

A jackknife approach to receptor modeling uncertainty estimation

Cliff Spiegelman (TAMU and TTI)
Eun Sug Park (TTI)

Basic model

$$\mathbf{Y} = \mathbf{A}\mathbf{P} + \boldsymbol{\varepsilon}$$

subject to $(\mathbf{A}, \mathbf{P}) \in \Omega$

Y : $n \times p$ data matrix,

A : $n \times q$ source contribution matrix,

P : $q \times p$ source composition matrix,

ε : $n \times p$ error matrix,

Ω : a set of feasible parameter values.

Existing approaches to uncertainty evaluation

- Propagation of error (delta method)
 - B.A. Roscoe and P.K. Hopke (1981); Park (1997) Dissertation
- Bootstrap
 - Henry et al. (1997), PNAS; Park et al. (2002), Environmetrics; Gajewski and Spiegelman (2004), Environmetrics
- Bayesian approaches
 - Park et al. (2001), JASA; Billheimer (2000), NRCSE Tech report
- Blocked bootstrap
 - Christiansen and Sain (2002), Technometrics
- Simulation experiments (parametric bootstrap)
 - Park et al. (2002), Environmetrics; Gajewski (2000), Dissertation

Our new approach

- One important feature of a bilinear model is that the transpose of the model has a bilinear form. This is important because it suggests the sampling of columns in addition to or in place of rows of the data matrix. The dependence structure of the columns of the observation matrix is different from the typical nonstationary time series structure found in the row dependence.
- Our reason for using a jackknife estimator rather than a bootstrap estimator is that the typical jackknife estimator deletes only one row or column of the data at a time. This saves computational effort. In addition typically identifiability of the parameters may be maintained whereas a typical bootstrap approach to this type of a problem would lose identifiability of the parameters. (This is particularly true if the identifiability constraints are on the contributions (\mathbf{A}).)

Notation

Let $\hat{\mathbf{A}}_{-j}$ and $\hat{\mathbf{P}}_{-j}$ denote the estimates of \mathbf{A} and \mathbf{P} obtained from the dataset with the j th column deleted as j goes from 1 to q . \square

The j th pseudo residuals defined as $r(\hat{\mathbf{A}}, j) = p\hat{\mathbf{A}} - (p-1)\hat{\mathbf{A}}_{-j}$ and $r(\hat{\mathbf{P}}, j) = p\hat{\mathbf{P}} - (p-1)\hat{\mathbf{P}}_{-j}$. In many papers the rows of \mathbf{P} and $\hat{\mathbf{P}}$ are normalized to sum to 1. When such normalization is used then an adjustment needs to be made to $\hat{\mathbf{P}}_{-j}$ before constructing pseudo residuals. In this case each element of the i th row of $\hat{\mathbf{P}}_{-j}$ should be multiplied by $\frac{1}{\sum_{l \neq j} \hat{\mathbf{P}}_{il}}$ in order to account for the deletion of the j th column. \square

Degrees of freedom

Typically the matrix \mathbf{A} is full column rank and if the measurement error is not the major source of variation then $\text{rank}(\mathbf{Y}) \approx \text{rank}(\mathbf{A}\mathbf{P}) \approx \text{rank}(\mathbf{P})$, and since the modeler knows the rank of \mathbf{P} the degrees of freedom should be easy to set.

To protect against the over parameterize case we estimate the rank of our model using the NUMFACT algorithm presented in Henry et al. (1999), and Park et al. (2000). The NUMFACT method is a way to estimate the underlying rank of matrix that is subject to error. We denote the estimated rank of the data determined by the NUMFACT algorithm by *NUMrank*. Then we protect against the effect of over parameterization by using $\text{MIN}(\text{rank}(\mathbf{P}), \text{NUMrank})$ as our degrees of freedom for the jackknifed estimates of variance. We are less worried about too few degrees of freedom in the case of under parameterization as that leads to model bias as being important and we use a jackknife estimate of bias to assess this issue.

NUMFACT and Related Statistics

- Notation:
- R : sample correlation matrix
- l_1, \dots, l_p : eigenvalues of R
- b_1, \dots, b_p : eigenvectors of R
- N : the number of independent resamples.
- $b_{1j}^*, \dots, b_{pj}^*$: eigenvectors of the correlation matrix R_j^* of the j th resample ($j = 1, \dots, N$).

$$W_i = \frac{\sum_{j=1}^N b_{ij}^{*t} (b_1 b_1^t + b_2 b_2^t + \dots + b_i b_i^t) b_{ij}^* / N}{1 - \sum_{j=1}^N b_{ij}^{*t} (b_1 b_1^t + b_2 b_2^t + \dots + b_i b_i^t) b_{ij}^* / N},$$

$i = 1, \dots, p-1$, and $\overline{W_p} \equiv 0$.

The Numfact statistic

$$S_i = \frac{\text{signal}_i}{\text{noise}} = \frac{\frac{l_i \sqrt{W_i}}{1 + \sqrt{W_i}}}{\left(\sum_{k=1}^{p-1} \frac{l_k}{1 + \sqrt{W_k}} \right) / (p-1)},$$

$i = 1, \dots, p-1$, and $S_p \equiv 0$.

Some Numfact examples

Table 3. Description of the Data Sets

Data Set	No. Vars.	No. Cases	Putative Number of Factors	References
Air Pollution Composition	13	2804	7	Henry et al.
Air Pollution Spatial	11	53	3	Henry
IR Spectra	17	81	3	Windig
Train Timing	10	40	3	
Word Count	26	54	5	
Ceramic Melter Temperatures	16	300	6	Wise et al. 1991

- Numfact seems to work well in receptor models and that was the motivating example
- Available from the PLS_Toolbox the main chemometric software package

Table 4. Estimates the Number of Factors by Different Methods

Data Set	Rule-of-One	90 Percent Var.	Scree Plot	Bartlett Correlation	Bartlett Covariance	Numfact
Air Pollution Compositional	3	4	4	6	9	7
Air Pollution Spatial	3	4	3	2	1	3
IR Spectra	2	2	3	9	8	3
Train Timing	2	3	2	3	2	3
Word Count	4	3	8	24	24	4
Ceramic Melter Temperatures	5	5	6	9	8	6

Usual simulation model

Simulated Example

We first consider a simulated example to illustrate (1). The data is generated based on the basic model with true source composition matrix P_0 given in Table 1 where $n = 50$, $p = 9$, and $q = 3$. The source contribution matrix A is generated from a truncated multivariate normal distribution $N_3(3, I_3)$ where μ and I_3 represents a 3×3 identity matrix. The errors associated with n observations are independently generated from the normal distribution with mean 0 and a diagonal covariance matrix so that the proportions of the error standard deviations to the model standard deviations are about 12-33%. The resulting data matrix Y consists of nonnegative numbers.

1	2	3	4	5	6	7	8	9
0.132	0	0	0.263	0	0.159	0.079	0.105	0.263
0.135	0.000	0.000	0.269	0.000	0.155	0.080	0.104	0.256
0.005	0.000	0.000	0.003	0.000	0.004	0.004	0.005	0.002
0.005	0.000	0.000	0.004	0.000	0.004	0.003	0.004	0.004
0.119	0.000	0.000	0.260	0.000	0.141	0.068	0.090	0.249
0.152	0.000	0.000	0.278	0.000	0.169	0.092	0.119	0.264
0.126	0.000	0.000	0.262	0.000	0.146	0.074	0.096	0.248
0.144	0.000	0.000	0.277	0.000	0.164	0.086	0.111	0.263
0	0.214	0.119	0	0.167	0	0.143	0.119	0.238
0.000	0.217	0.121	0.000	0.159	0.000	0.143	0.114	0.247
0.000	0.003	0.019	0.000	0.013	0.000	0.009	0.011	0.004
0.000	0.004	0.007	0.000	0.005	0.000	0.004	0.005	0.005
0.000	0.208	0.080	0.000	0.117	0.000	0.113	0.078	0.234
0.000	0.225	0.182	0.000	0.201	0.000	0.172	0.130	0.259
0.000	0.208	0.105	0.000	0.150	0.000	0.135	0.105	0.238
0.000	0.223	0.134	0.000	0.169	0.000	0.149	0.124	0.257
0.146	0.146	0.229	0.083	0	0	0.188	0	0.208
0.152	0.139	0.233	0.081	0.000	0.000	0.191	0.000	0.204
0.004	0.005	0.005	0.008	0.000	0.000	0.002	0.000	0.003
0.006	0.007	0.008	0.010	0.000	0.000	0.005	0.000	0.006
0.139	0.122	0.217	0.056	0.000	0.000	0.184	0.000	0.196
0.165	0.155	0.249	0.106	0.000	0.000	0.199	0.000	0.213
0.139	0.125	0.218	0.062	0.000	0.000	0.182	0.000	0.191
0.164	0.151	0.250	0.099	0.000	0.000	0.200	0.000	0.213

Standard errors are obtained based on a column deletion (from species 1 to 9).

Standard errors are obtained based on 200 bootstrap samples for which resampling is done over the rows.

Cl_{lower} represent the lower limit and the upper limit of the 95% Jackknife confidence intervals, respectively.

Cl_{upper} represent the lower limit and the upper limit of the 95% Bootstrap confidence intervals, respectively.

Autocorrelated example

- Same profiles
- Autocorrelation for errors = .7

Table 2. True source composition profiles (P_D), estimated source composition profiles (\hat{P}), Jackknife standard errors (JKSE) and bootstrap standard errors (BSE), Jackknife confidence intervals (JKCI), and Bootstrap confidence intervals (BCI) for P when the observations are temporally correlated

	1	2	3	4	5	6	7	8	9
True	0.132	0	0	0.263	0	0.159	0.079	0.105	0.263
Estimate	0.124	0.000	0.000	0.265	0.000	0.171	0.075	0.113	0.251
JKSE	0.009	0.000	0.000	0.011	0.000	0.016	0.009	0.010	0.026
BSE	0.005	0.000	0.000	0.003	0.000	0.004	0.004	0.003	0.005
JKCI _L	0.099	0.000	0.000	0.230	0.000	0.122	0.046	0.081	0.169
JKCI _U	0.151	0.000	0.000	0.300	0.000	0.221	0.104	0.144	0.334
BCI _L	0.114	0.000	0.000	0.259	0.000	0.164	0.069	0.107	0.242
BCI _U	0.134	0.000	0.000	0.271	0.000	0.179	0.083	0.118	0.260
True	0	0.214	0.119	0	0.167	0	0.143	0.119	0.239
Estimate	0.000	0.216	0.110	0.000	0.172	0.000	0.141	0.127	0.234
JKSE	0.000	0.004	0.012	0.000	0.017	0.000	0.009	0.013	0.007
BSE	0.000	0.002	0.004	0.000	0.003	0.000	0.003	0.003	0.003
JKCI _L	0.000	0.204	0.072	0.000	0.119	0.000	0.111	0.087	0.213
JKCI _U	0.000	0.229	0.147	0.000	0.225	0.000	0.170	0.169	0.255
BCI _L	0.000	0.211	0.101	0.000	0.165	0.000	0.135	0.120	0.229
BCI _U	0.000	0.220	0.118	0.000	0.179	0.000	0.145	0.134	0.240
True	0.148	0.148	0.229	0.083	0	0	0.189	0	0.209
Estimate	0.148	0.127	0.217	0.091	0.000	0.000	0.190	0.000	0.229
JKSE	0.005	0.015	0.011	0.010	0.000	0.000	0.009	0.000	0.025
BSE	0.005	0.005	0.005	0.005	0.000	0.000	0.003	0.000	0.005
JKCI _L	0.130	0.079	0.182	0.059	0.000	0.000	0.164	0.000	0.149
JKCI _U	0.162	0.175	0.252	0.123	0.000	0.000	0.217	0.000	0.309
BCI _L	0.138	0.119	0.207	0.080	0.000	0.000	0.184	0.000	0.219
BCI _U	0.155	0.136	0.227	0.101	0.000	0.000	0.197	0.000	0.240

Jackknife standard errors are obtained based on a column deletion (from species 1 to 9).

Bootstrap standard errors are obtained based on 200 bootstrap samples for which resampling is done over the rows.

JKCI_L and JKCI_U represent the lower limit and the upper limit of the 95% Jackknife confidence intervals, respectively.

BCI_L and BCI_U represent the lower limit and the upper limit of the 95% Bootstrap confidence intervals, respectively.

Intervals not capturing the true parameter value are shown in bold.

Bootstrap Vs Jackknife CIs

Jackknife captured the true value 95% of the time (missed 1 out of 18)

Bootstrap missed 8 out of 18 cases

Clinton Drive Example

The original data consists of 2,541 hourly observations (after initial screening of the outliers) on 54 volatile organic compounds (VOC) and total nonmethane organic carbon (TNMOC) - Henry, Spiegleman, Collins, and Park (1997).

Ten species are selected for the current study.

PC plots suggesting 3 main sources

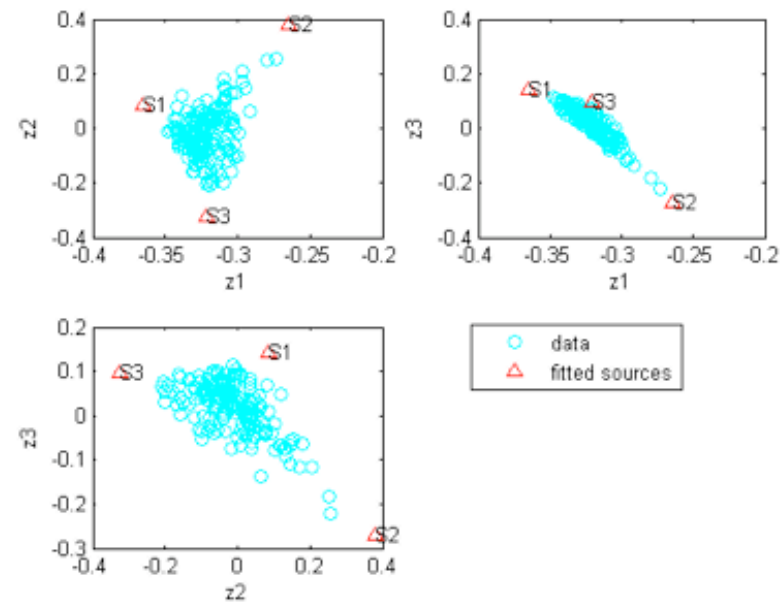


Figure 1. Principal component plots of the VOC data (o) and the fitted sources by CNLS (Δ)

Comparison of uncertainty limits

Table 4. Estimated source composition profiles (\hat{P}), Jackknife standard errors (JKSE), Bootstrap standard errors (BSE), Jackknife confidence intervals (JKCI), and Bootstrap confidence intervals (BCI) for P

Species		1	2	3	4	5	6	7	8	9	10
Source 1	Estimate	0.029	0.000	0.035	0.098	0.847	0.000	0.107	0.062	0.000	0.022
	JKSE	0.042	0.000	0.057	0.064	0.081	0.000	0.035	0.090	0.000	0.010
	BSE	0.020	0.000	0.007	0.004	0.021	0.000	0.008	0.011	0.000	0.005
	JKCI _L	0.000	0.000	0.000	0.000	0.389	0.000	0.000	0.000	0.000	0.000
	JKCI _U	0.162	0.000	0.215	0.303	0.904	0.000	0.220	0.348	0.000	0.055
	BCI _L	0.000	0.000	0.021	0.088	0.801	0.000	0.091	0.037	0.000	0.013
	BCI _U	0.072	0.000	0.049	0.107	0.882	0.000	0.121	0.082	0.000	0.032
Source 2	Estimate	0.028	0.000	0.000	0.009	0.000	0.034	0.043	0.228	0.838	0.021
	JKSE	0.051	0.000	0.000	0.040	0.000	0.028	0.032	0.079	0.077	0.014
	BSE	0.017	0.000	0.000	0.003	0.000	0.004	0.009	0.015	0.022	0.004
	JKCI _L	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.393	0.000
	JKCI _U	0.191	0.000	0.000	0.138	0.000	0.122	0.146	0.477	0.883	0.087
	BCI _L	0.000	0.000	0.000	0.000	0.000	0.028	0.021	0.200	0.587	0.011
	BCI _U	0.083	0.000	0.000	0.015	0.000	0.042	0.060	0.255	0.882	0.028
Source 3	Estimate	0.514	0.335	0.015	0.018	0.000	0.038	0.055	0.000	0.000	0.025
	JKSE	0.033	0.040	0.051	0.030	0.000	0.008	0.032	0.000	0.000	0.011
	BSE	0.012	0.011	0.007	0.003	0.000	0.003	0.007	0.000	0.000	0.004
	JKCI _L	0.407	0.207	0.000	0.000	0.000	0.012	0.000	0.000	0.000	0.000
	JKCI _U	0.820	0.483	0.178	0.112	0.000	0.085	0.158	0.000	0.000	0.039
	BCI _L	0.488	0.320	0.001	0.012	0.000	0.034	0.043	0.000	0.000	0.018
	BCI _U	0.535	0.383	0.028	0.025	0.000	0.044	0.068	0.000	0.000	0.031

Notes. 1. Jackknife standard errors are obtained based on a column deleted (from species 1 to 10).

2. Bootstrap standard errors are obtained based on 200 bootstrap samples for which resampling is done over the rows.

3. JKCI_L and JKCI_U represent the lower limit and the upper limit of the 95% Jackknife confidence intervals, respectively.

4. BCI_L and BCI_U represent the lower limit and the upper limit of the 95% Bootstrap confidence intervals, respectively.

Questions?

Further Discussion

- No model is exact but some are useful
 - Receptor models are often at least in part exploratory
 - Missing sources show up differently as species are dropped the projections on the remaining source profile columns may change dramatically
 - This is less likely to occur with the bootstrap approach particularly when the sample sizes are large
- Cliff is working on a bootstrap approach with Phil Hopke and Ron Henry may be working on his own. Both approaches take dependence structure into account. Cliff and Phil require short-midrange range dependence and use a variant of block bootstrapping. The issue to be addressed is bias adjustment.