this analysis does not consider the effect of Högbom deconvolution *within* cluster centers, where the recursive identification of morphology may introduce other sources of variance.

4 Experimental Investigation at the Sparse/Dense Boundary

In order to quantify these effects further, we perform an experimental implementation using the shot noise model. In particular, we wish to evaluate more typically borderline cases, in which the sparseness of distribution centroids is reduced to the point of dense overlap.

To do this, we distribute randomly-parameterized Gaussian distributions within the composite (i.e. non-feature selected) pattern-space, and form unidimensional marginal histograms to act as classifiers of the overall distribution. We hence consider a two class case, in which each class consists of a density function so specified. The random distributions are obtained according to the standard multivariate Gaussian distribution,

$$f(X) = \sum_e \frac{|covmat_e|^{-\frac{1}{2}}}{2\pi^{d/2}}.A_e.e^{-\frac{1}{2}(X-M_e)^T covmat_e^{-1}(X-M_e)} \tag{14}$$

d is the dimensionality of the problem (in this case 2); $P(A) = const$ for $A \in [0,1]$

The covariance matrix *covmat* is derived via its Eigendecomposition; i.e. $covmat = U\Lambda U^T$, such that U is considered a rotation matrix over arbitrarily chosen θ thus:

$P(\theta) = const$ for $0 < \theta < 360$, $P(\theta) = 0$ otherwise.

$$U = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}, \Lambda = \begin{pmatrix} R_x & 0 \\ 0 & R_y \end{pmatrix}$$

$P(R_x), P(R_y) = const$ for $0 < R_x, R_y < 1$, $P(R_x), P(R_y) = 0$ otherwise.

Sampling of this distribution is achieved via Cholesky factorization of *covmat* and multiplication with randomly sampled vector uniform distribution over the domain bounded by the Δ_i's. As a proxy for sparseness variation, we keep the number of marginal histogram bins fixed, and range over Gaussian number, width and sampling parameters (we choose $\sigma^{-1} = [1:10]*const$, Gauss n$^o = [1:5]$ and max sample n$^o = 9375$ so as to nominally straddle the sparse/dense border at $\sigma^{-1} \approx 6$, when σ is of the order of the histogram bin width). Bias and variance are then evaluated with respect to the two fusion methodologies; the density-normalised Sum Rule and Högbom filtered Sum Rule (since we evaluate bias and variance with respect to the final fused classifiers in the composite space, it is not appropriate to consider intra-ensemble covariance). We also consider a histogram binning classifier in the original space with identical bin-width characteristics to the marginal histograms. Two separate distributions are created for each sampling of the parameters and designated as class 1 and 2. A misclassification rate is also calculated. Results are as depicted in figures 1-3.

5 Discussion and Conclusions

We find that, under the test conditions of borderline sparse/dense shot-noise distribution, the Högbom method retains its low bias but develops a significantly higher variance than the Sum rule despite backprojection. However, this does not appear to affect Bayes error rate adversely. Hence the "boundary bias" [2] (i.e. $bias(f, E(\bar{f})) = \text{sign}(1/2 - f)(E(\bar{f}) - 1/2))$ that typically favors *generalized*

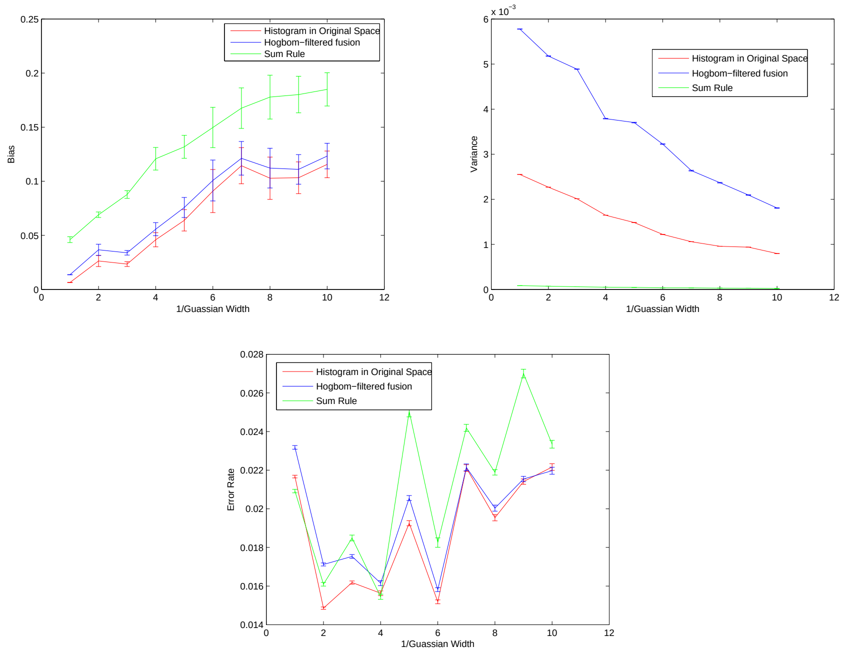

Fig. 1. Bias, Variance and Error Rate Per Histogram Bin vs Sparsity

low variance over low bias in terms of the Bayes error rate does not apply in this case. This would suggest that the Högbom method experiences low bias and low variance at the most classification-critical regions.

In conclusion, we have shown theoretically that under sparse conditions, filtered classifier fusion can decrease both bias and variance. In experimental conditions with more marginal distributions, this decreased bias is retained only at the expense of increased variance with respect to the linear decision rules. This would appear to be a side effect of seeking morphological correlation within the classifiers. However this does not appear to adversely effect misclassification rate.

Should this increased variance prove to be problematic in more general scenarios, bootstrap re-sampling (i.e. bagging) should mitigate the effect. In this way we can simultaneously reduce ensemble bias and variance (cf eg [6], [5]).

Acknowledgement: The research leading to these results has received funding from the EC 7^{th} Framework Programme FP7/2007-2013 under grant agreement n^{o} 215078.

6 Appendix 1: Procedural Implementing of Post-filtered Tomographic Classifier Combination

1. Assemble the combiners as a series of estimators ranging over n discrete feature spaces of respective dimensionality; $a_1, a_2 \ldots a_n$ for the class set; $\omega_1, \omega_2, \ldots\omega_m$; label these $P_n(\boldsymbol{X}_n)$, where $\boldsymbol{X}_n$ ranges over the vector space of dimensionality a_n.

2. Select the first class of the series, ω_1, and establish peak probability density value(s), $P_n^{\max}$, for of each expert's individual representation of that class.

3. Specify a pair of accuracy parameters, Δz and Δx, that respectively denote the probability density and feature-space resolutions.

4. Establish the 'hyper-area' between the probability density ordinates representing the peak value and (peak value$-\Delta z$) for each of the classifier PDFs: ie, the *scalar* number of $(\Delta x)^{a_i} \times \Delta z$ units between the two probability density values for each of the classifiers in the fusion. Vectors within these bounds are designated $\boldsymbol{X}'_n$.

5. Specify a matrix of dimension; $a_1 + a_2 + \ldots + a_n$ with each element designating an (initially zero) probability density value attributable to every $(\Delta x)^{a_1+a_2+\ldots+a_n}$ unit of the composite feature-space. Add a value, N, to those points representing all combinations of n concatenations of the respective (co-)ordinates established in **4**: That is, the Cartesian product $\{\boldsymbol{X}'_1\} \times \{\boldsymbol{X}'_2\} \times \{\boldsymbol{X}'_3\} \times \ldots \times \{\boldsymbol{X}'_n\}$. ($N$ must be $> \sum_{i=1}^{n} P_n^{\max}$).

6. Subtract the resolution parameter Δz from each peak value $P_n^{\max}$; $\forall i$, and set an iteration parameter (say, t) to zero.

7. Subtract a quantity $|X'_1| \times |X'_2| ... \times |X'_{i-1}| \times |X'_{i+1}| \times ... \times |X'_n| \times dz$ from the *current* peak value of each classifier, $P_n^{\max}$; $|X'_j|$ being the scalar values derived in **5**, ie: the number of coordinate vectors $\{\boldsymbol{X}'_i\}$ of dimensionality a_i counted by the PDF hyper-area establishing procedure above. Note, especially, the absence of $|X'_i|$ in the product entity.

8. Establish the *new* hyper-area value associated with subtraction **7**, ie: the hyper-area between the probability density ordinates representing the previous and current peak-values (as per **4**).

9. Allocate a value $N - t.\Delta z$ to those points in the deconvolution matrix representing *novel* coordinates established after the manner of **4.** That is, the Cartesian product <u>*difference*</u>:
$[(\{\boldsymbol{X}'_1\}_{old} \cup \{\boldsymbol{X}'_1\}_{new}) \times (\{\boldsymbol{X}'_2\}_{old} \cup \{\boldsymbol{X}'_2\}_{new}) \times \ldots \times (\{\boldsymbol{X}'_n\}_{old} \cup \{\boldsymbol{X}'_n\overline{\}_{new})]} - [\{\boldsymbol{X}'_1\}_{old} \times \{\boldsymbol{X}'_2\}_{old} \ldots \{\boldsymbol{X}'_n\}_{old}]$
(t the cycle count number, N as above).

10. Increment the cycle counter, t, by 1 and go to **7** while $P_n^{\max} > 0, \ \forall i$.

11. After termination of the major cycle **7-11**, subtract a value $t.\Delta z$ from each point of the deconvolution matrices to establish true PDFs, if required (see footnote 5).

12. Repeat from **2** for the remaining classes in the sequence $\omega_1, \omega_2 \ldots \omega_m$.

References

1. G. Brown, J. L. Wyatt, and P. Tiňo. Managing diversity in regression ensembles. *J. Mach. Learn. Res.*, 6:1621–1650, 2005.
2. J. H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Min. Knowl. Discov.*, 1(1):55–77, 1997.
3. M. A. Munson and R. Caruana. On feature selection, bias-variance, and bagging. In *ECML/PKDD (2)*, pages 144–159, 2009.
4. U. N and R. Nakano. Generalization error of ensemble estimators. *In Proceedings of International Conference on Neural Networks*, pages 90–95, 1996.
5. R. S. Smith and T. Windeatt. The bias variance trade-off in bootstrapped error correcting output code ensembles. In *MCS '09: Proceedings of the 8th International Workshop on Multiple Classifier Systems*, pages 1–10, Berlin, Heidelberg, 2009. Springer-Verlag.
6. Y. L. Suen, P. Melville, and R. J. Mooney. Combining bias and variance reduction techniques for regression trees. In *ECML*, pages 741–749, 2005.
7. K. Tumer and J. Ghosh. Theoretical foundations of linear and order statistics combiners for neuralpattern classifiers. *Technical Report TR-95-02-98, Computer and Vision Research Center, University of Texas, Austin*, 1995.
8. G. Valentini and T. G. Dietterich. Bias-variance analysis of support vector machines for the development of svm-based ensemble methods. *J. Mach. Learn. Res.*, 5:725–775, 2004.
9. D. Windridge and J. Kittler. A morphologically optimal strategy for classifier combination: Multiple expert fusion as a tomographic process. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(3):343–353, 2003.
10. D. Windridge and J. Kittler. Performance measures of the tomographic classifier fusion methodology. *Intern. Jrnl. of Pattern Recognition and Artificial Intelligence*, 19(6), 2005.