Measures of Location and Variation for the Column Space of a Random Matrix*

Mattias Villani

Abstract

Parameter matrices which are only uniquely identified up to arbitrary linear combinations of their columns occur frequently in multivariate analysis, e.g. the matrix of factor loadings in the common factor model. This note discusses how this special feature of the model affects the choice of summarizing measures for distributions of such matrices and new measures of location and variation are proposed. Possible applications include the reporting of a Bayesian posterior distribution of the loading matrix in factor analysis and the construction of new estimators based on the likelihood function.

Keywords: Bayesian inference, Factor analysis, Cointegration, Inference reporting.

1. Introduction

Many models in multivariate analysis contain parameter matrices for which only the space spanned by their columns, the column space, is identified; that is, two different parameter matrices with the same column space give identical probability distributions for the observed data. Two well-known examples of this non-identification are the matrix of loadings in factor analysis (Anderson, 1984) and the matrix of long run relations in cointegration analysis in econometrics (Johansen, 1995). This indeterminateness is resolved in the estimation phase by

*Department of Statistics, Stockholm University, S-106 91 Stockholm, Sweden. E-mail: mattias.villani@stat.su.se. The author is grateful for comments from Rolf Larsson and Daniel Thorburn. Financial support was provided by the Swedish Council of Research in Humanities and Social Sciences (HSFR).

restricting some of the parameters so that every possible column space corresponds to a unique set of parameter values.

Recent developments in computing technology and numerical algorithms have made Bayesian inference in complex multivariate models practically feasible and the benefits from such an approach are becoming widely acknowledged, see e.g. Arminger and Muthén (1998) for a recent contribution. The basic output from a Bayesian approach is the posterior distribution of all unknown parameters conditional on the observed data. This distribution is often summarized by a few low-dimensional quantities, e.g. the mean, mode or median for location, and the variance or interquartile range for spread. This note discusses how the aforementioned non-identification introduces special considerations for the summarizing quantities of the posterior distribution and new measures of location and variation are proposed.

Although our results are likely to find most of their applications within the realm of Bayesian statistics, non-Bayesians may find something of interest here too, e.g. those who prefer to base their inferences directly on the likelihood function, as the posterior distribution under uniform priors is simply the likelihood normalized to a density. This opens up the possibility to use other summarizing measures of the likelihood function than the usual mode and Hessian matrix.

2. The basic indeterminacy

As an example of a model where only the space spanned by parameter vectors is determinable, consider the common factor model (Anderson, 1984)

$$\mathbf{x} = \mathbf{\Lambda f} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\mathbf{x}$ is a p-dimensional (column) vector of observed measurements, $\mathbf{f}$ is a m-dimensional vector of latent, unobservable, factors, $\mathbf{\Lambda}$ is $p \times m$ matrix of factor loadings and $\boldsymbol{\epsilon}$ is a p-dimensional vector of unique errors. The indeterminacy of model (2.1) is easily demonstrated by introducing an arbitrary non-singular $m \times m$ matrix $\mathbf{C}$ in the following manner

$$\mathbf{\Lambda f} = \mathbf{\Lambda C C}^{-1}\mathbf{f} = \mathbf{\Lambda}^*\mathbf{f}^*, \tag{2.2}$$

where $\mathbf{\Lambda}^* = \mathbf{\Lambda C}$ and $\mathbf{f}^* = \mathbf{C}^{-1}\mathbf{f}$. From the likelihood function of model (2.1), it is easily seen that (2.2) implies that data are only informative regarding the column space of $\mathbf{\Lambda}$ ($\operatorname{sp}\mathbf{\Lambda}$), i.e. we cannot distinguish between different matrices within this subspace.

Other examples of models where only spaces spanned by parameter matrices can be determined are the reduced rank regression (Anderson, 1951), simultaneous equations models (Anderson, 1984) and the immensely popular cointegration models in time series econometrics (Johansen, 1995).

3. The choice of metric

To get a clear picture of why the usual summarizing measures are doubtful to use if only the column space of a matrix can be determined, consider the common factor model with two observable variables ($p = 2$) as indicators of a single common factor ($m = 1$). In this model, the matrix of factor loadings in (2.1) is a two-dimensional vector $\mathbf{\Lambda} = (\lambda_1, \lambda_2)'$ and the indeterminacy described in section 2 means that only the line spanned by $\mathbf{\Lambda}$ in R^2 can be uniquely determined. Another way of saying this is that only the ratio λ_1/λ_2 is estimable, λ_1 and λ_2 cannot be individually estimated from data. Let us use the restriction $\lambda_1 = 1$ to identify $\mathbf{\Lambda}$, any other restriction leads to similar results.

The usual way to obtain a location estimate for $\mathbf{\Lambda}$ is by inserting the mean or median of λ_2 into $\mathbf{\Lambda} = (1, \lambda_2)'$, see e.g. Arminger and Muthén (1998). This practice ignores the fact that only the line spanned by $\mathbf{\Lambda}$ is determinable. As an example, the location estimate based on the two vectors $(1, \kappa)'$ and $(1, -\kappa)'$ is $(1, 0)'$, for any value of κ, both when the mean and median are used (using the usual convention for the median). But $\mathrm{sp}\,(1, \kappa)'$ and $\mathrm{sp}\,(1, -\kappa)'$ both approach $\mathrm{sp}\,(0, 1)'$, as $\kappa \to \infty$, and $(0, 1)'$ would therefore be a more acceptable measure of location for large κ. In addition, if a uniform prior distribution is used for λ_2, the mean of λ_2 does not even exist for certain models, see Kleibergen and van Dijk (1994) and Bauwens and Lubrano (1996) for proofs of this statement in cointegration models.

The odd behavior of the mean and median estimator is caused by their implicit reliance on the Euclidean metric, which is inappropriate as a distance measure between spaces. To this latter point clearly, consider the Euclidean distance between the spaces spanned by two vectors of unit length, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$. Let θ $(0 \leq \theta < \pi)$ denote the angle between $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$. It is easily shown that

$$m^2(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2) = 2(1 - \cos\theta),$$

where $m(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2) = (\mathrm{tr}\,[(\mathbf{\Lambda}_1 - \mathbf{\Lambda}_2)'(\mathbf{\Lambda}_1 - \mathbf{\Lambda}_2)])^{1/2}$ is the usual Euclidean metric for matrices. $m(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$ is thus strictly increasing in θ and attains its maximum as $\theta \to \pi$, which is absurd since then $\mathrm{sp}\,\mathbf{\Lambda}_1 \to \mathrm{sp}\,\mathbf{\Lambda}_2$.

In order to define sensible summaries of the posterior distribution, the structure of the set of allowable parameters must be carefully examined and a suitable metric over this set chosen. Since only the column space of $\mathbf{\Lambda}$ is unique, $\mathbf{\Lambda}$ belongs to a smaller space than the space of all real $p \times m$ dimensional matrices. The space of $\mathbf{\Lambda}$ is the space of all m-dimensional subspaces of R^p, or the *Grassman manifold* (James, 1954). The relevant aspect to measure is therefore the distance between subspaces rather than the distance between the elements of the matrices themselves.

A distance measure between two subspaces $\operatorname{sp}\mathbf{\Lambda}_1$ and $\operatorname{sp}\mathbf{\Lambda}_2$ is described in Larsson and Villani (2000). This metric is the Frobenius norm ($\|\mathbf{A}\|_F = \operatorname{tr}(\mathbf{A}'\mathbf{A})^{1/2}$) of $\mathbf{\Lambda}'_{1\perp}\mathbf{\Lambda}_2$, i.e.

$$d(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2) = \operatorname{tr}(\mathbf{\Lambda}'_2\mathbf{\Lambda}_{1\perp}\mathbf{\Lambda}'_{1\perp}\mathbf{\Lambda}_2)^{1/2}, \tag{3.1}$$

where all involved matrices have been made orthonormal and $\mathbf{\Lambda}_{1\perp}$ denotes the orthonormal complement of $\mathbf{\Lambda}_1$. The idea is that if the norm of $\mathbf{\Lambda}'_{1\perp}\mathbf{\Lambda}_2$ is large then $\operatorname{sp}\mathbf{\Lambda}_2$ is close to $\operatorname{sp}\mathbf{\Lambda}_{1\perp}$ and therefore far from $\operatorname{sp}\mathbf{\Lambda}_1$. In the special case where $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ are vectors, it is easily seen (Larsson and Villani, 2000) that

$$d(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2) = \sin\theta,$$

where θ is the angle between $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$. Thus, $d(\mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$ approaches zero when θ approaches either zero or π, i.e. when $\operatorname{sp}\mathbf{\Lambda}_1 \to \operatorname{sp}\mathbf{\Lambda}_2$.

4. Location and variation measures

4.1. Location measure

The metric in (3.1) can be used to define an alternative location measure. For this purpose, note the following characterization of the expected value of $\mathbf{\Lambda}$

$$\bar{\mathbf{\Lambda}} = \arg\min_{\tilde{\mathbf{\Lambda}}} E\left[m^2(\mathbf{\Lambda}, \tilde{\mathbf{\Lambda}})\right], \tag{4.1}$$

where $E(\cdot)$ refers to the distribution of $\mathbf{\Lambda}$. By analogy with (4.1), we propose the following location measure for $\operatorname{sp}\mathbf{\Lambda}$.

Definition 4.1. *The span location measure is defined as*

$$\hat{\mathbf{\Lambda}} \stackrel{def}{=} \arg\min_{\tilde{\mathbf{\Lambda}}} E\left[d^2(\mathbf{\Lambda}, \tilde{\mathbf{\Lambda}})\right]. \tag{4.2}$$

Note that, as the metric in (3.1) is defined for orthonormal matrices, $\hat{\mathbf{\Lambda}}$ is necessarily orthonormal. A more interpretable location measure can of course be obtained by a simple rotation of $\hat{\mathbf{\Lambda}}$.

We have the following result.

Theorem 4.2.

$$\hat{\mathbf{\Lambda}} = (\mathbf{v}_1, ..., \mathbf{v}_m),$$

where $\mathbf{v}_i$ *is the eigenvector of* $E(\mathbf{\Lambda}\mathbf{\Lambda}')$ *corresponding to the* i*th largest eigenvalue.*

Proof. Using that $\mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Lambda}_\perp\mathbf{\Lambda}_\perp' = \mathbf{I}_p$ (Johansen, 1995), we can write

$$d(\mathbf{\Lambda}, \tilde{\mathbf{\Lambda}}) = [m - \mathrm{tr}(\tilde{\mathbf{\Lambda}}'\mathbf{\Lambda}\mathbf{\Lambda}'\tilde{\mathbf{\Lambda}})]^{1/2}.$$

Thus,

$$\hat{\mathbf{\Lambda}} \stackrel{def}{=} \arg\min_{\tilde{\mathbf{\Lambda}}} E\left[d^2(\mathbf{\Lambda}, \tilde{\mathbf{\Lambda}})\right] = \arg\max_{\tilde{\mathbf{\Lambda}}} E\left[\mathrm{tr}(\tilde{\mathbf{\Lambda}}'\mathbf{\Lambda}\mathbf{\Lambda}'\tilde{\mathbf{\Lambda}})\right] = \arg\max_{\tilde{\mathbf{\Lambda}}} \mathrm{tr}\left[\tilde{\mathbf{\Lambda}}' E(\mathbf{\Lambda}\mathbf{\Lambda}')\tilde{\mathbf{\Lambda}}\right]. \tag{4.3}$$

>From Lütkepohl (1991, Section A.14, Proposition A.4), the minimum is reached for $\hat{\mathbf{\Lambda}} = (\mathbf{v}_1, ..., \mathbf{v}_m)$. ■

A closed form expression for $E(\mathbf{\Lambda}\mathbf{\Lambda}')$ may not be available, but a numerical approximation may be used in its place. For example, importance sampling (Kloek and van Dijk, 1978) or the Gibbs sampler (Tierney, 1994) can be used to generate N draws from the distribution of $\mathbf{\Lambda}$, denoted by $\mathbf{\Lambda}^{(1)}, ..., \mathbf{\Lambda}^{(N)}$. These generated matrices can subsequently be made orthonormal and the following well-known result (Tierney, 1994) can be used to estimate $E(\mathbf{\Lambda}\mathbf{\Lambda}')$

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{\Lambda}^{(i)}\mathbf{\Lambda}^{(i)\prime} \stackrel{a.s.}{\to} E(\mathbf{\Lambda}\mathbf{\Lambda}'),$$

where $\mathbf{\Lambda}^{(i)}$ denotes the ith sampled matrix after the transformation to orthonormality and $\stackrel{a.s.}{\to}$ denotes almost sure convergence.

4.2. Variation measures

Although the variances of the free coefficients in $\mathbf{\Lambda}$ are often easily computed numerically by sampling from the distribution of $\mathbf{\Lambda}$, they may be of little help in assessing the variation of $\operatorname{sp}\mathbf{\Lambda}$, at least for weakly informative distributions of $\operatorname{sp}\mathbf{\Lambda}$.

A quite different measure of variation suggests itself from Theorem 4.2. Let $l_1, ..., l_p$ denote the eigenvalues of $E(\mathbf{\Lambda}\mathbf{\Lambda}')$ in descending order. Since l_i measures the variation of $\operatorname{sp}\mathbf{\Lambda}$ in the direction determined by $\mathbf{v}_i$, $l_1, ..., l_m$ can be used to assess the uncertainty regarding $\operatorname{sp}\mathbf{\Lambda}$.

A natural suggestion for an *overall* measure of variation of $\operatorname{sp}\mathbf{\Lambda}$ based on the $d(\cdot,\cdot)$ metric is given in the following definition.

Definition 4.3. *The overall span variation measure is defined as*

$$\tau^2_{\operatorname{sp}\mathbf{\Lambda}} \stackrel{def}{=} \frac{\mathrm{E}\left[d^2(\mathbf{\Lambda}, \hat{\mathbf{\Lambda}})\right]}{m(p-m)/p},$$

where $\hat{\mathbf{\Lambda}}$ *was defined in Definition 4.1.*

Theorem 4.4.

$$0 \leq \tau^2_{\operatorname{sp}\mathbf{\Lambda}} \leq 1.$$

Proof. The non-negativity of $\tau^2_{\operatorname{sp}\mathbf{\Lambda}}$ follows directly from Definition 4.3 and the non-negativity of the d-metric. To obtain the upper bound of $\tau^2_{\operatorname{sp}\mathbf{\Lambda}}$, Proposition A.4 in Lütkepohl (1991, Section A.14) can be used to show that $\tau^2_{\operatorname{sp}\mathbf{\Lambda}}$ can be written

$$\tau^2_{\operatorname{sp}\mathbf{\Lambda}} = \frac{m - \sum_{i=1}^{m} l_i}{m(p-m)/p}, \tag{4.4}$$

where l_i is the ith largest eigenvalue of $E(\mathbf{\Lambda}\mathbf{\Lambda}')$. Note also that

$$\sum_{i=1}^{p} l_i = \operatorname{tr}[E(\mathbf{\Lambda}\mathbf{\Lambda}')] = E[\operatorname{tr}(\mathbf{\Lambda}\mathbf{\Lambda}')] = E[\operatorname{tr}(\mathbf{\Lambda}'\mathbf{\Lambda})] = E[\operatorname{tr}(\mathbf{I}_m)] = m.$$

It is easy to see that

$$l_i = \frac{m}{p}, \text{ for } i = 1, ..., m.$$

minimizes $\sum_{i=1}^{m} l_i$ subject to the ordering constraint and the constraint $\sum_{i=1}^{p} l_i = m$. Thus, from (4.4),

$$\max \tau^2_{\mathrm{sp}\,\mathbf{\Lambda}} = \frac{m - \sum_{i=1}^{m} m/p}{m(p-m)/p} = 1.$$

■

Note that (4.4) in the proof of Theorem 4.4 can be used to compute $\tau^2_{\mathrm{sp}\,\mathbf{\Lambda}}$ efficiently.

5. Concluding remarks

This note has introduced summary measures for the distribution of matrices for which only the column space is uniquely determined. We have shown that the usual mean or median estimators of such matrices may behave badly.

The mode, on the other hand, does not suffer from the same difficulties. The mean is usually preferred over the mode, however, and the location measure defined here can be seen as a mean estimator based on a more appropriate metric than the Euclidean metric. Indeed, in a small simulation study in Villani (2000) our location estimator (applied to the normalized likelihood) was up to 30-35% more efficient than the maximum likelihood estimator for some parameter values and never more than 5-6% less efficient for any parameter value.

References

[1] Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions, *Ann. of Math. Stat.*, **22**, 327-351.

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

[3] Arminger, G. and Muthén, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis-Hastings algorithm, *Psychometrika*, **63**, 271-300.

[4] Bauwens, L. and Lubrano, M. (1996). Identification restrictions and posterior densities in cointegrated Gaussian VAR systems. In *Advances in Econometrics*, Volume **11**, Part B, JAI Press, 3-28.

[5] James, A. T. (1954). Normal multivariate analysis and the orthogonal group, *Ann. Math. Statist.* **25** (1954), 40-74.

[6] Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

[7] Kleibergen, F. and van Dijk, H. K. (1994). On the shape of the likelihood/posterior in cointegration models, *Econometric Theory*. **10**, 514-51.

[8] Kloek, T. and van Dijk, H. K. (1978). Bayesian estimates of system equation parameters; an application of integration by Monte Carlo, *Econometrica,* **46**, 1-19.

[9] Larsson, R. and Villani, M. (2000). A distance measure between cointegration spaces, *Economics Letters*, forthcoming.

[10] Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*, Springer-Verlag,Berlin.

[11] Tierney, L. (1994). Markov Chains for exploring posterior distributions (with discussion). *Ann. Statist.*, **22**, 1701-1762.

[12] Villani, M. (2000). *Aspects of Bayesian Cointegration*. Ph.D. Thesis, Department of Statistics, Stockholm University.