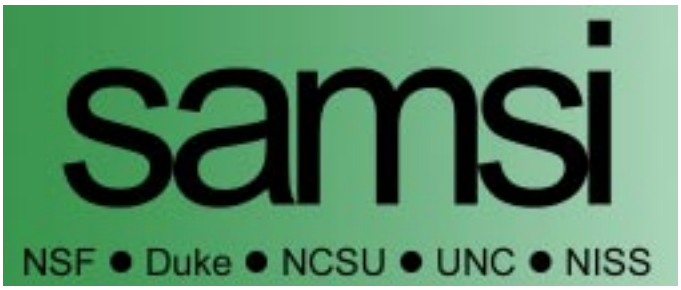

Apportioning Multiple Chemical Species to Sources

James V. Zidek

Technical Report #2003-18
January 3, 2005

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

Apportioning Multiple Chemical Species to Sources

James V Zidek*

University of British Columbia

January 3, 2005

Abstract

This paper adapts a multivariate empirical Bayes hierarchical predictor for use in imputing concentrations of chemical species measured over time, to their sources. The monitoring sites where these measurements are taken need not all measure the same suite of concentrations so that some data can be *missing-by-design*. These too may be imputed by the posterior predictive distribution derived in the paper. Moreover, simultaneous credibility regions can be constructed around the imputed values. Procedures for estimating the hyperparameters are presented along with posterior distributions for level one model parameters.

Keywords : Bayesian Interpolation; Co-Kriging; Matric-t Distribution; Predictive Distribution; Spatial Interpolation; Source apportionment; CO.

1 Introduction

This paper adapts the work of Sun, Le and Zidek (1997), itself an extension of Le and Zidek (1992 hereafter LZ) as well as Brown, Le and Zidek (1994a hereafter BLZ). to obtain a theory for imputing responses in multivariate random spatial fields when the monitoring sites have data *missing-by-design*. That field includes the unmeasured responses at "sources" of the concentrations of chemical species being measured. Like its predecessors that extension incorporates

*The work reported here was supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

uncertainty about model parameters including those in the spatial covariance matrix. Although we are forced to make empirical compromises in our Bayesian approach, our method allows fully arbitrary spatial covariance structures.

Our approach yields a multivariate-t posterior predictive distribution for the unmeasured responses both at certain monitoring sites where data are "missing-by-design" but as well, responses at the sources. That distribution can be dynamically updated with the incoming longitudinal data. Moreover it can be used in conjunction with nonlinear regression analyses to deal with structural errors-in-variables.

We state our main results in Section 2. Procedures for parameter estimation in Section 3. We discuss the strengthes and weaknesses of our method in Section 4.

2 Predicting Source Concentrations.

Suppose s_y monitoring (or "gauged") sites have been set up within a geographical area. With each site $i = 1, \ldots, s_y$, we associate a vector of concentrations of chemical species. Not all of these species need be measured at every site at every observation time $t = 1, \ldots, n$. However, we assume the subset to be measured at each site remains fixed site by design throughout the measurement period.

The monitored responses derive from s_x sources whose dynamically varying concentrations cannot be measured. Their prediction (or "imputation") is the primary purpose of the methodology described in this paper. That methodology can also be used to impute the responses that have been unmeasured by design. These imputations are based on the measurements that have been taken, assuming that none are missing at random. In all, we have $s = s_y + s_x$ sites and sources.

Our Bayesian approach to imputation involves two steps. In the first, we derive the required Bayesian predictor when the hyperparameters are known. In the second, described next, we fit those hyperparameters. To present that approach, we use the notation of Dawid (1981) as presented by Brown (1993) with requisite theory. Thus for a random matrix, $U : p \times q$, $U \sim \mu + \mathcal{N}(A, B)$ would mean that U has a vector-normal distribution with matrix mean μ; each row of U would have covariance matrix A, each column, covariance matrix B. Furthermore for a random matrix $V : q \times q$, $V \sim \mathcal{IW}(m^*; Q)$ means V^{-1} has a Wishart

distribution with $m^* + q - 1$ degrees of freedom and scale matrix Q^{-1}. Finally for a random matrix $T : p \times q$, $T \sim \mathcal{T}(m^*; P, Q)$ means that $T \mid V \sim \mathcal{N}(P, V)$ and $V \sim \mathcal{IW}(m^*; Q)$.

For each time point $t = 1, \ldots, n$, let W_t represent the $1 \times sk$ dimensional random vector of all random concentrations at both sources and sites, both measured and unmeasured. Let $W' = (W_1', \ldots, W_n')$. Given B and Σ, assume

$$W \sim ZB + \mathcal{N}(I_n, \Sigma). \tag{1}$$

The covariate matrix $Z(n \times h)$ determines the mean of W through the $h \times sk$ matrix of regression coefficients B.

A joint conjugate prior is adopted for B and Σ:

$$B \sim B^o + \mathcal{N}(F^{-1}, \Sigma) \tag{2}$$

and

$$\Sigma \sim IW(\delta, \Phi). \tag{3}$$

Here F^{-1} scales the sampling covariance matrix, Σ so that the result represents the uncertainty in the regression coefficients.

To clarify these assumptions, partition B as $B = (B^1, \ldots, B^{sk})$ and B^o in a similar way. Then our assumptions imply $E(B^i - B^{oi})(B^j - B^{oj})^T = \Sigma_{ij} F^{-1}$ for all i and j.

Partition W_t as $W_t = (X_t, Y_t)$, X_t $(1 \times s_x k)$ being the concatenation of the concentration vectors of all the species and sources at time t and Y_t $(1 \times s_y k)$, that of the gauged sites. After appropriately rearranging its elements, Y_t can further be partitioned into $Y_t^{(1)}(1 \times l)$, and $Y_t^{(2)}(1 \times (s_y k - l))$, corresponding, respectively, to the items unmeasured- and measured-by-design at the gauged sites. Since the same l items remain unmeasured during the whole monitoring period and they comprise l fixed columns in the matrix W, we call them *missing columns* in the sequel for expository simplicity.

Represent the unmeasured and measured column numbers of $Y^{(g)}$ by $i_1, \ldots, i_l$ and $i_{l+1}, \ldots, i_{s_y k}$, respectively. Let r_j, $j = 1, \ldots, s_y k$ be an $s_y k \times 1$-dimensional vector with j^{th} element one and the rest zero. We can then construct indicator matrices, R_1 and R_2 which "mark" the position of missing and present columns, respectively. More precisely, $R_1 = (r_{i_1}, \ldots, r_{i_l})$ and $R_2 = (r_{i_{l+1}}, \ldots, r_{i_{s_y k}})$. Finally R denotes the orthogonal matrix $R = (R_1, R_2)$. Observe that $Y^{(1)} = YR_1$ and $Y^{(2)} = YR_2$ consist of just the missing and present columns, respectively for the gauged sites. Because the response vector, W_t, has been partitioned into

three parts, we partition B, Σ, B^o and Φ accordingly. For example, we first partition Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix},$$

where Σ_{xx} and Σ_{yy} are $s_x k \times s_x k$, $s_y k \times s_y k$ matrices respectively. $R^{'}\Sigma_{yy}R$ can be further partitioned as

$$R^{'}\Sigma_{yy}R = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} R_1^{'}\Sigma_{yy}R_1 & R_1^{'}\Sigma_{yy}R_2 \\ R_2^{'}\Sigma_{yy}R_1 & R_2^{'}\Sigma_{yy}R_2 \end{pmatrix},$$

where Σ_{11} and Σ_{22} are $l \times l$, $(s_y k - l) \times (s_y k - l)$ matrices respectively. Further, let

$$\Psi_{yy} = R'\Phi_{yy}R = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix} = \begin{pmatrix} R_1'\Phi_{yy}R_1 & R_1'\Phi_{yy}R_2 \\ R_2'\Phi_{yy}R_1 & R_2'\Phi_{yy}R_2 \end{pmatrix},$$

with Ψ_{11} being $l \times l$ and Ψ_{22} being $(s_y k - l) \times (s_y k - l)$. Let $\Psi_{1|2} = \Psi_{11} - \Psi_{12}\Psi_{22}^{-1}\Psi_{21}$ and $\gamma = \Psi_{22}^{-1}\Psi_{21}$ for use in the sequel. Finally, let

$$(B_1^o, B_2^o) = B_y^o R = (B_y^o R_1, B_y^o R_2).$$

Conditional on $Y^{(2)} = y^{(2)}$, the theory given by Brown (1993, Appendix A.5) yields the predictive distribution of the levels of the unmeasured items at the sites:

$$Y^{(1)} \sim ZB_1^o + (y^{(2)} - ZB_2^o)\gamma + \mathcal{T}(\delta + s_y k - l, P_{1|2}, \Psi_{1|2})$$

where $P_{1|2} = I_n + ZF^{-1}Z' + (y^{(2)} - ZB_2^o)\Psi_{22}^{-1}(y^{(2)} - ZB_2^o)')$ and $\Psi_{1|2} = \Psi_{11} - \Psi_{12}\Psi_{22}^{-1}\Psi_{21}$.

Imputing the sources conditional on $Y^{(2)} = y^{(2)}$, requires another result which follows immediately from general theory:

$$X \sim ZB_x^o + \left(y^{(2)} - ZB_2^o\right)\Psi_{22}^{-1}R_2'\Phi_{yx} + \mathcal{T}\left(\delta + s_y k - l, P_{x|2}, \Phi_{x|2}\right)$$

where

$$P_{x|2} = I_n + ZF^{-1}Z' + \left(y^{(2)} - ZB_2^o\right)\Psi_{22}^{-1}\left(y^{(2)} - ZB_2^o\right)'$$

and

$$\Phi_{x|2} = \Phi_{xx} - \Phi_{xy}R_2\Psi_{22}^{-1}R_2'\Phi_{yx}.$$

The joint conditional predictive distribution of $(X, Y^{(1)}) \mid Y^{(2)} = y^{(2)}$ can be derived in the same way, but we omit the details for brevity.

By applying these results, we find $X_t \mid Y^{(2)} = y^{(2)}$ to be a multivariate-t distribution, a special case of a matric-t distribution:

$$X_t \sim Z_t B_u^o + \left(y_t^{(2)} - Z_t B_2^o\right)\Psi_{22}^{-1}R_2'\Phi_{yx} + \mathcal{T}\left(\delta + s_y k - l, P_{t|2}, \Phi_{x|2}\right), \qquad (4)$$

where

$$P_{t|2} = 1 + Z_t F^{-1} Z_t' + (y_t^{(2)} - Z_t B_2^o)\Psi_{22}^{-1}(y_t^{(2)} - Z_t B_2^o)'$$

and $\Phi_{x|2}$ is defined above. A similar result can be obtained for $Y_t^{(1)} \mid Y^{(2)} = y^{(2)}$. We can now find the conditional predictive distributions:

$$E(Y_t^{(1)} \mid Y^{(2)} = y^{(2)}) = Z_t B_1^o + \left(y_t^{(2)} - Z_t B_2^o\right)\gamma, \quad (5)$$

$$E(X_t \mid Y^{(2)} = y^{(2)}) = Z_t B_x^o + \left(y_t^{(2)} - Z_t B_2^o\right)\Psi_{22}^{-1} R_2' \Phi_{yx}. \quad (6)$$

The posterior distribution of $X_t \mid Y^{(2)} = y^{(2)}$ in (4) yields a simultaneous credibility region. To state our result for the sources, let $\hat{x}_t$ denote the Bayesian predictor of X_t when data are missing-by-design. More precisely, $\hat{x}_t = Z_t B_x^o + (y_t^{(2)} - Z_t B_2^o)\Psi_{22}^{-1} R_2' \Phi_{yx}$. Then given $Y^{(2)} = y^{(2)}$, the $1-\alpha$ level $(0 < \alpha < 1)$ simultaneous posterior credibility region is $\{X_t : (X_t - \hat{x}_t)\Phi_{x|2}^{-1}(X_t - \hat{x}_t)' < b\}$ where, $b = (s_x k * P_{t|2} * F_{1-\alpha, s_x k, \delta-(s_x-s_y)k-l}) * (\delta - (s_x - s_y)k - l)^{-1}$.

3 Specifying the Hyperparameters

To use the imputater described above, the hyperparameters Φ, δ, F, B^o must be specified. Because they are "level 2" parameters we expect their choice to be less critical than would be the specification of "level 1" parameters such as Σ. Moreover, uncertainty about them has already been incorporated to a considerable extent in the hierarchical model so as a result the posterior distribution is a matric-t rather than normal distribution. Finally, as a beneficial by-product of the Bayesian approach we have adopted, the data are expected to update the choices we made based on prior knowledge.

Thus, we elect to specify the hyperparameters in an empirical Bayes approach rather than to attempt a further layer of prior modelling. Although we thereby pay the price of underestimating the model's uncertainty, we gain computational simplicity. As well, the analytical form of the resulting model offers transparency and interpretability.

3.1 Estimating Φ and δ

Relying on the work of Chen (1979) Sun, Le and Zidek (1997, hereafter SLZ) develop an EM algorithm for finding type II MLE's of Φ and δ. To reduce

the number of parameters, they assume a Kronecker covariance structure, $\Phi = \Lambda \otimes \Omega$, Λ corresponding to covariation between sites/sources, on the one hand, and Ω, between species concentrations, on the other. Estimating Φ requires two steps: (i) Λ_y, Ω and δ are estimated by the EM algorithm, where Λ_y denotes the submatrix of Λ corresponding to the monitoring sites; (ii) the extension of Λ_y to Λ.

Data missing-by-design, force SLZ to modify the EM method of BLZ. (Refer to BLZ and Sun (1994) for a general description of the procedure.) The resulting algorithm may be summarized as follows:

E-STEP: Given the current values of Ψ_{yy}, δ,

$$E(\Sigma_{yy}^{-1} \mid Y^{(2)} = y^{(2)}) = R \begin{pmatrix} (\delta + (s_y - s_x)k - 1)\Psi_{1|2}^{-1} & -(\delta + (s_y - s_x)k - 1)\Psi_{1|2}^{-1}\gamma' \\ -(\delta + (s_y - s_x)k - 1)\gamma\Psi_{1|2}^{-1} & d_1 \end{pmatrix} R'$$

where

$$d_1 = [\delta + (s_y - s_x)k + n - l - h - 1]\hat{\Psi}_{22}^{-1} + [\delta + (s_y - s_x)k - 1]\gamma\Psi_{1|2}^{-1}\gamma' + l\Psi_{22}^{-1}$$

and

$$E(\log|\Sigma_{yy}| \mid Y^{(2)} = y^{(2)}, \Phi_{yy}, \delta) = -s_y k log(2) - \sum_{i=1}^{l} \Psi(\frac{\delta + (s_y - s_x)k - i}{2}) -$$

$$\sum_{i=1}^{s_y k - l} \Psi(\frac{\delta + n + (s_y - s_x)k - l - h - i}{2}) + \log|\Psi_{1|2}| + \log|\hat{\Psi}_{22}|.$$

M-STEP: Given the current values of Σ_{yy}^{-1} and $\log \mid \Sigma_{yy} \mid$, find the MLE of $\Phi_{yy} = \Lambda_y \otimes \Omega$, δ by repeating the following steps one and two until convergence and then go on to step three:

1. given the current $\Lambda_y^{(p)}$ and $\Omega^{(p)}$, represent $tr[(\Lambda_y^{(p)} \otimes \Omega^{(p)})\Sigma_{yy}^{-1}]$ as $tr(\Omega^{(p)}G)$ and set $\Omega^{(p+1)} = s_y(\delta + (s_y - s_x)k)G^{-1}$;
2. given current $\Lambda_y^{(p)}$ and $\Omega^{(p+1)}$, represent $tr[(\Lambda_y^{(p)} \otimes \Omega^{(p+1)})\Sigma_{yy}^{-1}]$ as $tr(\Lambda_y^{(p)}Q)$ and set $\Lambda_y^{(p+1)} = k(\delta + (s_y - s_x)k - 1)Q^{-1}$.
3. given the current Ψ_{yy}, estimate δ by solving the equation,

$$\sum_{i=l+1}^{s_y k} \left[\Psi(\frac{\delta + n + (s_y - s_x)k - h - i}{2}) - \Psi(\frac{\delta + (s_y - s_x)k - i}{2})\right]$$

$$= \log|\hat{\Psi}_{22}| - \log|\Psi_{22}|. \qquad (7)$$

The algorithm may be modified to accommodate a parametric covariance structure.

3.2 Estimating B^o and F^{-1}:

Specifying B^o requires prior knowledge and associated modelling. An illustration appears below in the example. We assume this has been done and we turn to the estimation of F^{-1}.

Since $Y^{(2)} = YR_2$, $Y^{(2)} \sim ZB_yR_2 + \mathcal{N}(I_n, \Sigma_{22})$. Define $\hat{B}_2 = (Z'Z)^{-1}Z'Y^{(2)}$ and $S_2 = Y^{(2)\prime}(I - Z(Z'Z)^{-1}Z')Y^{(2)}$. Anderson (1984, p291) shows that given B_y and Σ_{yy},

$$\begin{aligned} \hat{B}_2 &\sim B_2 + \mathcal{N}((Z'Z)^{-1}, \Sigma_{22}) \quad (8) \\ S_2 &\sim W_{s_yk-l}(\Sigma_{22}, n-h), \quad (9) \end{aligned}$$

where $B_2 = B_yR_2$. Since $B_2 = B_2^o + \mathcal{N}(F^{-1}, \Sigma_{22})$, Equation (10) implies

$$\hat{B}_2 \sim B_2^o + \mathcal{N}(F^{-1} + (Z'Z)^{-1}, \Sigma_{22}). \quad (10)$$

Let

$$\hat{B}_2 = \left(\hat{\beta}_1^{j_1}, \ldots, \hat{\beta}_{s_yk-l}^{j_{s_yk-l}}\right),$$

where $j_v \in \{1, \ldots, k\}$, $v = 1, \ldots, s_yk - l$. Obviously $(\hat{\beta}_v^{j_v})' = (Z'Z)^{-1}Z'Y_v^{(2)}$ and j_v marks the species type of $Y_v^{(2)}$. Represent B_2^o in a like manner. Then, for all v,

$$\frac{E(\hat{\beta}_2^{j_v} - \hat{\beta}_2^{oj_v})(\hat{\beta}_2^{j_v} - \hat{\beta}_2^{oj_v})^T}{\Sigma_{(22)(j_vj_v)}} = F^{-1}. \quad (11)$$

This suggests an unbiased estimator for F^{-1}:

$$\hat{F}^{-1} = \frac{n-h-2}{s_2k-l} \sum_{v=1}^{s_2k-l} \frac{(\hat{\beta}_2^{j_v} - \hat{\beta}_2^{oj_v})(\hat{\beta}_2^{j_v} - \hat{\beta}_2^{oj_v})^T}{S_{2(j_vj_v)}} - (Z'Z)^{-1}.$$

Example. Assume: (1) t=1; (2) no systematically missing data so $\Psi = \Phi$; (3) just one species so k=1; (4) $\Psi_{yy} = \Psi_{22} = S_\varepsilon + KS_aK'$, $\Psi_{yx} = KS_a$, and $\Psi xx = S_a$; (8) $\Omega = 1$. As well, we will assume that the expectations of X_t and Y_t are known to be x_{at} and $x_{at}K$, respectively, for all t. Furthermore, assume they have been subtracted.

Results from the previous section imply that

$$X \sim ZB_x^o + yK + T(\delta + s_y, P_{x|y}, \Psi_{x|y}),$$

where $P_{x|y} = I_n + ZF_{-1}Z' + y\Psi_{yy}^{-1}y'$ and $\Psi_{x|y} = S_a - S_aK\Psi_{yy}^{-1}K'S_a = (S_a^{-1} + K'S_\varepsilon^{-1}K)$. F_{-1} can be estimated as described above. As well, δ can be estimated by the EM algorithm above since Ψ_{yy} has been specified.

4 Discussion.

The methodology developed in this paper has a strengths and shortcomings, the latter to be addressed in future work. We summarize both below.

4.1 Strengths.

Str.1 The method attempts to fully reflect uncertainty about the underlying processes and hence model parameters through a hierarchical empirical Bayes approach. The empirical Bayes compromise offers computational simplicity. The resulting predictive matric-t distribution can have almost arbitrarily heavy tails and so provides a good deal of modelling flexibility.

Str.2 The method allows "strength to be borrowed" across space, time and species.

Str.3 Monitoring networks are often a concatenation of subnetworks established for a variety of purposes. Hence, not all monitors will measure the same set of chemical species concentrations, a feature that is allowed in the method presented in this paper.

Str.4 Our method does include covariates to represent temporal patterns and adjust for extraneous variation as long as the latter is measurement.

Str.5 The method does admit an associated theory of design. Thus, we could determine where to put additional monitoring sites so as to best predict the unknown source emissions. We could even contemplate the possibility of measuring some of the sources, albeit at considerable expense.

Weaknesses.

Sht.1 While the empirical Bayes approach captures much of the model's uncertainty, it fails to capture all that uncertainty since it estimates higher level hyperparameters, treats these estimates as certain and does not fully complete the model building process.

Sht.2 Responses are assumed to have a joint conditionally normal distribution. Data may need to be transformed to validate that assumption but finding the appropriate transformation can be difficult. More substantively, the underlying physical models may not be meaningfully on the scale of the transformed scale.

Sht.3 The responses are assumed to be temporally unautocorrelated, an assumption that will not generally obtain, especially for short time aggregates. Some of that correlation can be removed by incorporating systematic components in the covariate matrix, including temporal structure and covariates such as meteorology. Even so, the data will generally need to be prefiltered to validate that assumption.

However, prefiltering will lead to a loss of information, a big loss when the autocorrelation is large. The latter poses difficult modelling problems in any case. Accurately representing large autodependence is critical as failure to do so can have deleterious implications for inference with or without prefiltering.

Sht.4 Dynamically changing hyperparameters cannot be included in level 2 of the hierarchical model we have developed. Hence these will not be "tuned" by the data. However, our model does allow one time only adjustments of level for each site, source and species individually. Thus, for example a global bias adjustment of model outputs could be made.

Sht.5 Monitoring stations have different start-up times but the method in this paper cannot include the resulting monotone data structures. However, the generalization of the paper of Sun, Le and Zidek (1997) given by Kibria et al (2002) addresses this need. However, that generalization comes at a considerable cost in notational and other complexity. We have chosen to present the simpler alternative above.

Acknowledgements.

The work in this paper was done while the first author was a visitor in the Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, North Carolina. He is indebted to the Institute for generously providing facili-

ties. I am indebted to Dr Prasad Kasibhatla for introducing me to the topic of this paper.

References

[1] Anderson,T.W. (1984) *An Introduction to Multivariate Statistical Analysis*. New York: Wiley.

[2] Brown, P.J. (1993) *Measurement, Regression, and Calibration*. Clarendon Press, Oxford.

[3] Brown, P.J., Le, N. D. and Zidek, J.V. (1994a) Multivariate Spatial Interpolation and Exposure to Air Pollutants. *Canadian Journal of Statistics*, **22**, 489-509.

[4] Chen, C.F. (1979) Bayesian Inference for a Normal Dispersion Matrix and its Application to Stochastic Multiple Regression Analysis. *J. R. Statist. Soc.,* **B, 41**, 235-248.

[5] Dawid, A.P. (1981) Some Matrix–variate Distribution Theory. *Biometrika,* **68**, 265-74.

[6] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *J. R. Statist. Soc,* **B, 39**, 1-38.

[7] Le, N. D. and Zidek, J.V. (1992) Interpolation with Uncertain Spatial Covariance: A Bayesian Alternative to Kriging. *J. Mult. Anal,* **43**, 351-74.

[8] Press, S. J. (1982) *Applied Multivariate Analysis-Using Bayesian and Frequentist Methods of Inference*. New York: Holt, Rinehart & Winston

[9] Sun, W. (1994) "Bayesian Multivariate Interpolation with Missing Data and Its Applications." Unpublished Ph.D. Thesis, Department of Statistics, University of British Columbia, Vancouver, Canada.