APPENDIX L:

THE USE OF THE BIC FOR SELECTION OF THE NUMBER OF SOURCES

Appendix L: Use of the BIC for Selection of the Number of Sources

When using positive matrix factorization to perform source apportionment, it is important to choose the correct number of factors into which the data are decomposed. Each factor represents a different type of source, so selecting too few or too many factors can lead to nonsensical source profiles. An adaptation of the Bayesian Information Criterion (BIC) was used to aid in selecting the correct number of factors.

The Bayesian Information Criterion is intended to help select the best model from among several competing models. A BIC value is calculated for each model under consideration, and the model with the smallest BIC value is chosen as the best model. Essentially, the BIC is designed to choose a model that describes the data adequately without using too many parameters. The formula for BIC for a single model is

$$\mathrm{BIC} = -2(l(\psi; y) - l(\psi^*; y)) + 2p\log(\sqrt{n}) \qquad \text{(Eq. L-1)}$$

where $l(\psi; y)$ is the log-likelihood of the model under consideration, $l(\psi^*; y)$ is the log-likelihood of the most likely model in the subset of models considered, p is the number of parameters fit in the model, and n is the number of observations. The first term in the sum, often called the deviance, measures the difference between the log-likelihood of the best fitting model and the log-likelihood of the model under consideration. This term gets larger as the model under consideration gets farther from the best fitting model. The second term, the penalty term, penalizes models for the number of parameters used. This term gets larger as more parameters are included in the model. The combination of these two terms allows the BIC to assign large values to models with too few or too many parameters. Models that use too few parameters and fit the data poorly will have an inflated deviance while models that use an excessive number of parameters to fit the data will have an inflated penalty term. The best model should strike a balance between fitting the data well and using only a few parameters.

For the speciated $PM_{2.5}$ data in each city, PMF decompositions with 4, 5, 6, 7, 8, 9, and 10 factors were considered. For each of these decompositions, we obtain a Q value that is essentially the sum of squared errors in the model. Multiplying the Q value by -0.5 could be considered an approximate log-likelihood for the model. As a result, the value of $-0.5\times Q_{10}$ (the Q value for the PMF decomposition with 10 factors) could be used as $l(\psi^*; y)$ since it is the model with the highest likelihood of all the models considered. Similarly, $-0.5\times Q_i$ could be used as $l(\psi; y)$ in calculating the BIC for model i. For p we use the total number of values estimated in the factorization. In other words,

$$p = (T \times F) + (F \times S) \qquad \text{(Eq. L-2)}$$

where T is the number of time steps, F is the number of factors, and S is the number of species of $PM_{2.5}$ modeled. For n, we use the total number of observations: $T \times S$. Note that since the number of factors changes, p takes a different value for each model considered. In contrast, the value of n is the same for all models in a single city. With all of the parameters defined in this way, it is possible to calculate a BIC value for each model (each number of factors) in each city. Within each city, whichever number of factors produces a model that results in the lowest BIC value can be considered the correct number of factors.

In practice, the model with the lowest BIC is not always chosen. Sometimes, two consecutive models will have approximately the same small BIC value. In this case, the model with the best agreement in the apportionment of the FRM mass and the $PM_{2.5}$ mass was chosen.