GEOPHYSICAL RESEARCH LETTERS, VOL. 22, NO. 3, PAGES 307-310, FEBRUARY 1, 1995

The bootstrap method in space physics: Error estimation for the minimum variance analysis

H. Kawano[1]

Department of Earth and Planetary Physics, University of Tokyo, Tokyo, Japan

T. Higuchi

The Institute of Statistical Mathematics, Tokyo, Japan

Abstract. The minimum variance analysis technique introduced by Sonnerup and Cahill (1967) is a useful tool in space physics. However the statistical errors appearing in this method are difficult to estimate accurately because of the complicated form of the eigenvalue decomposition. To deal with this problem, this paper introduces the bootstrap method (Efron, 1979) which replaces analytical solutions with repeated simple calculations. We apply this method to the estimation of the statistical errors in the minimum variance direction and the average component in the minimum variance direction, and show that this method accurately estimates the errors.

Introduction

The technique referred to as the minimum variance analysis (MVA) or as the principal axis analysis (PAA) has been widely used to determine the normal directions of discontinuities in space [e.g., *Sonnerup and Cahill*, 1967; *Aubry et al.*, 1971; *Neugebauer et al.*, 1984], to determine wave normal directions [e.g., *Thorne et al.*, 1973; *Smith and Tsurutani*, 1976], and to examine the topology of plasmoids [e.g., *Slavin et al.*, 1993; *Moldwin and Hughes*, 1994]. The statistical errors (i.e., the accuracy) of the direction obtained and that of the magnetic field normal component have been estimated by *Sonnerup* [1971]. *Hoppe et al.* [1981] also presented a qualitative discussion of the statistical errors. The equations by *Sonnerup* [1971] give upper bounds for the errors as a result of the assumption that all terms contributing to the errors are independent. A brief introduction to MVA is shown in the Appendix with a minor modification to Sonnerup's formulas. One of the purposes of this paper is to provide a method for a more accurate estimation of the errors.

The mathematical procedures involved in MVA are identical to those in a principal component analysis (PCA) which is an old but useful method in multivariate analysis [*Hotelling*, 1933; *Anderson*, 1958]. Most of theoretical analysis in PCA are obtained by assuming that the data set are observations from a multivariate Gaussian distribution [*Beran and Srivastava*, 1985]. This special assumption is required mainly for mathematical tractability (in other words, for convenience). The problems in the application of PCA to actual data sets are as follows:

- Unless the actual data can be satisfactorily approximated to follow the multivariate Gaussian distribution, the method brings about serious biases.
- Even with the assumption of Gaussian distribution, it is quite difficult to get analytic equations for this relatively complicated statistical system.

In fact many kinds of data in space physics show non-Gaussian distributions [e.g., *Tsurutani et al.*, 1990]. In order to overcome these problems, a new statistical method, called the Bootstrap method, has been proposed [*Efron*, 1979]. The bootstrap method is a general methodology for non-parametrically estimating the statistical errors, such as the bias and standard error. This method requires much fewer distribution assumptions than predecessors, but demands high computing power for its realization. Wide-ranging introductions to the bootstrap method with its applications to various statistical problems are provided by *Efron* [1982], *Efron and Tibshirani* [1986, 1993], and *Kubokawa et al.* [1993]. Related ideas and improvements of the bootstrap method for efficient calculation are also found in these references. The papers by *Diaconis and Efron* [1983] and *Efron and Tibshirani* [1991] are appropriate for the statistical practitioner. Descriptions related to PCA in the bootstrap method, namely, a discussion of a covariance matrix and its interesting functions can be found in *Diaconis and Efron* [1983], *Efron and Tibshirani* [1993], and particularly, in *Beran and Srivastava* [1985]. Specific applications along this line to paleomagnetic analyses are given by, e.g., *Tauxe et al.* [1991].

The major object of this paper is to introduce the bootstrap method which provides more information about the statistical error of the estimated parameters in MVA, with fewer assumptions concerning the underlying distribution functions of the data. We apply this method to estimate the statistical errors of the minimum variance direction and the magnetic field normal component with the objective of demonstrating both its usefulness and its wide applicability to data analysis in the geophysical field.

Methodology

We begin with a description of the basic algorithm of the bootstrap method in terms of MVA. The multivariate data examined by means of MVA is usually a three-dimensional vector time series, $B(k)$ $(k = 1, \ldots, K)$, where $B(k) = [Bx(k), By(k), Bz(k)]^T$ and T denotes the transposition operation. It is convenient to express the observed data by an $3 \times K$ data matrix, X^O, the k-th column of X^O being $B(k)$:

$$X^O = [B(1), B(2), \ldots, B(K)]. \tag{1}$$

The superscript O is used to explicitly indicate that X^O is the original (observed) data set.

The bootstrap algorithm begins by generating a simulated data matrix, called the *bootstrap sample*, based on the simple assumption that each data has equal probability $1/K$. The procedure for its implementation on the computer is as follows; a random number device selects integers $j_1, j_2, \ldots, j_K$, each of which equals any value between 1 and K with probability $1/K$. Then, we get the *bootstrap sample* X^{*i} $(i = 1, \ldots, N)$ as

$$\begin{aligned} X^{*i} &= [B(j_1), B(j_2), \ldots, B(j_K)] \\ &= [B^{*i}(1), B^{*i}(2), \ldots, B^{*i}(K)]. \end{aligned} \tag{2}$$

[1] Now at Institute of Geophysics and Planetary Physics, University of California, Los Angeles.

Paper number 94GL02969
0094-8534/95/94GL-02969$03.00

The star notation indicates that X^{*i} is not the actual data matrix X^{O}, but a randomized version of X^{O}. N is the number of trials, an *ad-hoc* parameter for an actual application of the bootstrap method, as discussed later. Note that any chosen $B(k)$ may appear in X^{*i} zero times, once, twice, etc..

The next step in the bootstrap method is to estimate a parameter of interest, Θ, from X^{*i}. $\widehat{\Theta}$ usually indicates the *estimate* for Θ in the statistical literature. Then, the *estimate* for Θ on the basis of the *bootstrap sample* X^{*i}, called the *bootstrap replication*, is henceforth denoted as $\widehat{\Theta}^{*i}$. An example of $\widehat{\Theta}^{*i}$ is the mean vector, given by $\widehat{\mu}^{*i} = \sum_{k=1}^{K} B^{*i}(k)/K$. Another example is eigenvectors calculated by MVA , denoted by $\widehat{e}_j^{*i}$ $(j = 1, 2, 3)$.

The eigenvector corresponding to the smallest eigenvalue, or the minimum variance, $\widehat{e}_3^{*i}$, plays an important role in MVA, and is written below as $\widehat{n}^{*i} = [\widehat{n}_x^{*i}, \widehat{n}_y^{*i}, \widehat{n}_z^{*i}]^T$. Also important is the mean vector component along the minimum variance eigenvector; it is written below as $\widehat{B}_n^{*i} = \widehat{\mu}^{*i} \cdot \widehat{n}^{*i}$. In this paper we focus on $\widehat{n}^{*i}$ and $\widehat{B}_n^{*i}$.

The final step in the bootstrap method is to give a measure of statistical accuracy for the above quantities. In this study we use the standard error, and introduce two ways to calculate it. There is another, fundamentally more ambitious measure of statistical accuracy than standard errors, namely confidence intervals. Much interest is attached to bootstrap confidence intervals [*Efron and Tibshirani*, 1986; 1993; *Hall*, 1992; *Konishi*, 1990], but we omit this subject here.

A direct way for estimation of the standard error is given by the sample standard deviation of the N *replications*

$$\widehat{\Delta}_N(\Theta^{*}) = \sqrt{\frac{\sum_{i=1}^{N}\left(\widehat{\Theta}^{*i} - \bar{\Theta}^{*}\right)^2}{N-1}} \qquad (3)$$

where $\bar{\Theta}^{*} = \sum_{i=1}^{N} \widehat{\Theta}^{*i}/N$. Note that the notation $\widehat{\Delta}_N$ is usually replaced by $\widehat{se}_N$ in the statistical literature. Another way to obtain the bootstrap standard error is based on the 0.1587 and/or 0.8413 quantiles which correspond to -1 and 1 sigma points of normal density, respectively. It is obvious that 0.5000 quantile is the median. The latter way is useful for non-Gaussian distributions, but in this study we adopt $\widehat{\Delta}_N$. We also note that, in an actual data analysis, we should not rely entirely on a single statistic like $\widehat{\Delta}_N$, but should show *bootstrap replications* graphically [*Efron and Tibshirani*, 1993]: we always prefer to display a histogram of a *bootstrap replication* $\widehat{\Theta}^{*i}$.

There only remains one *ad-hoc* parameter for an actual application of the bootstrap method, namely the number of the *bootstrap samples* N. It is obvious that any bootstrap estimation can be improved as $N \to \infty$. Consequently, a bootstrap with a large number N is desired, but it requires a large computational effort. We therefore need a criterion which tells us a rough minimum number necessary for satisfactory results. Several rules gathered from many experiences with various situations are available [*Efron*, 1987; *Efron and Tibshirani*, 1993; *Kubokawa et al.*, 1993]; usually the necessary N does not exceed 2000, in particular for an estimation of $\widehat{\Delta}_N$. However in this study we set $N = 10^5$, which is larger than values suggested by *Efron and Tibshirani* [1993], in order to get very reliable results and in order to show the distributional shape of $\widehat{n}^{*i}$ and $\widehat{B}_n^{*i}$ clearly. Even with this enormous number, the total calculation time does not exceed 10 min for the MVA applications shown below, thanks to current computer speed.

Simulation Settings

Here we simulate an observed magneticfield data set of a rotational-discontinuity crossing, X^{O}, as follows:

$$\begin{aligned} Bx(k) &= 50\,sin(\theta_k) + \varepsilon_{3k-2} \\ By(k) &= 50\,cos(\theta_k) + \varepsilon_{3k-1} \\ Bz(k) &= 2 + \varepsilon_{3k} \ \ \text{[nT]} \\ \theta_k &= -60 + \frac{120}{K-1}(k-1) \ \ \text{[degree]} \\ k &= 1, 2, \ldots, K \end{aligned} \qquad (4)$$

where ε_j $(j = 1, 2, \ldots, 3K)$ are the independent and identically distributed (i.i.d.) Gaussian white noise sequences having a mean of 0 and a standard deviation of 2, thus the noise vectors $\varepsilon(k) = [\varepsilon(3k-2), \varepsilon(3k-1), \varepsilon(3k)]^T$ $(k = 1, \ldots, K)$ follow an i.i.d. isotropic three-dimensional Gaussian distribution. The x, y, and z axes correspond to the maximum, intermediate, and minimum variance directions without noise. That is, the "true" value of the minimum variance unit vector, n, is $n = [0, 0, 1]^T$, and the "true" value of the mean magnetic field component along the direction of n, B_n, is 2. Here K is set to 30.

From this X^{O} we generate N *bootstrap samples* X^{*1}, $X^{*2}, \ldots, X^{*N}$ with Eq. (2). As mentioned above, we use $N = 10^5$ in this study. By applying MVA to these *bootstrap samples*, we obtain 10^5 values of $\widehat{n}^{*i}$ and $\widehat{B}_n^{*i}$. On the basis of these *replications*, we can calculate the standard errors in them ($\widehat{\Delta}_N(\widehat{n}^{*})$ and $\widehat{\Delta}_N(\widehat{B}_n^{*})$).

As a comparison, we need to know the "true" distribution functions for n and B_n. Because this is a simulation study, we can find them by repeatedly creating datasets according to Eq. (4) but with different independent sets of noise. That is, we generate M datasets $X^{s,1}, X^{s,2}, \ldots, X^{s,M}$, each of which is obtained by Eq. (4). The superscript s is used to indicate that the dataset $X^{s,m}$ is neither the original dataset X^{O} nor the *bootstrap sample* X^{*i}. The ε_j $(j = 1, \ldots, 3 \times K \times M)$ values needed for this procedure are the i.i.d. Gaussian white noise sequences. We calculate $n^{s,m}$ and $B_n^{s,m}$ from $X^{s,m}$, and thus get M values of them and draw histograms. In order to conveniently compare these "true" histograms with the histograms of *bootstrap replications* $\widehat{n}^{*i}$ and $\widehat{B}_n^{*i}$, we set $M = N$.

Results and Discussion

Figures 1, 2, and 3 compare the histograms of the *bootstrap replications* $\widehat{n}_x^{*i}$, $\widehat{n}_y^{*i}$, and $\widehat{B}_n^{*i}$ (solid lines) with the "true" distributions (dashed lines). The vertical axis shows the number of cases for each bin out of a total of 10^5 $(= M = N)$ cases. The figures show that the bootstrap distribution approximates the true distribution very well, even though the bootstrap distribution is produced by using only K (= 30) data vectors. The small bias seen in this figure comes

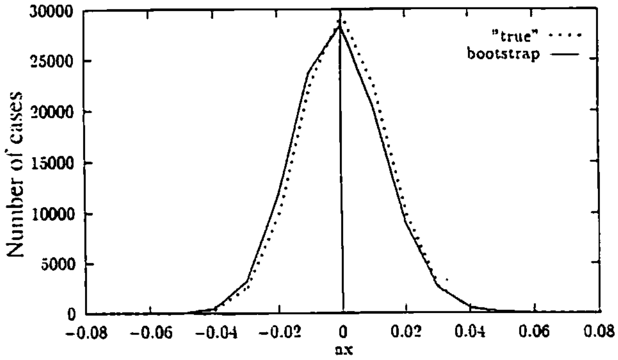

Figure 1. Distributions of bootstrap replications $\widehat{n}_x^{*i}$ (solid lines) and "true" n_x (dashed lines).

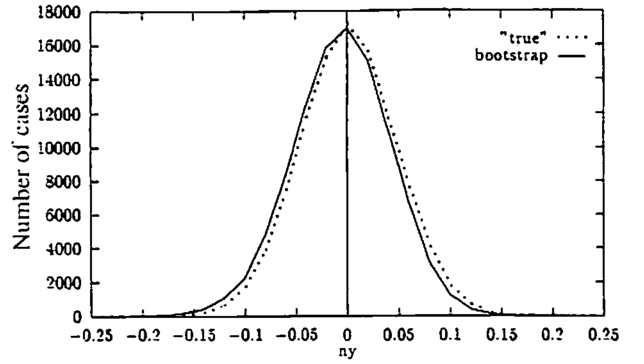

Figure 2. Distributions of bootstrap replications $\hat{n}_y^{*i}$ (solid lines) and "true" n_y (dashed lines).

from the fact that, for this X^O, the average of 30 $\varepsilon(k)$'s is close to but not exactly equal to a null vector.

Table 1 shows that standard deviations of the bootstrap distribution and the "true" distribution are very close, as a natural consequence of these distributions being almost the same, as shown in the figures. We have confirmed this good agreement for other X^O's with different parameters, such as different rotation angle in xy plane and different noise amplitude (not shown), in Eq. (4). We conclude that the bootstrap method provides a reasonable estimation of the statistical errors in n and B_n.

The table also lists error estimates obtained from the equations of *Sonnerup* [1971] for the original dataset X^O: the next to last column comes from his original equation, and the right-most column comes from equations with a small modification, as discussed in the Appendix. The table shows that usage of the original or the modified equations does not much affect the estimation (in the limit $K \rightarrow \infty$ they agree), but the modified equations are based on a more strict formulation and are simpler.

The table also shows that error estimates in the right-most two columns of the table are 2-3 times larger than the "true" values (this ratio seems to be common to other X^O's with different parameters in Eq. (4), but note that Eq. (4) represents the case of a planar magnetopause case only). In relation to this we remark that *Lepping and Behannon* [1980] stated that Sonnerup's formula underestimates the error, but we also note that, while Lepping and Behannon's numerical experiments appear valid, their formula does not include a dependency on the number of observed data (K in our notation), which is strange for a statistical error formula.

In this paper we have used the planar magnetopause model (Eq. 4). Because MVA implicitly assumes planarity, other kinds of "error" appear when one applies it to the case where the magnetopause is not planar and/or when temporal effects are present. The bootstrap method stated above will sense these error as well, because it treats any deviations from planarity in the same manner. This is currently being tested by using different models.

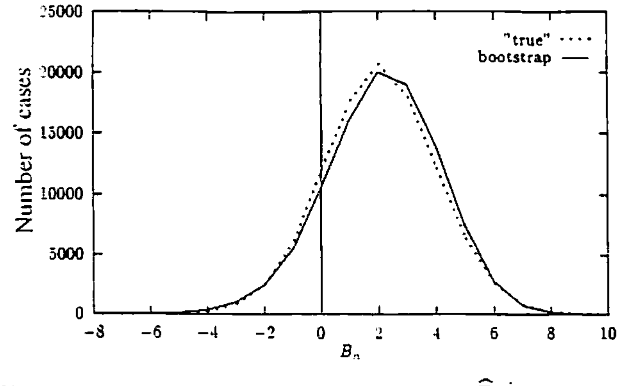

Figure 3. Distributions of bootstrap replications $\widehat{B}_n^{*i}$ (solid lines) and "true" B_n (dashed lines).

Table 1. Error estimates

	"true" errors	bootstrap error estimates	*Sonnerup*'s [1971] error estimates	estimates in Appendix
n_x	0.013	0.014	0.030	0.035
n_y	0.047	0.047	0.101	0.119
B_n	1.9	2.0	4.1	4.8

Finally, we comment on the other eigenvectors, that is, the maximum and intermediate variance directions. The statistical errors of these vectors can also be estimated by the bootstrap method: this should be important because the maximum variance direction for the convection electric field, $-\boldsymbol{v} \times \boldsymbol{B}$, is now being used as an estimate for the magnetopause normal direction *Sonnerup et al.* [1987] but no formulas have been given in the literature to estimate these errors. However care must be taken because the maximum and intermediate variance components of the field usually include an apparent trend. In this case, we need a modification of how to generate the *bootstrap sample* X^{*i}. This subject is beyond our major purpose here; it is open to future research (see however *Leger et al.* [1992]).

In summary, the bootstrap method is useful in estimating statistical errors in the minimum variance direction and in the average minimum variance component. Moreover, as stated in the introduction, this method is useful in a broad area of error estimations; it depends mainly on the computer speed, and needs few assumptions; in particular, it is applicable even if the population distribution is non-Gaussian.

Appendix: Modification of minimum variance error estimates by *Sonnerup* [1971]

Here the maximum, intermediate, and minimum variance axes are expressed by subscripts 1, 2, and 3, thus $n = e_3$ and $B_n = B_3 = B \cdot e_3$, where e_i denotes the unit vector. The variances λ_1, λ_2, and λ_3 are calculated as eigenvalues of a sample covariance matrix whose (i, j) element (in any coordinates) is expressed as

$$G_{ij} = \frac{1}{K-1} \sum_{k=1}^{K} (B_i(k) - \bar{B}_i)(B_j(k) - \bar{B}_j) \quad \text{(A1)}$$

where K is the number of data points, and the overhead bar denotes an average. Note the denominator is not K but $(K-1)$, in an attempt to get an unbiased variance.

The equations for the random error estimates are given as

$$\Delta e_3 = \pm e_1 \sqrt{\frac{\Delta\lambda}{\lambda_1 - \lambda_3}} \pm e_2 \sqrt{\frac{\Delta\lambda}{\lambda_2 - \lambda_3}} \quad \text{(A2a)}$$

$$\Delta \bar{B}_3 = \pm \sqrt{\frac{\lambda_3}{K} + \bar{B}_1{}^2 \frac{\Delta\lambda}{\lambda_1 - \lambda_3} + \bar{B}_2{}^2 \frac{\Delta\lambda}{\lambda_2 - \lambda_3}} \quad \text{(A2b)}$$

where the subscript denote minimum variance coordinates and $(\Delta\lambda)^2$ is the variance of λ_3. For a derivation of these equations, refer to the appendix of *Sonnerup* [1971].

When $B_3(k)$ (in the minimum variance coordinates) follows a Gaussian distribution with a population variance of V, a variable $\sum_{k=1}^{K} (B_3(k) - \bar{B}_3)^2$ has the χ-square distribution with $(K-1)$ degrees of freedom. Then, the value λ_3, corresponding to the sample variance of $B_3(k)$, can be estimated to have an average of V and variance of $2V^2/(K-1)$. By substituting λ_3 for V, we get

$$\Delta\lambda = \lambda_3\sqrt{\frac{2}{K-1}}\,. \qquad \text{(A3)}$$

This expression for $\Delta\lambda$ is the major modification we propose of the calculation by *Sonnerup* [1971]. A minor modification is that the denominator of the first term of (A2b) is not $(K-1)$ but K: the first term denotes the variance of $\bar{B}_3$, which is identical to V/K if $B_3(k)$ follows a Gaussian distribution.

Acknowledgments. The work at the University of Tokyo was supported by the Japanese Society for the Promotion of Science under grant 3577. This study was carried out under the ISM Cooperative Research Program (93-ISM.CRP-54).

References

Anderson, T. W., *An introduction to multivariate statistical analysis*, John Wiley & Sons, New York, 1958.

Aubry, M. P., M. G. Kivelson, and C. T. Russell, Motion and structure of the magnetopause, *J. Geophys. Res.*, *76*, 1673-1696, 1971.

Beran, R., and M.S. Srivastava, Bootstrap tests and confidence regions for functions of a covariance matrix, *The Annals of Statistics*, *13*, 95-115, 1985.

Diaconis, P., and B. Efron, Computer-intensive methods in statistics, *Scientific American*, *248*, 116-130, 1983.

Efron, B., Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*, 1-26, 1979.

Efron, B., *The Jackknife, the bootstrap and other resampling plans*, SIAM (Society for Industrial and Applied Mathematics), Philadelphia, 1982.

Efron, B., Better bootstrap confidence intervals (with discussion), *Journal of the American Statistical Association*, *82*, No. 397, 171-200, 1987.

Efron, B., and R. J. Tibshirani, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical Science*, *1*, 54-77, 1986.

Efron, B., and R. J. Tibshirani, Statistical data analysis in the computer age, *Science*, *253*, 390-395, 1991.

Efron, B., and R. J. Tibshirani, *An introduction to the bootstrap*, Chapman & Hall, New York, 1993.

Hall, P., *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York, 1992.

Hoppe, M. M., C. T. Russell, L. A. Frank, T. E. Eastman, and E. W. Greenstadt, Upstream hydromagnetic waves and their association with backstreaming ion populations: ISEE 1 and 2 observations, *J. Geophys. Res.*, *86*, 4471-4492, 1981.

Hotelling, H., Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, *24*, 417-441 and 498-520, 1933.

Konishi, S., Bootstrap methods and confidence intervals, *Japanese Journal of Applied Statistics*, *19*, 137-162, 1990 (in Japanese).

Kubokawa, T., S. Eguchi, A. Takemura, and S. Konishi, Recent developments of the theory of statistical inference, *Journal of Japan Statistical Society*, *22*, No.3, 257-312, 1993 (in Japanese with English Abstract).

Leger, C., D. N. Politis, and J. P. Romano, Bootstrap technology and applications, *Technometrics*, *34*, No. 4, 378-398, 1992.

Lepping, R. P., and K. W. Behannon, Magnetic field directional discontinuities. I. Minimum variance errors, *J. Geophys. Res.*, *85*, 4695-4703, 1980.

Moldwin, M. B., and W. J. Hughes, Observations of earthward and tailward propagating flux rope plasmoids: Expanding the plasmoid model of geomagnetic substorms, *J. Geophys. Res.*, *99*, 183-198, 1994.

Neugebauer, M., D. R. Clay, B. E. Goldstein, B. T. Tsurutani, and R. D. Zwickl, A reexamination of rotational and tangential discontinuities in the solar wind, *J. Geophys. Res.*, *89*, 5395-408, 1984.

Slavin, J. A., M. F. Smith, E. L. Mazur, D. N. Baker, E. W. Hones, Jr., T. Iyemori, and E. W. Greenstadt, ISEE 3 observations of traveling compression regions in the earth's magnetotail, *J. Geophys. Res.*, *98*, 15,425-15,446, 1993.

Smith, E. J., and B. T. Tsurutani, Magnetosheath lion roars, *J. Geophys. Res.*, *81*, 2261-2266, 1976.

Sonnerup, B. U. Ö., and L. J. Cahill Jr., Magnetopause structure and altitude from Explorer-12 observations, *J. Geophys. Res.*, *72*, 171-183, 1967.

Sonnerup, B. U. Ö., Magnetopause structure during the magnetic storm of September 24, 1961, *J. Geophys. Res.*, *76*, 6717-6735, 1971.

Sonnerup, B. U. Ö., I. Papamastorakis, G. Paschmann, and H. Lühr, Magnetopause properties from AMPTE/IRM observations of the convection electric field: Method development, *J. Geophys. Res.*, *92*, 12,137-12,159, 1987.

Tauxe, L., N. Kylstra, and C. Constable, Bootstrap statistics for paleomagnetic data, *J. Geophys. Res.*, *96*, 11,723-11,740, 1991.

Thorne, R. M., E. J. Smith, R. K. Burton, and R. E. Holzer, Plasmaspheric hiss, *J. Geophys. Res.*, *78*, 1581-1596, 1973.

Tsurutani, B. T., T. Gould, B. E. Goldstein, W. D. Gonzalez, and M. Sugiura, Interplanetary Alfvén waves and auroral (substorm) activity: IMP 8, *J. Geophys. Res.*, *95*, 2241-2252, 1990.

H. Kawano, Institute of Geophysics and Planetary Physics, University of California, Los Angeles, 90024-1567
hkawano@igpp.ucla.edu
T. Higuchi, Institute of Statistical Mathematics, Tokyo 106, Japan
higuchi@ism.ac.jp

(received July 1, 1994;
revised September 9, 1994; accepted October 24, 1994.)