Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC

Eun Sug Park Peter Guttorp Ronald C. Henry

NRCSE

Technical Report Series

NRCSE-TRS No. 043

March 9, 2000

The **NRCSE** was established in 1996 through a cooperative agreement with the United States Environmental Protection Agency which provides the Center's primary funding.

Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC

Eun Sug Park[1], Peter Guttorp[1], and Ronald C. Henry[2]

[1]National Research Center for Statistics and the Environment
University of Washington
Seattle, WA 98195

[2]Civil and Environmental Engineering
University of Southern California
Los Angeles, CA 90089.

Author’s Footnote

Eun Sug Park is Research Associate, National Research Center for Statistics and the Environment, University of Washington, Seattle, WA 98195. Peter Guttorp is Professor of Statistics and Director of the National Research Center for Statistics and the Environment, University of Washington, Seattle, WA 98195. Ronald C. Henry is Associate Professor of Civil and Environmental Engineering, University of Southern California, Los Angeles, CA 90089. Although the research described in this article has been funded by the United States Environmental Protection Agency through agreement CR825173-01-0 to the University of Washington, it has as not been subjected to the Agency’s required peer and policy review and therefore does not necessarily reflect the views of the Agency and no official endorsement should be inferred.

Abstract and Key Words

Multivariate receptor modeling aims to estimate pollution source profiles and the amounts of pollution based on a series of ambient concentrations of multiple chemical species over time. Air pollution data often show temporal dependence due to meteorology and/or background sources. Previous approaches to receptor modeling do not incorporate this dependence. We model dependence in the data using a time series approach so that we can incorporate extra sources of variability in parameter estimation and uncertainty estimation. We estimate parameters using the Markov chain Monte Carlo method, which makes simultaneous estimation of parameters and uncertainties possible. The methods are applied to simulated data and 1990 Atlanta air pollution data. The results show promise towards the goal of accounting for the dependence in the data.

KEY WORDS: Dynamic models; Kalman filter; Gibbs sampler; Metropolis-Hastings algorithm; Compositions; Air pollution.

1. INTRODUCTION

The goal of receptor modeling is to identify the pollution sources and assess the amounts of pollution based on observations collected at a particular site, and from that information to develop an effective air quality management plan. The basic mathematical model can be written as follows based on chemical mass balance assumptions (see, e.g., Hopke, 1985, 1991, 1997; Gleser 1997):

$$y_t = \sum_{k=1}^{q} \alpha_{tk} P_k + \varepsilon_t, \qquad t = 1, \cdots, n \tag{1}$$

where $y_t = \left(y_{t1}, y_{t2}, \cdots, y_{tp}\right)$ is the *t*th observation, q is the number of sources, $P_k = \left(p_{k1}, p_{k2}, \cdots, p_{kp}\right)$ is the *k*th source composition (consisting of the fractional amount of each species in the emissions from the *k*th source), α_{tk} is the contribution from the *k*th source on the *t*th day, and $\varepsilon_t = \left(\varepsilon_{t1}, \varepsilon_{t2}, \cdots, \varepsilon_{tp}\right)$ is the measurement error associated with the *t*th observation. In matrix terms, the model (1) can be written as

$$Y = AP + E \tag{2}$$

where A is $n\times q$ source contribution matrix, P is $q\times p$ source composition matrix, and E is $n\times p$ error matrix. The model (1) may be viewed as a factor analysis model in the sense that Y is the only observable quantity while q, P, and A are all unknown quantities that need to be estimated (or predicted). Early approaches to multivariate receptor modeling include exploratory factor analysis, principal component analysis, target transformation factor analysis, and others (see, e.g., Henry 1991). It is well known that, without imposing additional constraints on the parameters, the factor analysis model is not identifiable even with known number of sources, q. There have been several attempts to avoid this problem by imposing more restrictive constraints on either the P or the A matrix (see Henry and Kim 1990; Henry, Lewis, and Collins 1994; Yang 1994; Park 1997). As a matter of fact, there could be many different sets of identifiability conditions, each making sense in its own context. Park, Spiegelman, and Henry (1999) discuss identifiability conditions that are meaningful in receptor models.

The assumption of independence among the observations y_t has been made either implicitly or explicitly in all previous approaches to multivariate receptor modeling, see, for instance, Hopke (1991), Henry (1991), Yang (1994), Gleser (1997), Park (1997), and Park et al. (1999). Air pollution data, however, are usually obtained as a series of measurements on concentrations of aerosols over time, and meteorology often induces some degree of dependence in the data. Observations closer in time tend to be more correlated than observations farther apart in time (e.g., Figure 1).

{Insert Figure 1}

In some cases the assumption of independence may not be grossly wrong because environmental data usually contains many missing values or erroneous observations, and after initial screening of the data, time separation between any pair of measurements may become large enough so that serial correlation can be ignored in the screened data. This, of course, is not always the case. The research in this paper was motivated by a 1990 Atlanta air pollution composition data set consisting of hourly measurements of volatile hydrocarbon (VHC) species. This data set was used in Henry et al. (1994) to derive vehicle-related hydrocarbon source compositions from the ambient data. In that study, three types of measured source profiles specific to Atlanta in the summertime of 1990 were also available: roadway emissions, whole gasoline, and gasoline headspace (see Henry et al. 1994). The compositions of those three sources for nine selected vehicle-related species are provided in Table 1.

{Insert Table 1}

It is worthwhile to mention that those direct source measurements were obtained, under rather restricted conditions, independently of the data (e.g., roadway compositions were obtained as highway tunnel measurements during morning rush hour). Thus it is not unlikely that the measured source compositions could be different from the true source compositions P_0 for the data due to pollutant transport (between source and receptor) and reactions (and also to measurement errors, variations in source compositions, and the contribution of minor sources). Nonetheless, the measured source compositions may serve as a guideline for the true source compositions.

Assuming that the measured compositions in Table 1 are the true source compositions, i.e., P is known in model (2), A can be estimated easily, for instance, as an ordinary least squares (OLS) solution, $\hat{A}_{OLS} = YP'(PP')^{-1}$, if we ignore dependence structure in the data (and vice versa, i.e., $\hat{P}_{OLS} = (A'A)^{-1} A'Y$ if A is known or estimated first. This was done in almost all previous works without checking the independence assumption). Figures 1 and 2 show the autocorrelation function (ACF) plot of the raw data Y and residuals calculated as $Y - \hat{A}_{OLS}P$ for each of nine species, respectively.

{Figure 2 about here}

Figure 3 shows ACF plots of OLS estimates of source contributions, $\hat{A}_{OLS}$.

{Figure 3 about here}

All three plots reveal significant serial correlation in the data. It is well known in time series literature that in the presence of the correlated residuals, the standard error (not adjusting for the correlation in the residuals) of OLS estimate of the trend (which may be regarded as P in our model) in the regression is often grossly wrong. Although the correct standard error of OLS estimate may be obtained by adjusting for the correlation, it is still not the best estimate since the generalized least squares estimate, taking the correlation into account in the estimation procedure, has smaller standard error. The goal of this article is to extend receptor models to account for temporal dependence in the data so that we can incorporate that source of variability in estimation of parameters and uncertainties. In Section 2, we introduce models accounting for time dependence in the observations. Estimation of parameters is discussed in Section 3. Sections 4 and 5 contain examples from simulated data and the Atlanta air pollution data, respectively. Finally, concluding remarks are made in Section 6.

2. MODEL

Assume that the y_t in (1) are dependent. We first need to decide how to model this dependence. It seems reasonable to assume that the source contribution on time *t* depends on the past source contributions (as Figure 3 indicates). Also, it is often the case that ε contains not only pure measurement error but also all the remaining sources of variability that is not explained by the systematic part of our model such as background sources (unmodeled minor sources) and meteorology, etc. Then it is likely that the ε_t are also correlated in time due to the effect of meteorology and unmodeled sources (see Figure 2). We may decompose ε_t into two terms $\varepsilon_t = \eta_t + \delta_t$ where η_t represents variability correlated in time due to meteorology or background sources, and δ_t represents residual, unpredictable variability due to pure measurement error, independent over time.

We consider the model

$$y_t = \alpha_t P + \eta_t + \delta_t$$

where $\alpha_t = \left(\alpha_{t1}, \alpha_{t2}, \ldots, \alpha_{tq}\right)$ is a stationary vector AR(1) process centered at $\xi = \left(\xi_1, \xi_2, \ldots, \xi_q\right)$, $\eta_t = \left(\eta_{t1}, \eta_{t2}, \ldots, \eta_{tp}\right)$ is a stationary vector AR(1) process centered at **0**, and $\delta_t = \left(\delta_{t1}, \delta_{t2}, \ldots, \delta_{tp}\right) \sim N_p(\mathbf{0}, \Sigma)$ where $\Sigma = diag\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2\right)$. We use '$N_k(\cdot,\cdot)$' to denote k-dimensional multivariate normal distribution throughout the paper. This model may be written in Dynamic Linear Model (DLM) form (West and Harrison, 1997) as

Observation equation: $\quad y_t = \alpha_t P + \eta_t + \delta_t, \quad \delta_t \sim N_p(\mathbf{0}, \Sigma)$

Evolution equation: $\quad \alpha_t = \xi + \left(\alpha_{t-1} - \xi\right)\Phi + u_t, \quad u_t \sim N_q(\mathbf{0}, U)$

$$\eta_t = \eta_{t-1}\Theta + \upsilon_t, \quad \upsilon_t \sim N_p(\mathbf{0}, V) \qquad (3)$$

where $u_t = \left(u_{t1}, u_{t2}, \ldots, u_{tq}\right)$, $\Phi = diag\left(\phi_1, \phi_2, \ldots, \phi_q\right)$, ϕ_k is an AR coefficient for the *k*th source contribution, $\upsilon_t = \left(\upsilon_{t1}, \upsilon_{t2}, \ldots, \upsilon_{tp}\right)$, $\Theta = diag\left(\theta_1, \theta_2, \ldots, \theta_p\right)$, and θ_j is an AR coefficient for *j*th element of η_t. Note that marginal distribution for each α_t is

$$\alpha_t \sim N_q\left(\xi, W\right), \qquad W = \Phi W \Phi + U \tag{4}$$

and for each η_t is

$$\eta_t \sim N_p\left(\mathbf{0}, M\right), \qquad M = \Theta M \Theta + V\,. \tag{5}$$

3. ESTIMATION

As the model gets complicated by inclusion of more parameters, Markov chain Monte Carlo (MCMC) simulation (Tierney 1994; Chib and Greenberg 1995; Besag, Green, Higdon, and Mengersen 1995; Gilks, Richardson, and Spiegelhalter 1996) seems to be an attractive approach for parameter estimation. Note also that the parameters of the models (1) or (3) are all unknown, and the problem of parameter estimation is essentially nonlinear, but the Markov chain Monte Carlo method makes the problem linear by use of conditional distributions. We introduce a Bayesian framework to employ an MCMC method (constraints and identifiability conditions can be used as a part of the prior distribution). As mentioned in Section 1, the receptor model can be viewed as a special type of a factor analysis model (with the constraints that the elements of factor loading matrix should be all nonnegative). For identifiability of the model we borrow conditions from the confirmatory factor analysis model (Anderson 1984).

C1. There are at least $q-1$ zero elements in each row of *P*,

C2. The rank of $P^{(k)}$ is $q-1$, where $P^{(k)}$ is the matrix composed of the columns containing the assigned 0's in the *k*th row with those assigned 0's deleted.

Under the above conditions the source profiles, *P*, are identified up to normalization, which is enough for the purpose of receptor model. (As long as the relative amount of each species in a source is determined, a source can be identified.) The conditions C1 and C2 (and nonnegativity constraints on the elements of *P*) are absorbed into prior distribution for *P*.

Under the normal error assumption on δ, the likelihood $f\left(Y|\ldots\right)$ is written as

$$f\left(Y|\ldots\right) = |2\pi\Sigma|^{-\frac{n}{2}} \exp\left\{-\tfrac{1}{2} tr\Sigma^{-1}\left(Y - AP - \eta\right)'\left(Y - AP - \eta\right)\right\} \tag{6}$$

where η is $n \times p$ matrix of which rows are η_t, $t = 1,\ldots,n$. We use '$|\ldots$' to denote conditioning on all other variables. For a prior distribution $p(\cdot)$, we assume that

$$\begin{aligned}&p\left(P,\Sigma,\Phi,U,\alpha_1,\ldots,\alpha_n,\Theta,V,\eta_1,\ldots,\eta_n\right)\\ &= p(P)p(\Sigma)p(\Phi)p(U)p\left(\alpha_1,\ldots,\alpha_n|\Phi,U,\xi_0\right)p(\Theta)p(V)p\left(\eta_1,\ldots,\eta_n|\Theta,V\right).\end{aligned}$$

For the sake of brevity, ξ is assumed known to be $\xi = \xi_0$. Note that (3) implies

$$p\left(\alpha_1,\ldots,\alpha_n|\Phi,U,\xi\right) = (2\pi)^{-\frac{n}{2}}|W|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}\gamma_1 W^{-1}\gamma_1'\right)|U|^{-\frac{n-1}{2}}\exp\left[-\tfrac{1}{2}tr\left(U^{-1}\sum_{t=2}^{n}\left(\gamma_t - \gamma_{t-1}\Phi\right)'\left(\gamma_t - \gamma_{t-1}\Phi\right)\right)\right]$$

where $\gamma_t = \alpha_t - \xi_0$ and

$$p\left(\eta_1,\ldots,\eta_n|\Theta,\eta\right) = (2\pi)^{-\frac{n}{2}}|M|^{-\frac{1}{2}}\exp\left(-\tfrac{1}{2}\eta_1 M^{-1}\eta_1'\right)|V|^{-\frac{n-1}{2}}\exp\left[-\tfrac{1}{2}tr\left(V^{-1}\sum_{t=2}^{n}\left(\eta_t - \eta_{t-1}\Theta\right)'\left(\eta_t - \eta_{t-1}\Theta\right)\right)\right].$$

Based on a series of observations y_1, $\ldots$, y_n, we are interested in sampling the full posterior $\pi\left(P,\Sigma,\Phi,U,\alpha_1,\ldots,\alpha_n,\Theta,V,\eta_1,\ldots,\eta_p\middle|Y\right)$. We use "block-at-a-time" Metropolis-Hastings algorithm (Chib and Greenberg, 1995). We shall make use of seven move types in implementing MCMC:

(a) updating *P*

(b) updating Σ

(c) updating Φ

(d) updating U

(e) updating Θ

(f) updating V

(g) updating α and η.

Letting $\tilde{P} = (A'A)^{-1} A'(Y-\eta)$ and $S = \left(Y - \eta - A\tilde{P}\right)' \left(Y - \eta - A\tilde{P}\right)$, and using the orthogonality properties associated with $\tilde{P}$ (see Press 1982), (6) can be written as

$$|2\pi\Sigma|^{-\frac{n}{2}} \exp\left\{-\tfrac{1}{2} tr\Sigma^{-1}S\right\} \exp\left\{-\tfrac{1}{2} tr\Sigma^{-1}\left(P-\tilde{P}\right)'(A'A)\left(P-\tilde{P}\right)\right\}$$

$$\propto \exp\left\{-\tfrac{1}{2}\left(vecP - vec\tilde{P}\right)'\left(\Sigma^{-1} \otimes A'A\right)\left(vecP - vec\tilde{P}\right)\right\}.$$

Let the prior distribution for *P* be

$$p(P) = p(vecP) \sim N\left(m_0, C_0\right)\mathbf{I}\left(P_{kj} \geq 0, \quad k = 1,\ldots,q, \quad j = 1,\ldots,p\right)$$

where m_0 is a pq-dimensional vector and C_0 is a $pq \times pq$-dimensional diagonal matrix. Enforcing the constraints C1-C2 is equivalent to using a degenerate point prior for some of the elements of *P*. We set $q \times (q-1)$ elements of m_0 and the corresponding elements of C_0 to be zero, which makes the prior distribution for P a truncated singular normal distribution (though still proper). Then the resulting full conditional posterior distribution $\pi\left(P|\ldots\right)$ is again a truncated singular normal distribution, which can be written as

$$vecP|\ldots \sim N_q\left(m, \quad C\right)\mathbf{I}\left(P_{kj} \geq 0, \quad k = 1,\ldots,q, \quad j = 1,\ldots,p\right)$$

where $m = C\left\{\left(\Sigma^{-1} \otimes A'\right)vec(Y-\eta) + C_0^{-} m_0\right\}$, $C = \left(\Sigma^{-1} \otimes A'A + C_0^{-}\right)^{-1}$ where C_0^{-} is a generalized inverse of C_0. Since both of Σ and C_0 are diagonal, for the columns of P with no zero elements, we have

$$P_j\big|\ldots \sim N_q\left(m_j, \quad C_j\right)\mathbf{I}\left(P_{kj} \geq 0, \quad k = 1,\ldots,q\right)$$

where $m_j = C_j\left\{\sigma_j^{-2} A'\left(\underline{y}_j - \underline{\eta}_j\right) + C_{0j}^{-1} m_{0j}\right\}$, $C_j = \left(\sigma_j^{-2} A'A + C_{0j}^{-1}\right)^{-1}$, m_{0j} is a *q*-dimensional prior mean vector of P_j, C_{0j} is a corresponding submatrix of C_0, $\underline{y}_j$ is the *j*th column of *Y*, and $\underline{\eta}_j$ is the *j*th column of η. For the columns of *P* containing zero elements, let q^* be the number of nonzero elements for that column and P_j^* be a column vector consisting of those q^* elements. Then

$$P_j^* \big| \ldots \sim N_{q^*}\left(m_j^*, \quad C_j^*\right)\mathbf{I}\left(P_{kj}^* \geq 0, \quad k = 1, \ldots, q\right)$$

where $m_j^* = C_j^* A^{*\prime}\left\{\sigma_j^{-2} A^{*\prime}\left(\underline{y}_j - \underline{\eta}_j\right) + C_{0j}^{*-1} m_{0j}^*\right\}$, $C_j^* = \left(\sigma_j^{-2} A^{*\prime} A^* + C_{0j}^{*-1}\right)^{-1}$, m_{0j}^* is a q^*-dimensional prior mean vector of nonzero elements of P_j, C_{0j}^* is a corresponding submatrix of C_0, and A^* consists of the columns of *A* corresponding to nonzero elements of P_j.

If there is no prior information about the source compositions but the zero elements, we may use a noninformative prior $p(P) = \prod_{k=1}^{q}\left[\prod_{j=1}^{p} \mathbf{I}\left(P_{kj} \geq 0\right)\mathbf{I}\left(P_{kj} = 0, j \in J_0\right)\right]$ where J_0 is the index set for which $P_{kj} = 0$, which takes into account the conditions C1-C2 and nonnegativity only. Under this prior, we have, for the columns of P with no zero element,

$$P_j \big| \ldots \sim N_q\left(m_j, \quad C_j\right)\mathbf{I}\left(P_{kj} \geq 0, \quad k = 1, \ldots, q\right)$$

where $m_j = (A'A)^{-1} A'\left(\underline{y}_j - \underline{\eta}_j\right)$, $C_j = \sigma_j^2 (A'A)^{-1}$. For the columns of *P* containing zero elements, we get

$$P_j^* \big| \ldots \sim N_{q^*}\left(m_j^*, \quad C_j^*\right)\mathbf{I}\left(P_{kj}^* \geq 0, \quad k = 1, \ldots, q\right)$$

where $m_j^* = \left(A^{*\prime} A^*\right)^{-1} A^{*\prime}\left(\underline{y}_j - \underline{\eta}_j\right)$, $C_j^* = \sigma_j^2\left(A^{*\prime} A^*\right)^{-1}$.

Hence move (a) can be performed using either a Gibbs sampler or a simple Metropolis-Hastings algorithm.

Under a usual inverse gamma prior distribution for σ_j^2, $\sigma_j^{-2} \sim \Gamma(\alpha, \beta)$, $j = 1, \ldots, p$, with the parameterization in which the mean and variance are α/β and α/β^2, respectively, the full conditional for $\left\{\sigma_j^2\right\}$ are

$$\sigma_j^{-2} | \ldots \sim \Gamma\left(\alpha + \tfrac{1}{2} n, \beta + \tfrac{1}{2} d_j\right)$$

where $d_j = \left(\underline{y}_j - \underline{\eta}_j - AP_j\right)'\left(\underline{y}_j - \underline{\eta}_j - AP_j\right)$. This can be easily sampled using a Gibbs sampler.

Moves (c) - (g) require Metropolis-Hastings steps. We use the same strategy as those given in Chib and Greenberg (1995) and West and Harrison (1997) to update Φ and *U*, respectively. Let $\gamma_t = \alpha_t - \xi_0$. Under uniform priors for ϕ_k, writing $\phi = \left(\phi_1, \ \ldots, \ \phi_q\right)$ for the diagonal of Φ, and $D = diag\left(\gamma_{t-1}\right)$, the full conditional posterior density for Φ, $\pi\left(\phi \middle| \ldots\right)$, is proportional to

$$c(\Phi)\, f_{nor}\left(\phi \middle| b, B\right) I\left(0 < \phi < 1\right)$$

where f_{nor} is the *q*-variate normal density function, $B^{-1} = \sum_{t=2}^{n} D' U^{-1} D$, $b = B \sum_{t=2}^{n} \gamma_t U^{-1} D'$, $c(\Phi) = |W|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\gamma_1 W^{-1} \gamma_1'\right)$, $W = \Phi W \Phi + U$ and $I\left(0 < \phi < 1\right) = \prod_{k=1}^{q} I\left(0 < \phi_k < 1\right)$. We use $N_q(b, B)$ as a proposal distribution for ϕ (independent proposal). That is, we sample a candidate ϕ_i^* from $N_q(b, B)$, compute the corresponding diagonal matrix Φ^* and variance matrix W^* such that $W^* = \Phi^* W^* \Phi^* + U$, and accept new ϕ vector with probability $\min\left\{1, \ \dfrac{c\left(\Phi^*\right) I\left(0 < \phi^* < 1\right)}{c(\Phi) I\left(0 < \phi < 1\right)}\right\}$.

The full conditional posterior for U, $\pi\left(U \middle| \ldots\right)$, is proportional to

$$p(U) a(U) |U|^{-\frac{n-1}{2}} \exp\left[-\tfrac{1}{2} trace\left(U^{-1} G\right)\right]$$

where $G = \sum_{t=2}^{n} \left(\gamma_t - \gamma_{t-1}\Phi\right)'\left(\gamma_t - \gamma_{t-1}\Phi\right)$ and $a(U) = |W|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\gamma_1 W^{-1} \gamma_1{}'\right)$. Note that *G* follows a Wishart distribution with parameters *U* and $n-1$, i.e.,

$$G \sim W_q(U, n-1)$$

where $f(G) = \dfrac{|G|^{\frac{1}{2}(n-k-2)} \exp\left[-\frac{1}{2} trace\left(U^{-1} G\right)\right]}{2^{\frac{1}{2}k(n-1)} |U|^{\frac{1}{2}(n-1)} \Gamma_k\left(\frac{1}{2}(n-1)\right)}$. Under an inverse Wishart prior

$$U \sim W_q^{-1}(\Psi_0, m_0)$$

where the density is given by

$$p(U) = \frac{|\Psi_0|^{\frac{1}{2}m_0} |U|^{-\frac{1}{2}(m_0+k+1))} \exp\left[-\frac{1}{2} trace\left(\Psi_0 U^{-1}\right)\right]}{2^{\frac{1}{2}m_0 k} \Gamma_k\left(\frac{1}{2} m_0\right)},$$

the conditional distribution of U given G is $U|G \sim W_q^{-1}(\Psi_0 + G, m_0 + n - 1)$, and so the full conditional posterior for *U* is proportional to

$$a(U) f_{Wishart^{-1}}(U|\Psi_0 + G, m_0 + n - 1)$$

where $f_{Wishart^{-1}}$ is the inverse Wishart density function. We use this inverse Wishart distribution $W_q^{-1}(\Psi_0 + G, m_0 + n - 1)$ as a proposal distribution for *U*. The acceptance probability in this case is given by

$$\min\left\{1, \frac{a(U^*)}{a(U)}\right\}$$

where $W^* = \Phi W^* \Phi + U^*$.

Move types (e)-(f) are essentially the same as move types (c)-(d) with substitution of Θ, V, M and η for Φ, U, W, and γ, respectively.

Move (g), updating α (equivalently, updating $\gamma_t = \alpha_t - \xi_0$) and η, can be implemented by forward-filtering, backward-sampling algorithm (West and Harrison 1997) applied to $y_t - \mu_0$ where $\mu_0 = E(y_t)$. Note that the assumption that μ_0 is known is not a strong assumption. Model (3) can be rewritten as

$$y_t - \mu_0 = \lambda_t \mathbf{F} + \delta_t \quad \text{and} \quad \lambda_t = \lambda_{t-1}\mathbf{G} + \rho_t, \tag{7}$$

where $\lambda_t = [\gamma_t \quad \eta_t]$ is the state vector at time *t*, $\mathbf{F} = \begin{bmatrix} \mathbf{P} \\ \mathbf{I}_{p \times p} \end{bmatrix}$, $\mathbf{G}$ is the *(k+p)×(k+p)* matrix,

$\mathbf{G} = \begin{bmatrix} \Phi & \mathbf{0} \\ \mathbf{0} & \Theta \end{bmatrix}$, and $\rho_t = [u_t \quad \upsilon_t]$ with variance matrix $\Omega = \begin{bmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & V \end{bmatrix}$. To sample from the full conditional posterior $\pi\left(\lambda_1, \lambda_2, \ldots, \lambda_n \middle| \ldots\right)$, we sequentially simulate the individual vectors $\lambda_n, \lambda_{n-1}, \ldots, \lambda_1$ as follows:

1) Sample λ_n from $N_q\left(m_n, C_n\right)$ where m_n and C_n are obtained from the Kalman filtering recurrences

$$\begin{aligned}
&m_{t+1} = m_t\mathbf{G} + e_{t+1}K_{t+1},\\
&e_{t+1} = y_{t+1} - \mu_0 - m_t\mathbf{GF},\\
&K_{t+1} = \left(\mathbf{\Sigma} + \mathbf{F}^t R_{t+1}\mathbf{F}\right)^{-1}\mathbf{F}^t R_{t+1},\\
&C_{t+1} = R_{t+1} - R_{t+1}\mathbf{F}K_{t+1},\\
&R_{t+1} = \mathbf{G}C_t\mathbf{G}^t + \Omega.
\end{aligned}$$

2) For each $t = n-1, n-2, \ldots, 1$, sample λ_t from $N_q\left(h_t, H_t\right)$ where $h_t = m_t + \left(\lambda_{t+1} - a_{t+1}\right)B_t$, $H_t = C_t - B_t{}' R_{t+1}B_t$, $B_t = R_{t+1}^{-1}\mathbf{G}C_t$, $a_{t+1} = m_t\mathbf{G}$, and λ_{t+1} is the value just sampled.

Note that the likelihood (6) is invariant with respect to changes in scale of *A* or *P* (even after identifiability conditions C1-C2 are taken into account), and the parameters *A* (and so ξ and *U*) and *P* are identified except for multiplication by a diagonal matrix (consisting of scale constants), i.e., we would estimate AD^{-1} ($D^{-1}\xi$, $D^{-1}UD^{-1}$) and DP unless we use a very precise informative prior. As already mentioned, knowing (estimating) *P* up to a normalizing constant fulfills the objective of receptor modeling. It can also be shown that a scale constant matrix *D* (although it is unknown and depends on the initial value of the parameters) does not vary from iteration to iteration within an MCMC run. In this sense our MCMC scheme is self-consistent, and so the adjustment for the scale constant matrix does not need to be made at each step. If the scale constant (the matrix *D*) is ever known (e.g., the total mass of pollutant particle is known), the adjustment can be directly applied to the posterior summaries simply by multiplying (or dividing) by *D*. Care must be taken though in specifying the initial values for the parameters or hyperparameters for the prior distributions to ensure that at least they are approximately on the same scale or in a consistent fashion (e.g., ξ, hyperparameters for *U*, and initial value for *A* or *P*).

Finally, the posterior probability statements can directly be made on the identifiable quantities such as the normalized *P* or the scaled matrix of *U* (i.e., the correlation matrix of *A*) as discussed in Besag et al. (1995).

Remark 1. When α_t and ε_t are assumed to be independent, it can easily be shown that under a normal prior distribution $\alpha_t \sim N_q(\xi_0, \Xi_0)$, the full conditional distribution for α_t, $\pi(\alpha_t | \ldots)$, is a normal distribution through conjugacy, i.e.,

$$\alpha_t | \ldots \sim N_q\Big(\big(y_t \Sigma_\varepsilon^{-1} P' + \xi_0 \Xi_0^{-1}\big)\big(P \Sigma_\varepsilon^{-1} P' + \Xi_0^{-1}\big)^{-1}, \big(P \Sigma_\varepsilon^{-1} P' + \Xi_0^{-1}\big)^{-1} \Big)$$

where $\Sigma_\varepsilon = \operatorname{cov}(\varepsilon_t) = diag(\sigma_{\varepsilon 1}^2, \ \ldots, \ \sigma_{\varepsilon p}^2)$. This can be updated using a Gibbs sampler, and with moves (a) and (b) where $\underline{y}_j - \underline{\eta}_j$ and σ_j^2 are replaced by $\underline{y}_j$ and $\sigma_{\varepsilon j}^2$, respectively, it completes one cycle of MCMC when the observations are treated as independent. In Section 4, this approach is also compared to our time series approach when the observations are actually dependent.

4. SIMULATION

The data are generated by the model (3) with $p = 7$, $n = 200$, $q = 3$, $\sigma_1^2 = \ldots = \sigma_7^2 = 3$, $\phi_1 = \phi_2 = \phi_3 = 0.8$, $\xi_0 = (10, \ 12, \ 14)$, $U = \sigma_u^2 \mathrm{I}_{k \times k}$ where $\sigma_u^2 = 3$, $\theta_1 = \ldots = \theta_7 = .7$, $V = \sigma_\upsilon^2 \cdot \mathrm{I}_{7 \times 7}$ where $\sigma_\upsilon^2 = 3$. The initial values of α and η are given by $\alpha_{1k} = \xi_0 + \sqrt{\dfrac{\sigma_u^2}{1 - \phi_k^2}} Z_k$, where $Z_k \sim N(0,1)$, $k = 1,2,3$ and $\eta_{1j} = \sqrt{\dfrac{\sigma_\upsilon^2}{1 - \theta_j^2}} Z_j$, $j = 1, \ldots, 7$, respectively. The true source composition matrix P_0 (normalized to sum to 1) is given in Table 2. It follows from (4) and (5) that $W = 8.333 \cdot \mathbf{I}_{3 \times 3}$ and $M = 5.882 \cdot \mathbf{I}_{7 \times 7}$.

In implementing MCMC, we take $\alpha = 3$ and $\beta = 8$ for the prior on $\sigma_j^2,\ j = 1,\ldots,7$ (yielding the prior mean 4), $m_0 = 7$ and $\Psi_0 = 9 \cdot \mathbf{I}_{3\times3}$ for the prior on *U* (yielding the prior mean $3 \cdot \mathbf{I}_{3\times3}$), and set the scale matrix for the prior on *V* equal to $9 \cdot \mathbf{I}_{7\times7}$ and the degrees of freedom equal to 11 (yielding the prior mean $3 \cdot \mathbf{I}_{7\times7}$), each ensuring a proper but relatively diffuse prior. We use a noninformative prior distribution for the nonzero elements of *P* throughout simulation.

The posterior summaries for the model parameters, P, Σ, Φ, U, Θ, and V, based on 2,000 values subsampled from 20,000 iterations following a 20,000 burn-in period are reported in Tables 3-5. For the source composition matrix *P* and the variance matrix *U*, those summaries are obtained in terms of normalized *P* (sum to 1) and the scaled variance matrix $\mathbf{R}_\mathbf{U}$ (the correlation matrix) since they are identified only up to a constant multiplier.

{Tables 3-5 about here}

We also report the posterior summaries obtained from the approach for independent observations (see Remark 1) in Table 6. Since this approach does not decompose the error variances into Σ and *M*, we treat the estimates of the error variances as the estimates for $\Sigma_\varepsilon^2 = diag\left(\sigma_{\varepsilon 1}^2,\ \ldots,\ \sigma_{\varepsilon p}^2\right) = \Sigma + M$. The prior mean and the covariance matrix of α_t are set to be $\xi_0 = \left(10 \quad 12 \quad 14\right)$ and $\Xi_0 = 100 \cdot \mathbf{I}_{3\times3}$, respectively, and the hyperparameters of the priors on $\sigma_{\varepsilon j}^2$ $(j = 1,\ldots,7)$ are taken as $\alpha = 4$ and $\beta_j = 27$, $j = 1,\ldots,7$, (yielding the prior mean 9). The results are based on a posterior sample of size 2,000 obtained by subsampling every 10th from 20,000 values following a 20,000 burn-in period.

{Table 6 about here}

By comparing Table 3 and Table 6, it can be noted that the approach accounting for dependence in the data yields much better result in terms of posterior inferences than the approach not accounting for dependence. In Table 3 only 2 of the 15 (nonzero) elements of P_0 lie outside the 95% credible intervals (all are within the 99% credible intervals though we do not report them in the table) whereas in Table 6 ten elements of P_0 fall ouside the 95% credible intervals (9 of them are not captured even by the 99% credible intervals). Simultaneous credible regions for the whole matrix P_0 can also be constructed using the method (based on order statistics) suggested in Besag et al. (1995). Table 3 includes the 80% credible regions and these contain all elements of P_0 (The same holds for the 70% credible regions). In Table 6, nine elements of P_0 are still outside the 80% credible regions (7 of them are not captured even by the 90% credible regions). This is a natural consequence of not taking into account the correlation in the errors into the calculation of standard errors (posterior standard deviations here). In fact, the posterior standard deviations in Table 6 are much smaller than they should have been. Figure 4 shows the side-by-side barplots of the true source compositions (P_0) and the posterior mean of P from two different approaches, time series approach ($\hat{P}_{ts}$) and approach ignoring dependence ($\hat{P}_{indep}$), with R^2 values between P_0 and estimates. Again it can be seen that $\hat{P}_{ts}$ gives a much better approximation to the true source composition matrix P_0 than $\hat{P}_{indep}$ does.

5. APPLICATION TO ATLANTA DATA

The 1990 Atlanta data described in Section 1 has two types of temporal dependence structure, correlation in α and correlation in ε (see figures 2 and 3). We use model (3) with $q = 3$ to analyze this data set consisting of 538 measurements on 9 chemical species. For identifiability conditions, zeros are preassigned for CyHx+2MHx (cyclohexane+2-methylhexane) and 2,3-DMP (2,3-dimethylpentane) of source 1 (Roadway), acetylene and propene of source 2 (Gasoline), acetylene and 2,2,4-TMP (2,2,4-trimethylpentane) of source 3 (Headspace) since the relative concentrations of those species in each source are observed to be very low from Table 1. An OLS estimate $\hat{A}^*_{OLS} = YP^t_{measured}\left(P_{measured}P^t_{measured}\right)^{-1}$ where $P_{measured}$ is the measured source compositions (with zeros preassigned and each row normalized to sum to 100) was used as an initial value for *A*. The mean source contribution was set to $\xi_0 = (.37 \quad .14 \quad .03)$, which is the arithmetic mean of $\hat{A}^*_{OLS}$. Note that the specification of the value of ξ_0 is somewhat arbitrary due to the scale invariance property mentioned in Section 3. We only need to ensure that ξ_0 and the initial value of *A* are on the same scale. Since the measured source compositions ($P_{measured}$) can be regarded as prior information, we use as a prior distribution for *P* a truncated singular normal distribution with the mean $P_{measured}$ and the variance 900 for the nonzero elements of *P*, which ensures a fairly vague prior (the elements of $P_{measured}$ have the values between 0 and 100). The scale matrix for an inverse Wishart distribution for *U* was set to $\Psi_0 = 16 \cdot diag(1, \quad 0.7, \quad 0.08)$ with the degrees of freedom $m_0 = 20$, yielding the prior mean of $\Psi_0 / 16 = diag(1, \quad 0.7, \quad 0.08)$. This choice of the hyperparameter values was made to ensure that the prior distribution is moderately informative but flexible enough to cover the range of possible values of *U*. For the hyperparameters of the priors on σ_j^2, $j = 1, \text{L} \ , 9$, we take $\alpha = 5$ and $\beta_j = 48$ (the prior mean 12), and for the hyperparameters of prior on *V* we set the scale matrix equal to $27 \cdot \mathbf{I_p}$ and the degrees of freedom equal to 13 (so that the prior mean is $9 \cdot \mathbf{I_p}$), ensuring a proper but relatively diffuse prior. For each parameter, a posterior sample of size 1,000 was obtained by subsampling every 10th from 10,000 values following a 10,000 burn-in period. Tables 7-9 contain posterior summaries for some model parameters.

{Tables 7 and 8 about here}

The AR coefficients ϕ_k are estimated to be $\hat{\phi}_1 = .78$, $\hat{\phi}_2 = .68$, and $\hat{\phi}_3 = .48$, respectively, suggesting that there is substantial autocorrelation in roadway contribution and moderate autocorrelation in gasoline contribution and headspace contribution.

The side-by-side barplots of the measured source compositions (in Table 1) and estimated compositions are given in Figure 5 with R^2 values between measured and estimated compositions. In general, there seems to be good agreement between them.

{Figure 5 about here}

As mentioned in Section 1, the measured compositions are not the true source compositions in the sense of Section 4 for the data though they are expected to be generally close to the true compositions. For the Headspace composition profile (for which the measured and the estimated compositions show the best agreement), all but one (2MPentan) of the measured values fall in the 99% credible intervals. The 80% simultaneous credible regions (constructed by the method of Besag et al. 1995) are also reported in Table 7 and these capture all of the measured Headspace composition.

6. CONCLUSIONS AND DISCUSSION

In this article we develop a time series extension of multivariate receptor modeling in order to capture in the estimation process extra variability due to temporal dependence in air pollution data. Recent developments in MCMC methodology make estimation of parameters of complex models possible. By modeling the dependence structure, we can get more reliable estimates for the source compositions and their uncertainties, which are of our primary interest. As a by-product we can assess the amount of variability and autocorrelation in the source contributions and the errors. It also makes it possible to forecast the level of pollutants $\left(y_{t+k}\right)$ and the amount of pollution $\left(\alpha_{t+k}\right)$, which has been regarded as one of the model limitations in previous receptor modeling approaches (see the EPA discussion at http://www.epa.gov/oar/oaqps/pams/analysis/receptor/rectxtsac.html).

Throughout the article we assume that the errors are normally distributed. Environmental data often contain many outliers, and it is sometimes more appropriate to use the lognormal distribution to describe the data. The usual transformation technique does not help especially in the context of receptor modeling. By log-transforming the data the chemical mass balance equation of the model no longer applies directly, and we need to deal with model identifiability using different conditions. Alternatively, we may consider a multivariate T-distribution or a mixture of normal distributions to describe the error distribution. In the application to Atlanta data, the histogram of the residuals for each species looks in general bell shaped, but shows a few outliers for some of the species. This might suggest a use of heavy-tailed distribution for errors though it was not pursued further in this article. Non-normal dynamic modeling is still an active research area (see West and Harrison 1997), and we expect that multivariate receptor modeling can be extended further using non-normal dynamic models.

Another assumption we have made is that the errors have mean 0. To be more realistic, it would be preferable to generalize this to include the unknown non-zero mean errors, corresponding to unknown sources. This again involves the development of new identifiability conditions.

Finally, air pollution data is often obtained from multiple receptors. How to incorporate spatial variability as well as temporal variability in modeling when multiple species are measured is a challenging problem. Even in the case of no temporal dependence, this problem remains open.

REFERENCES

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis* (2nd ed.), New York: Wiley.

Besag, J., Green, P., Higdon D., and Mengersen K. (1995),"Bayesian Computation and Stochastic Systems," *Statistical Science,* 10, 3-41.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis-Hastings Algorithm," *American Statistician,* 49, 331-335.

Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. (1996), *Markov chain Monte Carlo in practice,* Chapman & Hall.

Gleser, L.J. (1997), "Some Thoughts on Chemical Mass Balance Models," *Chemometrics and Intelligent Laboratory Systems,* 37, 15-22.

Henry, R.C. (1991), "Multivariate Receptor Models," in *Receptor Modeling for Air Quality Management* (ed. P. Hopke), pp.117-147. Amsterdam: Elsevier.

———— (1997), "History and Fundamentals of Multivariate Air Quality Receptor Models," *Chemometrics and Intelligent Laboratory Systems,* 37, 37-42.

Henry, R.C., and Kim, B.M. (1990), "Extension of Self-Modeling Curve Resolution to Mixtures of More than Three Components. part 1. Finding the Basic Feasible Region," *Chemometrics and Intelligent Laboratory Systems,* 8, 205-216.

Henry, R.C., Lewis, C.W., and Collins, J.F. (1994), "Vehicle-Related Hydrocarbon Source Compositions from Ambient Data: the Grace/Safer Method," *Environmental Science and Technology,* 28, 823-832.

Henry, R.C., Lewis, C.W., and Hopke, P.K. (1984), "Review of Receptor Model Fundamentals," *Atmospheric Environment,* 18, 1507-1515.

Hopke, P.K. (1985), *Receptor Modeling in Environmental Chemistry,* New York: Wiley.

———— (1991), "An Introduction to Receptor Modeling," *Chemometrics and Intelligent Laboratory Systems,* 10, 21-43.

——— (1997), "The Chemical Mass Balance as a Multivariate Calibration Problem," *Chemometrics and Intelligent Laboratory Systems*, 37, 5-14.

Park, E. S. (1997), "Multivariate Receptor Modeling from a Statistical Science Viewpoint," unpublished Ph.D. dissertation, Texas A&M University, Dept. of Statistics.

Park, E. S., Spiegelman, C. H., and Henry, R. C. (1999), "Bilinear Estimation of Pollution Source Profiles in Receptor Models," Technical Report 006, University of Washington, National Research Center for Statistics and the Environment.

Press, S. J. (1982), *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference* (2nd edition). New York: Krieger.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *Annals of Statistics.* 22, 1701-1762

West, M., and Harrison (1997), *Dynamic Linear Models,* New York: Springer-Verlag.

Yang, H. (1994), "Confirmatory Factor Analysis and its Application to Receptor Modeling," unpublished Ph.D. dissertation, University of Pittsburgh, Dept. of Mathematics and Statistics.

TABLES

Table 1. Measured source composition profiles

Source	acetylene	propene	nButane	2MPentan	3MPentan	benzene	CyHx +2MHx	2,3-DMP	2,2,4-TMP
roadway	0.181	0.094	0.197	0.116	0.069	0.132	0.049	0.043	0.120
gasoline	0	0.002	0.197	0.221	0.138	0.108	0.116	0.067	0.152
headspace	0	0.007	0.685	0.144	0.075	0.034	0.021	0.014	0.021

Note: Each source profile is normalized sum to one

Table 2. True source composition profiles (P_0)

	1	2	3	4	5	6	7
Source 1	0	0.248	0	0.102	0.306	0.128	0.216
Source 2	0.242	0	0.266	0	0.009	0.044	0.440
Source 3	0.311	0.250	0.039	0.302	0	0.099	0

Note: Each source profile is normalized sum to one

Table 3. Summaries of the posterior distribution for P when the data is generated by model (3) and the approach accounting for dependence is used

Param.	j	1	2	3	4	5	6	7
P_{1j}	Mean	0	0.234	0	0.087	0.339*	0.124	0.216
	SD	0	0.018	0	0.023	0.016	0.013	0.033
	LSCR	0	0.191	0	0.025	0.299	0.088	0.137
	LCI	0	0.205	0	0.049	0.313	0.101	0.160
	UCI	0	0.262	0	0.124	0.306	0.147	0.269
	USCR	0	0.279	0	0.145	0.378	0.158	0.293
P_{2j}	Mean	0.204*	0	0.253	0	0.044	0.043	0.456
	SD	0.026	0	0.017	0	0.029	0.013	0.016
	LSCR	0.137	0	0.214	0	0.001	0.009	0.416
	LCI	0.157	0	0.225	0	0.004	0.021	0.430
	UCI	0.241	0	0.282	0	0.100	0.065	0.484
	USCR	0.256	0	0.295	0	0.127	0.075	0.502
P_{3j}	Mean	0.298	0.264	0.029	0.304	0	0.106	0
	SD	0.009	0.010	0.011	0.009	0	0.008	0
	LSCR	0.278	0.237	0.003	0.284	0	0.085	0
	LCI	0.284	0.247	0.011	0.290	0	0.093	0
	UCI	0.313	0.279	0.046	0.319	0	0.118	0
	USCR	0.320	0.288	0.056	0.328	0	0.126	0

Note: 1. SD stands for the posterior standard deviation; 2. LCI and UCI stand for the lower limit and upper limit of the 95% credible interval; 3. Asterisk (*) indicates that the true parameter value is not captured by the 95% credible interval; 3. Asterisk (*) indicates that the true parameter value is not captured by the 95% credible interval; 4. LSCR and USCR stand for the lower limit and upper limit of the 80% simultaneous credible region.

Table 4. Posterior means and standard deviations of Φ and $R_{\mathbf{U}}$ (correlation matrix corresponding to *U*) when the data is generated by model (3) and the approach accounting for dependence is used

	ϕ_k	Correlations in $R_{\mathbf{U}}$		
$k = 1$	0.826 (0.044)	1		
$k = 2$	0.834 (0.042)	0.010 (0.133)	1	
$k = 3$	0.817 (0.040)	0.245 (0.108)*	-0.141 (0.102)	1

Note: 1. Posterior standard deviation is given in the parenthesis; 2. Asterisk (*) indicates that the true parameter value is not captured by the 95% credible interval.

Table 5. Posterior means and standard deviations of Θ, V, and Σ when the data is generated by model (3) and the approach accounting for dependence is used

	θ_j	Diagonal elements of V	σ_j^2
$j = 1$	0.379 (0.194)*	2.463 (1.295)	3.823 (1.238)
$j = 2$	0.628 (0.178)	2.777 (1.304)	2.908 (1.002)
$j = 3$	0.836 (0.100)	2.030 (0.924)	4.368 (1.010)
$j = 4$	0.801 (0.102)	2.470 (1.127)	4.072 (1.077)
$j = 5$	0.539 (0.207)	2.634 (1.431)	4.252 (1.509)
$j = 6$	0.609 (0.121)	2.485 (0.950)	3.279 (0.921)
$j = 7$	0.650 (0.191)	2.496 (1.457)	2.547 (1.029)

Note: 1. Posterior standard deviation is given in the parenthesis; 2. Asterisk (*) indicates that the true parameter value is not captured by the 95% credible interval.

Table 6. Summaries of the posterior distribution for the parameters P and Σ_{ε} when the data is generated by model (3) but the approach ignoring dependence (given in Remark 1) is used

Param.	j	1	2	3	4	5	6	7
P_{1j}	Mean	0	0.214*	0	0.084	0.339*	0.125	0.239
	SD	0	0.014	0	0.014	0.011	0.008	0.022
	LSCR	0	0.180	0	0.050	0.314	0.105	0.189
	LCI	0	0.190	0	0.060	0.322	0.112	0.205
	UCI	0	0.236	0	0.106	0.357	0.137	0.277
	USCR	0	0.246	0	0.115	0.365	0.144	0.297
P_{2j}	Mean	0.123*	0	0.201*	0	0.154*	0.063*	0.459*
	SD	0.012	0	0.008	0	0.011	0.007	0.009
	LSCR	0.096	0	0.182	0	0.125	0.045	0.439
	LCI	0.104	0	0.187	0	0.136	0.051	0.445
	UCI	0.142	0	0.214	0	0.172	0.074	0.474
	USCR	0.154	0	0.221	0	0.179	0.080	0.482
P_{3j}	Mean	0.292*	0.282*	0.036	0.286*	0	0.103	0
	SD	0.005	0.005	0.007	0.004	0	0.005	0
	LSCR	0.281	0.269	0.021	0.276	0	0.092	0
	LCI	0.284	0.274	0.026	0.278	0	0.096	0
	UCI	0.300	0.291	0.047	0.293	0	0.111	0
	USCR	0.304	0.296	0.054	0.297	0	0.115	0
$\sigma^2_{\varepsilon j} = 8.882$	Mean	5.565*	8.648	10.415	11.375	8.275	7.873	7.255
	SD	1.453	1.853	1.403	1.621	2.246	0.840	2.768

Note: 1. SD stands for the posterior standard deviation; 2. LCI and UCI stand for the lower limit and upper limit of the 95% credible interval; 3. Asterisk (*) indicates that the true parameter value is not captured by the 95% credible interval; 4. LSCR and USCR stand for the lower limit and upper limit of the 80% simultaneous credible region.

Table 7. Summaries of the posterior distribution for P for the Atlanta data

Param.	*Species*	acetylene	propene	nButane	2MPentan	3MPentan	benzene	CyHx	2,3-DMP	2,2,4-TMP
								+2Mhx		9
	j	1	2	3	4	5	6	7	8	
roadway	Mean	0.275	0.115	0.279	0.086	0.049	0.126	0	0	0.069
	SD	0.008	0.004	0.013	0.004	0.003	0.004	0	0	0.005
	LSCR	0.257	0.107	0.247	0.076	0.042	0.117	0	0	0.057
	LCI	0.257	0.107	0.248	0.076	0.043	0.118	0	0	0.057
	UCI	0.295	0.124	0.305	0.095	0.056	0.135	0	0	0.081
	USCR	0.297	0.125	0.307	0.096	0.056	0.136	0	0	0.081
gasoline	Mean	0	0	0.172	0.191	0.113	0.088	0.123	0.098	0.217
	SD	0	0	0.019	0.005	0.003	0.004	0.005	0.004	0.008
	LSCR	0	0	0.127	0.179	0.104	0.077	0.112	0.089	0.200
	LCI	0	0	0.128	0.180	0.105	0.078	0.112	0.090	0.201
	UCI	0	0	0.214	0.202	0.121	0.097	0.134	0.107	0.236
	USCR	0	0	0.217	0.204	0.122	0.099	0.135	0.107	0.238
headspace	Mean	0	0.009	0.693	0.116	0.063	0.052	0.021	0	0.046
	SD	0	0.007	0.035	0.011	0.007	0.010	0.009	0	0.017
	LSCR	0	0.000	0.606	0.083	0.042	0.028	0.001	0	0.007
	LCI	0	0.001	0.609	0.087	0.045	0.029	0.002	0	0.008
	UCI	0	0.029	0.773	0.142	0.080	0.074	0.044	0	0.088
	USCR	0	0.034	0.776	0.145	0.081	0.076	0.046	0	0.093

Note: 1. SD stands for the posterior standard deviation; 2. LCI and UCI stand for lower limit and upper limit of the 99% credible interval; 3. LSCR and USCR stand for lower limit and upper limit of the 80% simultaneous credible region.

Table 8. Posterior means and standard deviations of Φ and $R_{\mathbf{U}}$ (correlation matrix corresponding to *U*) for the Atlanta data

	ϕ_k	Correlations in $R_{\mathbf{U}}$		
$k=1$	0.775 (0.036)	1		
$k=2$	0.677 (0.062)	0.207 (0.045)	1	
$k=3$	0.476 (0.114)	-0.069 (0.051)	-0.049 (0.047)	1

Note: Posterior standard deviation is given in the parenthesis.

Table 9. Posterior means and standard deviations of Θ, diagonal elements of V, and Σ for the Atlanta data

Species	θ_j	Diagonal elements of V	σ_j^2
Acetylene	0.512 (0.110)	1.039 (0.243)	1.148 (0.127)
Propene	0.550 (0.066)	0.405 (0.058)	0.506 (0.042)
nButane	0.400 (0.201)	2.929 (1.339)	3.683 (0.751)
2Mpentan	0.221 (0.086)	0.520 (0.102)	0.534 (0.045)
3Mpentan	0.162 (0.073)	0.280 (0.040)	0.349 (0.026)
Benzene	0.360 (0.092)	0.379 (0.055)	0.501 (0.040)
CyHx+2Mhx	0.237 (0.088)	0.341 (0.048)	0.448 (0.036)
2,3-DMP	0.269 (0.086)	0.261 (0.033)	0.360 (0.027)
2,2,4-TMP	0.643 (0.062)	0.681 (0.138)	0.758 (0.070)

Note: Posterior standard deviation is given in the parenthesis.

Figure Titles and Legends

Figure 1. Autocorrelation function (ACF) plots of Y for Atlanta data

Figure 2. Autocorrelation function (ACF) plots of the residuals for Atlanta data: $Y - \hat{A}_{OLS} P$ where P is the measured source compositions in Table 1

Figure 3. Autocorrelation function (ACF) plots of source contributions ($\hat{A}_{OLS}$) for Atlanta data

Figure 4. Side-by-side barplots of the true source compositions (P_0) and the estimated compositions obtained from two different approaches, time series approach and approach ignoring dependence

Figure 5. Side-by-side barplots of the measured source compositions and the estimated compositions for the Atlanta data

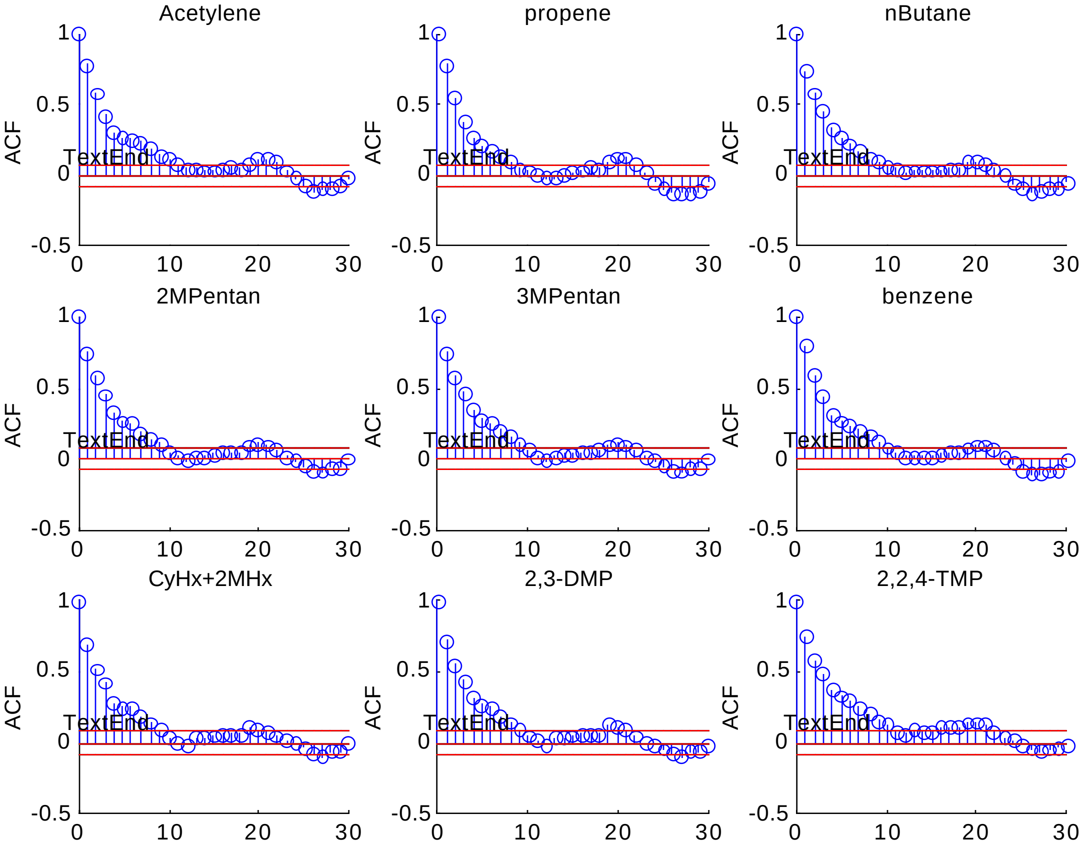

Acetylene
propene
nButane
2MPentan
3MPentan
benzene
CyHx+2MHx
2,3-DMP
2,2,4-TMP
ACF
TextEnd
1
0.5
0
-0.5
0
10
20
30

Figure 1

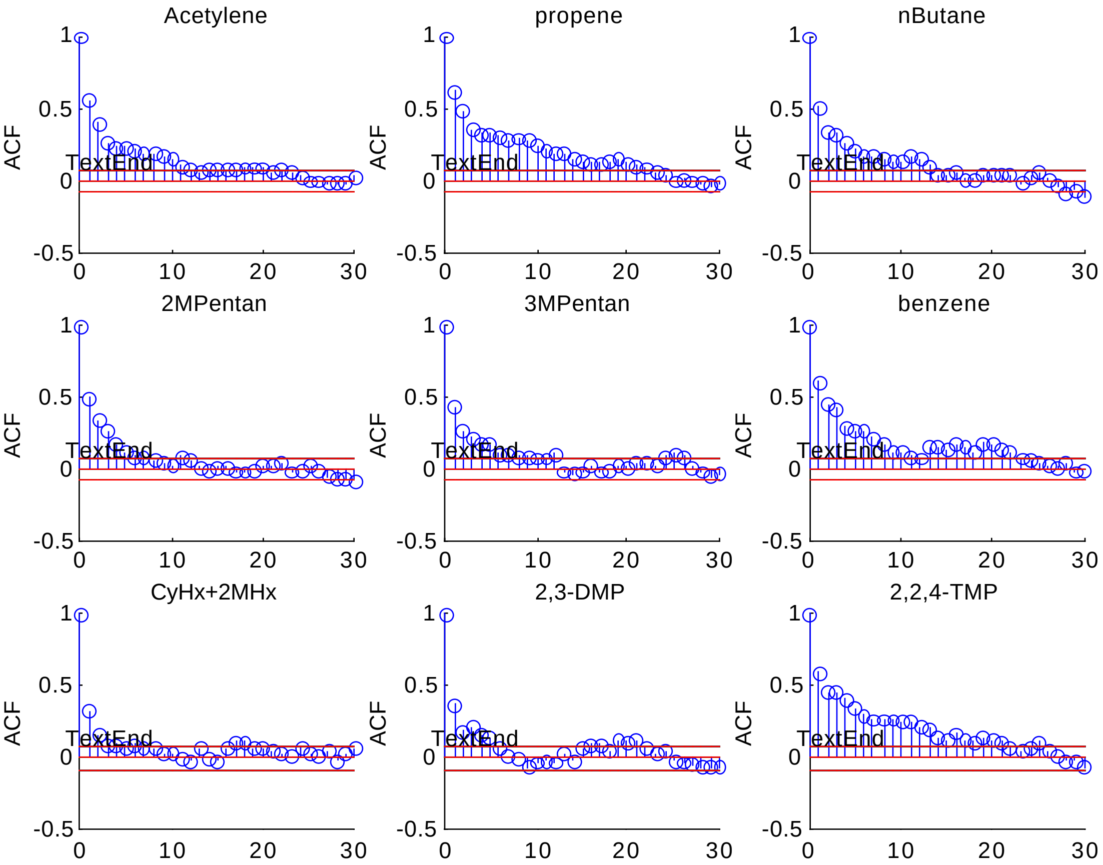
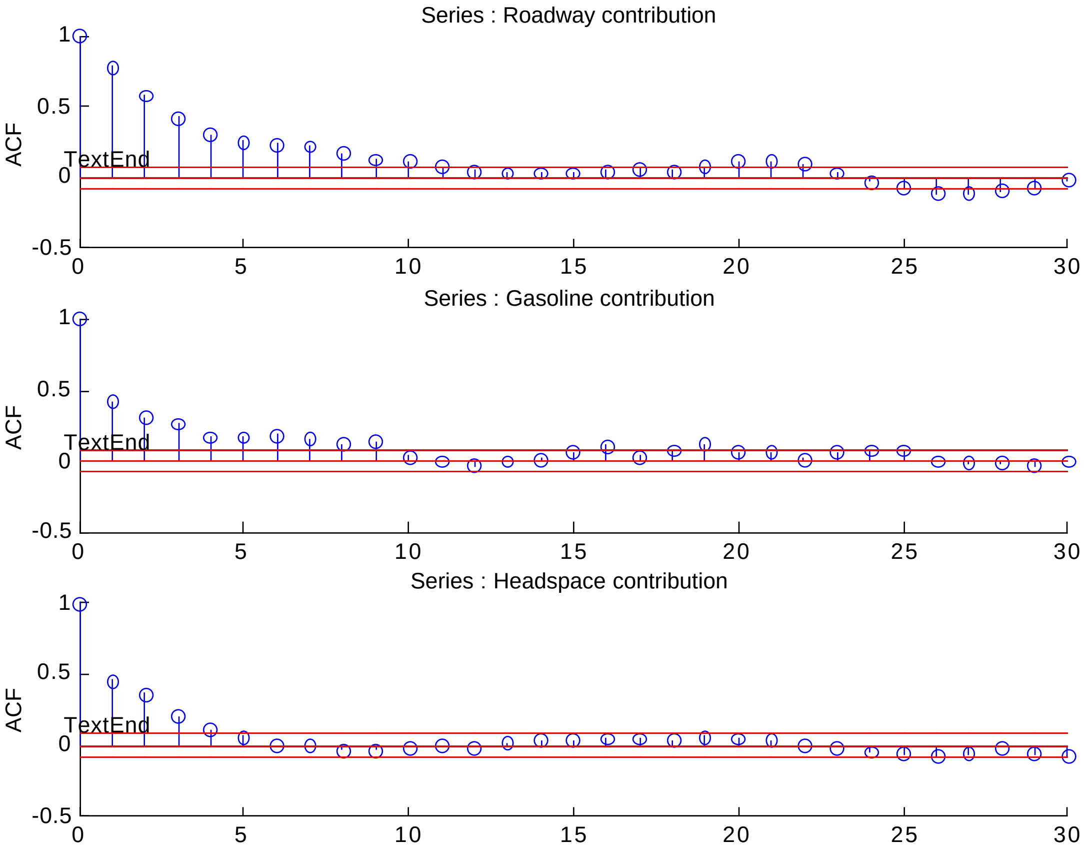

Acetylene
propene
nButane
2MPentan
3MPentan
benzene
CyHx+2MHx
2,3-DMP
2,2,4-TMP
ACF
TextEnd
1
0.5
0
-0.5
0
10
20
30

Figure 2



Series : Roadway contribution

Series : Gasoline contribution

Series : Headspace contribution

Figure 3

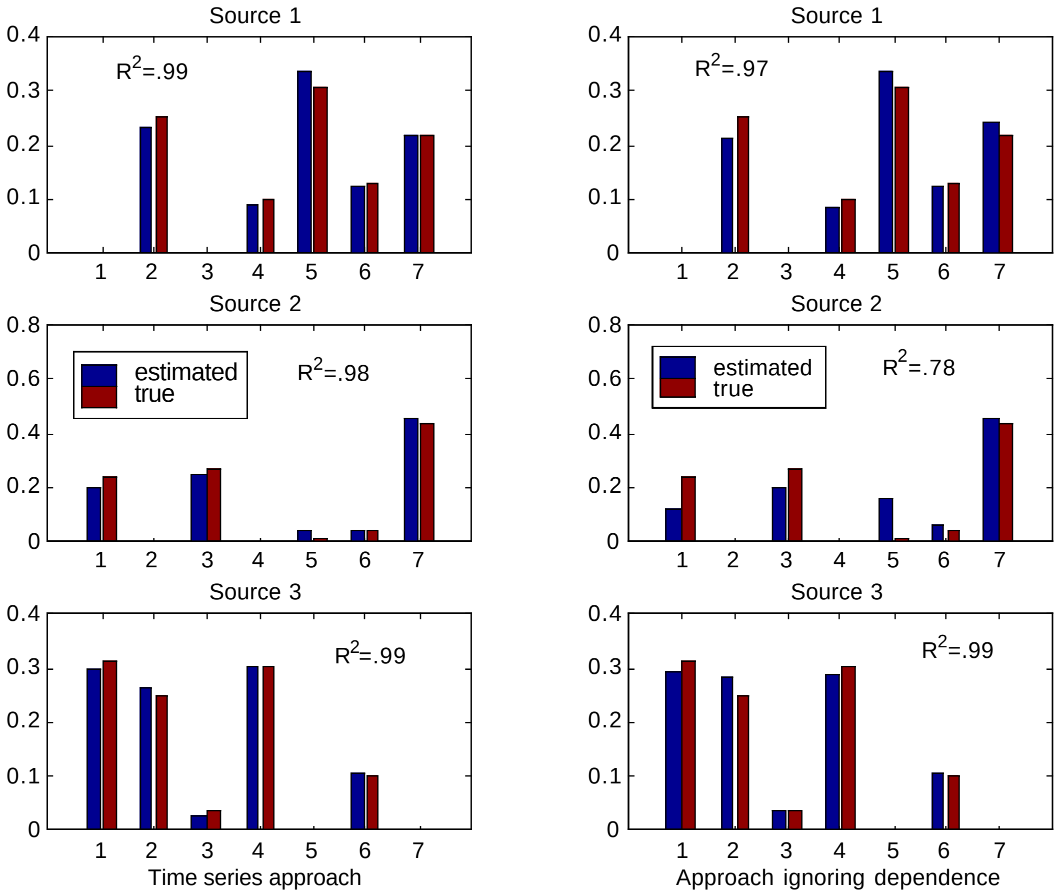

Source 1
R^2=.99
0.4
0.3
0.2
0.1
0
1 2 3 4 5 6 7
Source 1
R^2=.97
Source 2
estimated
true
R^2=.98
0.8
0.6
0.4
0.2
0
Source 2
R^2=.78
Source 3
R^2=.99
Source 3
R^2=.99
Time series approach
Approach ignoring dependence

Figure 4

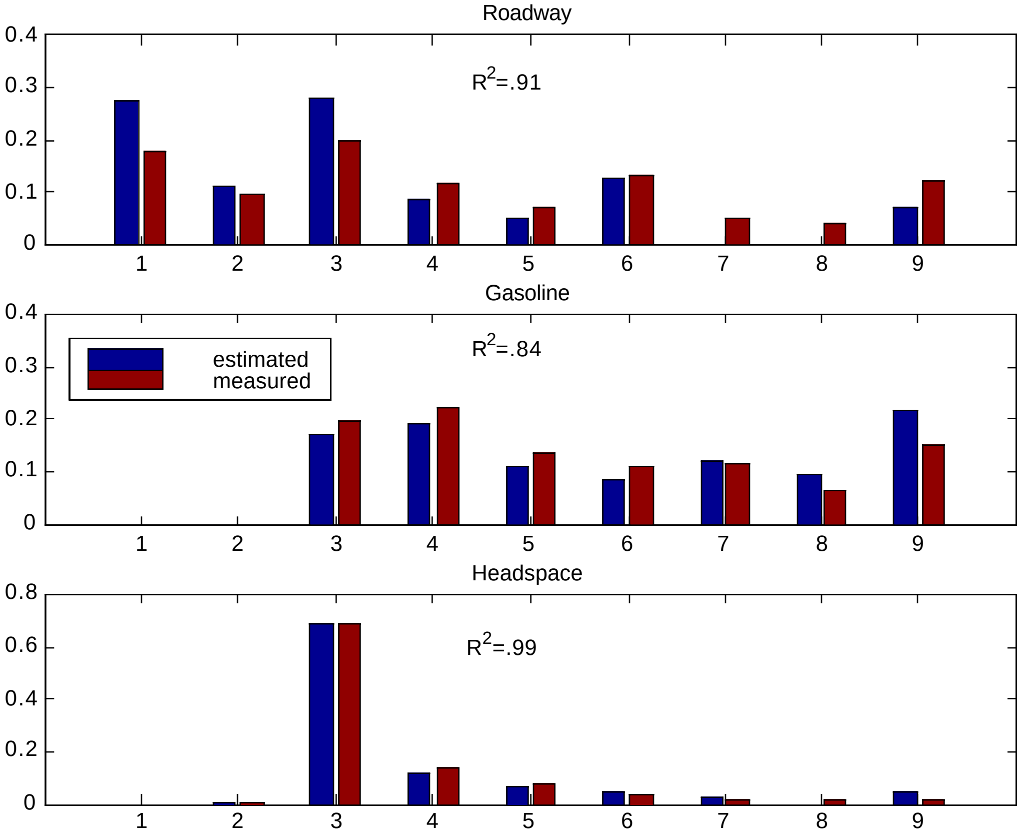

Roadway
R²=.91
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
Gasoline
R²=.84
estimated
measured
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
9
Headspace
R²=.99
0.8
0.6
0.4
0.2
0
1
2
3
4
5
6
7
8
9

Figure 5