LA-UR-02-7281

Approved for public release;
distribution is unlimited.

Title: Chemical Identification Using Bayesian Model Selection

Author(s):

Tom Burr
Herb Fry
Brian McVey
Eric Sander

Submitted to: Spring Research Conference on Statistics in Industry and Technology
University of Michigan, Ann Arbor Michigan
May 2002

Chemical Identification using Bayesian Model Selection

Tom Burr, Herb Fry, Brian McVey
Los Alamos National Laboratory
Eric Sander
National Nuclear Security Administration

Abstract

Remote detection and identification of chemicals in a scene is a challenging problem. We introduce an approach that uses some of the image's pixels to establish the background characteristics while other pixels represent the target for which we seek to identify all chemical species present. This leads to a generalized least squares problem in which we focus on "subset selection" to identify the chemicals thought to be present. Bayesian model selection allows us to approximate the posterior probability that each chemical in the library is present by adding the posterior probabilities of all the subsets which include the chemical. We present results using realistic simulated data for the case with 1 to 5 chemicals present in each target and compare performance to a hybrid of forward and backward stepwise selection procedure using the *F* statistic.

Introduction

We consider infrared hyperspectral image data in which we first locate chemical plumes and then characterize the chemical components of the plume (McVey et. al., 2002). Our focus is chemical identification from a library of tens to hundreds of chemicals using a passive infrared (IR) detector several kilometers above the ground. A typical scene is 1000 *x* 128 pixels, where the detected signal at each pixel depends on the ground radiance, atmospheric transmission, instrument noise, and whether a chemical plume lies between the ground and the detector. A prior analysis identifies the plume pixels (pixels that have a plume influencing the signal) from among the background pixels (pixels that do not have a plume influencing the signal). Simplifying assumptions lead to a general least squares (GLS) problem for which we choose which subset (up to 3 chemicals from approximately 100 chemicals for example) of chemicals is most likely to be present in the plume.

The next section presents a physical model and the assumptions to convert this to a GLS problem. Following sections present model selection approaches including Bayesian model averaging (BMA) and "pick the winner" (PW) using penalized likelihood (the *F* test is a special case). We present results with simulated data, and include cases where the errors do not have a Gaussian distribution and the predictors have nonnegligible measurement error. We conclude that the false positive and negative rates for BMA are similar to those for PW provided the appropriate penalty is used with PW.

Background

A hyperspectral detector detects photons emitted in the IR region from the ground. The signal from background pixel *i* is

$$S_i^b = \varepsilon_i(\nu_j)L_i^g(\nu_j)\tau(\nu_j) + N_i(\nu_j) \qquad (1),$$

where $\varepsilon_i(\nu_j)$ is the emissivity at wavelength (or frequency) ν_j, $L_i^g(\nu_j)$ is the Planck function at ground temperature, $\tau(\nu_j)$ is the atmospheric transmission, and $N_i(\nu_j)$ is the noise. The signal from plume pixel i is

$$S_i^p = \alpha_p(\nu_j)[L_i^p(\nu_j) - \varepsilon_i(\nu_j)L_i^g(\nu_j)]\tau(\nu_j) + S_i^b \quad (2),$$

where $\alpha_p(\nu_j)$ is the plume absorption and $L_i^p(\nu_j)$ is the Planck function at plume temperature. The plume has two effects: it emits in the IR region, but it also absorbs the radiation emitted from itself and from the ground.

The $\varepsilon_i(\nu_j)$ terms (emissitivites) depend on the properties of the background. Concrete, asphalt, buildings, grass, dirt, water, and other common background features each have their characteristic emissivity. Our synthetic background scenes are generated from various mixtures of approximately 100 typical background emissitivites. Our synthetic plumes are generated from a library of approximately 100 chemical species of interest. Typically there are 10,000 to 16,000 pixels in a scene with 1 to a few plumes. The number of pixels in a plume is 10's to 100's.

For each wavelength ν_j, we subtract the mean response $\bar{y}$ (over all pixels), from the responses so that the centered response y has mean 0. We then assume that $y \sim N(0,\hat{\Sigma})$, where $\hat{\Sigma}$ is the n-by-n sample covariance matrix, where n is the number of wavelengths, and $\hat{\Sigma}$ is estimated using all of the pixels. Alternatively, $\hat{\Sigma}$ is estimated using only the background (nonplume) pixels, but in practice there is little difference between using all or just the background pixels to estimate $\hat{\Sigma}$.

Next, the effect of the plume is assumed to be linear and additive in the chemicals present (assuming Beer's law, and a weak plume so that $1-e^{-x} \cong x$). Therefore, the model for the plume pixels is

$$y = X\beta + e, \text{ with } e \sim N(0,\hat{\Sigma}) \qquad (3).$$

Equation (3) is effectively solved using GLS in which we multiply both sides of (3) by $\hat{\Sigma}^{-1/2}$ to convert to ordinary least squares with transformed data. To simplify notation, we will continue to use y as the transformed response and X as the transformed predictor matrix.

We have a library of 100's of chemicals so we cannot usually fit the full model. Instead, we assume that the true model has 1 to 3 chemicals present, so the problem reduces to model selection (actually, to subset selection, which is a special case of model selection). Challenging issues include: (a) the error distribution is not necessarily well approximated by a Gaussian; (b) there are errors in X due to imperfect corrections for atmospheric transmission, and (c) a more realistic model is a mixture of Gaussians, each with different mean to represent a heterogeneous scene with concrete, grass, dirt, and other patches. We will consider issues (a) and (b) in this paper.

Model Selection

There are

$$M = \binom{p_{Lib}}{0} + \binom{p_{Lib}}{1} + \binom{p_{Lib}}{2} + \binom{p_{Lib}}{3} \text{ models}$$

if we allow the null model, all single-chemical models, all paired-chemical models, and all triple-chemical models. We will report results for two cases:

(1) p_{Lib} is too large for an exhaustive search over all possible models; and

(2) p_{Lib} is small enough that an exhaustive search is possible

If exhaustive search is not possible, several stochastic searches could be considered. We report results only for a strategy involving the leaps algorithm and Occam's window (Raftery et. al. 1997), as implemented in Splus6 (2002).

BMA description.

Following Raftery et. al. (1997), and Neath and Cavanaugh (1997), we approximate the Bayesian information criterion (BIC) using

$$\begin{aligned} BIC_i = {} & (n - r_i - 1) \times \ln(RSS_i / n) + \\ & \ln(|X^T X|) - r_i \times \ln(2\pi) - 2 \times \ln(\gamma(\mathrm{r_i})), \end{aligned} \qquad (4)$$

where r_j is the number of chemicals in model i, $RSS_i = \sum_{j=1}^{n} (y_j - x\hat{\beta})^2$ is the the residual sum of squares from model i, and γ is the prior probability for r_j chemicals. Then the probability of model i is approximated using $P(M_i) \propto e^{(-BIC_i/2)}$. The probability that chemical j is present is $P(C_j) = \sum_{i=1}^{M} I(C_j \in M_i)P(M_i)$, where $I(\square)$ equals 1 if its argument is true. We select a tunable threshold, such as $T = 0.90$ and predict that chemical j is present if $P(C_j)$ exceeds T. Although it is tempting to conclude that the false positive probability is 1-T, we show elsewhere that the false positive probability must be evaluated by considering the distribution of $P(C_j)$ given X, β, and Σ.

PW description

We use RSS_i from model i and the estimated residual variance using the best fitting model in the criterion L, where $L_i = \frac{RSS_i}{\hat{\sigma}^2_{best}} + Kr_i$ for model i. (George, 2000, except we use $\hat{\sigma}^2_{best}$ rather than $\hat{\sigma}^2_{Full}$). The model having smallest L_i is the chosen model. The value $K = 2$ corresponds to the Aikike information criterion (AIC) or Mallow's C_p and $K = \ln(n)$ corresponds to the BIC (George, 2000). The usual F test is a special case of a PW method, in which the model having the largest F test (compared to the null model) is the selected model. It is often implemented using a stepwise search, but the F statistic can be evaluated for all models if M is not too large.

Results on simulated scenes.

Example 1a

Stochastic search using leaps/Occam with n = 300 wavelengths, p = 116 chemicals for BMA and stepwise forward regression as implemented in HIP (McVey et al, 2002) for PW. The true model was chemicals (64,69,82, and 83). For 4 different signal-to-noise ratios (SNR), BMA and PW results using F as special case of the PW method. We assume there is no error in X. Figure 1 is a typical example of the results of a matched filter ($\hat{\beta}_i$) over all frequencies for chemical k on a synthetic image using the HIP software. In this case, the signal strength is defined empirically as the noise equivalent concentration length (NECL), where

$$NECL_k = \{\frac{1}{p}\sum_{i=1}^{p}(\hat{\beta}_i - \bar{\beta})^2\}^{-1/2}$$ where $\bar{\beta}$ is the average over all pixels of the estimate of the coefficient for chemical k and p is the number of plume pixels. Results are in Table 1 as applied to the "superpixel" which is the average over all plume pixels.

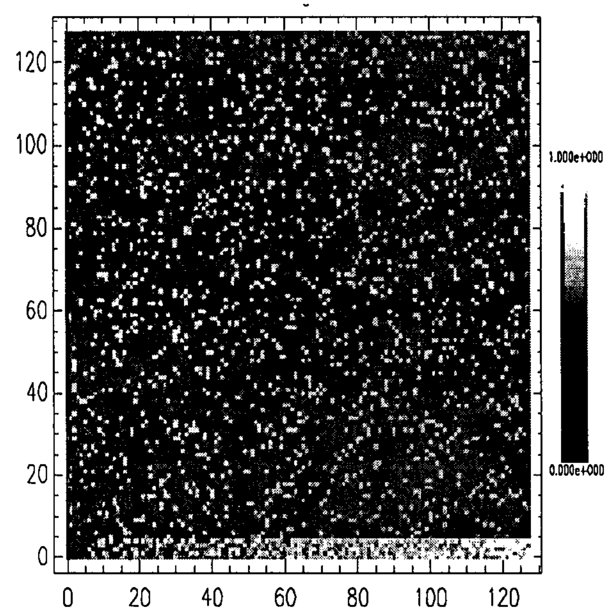

Figure 1. Synthetic image with the background emissitivitiess varying randomly across the scene. A plume is visible in the lower portion.

	BMA	PW
NECL1	64,69,82,83	64,69,82,83
NECL2	64,69,82	64,69,82
NECL3	64,82	64,82
NECL4	21,80	82

Table 1. The found chemicals for each of 4 NECLs. The true model is chemicals (64,69,82,83). The BMA threshold was 0.90 and the PW threshold was 0.001.

Note that BMA and PW give the same results except for NECL4 (worst case). For NECL4, BMA has 4 FNs and 2 FPs, and PW as 3 FNs.

Example 1b

Same as Example 1a using NECL1, but we scaled y and the X columns to have a variance equal to 1, and added additional error to y and introduced error in X. Results are in Table 2.

σ_X	σ_y	BMA	PW
		FP, FN	**FP, FN**
0.01	**0.1**	0,0	0,0
0.02	**0.2**	3,32	3,23
0.03	**0.3**	15,62	37,58
0.04	**0.4**	46,70	67,67
0.05	**0.5**	121,72	93,75
TOTALS		184,236	200,213
TOTAL ERRORS		420	413

Table 2. False positive (FP) and false negative (FN) counts for each of 5 σ_X and σ_y values for BMA and PW. Overall, the BMA and PW results are very similar. The BMA threshold was 0.90 and the PW threshold was 0.001.

Example 1c. True model is randomly selected subset of 0 to 4 of the 116 chemicals. Results are in Table 3. The BMA threshold was T = 0.99

σ_X	σ_y	BMA	PW
		FP, FN	**FP, FN**
0.01	**0.1**	0,0	13,0
0.02	**0.2**	0,0	51,0
0.03	**0.3**	0,0	80,0
0.04	**0.4**	79,0	86,0
0.05	**0.5**	238,0	86,0
TOTALS		317,0	316,0
TOTAL ERRORS		317	316

Table 3. False positive (FP) and false negative (FN) counts for each of 5 σ_X and σ_y values for BMA and PW. Overall, the BMA and PW results are very similar.

Note from Table 3 that for the smaller values of σ_X and σ_y, the FP rate is higher for PW. We have not yet thoroughly explored the operating characteristic curves that evaluate FP and FN rates as a function of the decision thresholds. Also, in this case, there is a threshold that the *F* test (a special case of the PW method) must exceed in order for a chemical to be included in the chosen model. This absolute threshold can be chosen empirically just as the threshold *T* for BMA.

Example 2. Exhaustive search, p_{Lib}=13, *n* =300. Here we restricted the library to only 13 chemicals so that exhaustive search would be feasible in our Splus code. Elsewhere (McVey et. al. 2002) we report on exhaustive search results using larger libraries with the search executed in C++.
We randomly generated artificial X matrices (artificial chemicals) having unit variance, zero mean, and 0.9 or 0.5 as the correlation (all off-diagonal entries of X^TX/n were either 0.9 or 0.5). Results are in Tables 4a and 4b for 2 BMA methods and 2 PW methods. The first BMA method uses only the first term in Eq. 4, which is the most common approximation to the Bayes factor. The second BMA method uses all terms in Eq. 4, which are the same as Neath and Cavaneaugh, 1997 plus the $r_j \ln(2\pi)$ term that appears to have been dropped in Neath and Caveneaug by mistake. The threshold $T = 0.9$ was used to decide whether a chemical was present. The first PW method using $K = \ln(n)$ and the second uses $K = 1.5 \ln(n)$, where the factor of 1.5 was chosen to reduce the FP rate. Although the results are not shown here, the value K = 2 (corresponds to the AIC) was too small, leading to a large FP rate of approximately 6 times the FP rate of the $K = \ln(n) = 5.7$. It is known that unless the true model size increases with *n*, the AIC selects models that are too large (George, 2002).

σ_X,σ_y	BMA:2	PW:2
	FP, FN	FP, FN
0.01,0.1	2,3 1,1	16,3 0,0
0.02,0.2	3,3 1,1	16,3 0,0
0.03,0.3	2,5 3,1	17,6 1,1
0.04,0.4	1,2 14,14	20,2 0,2
0.05,0.5	1,2, 25,25	21,5 6,10
Totals	9,15 44,42	90,19 7,13
Total Errors	53,57	97,32

Table 4a. False positive (FP) and false negative (FN) counts for each of 5 σ_X and σ_y values for BMA and PW for X^TX/n= 0.9 on the off-diagonal.

σ_X,σ_y	BMA:2	PW:2
	FP, FN	FP, FN
0.01,0.1	3,4 101,98	32,17 51,68
0.02,0.2	2,3 101,98	31,15 50,67
0.03,0.3	3,3 102,99	32,17 53,69
0.04,0.4	2,4 106,104	33,11 59,75
0.05,0.5	2,2 113,111	38,18 66,78
Totals	12,16 523,510	44,42 279,357
Total Errors	535,526	445,435

Table 4b. False positive (FP) and false negative (FN) counts for each of 5 σ_X and σ_y values for BMA and PW for X^TX/n= 0.5 on the off-diagonal.

Overall, there is little difference between BMA using only the first term in Eq. (4) and BMA using all terms in Eq. (4). This contrasts with the results in Neath and Cavanaugh which suggested that the extra terms in Eq. (4) would improve performance. Comparing BMA to PW, because BMA has a lower FP rate, it also had a higher FN rate. However, because the total error rate was much lower for the PW methods, we have little reason to suspect that the BMA performance can be better than the PW performance in any general sense.
Also note that the only change between Tables 4a and 4b is the value of the off-diagonal entry in X^TX. Because the performance was much worse in Table 4b with $X^TX/n = 0.5$, this shows that the FP and FN rates of BMA cannot be extracted from the threshold value of $T = 0.9$. Also, because of the very high false negative rate in the $X^TX/n = 0.5$ case, we expect that if we lowered the threshold we could reduce the error rate. However, the false positive rate is already as high or nearly as high as the $X^TX/n = 0.9$ case, so we expect to still do worse in the $X^TX/n = 0.5$ case. This is a surprising result that warrants further study.

Model Departure Results

We simulated data having Gaussian, Uniform, and Gaussian mixture (GM, to mimic the effect of outliers) distributions, each with the same variance. We used a bimodal GM to mimic having a main distribution f_1 from which most (more than 90%) of the errors arise and an outlying distribution f_2 from which the rest of the errors arise. The GM arises from a combining the two distributions resulting in a probability density of $f = \alpha_1 f_1 + (1- \alpha_2)f_2$, where where chose $\alpha_1 = 0.9$. We used $P_{Lib}=13$, $n=300$, and the true model had chemicals (1,2,3) present. First we evaluated $P(C_1)$, $P(C_2)$, …, $P(C_{13})$ for each of the 13 chemicals and compared these 13 probabilities in the Uniform errors case and in the GM case to the Gaussian errors case. Qualitatively, the conclusion is that there is mild disagreement in Uniform vs Gaussian, and moderate disagreement in GM vs Gaussian. More quantitatively, 65 *t*-scores (estimated probabilities for 13 chemicals, each with 5 error variances) with many degrees of freedom that should have been approximately N(0,1) distributed had several values exceeding 3 and a few exceeding 4 for Uniform vs G, and many values exceeding 4 for GM vs G.

The most relevant effect of model departure is the potential for the FP and FN rates to be impacted. Using the same three error models (Gaussian, Uniform, and GM), and repeating the analysis from Example 1c, we find that the FP and FN rates are not noticeably impacted with Uniform or GM errors compared to the Gaussian error case. This is tentative good news that suggests performance can be reasonably well evaluated assuming Gaussian errors even if the errors are rather badly non-Gaussian.

Summary

Overall, the BMA performance (FP and FN rates) is similar to the PW performance. Also, both methods exhibit similar robustness to model departure of the types evaluated. For both BMA and PW, it is important to select good thresholds, which is always possible using simulation. It would be useful to have analytical approximations to complement simulation results.

The main source of errors in the predictor matrix X is a transmission-through-the atmosphere correction step. Current work is aimed at quantifying how errors in transmission corrections propagate into errors in X. This work evaluated the impact of errors in X in a generic way, and simultaneously increased errors in X and y. A more systematic exploration of the impact of error sources is under way. Also, we are addressing issue (c) mentioned in the Introduction regarding whether the most appropriate data model is a mixture of Gaussians, each having a different mean vector. This would model the fact that each type of non-plume pixel falls into a category such as concrete, grass, water, etc. Similarly, each plume pixel has its own background effect. Therefore, model (3) does not physically represent any given plume pixel (unless there is only one type of background pixel, such as concrete) because the covariance matrix $\hat{\Sigma}$ is estimated by pooling all non-plume pixels. A more realistic model is a Gaussian mixture model on a per-plume-pixel basis. However, unless we could estimate the true background of the plume pixels, this approach would be infeasible. A compromise approach would be to cluster the non-plume pixels into categories such as concrete and grass, assume the plume is entirely over concrete and apply model (3), using $\hat{\Sigma}$ as estimated using the concrete-background pixels, then repeat by assuming the plume is entirely over grass, water, etc. Such a strategy would extend our current BMA or PW approaches regarding assessing confidence in the chosen solution.

References

Hoeting, J., Madigan, D., Raftery, A., and George, E. (2000), "The Variable Selection Problem," University of Texas-Austin Technical report available at http://bevo2.bus.utexas.edu/GeorgeE/

McVey, B., Burr, T., and Fry, H., "Distribution of Chemical False Positives for Hyperspectral Image Data," Official Use Only Los Alamos Report, 2002.

Neath, A., and Cavanaugh, J (1997), "Regression and Time Series Model Selection Using Variants of the Schwarz Information Criterion," *Communications in Statistics- Theory and Models*, 26, 559-580.

Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179-191.

Splus 6 for Linux, Insightful Corp., 2002.

Volinsky, C. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science* 14(4), 382-417.