Parameter Expansion for Data Augmentation

Jun S. LIU and Ying Nian WU

Viewing the observed data of a statistical model as incomplete and augmenting its missing parts are useful for clarifying concepts and central to the invention of two well-known statistical algorithms: expectation-maximization (EM) and data augmentation. Recently, Liu, Rubin, and Wu demonstrated that expanding the parameter space along with augmenting the missing data is useful for accelerating iterative computation in an EM algorithm. The main purpose of this article is to rigorously define a parameter expanded data augmentation (PX-DA) algorithm and to study its theoretical properties. The PX-DA is a special way of using auxiliary variables to accelerate Gibbs sampling algorithms and is closely related to reparameterization techniques. We obtain theoretical results concerning the convergence rate of the PX-DA algorithm and the choice of prior for the expansion parameter. To understand the role of the expansion parameter, we establish a new theory for iterative conditional sampling under the transformation group formulation, which generalizes the standard Gibbs sampler. Using the new theory, we show that the PX-DA algorithm with a Haar measure prior (often improper) for the expansion parameter is always proper and is optimal among a class of such algorithms including reparameterization.

KEY WORDS: Auxiliary variable; EM algorithm; Gibbs sampler; Group of transformations; Haar measure; Locally compact group; Markov chain Monte Carlo; Maximal correlation; Overparameterization; Rate of convergence; Reparameterization.

1. INTRODUCTION

The incomplete-data formulation is fundamental in Rubin's (1978) causal inference model and is also key to both the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977) and the data augmentation (DA) algorithm (Tanner and Wong 1987). In this formulation, the set of observed data is augmented to a set of "completed data" that typically follows a simpler model. Computation of the maximum likelihood estimates or the posteriors is accomplished by iterating between imputing the missing data and fitting the complete-data model. Recently, Liu, Rubin, and Wu (1998) noticed that along with imputing the missing data, which simplifies computation at the expense of iterations, expanding the parameter space for the complete-data model can be a useful technique for accelerating the convergence of the EM algorithm. More precisely, they observed that with the imputed missing data, some extra parameters can often be introduced without distorting the original observed-data model. A more precise definition appears in Section 3. In this case, the EM algorithm can be implemented for the expanded model (the PX-EM algorithm), and it converges with a monotone increase in observed-data likelihood. Liu et al. (1998) showed both empirically and theoretically that the PX-EM algorithm can often yield significant acceleration over the ordinary EM algorithm. As a precursor to the PX-EM algorithm, Meng and van Dyk (1997) proposed efficient augmentation that instead of finding an overparameterization as with the PX-EM algorithm identifies an optimal reparameterization among a class of reparameterizations indexed by a working parameter.

Because the DA algorithm, or the more general Gibbs sampling algorithms (Gelfand and Smith 1990), can be viewed as stochastic generalizations of the EM-type algorithms, it has been widely speculated that overparameterization (or auxiliary variable) and reparameterization (or efficient augmentation) methods should be similarly useful for designing more efficient stochastic algorithms. In fact, some efforts along this line have been made (see Meng and van Dyk 1997 and the discussions therein). Although success stories of using auxiliary variables in Markov chain Monte Carlo (MCMC) abound (see, e.g., Higdon 1998), theory indicates that extreme care must be taken to design a useful method (Geyer 1992; Liu 1994).

Following the parameter expansion idea, in this article we define a *parameter-expanded data augmentation* (PX-DA) algorithm. We explore the relationship between the PX-DA algorithm and overparameterization, reparameterization, and group of transformations. In proving optimality properties of the PX-DA algorithm, we establish a new theory of iterative conditional sampling. In particular, we show that a Gibbs-like step can be generalized as movement along an orbit of a transformation group acting on the space of interest. An explicit formula for drawing an element from the group, conditional on the current state, is given so that the target distribution of interest is invariant under such moves.

To illustrate the main benefits and important issues, Section 2 works through an intentionally simple example that nevertheless conveys most of the relevant ideas. Section 3 defines a general PX-DA algorithm and shows that reparameterization can often be viewed as a special PX-DA algorithm. This connection helps us prove an optimality result in Section 5. Section 4 introduces the data-transformation concept in a MCMC sampler and shows that any PX-

Jun S. Liu is Assistant Professor, Department of Statistics, Stanford University, Stanford, CA 94305. Ying Nian Wu is Assistant Professor, Department of Statistics, University of California, Los Angeles, CA 90095. Liu's research was partially supported by National Science Foundation grants DMS-9501570 and DMS-9803649, and the Terman fellowship. Wu thanks the Department of Statistics, Stanford University for partial support during the summer of 1997. Part of the manuscript was prepared while Liu was visiting Department of Statistics of UCLA. The authors are grateful to Persi Diaconis, Andrew Gelman, Hui-Nien Hung, Chuanhai Liu, Xiao-Li Meng, Jeff Wu, two anonymous referees, and the associate editor for their helpful discussions and important suggestions.

Journal of the American Statistical Association
December 1999, Vol. 94, No. 448, Theory and Methods

DA algorithm in which the expansion parameter indexes a data-transformation mechanism and has an independent prior converges no slower than the ordinary DA algorithm. Section 5 establishes a new theory that generalizes the regular Gibbs sampler, studies the use of Haar priors in the PX-DA algorithm, and investigates why such priors should be used. Connections between our results and fiducial distribution, maximal ancillary, and noninformative prior are noted. Section 6 gives a numerical example to illustrate the methodology, and Section 7 concludes with a discussion.

2. A SIMPLE EXAMPLE

To illustrate basic ideas of the PX-DA algorithm, we consider the complete-data model

$$\mathbf{y}|\theta,\mathbf{z} \sim \mathrm{N}(\theta+\mathbf{z},1), \qquad \mathbf{z}|\theta \sim \mathrm{N}(0,D), \tag{1}$$

where θ is an unknown parameter and D is a known constant. In this model $\mathbf{y}$ is the observed data and $\mathbf{z}$ is missing. Hence the observed-data model is $\mathbf{y}|\theta \sim \mathrm{N}(\theta, 1+D)$. This example can be viewed as a special case of the random-effects model. The observed data $\mathbf{y}$ and missing data $\mathbf{z}$ are multidimensional random variables in general, but are kept as one-dimensional in this simple example. With a flat prior on θ, the posterior distribution of θ is $\mathrm{N}(\mathbf{y}, 1+D)$.

Let us treat this example as a missing-data problem in which one uses the DA algorithm to iterate between the two conditional draws

$$\mathbf{z}|\theta,\mathbf{y} \sim \mathrm{N}\left(\frac{\mathbf{y}-\theta}{1+D^{-1}}, \frac{1}{1+D^{-1}}\right)$$

and

$$\theta|\mathbf{y},\mathbf{z} \sim \mathrm{N}(\mathbf{y}-\mathbf{z},1) \tag{2}$$

for simulating the posterior distribution of θ given $\mathbf{y}$. Liu, Wong, and Kong (1994, 1995) derived that the convergence rate (i.e., the second-largest eigenvalue of the induced Markov chain transition operator) of any DA scheme is equal to the square of the maximal correlation between $\mathbf{z}$ and θ, which in this case (i.e., with normality) happens to be the absolute value of the lag-1 autocorrelation of the draws of θ at stationarity. This rate (the larger its value, the slower the chain converges) can be computed as

$$r_0 = 1 - \frac{E(\mathrm{var}(\theta|\mathbf{y},\mathbf{z})|\mathbf{y})}{\mathrm{var}(\theta|\mathbf{y})} = \frac{1}{1+D^{-1}}.$$

The rate r_0 also corresponds to the Bayesian fraction of missing information of Rubin (1987).

It is seen from the ordinary DA iterations (2) that forcing $\mathbf{z}$ to have mean 0 causes relatively high "association" between θ and $\mathbf{z}$ and slows down the algorithm. We can overparameterize (1) to get

$$\mathbf{y}|\theta,\alpha,\mathbf{w} \sim \mathrm{N}(\theta-\alpha+\mathbf{w},1), \qquad \mathbf{w}|\theta,\alpha \sim \mathrm{N}(\alpha,D). \tag{3}$$

This expanded complete-data model preserves the observed-data model $\mathbf{y}|\theta,\alpha \sim \mathrm{N}(\theta,1+D)$ but has an expansion parameter α identifiable only from the complete data $(\mathbf{y},\mathbf{w})$. We call such a parameterization "overcentering." When α is fixed at 0, the expanded model (3) reduces to the original model (1).

To implement a DA algorithm for the expanded model, we need to assign a prior for α. For now we use a proper prior, $\alpha \sim \mathrm{N}(0,B)$, which is independent of the prior of θ. Then we have the following PX-DA algorithm:

1. Draw $(\mathbf{w},\alpha)$ conditional on $(\theta,\mathbf{y})$; that is, draw $\alpha|\theta,\mathbf{y} \sim \mathrm{N}(0,B)$ and draw
$$\mathbf{w}|\theta,\alpha,\mathbf{y} \sim \mathrm{N}\left(\frac{\mathbf{y}-\theta}{1+D^{-1}}+\alpha, \frac{1}{1+D^{-1}}\right).$$
2. Draw (θ,α) conditional on $(\mathbf{y},\mathbf{w})$; that is, draw
$$\alpha|\mathbf{y},\mathbf{w} \sim \mathrm{N}\left(\frac{B\mathbf{w}}{B+D}, \frac{1}{B^{-1}+D^{-1}}\right)$$
and draw $\theta|\mathbf{y},\mathbf{w},\alpha \sim \mathrm{N}(\mathbf{y}-\mathbf{w}+\alpha,1)$.

Because α is sampled in both of the foregoing steps, if we look only at θ and $\mathbf{w}$, the procedure is mathematically equivalent to a collapsed sampler (Liu 1994) with α integrated out from the joint posterior distribution. The only reason for carrying α in the procedure is to simplify computation. The rate of convergence of this PX-DA algorithm is

$$\begin{aligned} r_{\mathrm{px}} &= 1 - \frac{E(\mathrm{var}(\theta|\mathbf{y},\mathbf{w}))}{\mathrm{var}(\theta|\mathbf{y})} \\ &= \frac{D-(B^{-1}+D^{-1})^{-1}}{1+D} \le \frac{1}{1+D^{-1}}. \end{aligned}$$

Thus the new algorithm is never slower than the ordinary DA algorithm for (1).

As $B \to \infty$, the rate r_{px} goes to 0, meaning that the algorithm converges in one iteration. It is easy to check that as $B \to \infty$, the algorithm is still well defined by its limiting one-iteration transition despite the fact that the prior for α is improper. Computationally, this limiting transition can be realized by expressing the two steps of the PX-DA as a random mapping dependent of B and then letting B go to infinity. More precisely, if one starts at θ_0, then step 1 corresponds to setting $\alpha = \sqrt{B}Z_1$,

$$\mathbf{w} = \frac{\mathbf{y}-\theta_0}{1+D^{-1}} + \sqrt{B}Z_1 + \frac{Z_2}{\sqrt{1+D^{-1}}},$$

and the new θ drawn at step 2 is

$$\theta_1 = \mathbf{y} - \frac{D}{B+D}\left\{\frac{\mathbf{y}-\theta_0}{1+D^{-1}} + \sqrt{B}Z_1 + \frac{Z_2}{\sqrt{1+D^{-1}}}\right\} + Z_3\sqrt{1+\frac{1}{B^{-1}+D^{-1}}},$$

where Z_1, Z_2, and Z_3 are independent standard Gaussian random variables. Hence as $B \to \infty$, the limiting transition is $\theta_1 = \mathbf{y} + \sqrt{1+D}Z_3 \sim \mathrm{N}(\mathbf{y}, 1+D)$. In Section 4.3 we show that this limiting argument is generally applicable.

3. THE PARAMETER-EXPANDED DATA AUGMENTATION

3.1 A Formal Definition

For the rest of the article, we let $\mathbf{y}$ be the observed data, $\mathbf{z}$ be the missing data in the original model, and θ be the parameter of interest. We use the generic notation f to denote the probability densities related to the original model in a standard DA implementation. For example, $f(\mathbf{y},\mathbf{z}|\theta)$ represents the complete-data model, $f(\mathbf{y}|\theta)$ is the observed-data model, $f(\theta)$ denotes the prior, and $f(\mathbf{y},\mathbf{z}) = \int f(\mathbf{y},\mathbf{z}|\theta)f(\theta)\,d\theta$ is the marginal distribution of the complete data. We use p for distributions related to the expanded model in the PX-DA framework.

The DA algorithm (Tanner and Wong 1987) is as follows:

1. Draw $\mathbf{z} \sim f(\mathbf{z}|\theta,\mathbf{y}) \propto f(\mathbf{y},\mathbf{z}|\theta)$.
2. Draw $\theta \sim f(\theta|\mathbf{z},\mathbf{y}) \propto f(\mathbf{y},\mathbf{z}|\theta)f(\theta)$.

Suppose that we can find a hidden identifiable parameter α in the complete-data model $f(\mathbf{y},\mathbf{z}|\theta)$. Then we can expand this model to a larger model, $p(\mathbf{y},\mathbf{w}|\theta,\alpha)$, that preserves the observed-data model $f(\mathbf{y}|\theta)$. Mathematically, this means that the probability distribution $p(\mathbf{y},\mathbf{w}|\theta,\alpha)$ satisfies

$$\int p(\mathbf{y},\mathbf{w}|\theta,\alpha)\,d\mathbf{w} = f(\mathbf{y}|\theta).$$

We call α an expansion parameter throughout. For notational clarity, here we use $\mathbf{w}$ instead of $\mathbf{z}$ to denote the missing data under the expanded model. To implement the DA algorithm for the expanded model, we need to give a joint prior distribution $p(\theta,\alpha)$. It is straightforward to prove the following.

Proposition 1. The posterior distribution of θ is the same for both models if and only if the marginal prior distribution for θ from $p(\theta,\alpha)$ agrees with $f(\theta)$; that is, $p(\theta|\mathbf{y}) = f(\theta|\mathbf{y})$ if and only if $\int p(\theta,\alpha)\,d\alpha = f(\theta)$.

Therefore, we need only specify the conditional prior distribution $p(\alpha|\theta)$ while maintaining the marginal prior for θ at $f(\theta)$. It is clear that given θ and $\mathbf{y}$, the posterior distribution of $\alpha, p(\alpha|\mathbf{y},\theta)$, remains $p(\alpha|\theta)$, because α is not identifiable from $\mathbf{y}$. For the time being, we assume that $p(\alpha|\theta)$ is a proper probability distribution. The PX-DA algorithm can then be defined as iterating the following steps:

1. Draw $(\alpha,\mathbf{w})$ jointly from their conditional distribution given $(\mathbf{y},\theta)$. This can be achieved by first drawing α from $p(\alpha|\theta)$, and then drawing $\mathbf{w}$ according to
$$\mathbf{w}|\mathbf{y},\theta,\alpha \sim p(\mathbf{y},\mathbf{w}|\theta,\alpha).$$
2. Draw (θ,α) jointly according to
$$\theta,\alpha|\mathbf{y},\mathbf{w} \sim p(\mathbf{y},\mathbf{w}|\theta,\alpha)p(\alpha|\theta)f(\theta).$$

Mathematically, this algorithm iteratively draws $\mathbf{w}$ conditional on θ and $\mathbf{y}$, and then draws θ conditional on $\mathbf{w}$ and $\mathbf{y}$, with α marginalized out in both steps. So the convergence of this scheme to its target distribution follows directly from standard Markov chain theory (see, e.g., Liu et al. 1995). In Section 4 we identify two conditions under which the PX-DA algorithm can be shown to be superior to the DA algorithm. In Section 5 we show that assigning a Haar invariant prior to α is optimal under fairly reasonable conditions.

3.2 Degenerate Priors and Reparameterization

Because $p(\mathbf{y},\mathbf{w}|\theta,\alpha)$ is an extension of $f(\mathbf{y},\mathbf{z}|\theta)$, one often can find a constant α_{null} such that $f(\mathbf{y},\mathbf{z}|\theta) = p(\mathbf{y},\mathbf{z}|\theta,\alpha_{\text{null}})$; that is, the expansion parameter α is hidden at α_{null} in the original model. Therefore, if we let $p(\alpha|\theta) = \delta_{\alpha_{\text{null}}}$, then the corresponding PX-DA algorithm reduces to the original DA algorithm for $f(\mathbf{y},\mathbf{z}|\theta)$. For the simple example in Section 2, we have $\alpha_{\text{null}} = 0$.

If $p(\alpha|\theta) = \delta_{A(\theta)}$ (i.e., $\alpha = A(\theta)$, for some function A), then the corresponding PX-DA algorithm becomes the DA algorithm for $p(\mathbf{y},\mathbf{w}|\theta,A(\theta))$, which interestingly can be viewed as a reparameterization of $f(\mathbf{y},\mathbf{z}|\theta)$. Consider the example in Section 2. By letting $p(\alpha|\theta) = \delta_\theta$, we obtain the following recentering parameterization (Gelfand, Sahu, and Carlin 1995):

$$\mathbf{y}|\theta,\mathbf{w} \sim \mathrm{N}(\mathbf{w},1), \qquad \mathbf{w}|\theta \sim \mathrm{N}(\theta,D). \tag{4}$$

The DA algorithm for (4) has a rate of convergence $r_1 = 1/(1+D)$. Thus when $D > 1$, this algorithm converges faster than under the original parameterization, whereas when $D < 1$, the resulting algorithm is slower.

We now consider the problem of searching for the best degenerate prior $p(\alpha|\theta) = \delta_{A(\theta)}$ so as to induce an optimal reparameterization. For simplicity, we let $D = 1$ and consider only the class of scalar functions $\{A(\theta) = A\theta, A \in R^1\}$. Then for a fixed A, the corresponding reparameterization (which we call *partial centering*) is

$$\mathbf{y}|\theta,\mathbf{w} \sim \mathrm{N}((1-A)\theta+\mathbf{w},1), \qquad \mathbf{w}|\theta \sim \mathrm{N}(A\theta,1). \tag{5}$$

Model (5) leads to a DA algorithm with a rate of convergence $r(A) = 1 - 1/(2(A^2+(1-A)^2))$, which achieves minimum 0 at $A = .5$. Therefore, the prior $p(\alpha|\theta) = \delta_{\theta/2}$ leads to a PX-DA algorithm that converges in one iteration.

Finding the optimal parameterization in (5) among all those indexed by A is essentially the idea of efficient augmentation of Meng and van Dyk (1997), which has a broader statistical meaning as augmenting the least amount of missing information. It is interesting to realize that efficient augmentation is technically a special PX-DA algorithm where one wants to find a degenerate prior of the form $p(\alpha|\theta) = \delta_{A(\theta)}$ to make the dependence between the missing data and θ as small as possible. Because the set of all $A(\theta)$ is too large to optimize over, it must be parameterized [e.g., $A(\theta) = A\theta$] with the aid of intuition. An alternative strategy, which we explore in this article, is to go to the other extreme; that is, to make α and θ independent a priori and let the imputed data decide α at each iteration. Note that the noninformative prior $p(\alpha|\theta) \equiv 1$ in the simple example also leads to a PX-DA algorithm converging in one iteration. Section 5 gives a theoretical reason why it is generally favorable to use an invariant prior for α. In the next section we show that together with a data transformation mechanism, the independent prior of α always leads to a PX-DA algorithm with improved rate of convergence.

4. DATA TRANSFORMATION AND CONVERGENCE PROPERTIES

4.1 Parameter Expansion by Data Transformation

It is often the case that the expansion parameter α corresponds to a transformation of the missing data $\mathbf{z}$. Thus adjusting α can be intuitively understood as "analyzing" the missing data. Condition (a) as follows is fundamental in understanding the role of the expansion parameter. With Conditions (a) and (b), we can prove that the PX-DA algorithm outperforms the DA algorithm.

Condition (a). The parameter α indexes a "data-transformation" mechanism with $\mathbf{z} = t_\alpha(\mathbf{w})$. That is, for any fixed α, the function t_α induces a one-to-one and differentiable mapping (or, a C^1 diffeomorphism) between $\mathbf{z}$ and $\mathbf{w}$, and $\mathbf{w}$ plays the role of missing data in the new scheme. In other words, for any fixed value of α, the original model $f(\mathbf{y}, \mathbf{z}|\theta)$ determines the expanded model $p(\mathbf{y}, \mathbf{w}|\theta, \alpha)$ in the following way:

$$p(\mathbf{y}, \mathbf{w}|\theta, \alpha) = f(\mathbf{y}, t_\alpha(\mathbf{w})|\theta)|J_\alpha(\mathbf{w})|,$$

where $J_\alpha(\mathbf{w}) = \det\{\partial t_\alpha(\mathbf{w})/\partial \mathbf{w}\}$ is the Jacobian term evaluated at $\mathbf{w}$.

Condition (b). Parameters α and θ are independent a priori; that is, $p(\alpha|\theta) = p_0(\alpha)$.

Under these two conditions, the PX-DA algorithm becomes

Scheme 1:

1. Draw $\mathbf{z} \sim f(\mathbf{z}|\theta, \mathbf{y}), \alpha \sim p_0(\alpha)$, and compute $\mathbf{w} = t_\alpha^{-1}(\mathbf{z})$.
2. Draw (θ, α) jointly according to $\theta, \alpha|\mathbf{y}, \mathbf{w} \sim f(\mathbf{y}, t_\alpha(\mathbf{w})|\theta)|J_\alpha(\mathbf{w})|p_0(\alpha)f(\theta)$.

In many Bayesian computation problems, the marginal distribution of the complete data, $f(\mathbf{y}, \mathbf{z}) = \int f(\mathbf{y}, \mathbf{z}|\theta)f(\theta)\, d\theta$ can be obtained in closed form. Then step 2 of Scheme 1 can sometimes be accomplished by first drawing α from $[\alpha|\mathbf{y}, \mathbf{w}] \propto f(\mathbf{y}, t_\alpha(\mathbf{w}))|J_\alpha(\mathbf{w})|p_0(\alpha)$ (its marginal posterior distribution), and then drawing θ conditional on α. In this case, Scheme 1 can be rewritten as a minor modification of the DA algorithm:

Scheme 1.1

1. Draw $\mathbf{z} \sim f(\mathbf{z}|\theta, \mathbf{y})$, in the same way as the DA algorithm.
2. Draw $\alpha_0 \sim p_0(\alpha)$, and compute $\mathbf{w} = t_{\alpha_0}^{-1}(\mathbf{z})$. Draw $\alpha_1 \sim [\alpha|\mathbf{y}, \mathbf{w}] \propto f(\mathbf{y}, t_\alpha(\mathbf{w}))|J_\alpha(\mathbf{w})|p_0(\alpha)$. Compute $\mathbf{z}' = t_{\alpha_1}(t_{\alpha_0}^{-1}(\mathbf{z}))$.
3. Draw $\theta \sim f(\theta|\mathbf{y}, \mathbf{z}')$, in the same way as the DA algorithm.

Compared to the DA algorithm, Scheme 1.1 adjusts the missing data, from $\mathbf{z}$ to $\mathbf{z}'$, through using a set of transformations before the next θ is drawn. The convergence of Scheme 1.1 follows from the convergence of the general PX-DA algorithm. Moreover, as a result of the following theorem, step 2 of Scheme 1.1 leaves $f(\mathbf{z}|\mathbf{y})$ invariant.

Theorem 1. Suppose that (a) the random variable $\mathbf{z} \sim \pi(\mathbf{z})$; (b) $\{t_\alpha\colon \alpha \in \mathcal{A}\}$ is a set of transformations on $\mathbf{z}$; and (c) a probability measure $p_0(\alpha)$ can be defined on $\mathcal{A}$. Let α_0 be a random draw from the prior $p_0(\alpha)$ and let $\mathbf{w} = t_{\alpha_0}^{-1}(\mathbf{z})$. If

$$\alpha_1 \sim \pi(\alpha|\mathbf{w}) \propto \pi(t_\alpha(\mathbf{w}))|J_\alpha(\mathbf{w})|p_0(\alpha),$$

then $z' = t_{\alpha_1}(\mathbf{w})$ follows the distribution π.

Proof. Consider the joint distribution of α and $\mathbf{w}$,

$$p(\alpha, \mathbf{w}) = p_0(\alpha)\pi(t_\alpha(\mathbf{w}))|J_\alpha(\mathbf{w})|, \qquad (6)$$

under which the marginal distribution of α is $p_0(\alpha)$ and $t_\alpha(\mathbf{w}) \sim \pi$. Therefore, if $\alpha_0 \sim p_0(\alpha)$ and $\mathbf{z} \sim \pi$, then $\mathbf{w} = t_{\alpha_0}^{-1}(\mathbf{z})$ must follow the marginal distribution of $\mathbf{w}$ under (6). Because the new α_1 is drawn from the conditional distribution of α given $\mathbf{w}$, under (6), we easily see that the joint distribution of $(\alpha_1, \mathbf{w})$ must be the same as (6). Hence $\mathbf{z}' = t_{\alpha_1}(\mathbf{w})$ follows the distribution π.

If we let $\pi(\mathbf{z}) = f(\mathbf{z}|\mathbf{y})$ in Theorem 1, then it is clear that step 2 of Scheme 1.1 leaves $f(\mathbf{z}|\mathbf{y})$ invariant. Generally speaking, Theorem 1 provides a recipe for conducting a move, which leaves π invariant, along the trace $\mathcal{S} = \{t_\alpha(\mathbf{z}), \alpha \in \mathcal{A}\}$ of a set of transformations.

4.2 Rate of Convergence

Theorem 2. Suppose that Conditions (a) and (b) hold. Then Scheme 1.1, or, equivalently, Scheme 1, converges no slower than the DA algorithm defined in Section 3.1.

Proof. The movement in one PX-DA iteration can be represented by a simple directed graph (or conditional independence graph),

$$\theta_{\text{old}} \to \mathbf{z} \to \mathbf{z}' \to \theta_{\text{new}},$$

whereas the DA iteration can be represented as $\theta_{\text{old}} \to \mathbf{z} \to \theta_{\text{new}}$. The conditional independence between z' and θ_{old} is a consequence of step 2 of Scheme 1.1. To prove the theorem, we need only show that the complete-data variance of $h(\theta)$ given $\mathbf{z}$ in the PX-DA algorithm is no smaller than that in the DA algorithm for any square-integrable functions h. (This is because that the convergence rate of a DA algorithm is completely characterized by the infimum of this variance, with the infimum over all mean 0 and variance 1 functions; see Liu et al. 1994.)

The intuition behind this is that the extra variation caused by adjusting $\mathbf{z}$ to $\mathbf{z}'$ makes θ move more freely under the PX-DA algorithm. In the following, the subscript "DA" indicates the probability transition under the DA algorithm, the subscript "PX-DA" indicates that under the PX-DA algorithm, and the subscript "f" indicates distribution (or joint distribution) induced by the original model $f(\mathbf{y}, \mathbf{z}, \theta)$.

Because Scheme 1.1 and the DA algorithm produce the same joint distribution for $(\theta, \mathbf{z})$ (on convergence), we have for all $h(\theta)$, square-integrable functions with respect to

$f(\theta|\mathbf{y})$ that

$$\text{var}_{\text{PX-DA}}\{h(\theta)\} = \text{var}_{\text{DA}}\{h(\theta)\} = \text{var}_f\{h(\theta)\}.$$

Furthermore, the conditional variance of θ under the PX-DA algorithm satisfies

$$\begin{aligned} &E_{\text{PX-DA}}[\text{var}_{\text{PX-DA}}\{h(\theta)|\mathbf{z}\}] \\ &\quad = E_f[E_{\text{PX-DA}}[\text{var}_f\{h(\theta)|\mathbf{z}'\}|\mathbf{z}] \\ &\qquad + \text{var}_{\text{PX-DA}}[E_f\{h(\theta)|\mathbf{z}'\}|\mathbf{z}]] \\ &\quad = E_f[\text{var}_f\{h(\theta)|\mathbf{z}'\}] + E_f[\text{var}_{\text{PX-DA}}[E_f\{h(\theta)|\mathbf{z}'\}|\mathbf{z}]] \\ &\quad \geq E_f[\text{var}_f\{h(\theta)|\mathbf{z}\}]. \end{aligned}$$

The second equality holds because the relationship between θ and $\mathbf{z}'$ in the PX-DA algorithm is induced by the sampling of θ from $f(\theta|\mathbf{y},\mathbf{z}')$. On the other hand, Liu et al. (1994) showed that for the DA algorithm,

$$E_{\text{DA}}[\text{var}_{\text{DA}}\{h(\theta)|\mathbf{z}\}] = E_f[\text{var}_f\{h(\theta)|\mathbf{z}\}].$$

Thus the desired result follows from theorem 3.2 of Liu et al. (1994).

Remark 1. It is clear from the proof that the comparison remains valid if step 2 of Scheme 1.1 is generalized to any adjustment of $\mathbf{z}$ to $\mathbf{z}'$ that is independent of θ_{old} and leaves the marginal distribution of $\mathbf{z}$ invariant.

Remark 2. Conditions (a) and (b) are both important in the derivation of Theorem 2 and can be satisfied by all of the examples of Liu et al. (1998). For condition (b), we have seen in Section 3.2 that the dependent (degenerate) prior $p_0(\alpha) = \delta_{a(\theta)}$ can lead to a PX-DA algorithm with a slower rate of convergence than the original DA algorithm.

For condition (a), consider the parameter expansion scheme

$$[\mathbf{y}|\mathbf{w}] = \text{N}(\theta + \mathbf{w}, 1+\alpha), \qquad \mathbf{w} \sim \text{N}(0, D-\alpha),$$

where α is between $[-1, D]$. This scheme does not admit a data-transformation mechanism. It is obvious that if and only if $\alpha = D$, the corresponding DA algorithm converges in one iteration. Therefore, unless $p_0(\alpha)$ is a point mass on D, the PX-DA algorithm is always slower than the ordinary DA algorithm with α fixed at D.

Remark 3. A variation of Scheme 1, which we call the conditional PX-DA algorithm, is to let α_0 be the same as α_1 drawn in the previous iteration instead of a new draw from its prior. That is, this scheme iterates between sampling $\mathbf{w}$ given $(\theta, \alpha, \mathbf{y})$ and sampling (θ, α) given $(\mathbf{w}, \mathbf{y})$. Although the conditional PX-DA algorithm shares the same lag-1 autocorrelation for θ with the PX-DA algorithm, it is less efficient than the PX-DA algorithm because of the comparison theorem of Liu et al. (1994). Furthermore, the conditional PX-DA algorithm can sometimes be inferior to the ordinary DA algorithm. Take the example in Section 2. If step 1 of drawing α from $\text{N}(0, B)$ had been skipped, then the convergence rate of the scheme would have been

$$r_2 = 1 - \frac{E(\text{var}(\mathbf{w}|\mathbf{y},\alpha,\theta))}{\text{var}(\mathbf{w}|\mathbf{y})} = 1 - \frac{(1+D^{-1})^{-1}}{B+D},$$

which is always greater than $(1+D^{-1})^{-1}$ provided that $B > 0$. Hence the conditional PX-DA algorithm is slower than the ordinary DA algorithm in this case.

4.3 A Limiting Procedure for Using Improper Priors

The use of an independent prior for α can be intuitively understood as letting the imputed data, $\mathbf{z}$, decide which transformation (i.e., α) to use during the iterative computation. This suggests that one may want to use a very diffused prior for α, which often corresponds to an improper prior. In this case Scheme 1 cannot be implemented, because sampling from $p_0(\alpha)$ (step 2) is no longer feasible. If the improper prior is the limit of a sequence of proper priors, then we can use the following result to realize a PX-DA algorithm.

To fix notation, let $p_B(\alpha)$ be a proper prior for α with hyperparameter B, and suppose that $p_B(\alpha)$ converges to an improper prior $p_\infty(\alpha)$ as $B \to B_\infty$. Let $K_B(\mathbf{z}'|\mathbf{z})$ be the transition induced by step 2 of Scheme 1.1. If a limiting version of step 2 exists as $B \to B_\infty$, then a limiting PX-DA algorithm exists, and it should still have a better performance than the DA algorithm.

Theorem 3. Suppose that $K_B(\mathbf{z}'|\mathbf{z}) \to K_\infty(\mathbf{z}'|\mathbf{z})$ as $B \to B_\infty$ for almost all $\mathbf{z}$, where $K_\infty(\mathbf{z}'|\mathbf{z})$ is a proper probability transition function. Then a limiting PX-DA algorithm can be implemented as draw $\mathbf{z} \sim f(\mathbf{z}|\mathbf{y},\theta)$, draw $\mathbf{z}' \sim K_\infty(\mathbf{z}'|\mathbf{z})$, and draw $\theta \sim f(\theta|\mathbf{y},\mathbf{z}')$. This limiting PX-DA still converges to the target distribution.

Proof. This is a consequence of the following lemma.

Lemma 1. Suppose that $K_B(x,y)$ is a sequence of probability transition functions, all having $\pi(x)$ as invariant distribution. If $K_\infty(x,y) = \lim_{B\to B_\infty} K_B(x,y)$ is a proper transition function, then π is an invariant distribution of K_∞.

Proof. Because $\int \pi(x)K_B(x,y)\,dx = \pi(y)$ a.s., by Fatou's lemma we have

$$\begin{aligned} \int \pi(x)K_\infty(x,y)\,dx &= \int \lim_{B\to B_\infty} \pi(x)K_B(x,y)\,dx \\ &\leq \lim_{B\to B_\infty} \int \pi(x)K_B(x,y)\,dx = \pi(y), \qquad \forall y. \end{aligned}$$

Because $\int\int \pi(x)K_\infty(x,y)\,dx\,dy = \int \pi(y)\,dy = 1$, we have that $\int \pi(x)K_\infty(x,y)\,dx = \pi(y)$ a.s.

Deriving the computer code for $K_\infty(\mathbf{z}'|\mathbf{z})$ often involves only simple algebra.

5. GROUP STRUCTURE AND HAAR PRIOR FOR THE EXPANSION PARAMETER

5.1 Locally Compact Group and the Haar Measure

More specific results can be obtained if the set of the transformations $\{t_\alpha, \alpha \in \mathcal{A}\}$ is endowed with a finer structure; that is, if it forms a locally compact group. We show here that if the prior of α corresponds to a Haar measure, whether it is a proper probability distribution or not, there is always a well-defined PX-DA algorithm that is optimal in terms of the convergence rate.

More formally, a set $\mathcal{A}$ is called a *group* with respect to an operation "$\cdot$" if (a) for all $\alpha, \alpha' \in \mathcal{A}, \alpha \cdot \alpha' \in \mathcal{A}$; (b) there is an identity element $e \in \mathcal{A}$ so that $\alpha \cdot e = e \cdot \alpha = \alpha$, for all $\alpha \in \mathcal{A}$; and (c) for all $\beta \in \mathcal{A}$, we can find a unique $\beta^{-1} \in \mathcal{A}$ so that $\beta \cdot \beta^{-1} = \beta^{-1} \cdot \beta = e$. We assume that the set of transformations $\{t_\alpha, \alpha \in \mathcal{A}\}$ has the same group structure as $\mathcal{A}$; that is, $t_e(\mathbf{z}) = \mathbf{z}$, and for all $\alpha, \alpha' \in \mathcal{A}, t_\alpha(t_{\alpha'}(\mathbf{z})) = t_{\alpha \cdot \alpha'}(\mathbf{z})$. We call $\mathcal{A}$ a *locally compact group* or *topological group* if topologically $\mathcal{A}$ is a locally compact space and operations $(\alpha, \beta) \to \alpha \cdot \beta$ and $\alpha \to \alpha^{-1}$ are continuous (Rao 1987). If the operations are analytic (i.e., in C^∞), then $\mathcal{A}$ is called a *Lie group*. Group $\mathcal{A}$ is automatically a locally compact group if it is finite.

For any measurable subset $B \subset \mathcal{A}$ and element $\alpha_0 \in \mathcal{A}$, the notation $B \cdot \alpha_0$ defines a subset of $\mathcal{A}$ resulting from "acting" on every element of B by α_0. A *right Haar measure* $H(d\alpha)$ on $\mathcal{A}$ is defined as a measure that is invariant under the group acting on the right; that is, it satisfies

$$\begin{aligned} H(B) &= \int_B H(d\alpha) = \int_{B \cdot \alpha_0} H(d\alpha) \\ &= H(B \cdot \alpha_0), \qquad \forall \alpha_0 \in \mathcal{A}, \end{aligned}$$

and for all measurable subset $B \subset \mathcal{A}$. One can similarly define a *left Haar measure*. Under mild conditions, the right (or left) Haar measure is unique up to a positive multiplicative constant (Rao 1987). Saying that $\mathcal{A}$ is *unimodular* means that its right Haar measure is also a left Haar measure. When $\mathcal{A}$ is compact (e.g., a finite group) or abelian (i.e., $\alpha \cdot \beta = \beta \cdot \alpha$; e.g., the translation and scale groups), one can show that its right Haar measure is unimodular (see Rao 1987, prop. 4, p. 498). Otherwise, the right and left Haar measures may differ by a modular function.

If $\mathcal{A}$ is a compact group, then its unimodular Haar measure is the uniform probability measure. If $\mathcal{A}$ is the translation group (e.g., $t_\alpha(\mathbf{z}) = \mathbf{z} + \alpha$ as in the overcentering parameterization), then the unimodular Haar measure for α is simply the Lebesgue measure. If $\mathcal{A}$ is a scale group [i.e., for scalar $\alpha, t_\alpha(\mathbf{z}) = \alpha\mathbf{z}$) as in the probit regression example in Sec. 6], then the Haar measure for α is proportional to $|\alpha|^{-1} d\alpha$. If $\mathcal{A}$ is the group of nonsingular $k \times k$ matrices, then the unimodular Haar measure is $|\alpha|^{-k} d\alpha$. In the following, we assume that a density $H(\alpha)$ exists for the unimodular Haar measure with respect to the Lebesgue or counting measure; that is, $H(d\alpha) = H(\alpha) d\alpha$.

5.2 The Parameter-Expanded Data Augmentation Algorithm With the Haar Prior

Here we show that if the prior distribution of α is a unimodular Haar measure, then the following PX-DA scheme is always proper. We show its optimality in Section 5.4.

Scheme 2:

1. Draw $\mathbf{z} \sim f(\mathbf{z}|\theta, \mathbf{y})$, the same as in the DA algorithm.
2. Draw (θ, α) jointly according to $\theta, \alpha|\mathbf{y}, \mathbf{z} \sim f(\mathbf{y}, t_\alpha(\mathbf{z})|\theta)|J_\alpha(\mathbf{z})|H(d\alpha)f(\theta)$; that is, draw from the posterior distribution of the parameters in the expanded model.

The only difference between Scheme 2 and the DA algorithm is that step 2 of Scheme 2 draws from the posterior of (θ, α) in the expanded model, with a Haar prior on α. Compared with Scheme 1.1, Scheme 2 essentially fixes α_0 at $\alpha_{\text{null}} = e$ instead of drawing it from $p_0(d\alpha)$. Modeling Scheme 1.1, we rewrite Scheme 2 to conform with the ordinary DA algorithm:

Scheme 2.1

1. Draw $\mathbf{z} \sim f(\mathbf{z}|\theta, \mathbf{y})$.
2. Draw $\alpha \sim p(\alpha|\mathbf{y}, \mathbf{z}) \propto f(\mathbf{y}, t_\alpha(\mathbf{z}))|J_\alpha(\mathbf{z})|H(d\alpha)$. Compute $\mathbf{z}' = t_\alpha(\mathbf{z})$.
3. Draw $\theta \sim f(\theta|\mathbf{y}, \mathbf{z}')$.

As in Scheme 1.1, step 2 of Scheme 2.1 defines a transition from $\mathbf{z}$ to $\mathbf{z}'$. We show that this step not only leaves the distribution $f(\mathbf{z}|\mathbf{y})$ invariant, but also produces some optimality properties.

Let $\mathcal{Z}$ be the space of $\mathbf{z}$. For a given $\mathbf{z}$, the set $\{t_\alpha(\mathbf{z}): \alpha \in \mathcal{A}\} \subset \mathcal{Z}$ is called an *orbit*. Formally, we say that $\mathbf{z}$ and $\mathbf{z}'$ lie on the same orbit if and only if there exists a unique $\alpha \in \mathcal{A}$, such that $\mathbf{z}' = t_\alpha(\mathbf{z})$. Different orbits do not intersect, and the space $\mathcal{Z}$ can be partitioned into the union of all of the orbits, each of which has the same structure as $\mathcal{A}$. To illustrate, consider the following examples. For Example A, let $\mathbf{z} = (z_1, z_2) \in R^2$ and $\mathcal{A} = \{\alpha \in R^1: t_\alpha(\mathbf{z}) = (z_1 + \alpha, z_2 + \alpha)\}$. Figure 1(a) shows a set of orbits for this translation group. For Example B, let $\mathbf{z} = (z_1, z_2) \in R^2$ and $\mathcal{A} = \{\alpha > 0: t_\alpha(\mathbf{z}) = (\alpha z_1, \alpha z_2)\}$. Figure 1(b) shows a set of orbits for this scale group.

Suppose that this set of orbits can be represented by a smooth cross-section $\mathcal{Q}$ (Wijsman 1966), which is defined as a subset of $\mathcal{Z}$ that intersects with (almost) every orbit exactly once. In Example A, a cross-section is $\mathcal{Q} = \{(z_1, z_2): z_1 = 0\}$. In Example B, a cross-section is $\mathcal{Q} = \{(z_1, z_2): z_1^2 + z_2^2 = 1\}$. When $\mathcal{A}$ is finite or a compact Lie group, the existence of a smooth $\mathcal{Q}$ has been established (see Wijsman 1966 for references). Palais (1961) provided results on the existence of $\mathcal{Q}$ when $\mathcal{A}$ is not compact.

With the cross-section $\mathcal{Q}$, any $\mathbf{z} \in \mathcal{Z}$ can be described by its orbit $r \in \mathcal{Q}$ and its position $\beta \in \mathcal{A}$ on the orbit. Step 2 of Scheme 2.1 effectively generates a new position conditional on the orbit. To prove its properness, we need the following.

Condition (c). The transformation group $\mathcal{A}$ is locally compact and has a unimodular Haar measure $H(\alpha)\,d\alpha$. There exists a smooth cross-section $\mathcal{Q} \subset \mathcal{Z}$, and the mapping Z: $Z(\beta, r) = t_\beta(r)$ for $\mathcal{A} \times \mathcal{Q} \to \mathcal{Z}$ is one-to-one and continuously differentiable; that is, a diffeomorphism.

In all of the specific cases considered in this article, such as the translation and scale groups, the existence of a smooth $\mathcal{Q}$ and the diffeomorphism Z can be easily verified. Under the mapping Z, we have for all $\mathbf{z} \in \mathcal{Z}$ a new parameterization: there exists (β, r), such that $\mathbf{z} = Z(\beta, r)$, where r indicates the orbit on which $\mathbf{z}$ lies and β indicates its position on r. A simple property of this diffeomorphism

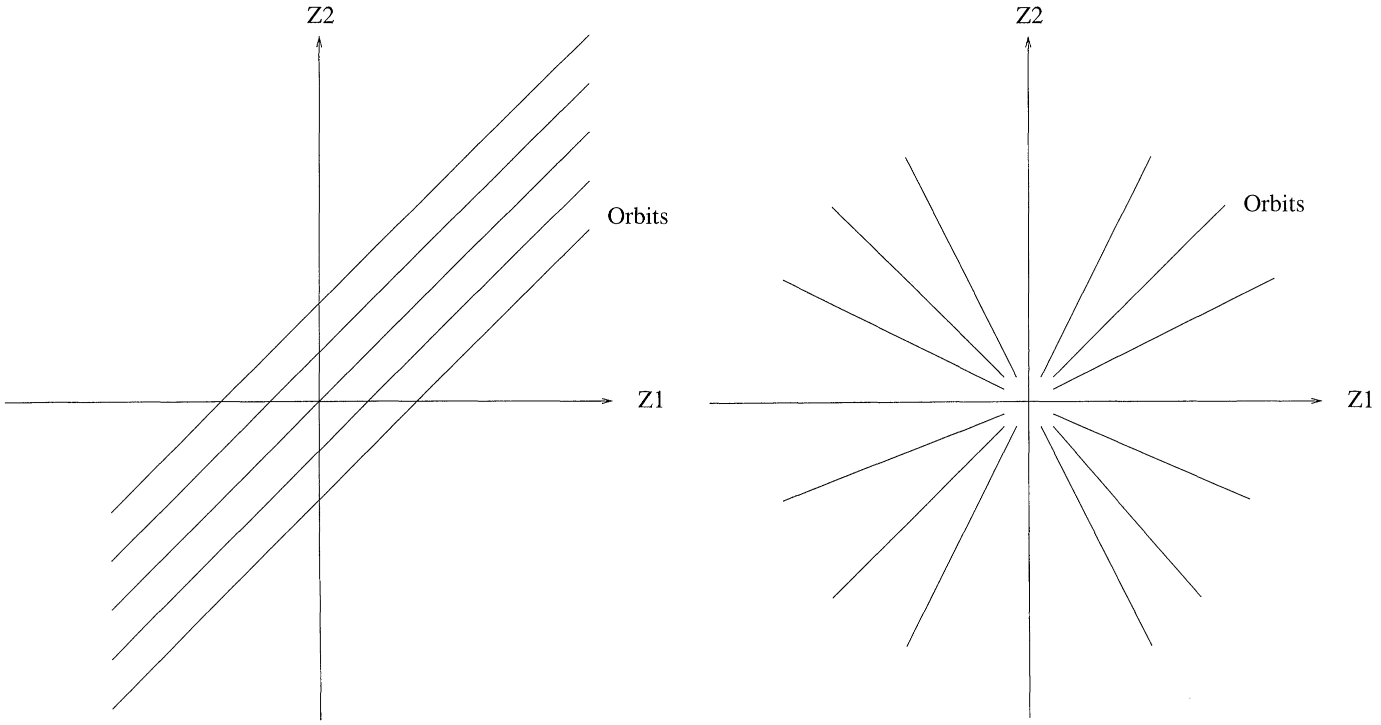

Figure 1. The Set of Orbits in R^2 for (a) the Translation Group and (b) the Scale Group.

is that for all $\alpha \in \mathcal{A}$ and $\mathbf{z} = Z(\beta, r)$,

$$t_\alpha(\mathbf{z}) = t_\alpha(Z(\beta, r)) = Z(\alpha \cdot \beta, r). \tag{7}$$

Based on this formulation, we can describe Scheme 2.1 by

$$\theta_{\text{old}} \to (\beta, r) \to (\beta', r) \to \theta_{\text{new}}, \tag{8}$$

where $\beta' = \alpha \cdot \beta$.

Lemma 2. Let

$$I_r(\alpha) = \partial Z(\alpha, r)/H(d\alpha) = \frac{\partial Z(\alpha, r)/\partial\alpha}{H(\alpha)}$$

and $K_\alpha(r) = \partial Z(\alpha, r)/\partial r$. Under Condition (c), the diffeomorphism Z satisfies

$$[I_r(\alpha), K_\alpha(r)] = J_\alpha(Z(e, r))[I_r(e), K_e(r)].$$

Proof. Because of left invariance of $H(d\alpha)$, we have $I_r(\alpha\cdot\beta) = \partial Z(\alpha\cdot\beta, r)/H(d\beta)$. By differentiating both sides of (7) with respect to β, we have

$$J_\alpha(Z(\beta, r))I_r(\beta) = I_r(\alpha \cdot \beta).$$

Differentiating both sides of (7) with respect to r gives rise to

$$J_\alpha(Z(\beta, r))K_\beta(r) = K_{\alpha\cdot\beta}(r).$$

The result then follows by letting $\beta = e$.

Theorem 4. Let π be an arbitrary probability measure on $\mathcal{Z}$, and suppose that Condition (c) holds. For $\mathbf{z} \in \mathcal{Z}$, we let

$$g(\mathbf{z}) = \int \pi(t_\alpha(\mathbf{z}))|J_\alpha(\mathbf{z})|H(d\alpha). \tag{9}$$

Then $g(\mathbf{z}) < \infty$. If we write $\mathbf{z} = Z(\beta, r)$ and let α be drawn from

$$\pi(\alpha|\mathbf{z}) = \pi(t_\alpha(\mathbf{z}))|J_\alpha(\mathbf{z})|H(d\alpha)/g(\mathbf{z}), \tag{10}$$

then $\alpha \cdot \beta$ follows distribution π_r, the conditional distribution of position β given orbit r induced by π, and $\alpha \cdot \beta$ is independent of β conditional on r. If $\mathbf{z} \sim \pi$, then $\mathbf{z}' = t_\alpha(\mathbf{z})$ follows π.

Proof. From Lemma 2, the joint distribution of (α, r) induced by π is

$$\pi(\alpha, r) = \pi(Z(\alpha, r))|J_\alpha(Z(e, r))|\|I_r(e), K_e(r)\|H(d\alpha)dr,$$

where $\|I_r(e), K_e(r)\|$ is the absolute value of the determinant of $[I_r(e), K_e(r)]$. The marginal distribution of r is then

$$\|I_r(e), K_e(r)\|h(r),$$

where

$$h(r) = \int \pi(Z(\alpha, r))|J_\alpha(Z(e, r))|H(d\alpha) < \infty.$$

The conditional distribution for the position is then

$$\pi_r(\alpha) = \pi(Z(\alpha, r))|J_\alpha(Z(e, r))|H(d\alpha)/h(r).$$

Because $J_{\alpha\cdot\beta}(\mathbf{z}) = J_\alpha(t_\beta(\mathbf{z}))J_\beta(\mathbf{z})$ for any $\alpha, \beta \in \mathcal{A}$, and $H(d\alpha)$ is a Haar measure, we have

$$\begin{aligned} g(\mathbf{z}) &= \int \pi(Z(\alpha\cdot\beta, r))\frac{|J_{\alpha\cdot\beta}(Z(e, r))|}{|J_\beta(Z(e, r))|}H(d\alpha\cdot\beta) \\ &= \frac{h(r)}{|J_\beta(Z(e, r))|} < \infty \end{aligned}$$

and

$$\pi(t_\alpha(\mathbf{z}))|J_\alpha(\mathbf{z})|H(d\alpha)/g(\mathbf{z}) \\ = \pi(t_{\alpha\cdot\beta}(Z(e,r)))|J_{\alpha\cdot\beta}(Z(e,r))|H(d\alpha\cdot\beta)/h(r).$$

Hence $\alpha\cdot\beta$ follows the conditional distribution π_r and is independent of β. As in the regular Gibbs sampler, $t_\alpha(\mathbf{z}) = Z(\alpha\cdot\beta, r)$ must follow π provided that $\mathbf{z}\sim\pi$.

Corollary 1. Step 2 of Scheme 2.1 (i.e., the complete-data posterior of α) is always proper and it leaves $f(\mathbf{z}|\mathbf{y})$ invariant. Thus Scheme 2.1 or, equivalently, Scheme 2 converges to the target distribution of interest.

Proof. Let $\pi(\mathbf{z}) = f(\mathbf{z}|\mathbf{y})$ in Theorem 4.

5.3 Some Remarks

Remark 4. A key assumption in proving Theorem 4 is the existence of a smooth cross-section and the diffeomorphism Z. In practice, we need not know what they are and often do not need to check on their existence; if the finiteness of (9) can be established by other means, then we can prove the invariance without using the change-of-variable technique used here. In particular, we can establish that for any measurable function $h, E_\pi h(t_\alpha(\mathbf{z})) = E_\pi h(\mathbf{z})$. Direct examination also reveals that (10) is invariant along the orbit of $\mathbf{z}$; that is, if $\mathbf{z}_1 = t_{\beta_1}(\mathbf{z})$ for some $\beta_1\in\mathcal{A}$ and $\alpha_1\sim\pi(\alpha|\mathbf{z}_1)$, then $t_{\alpha_1}(\mathbf{z}_1)$ has the same distribution as $t_\alpha(\mathbf{z})$, with $\alpha\sim\pi(\alpha|\mathbf{z})$. The only important formula from a practitioner's standpoint is (10). Recently, Liu and Sabatti (1999) extended the invariance result of Theorem 4 to the case when $H(d\alpha)$ is only a *left* Haar measure.

Remark 5. Theorem 4 shows that using a Haar measure enables us to generate a $\beta' = \alpha\cdot\beta$ conditionally independent of β with given orbit information. Mathematically, this means that the position variable is integrated out of the sampler. Hence, Scheme 2.1 actually induces a collapsed diagram compared to (8):

$$\theta_{\text{old}} \to r \to \theta_{\text{new}}. \tag{11}$$

Remark 6. The invariance property explored in this theorem has a close relationship with work by Bondar (1972, 1976), Fraser (1961), and Wijsman (1966) on structural distribution, fiducial inference, and maximal invariance. In fact, Bondar (1972) proved a very similar theorem under the condition that $J_\alpha(\mathbf{z})$ is independent of $\mathbf{z}$. Their results do not require that the Haar measure be unimodular. As mentioned in Remark 4, the invariance result of Theorem 4 also holds when $H(d\alpha)$ is not unimodular.

Remark 7. Consider the parametric family $\mathcal{P} = \{p_\alpha(\mathbf{z}) = \pi(t_\alpha(\mathbf{z}))|J_\alpha(\mathbf{z})|: \alpha\in\mathcal{A}\}$, where π is a known base distribution. Of interest is the inference on α with an observation $\mathbf{z}$. A natural pivotal quantity is $t_\alpha(\mathbf{z})$, whose distribution is π. When the space of $\mathbf{z}$ has the identical group structure as $\mathcal{A}$, the fiducial distribution of α derived by Fraser (1961) is equivalent to (10). With the same setting, the maximal ancillary statistic is the orbit of $\mathbf{z}$. Thus the inference of α conditional on maximal ancillary (Pitman estimator) is equivalent to (10).

Remark 8. With the foregoing parametric family setting, the noninformative nature of $H(d\alpha)$ can be seen in light of Theorem 4. If $\mathbf{z}$ is an observation from the base distribution π and α is drawn from the posterior (10), then the fact that $t_\alpha(\mathbf{z})$ follows π can be interpreted as that using $H(d\alpha)$ as a prior does not "disturb" the base distribution a posteriori. No other prior can achieve this.

5.4 Optimality of the Haar Measure in the Parameter-Expanded Data Augmentation Algorithm

With Conditions (a) and (c) [e.g., $\mathbf{z} = Z(\beta, r)$], we can rewrite the original model as

$$f(\mathbf{z},\theta|\mathbf{y})d\mathbf{z}d\theta = g(\beta,r,\theta)drd\theta H(d\beta), \tag{12}$$

where g denotes the density after transformation. The expanded model can be rewritten as

$$p(\mathbf{z},\theta,\alpha|\mathbf{y})d\mathbf{z}d\theta H(d\alpha) \\ = g(\alpha\cdot\beta,r,\theta)p(\alpha|\theta)H(d\alpha)H(d\beta)d\theta dr, \tag{13}$$

where $p(\alpha|\theta)$ is the conditional prior density of α with respect to the Haar measure $H(d\alpha)$. As a consequence, the PX-DA algorithm induces the iteration between the conditional draws $\alpha,\beta,r|\theta$ and $\alpha,\theta|\beta,r$ under the joint distribution (13), which is mathematically equivalent to iterating between the conditional draws $\beta,r|\theta$ and $\theta|\beta,r$ (with α integrated out) in (13). In comparison, the DA algorithm induces the same iteration based on (12). Note that (12) and (13) have the same marginal distribution for (r,θ) conditional on $\mathbf{y}$. The difference between the two algorithms is only in the conditional distribution of (β,r) given θ.

If we let $p(\alpha|\theta)\equiv 1$, then Theorem 4 shows that the new position $\beta' = \alpha\cdot\beta$ is independent of β given r. Thus the PX-DA algorithm with a Haar measure prior on α effectively induces the iteration between drawing $r|\theta$ and drawing $\theta|r$, with both α and β integrated out from (13). In this sense, parameter expansion is doing exactly efficient augmentation, where only the orbit of the missing data is augmented. Based on the collapsing theorem of Liu (1994), we have the following theorem.

Theorem 5. Under Conditions (a) and (c), Scheme 2 is as good as or better than any PX-DA algorithm with a proper prior $p(\alpha|\theta)H(d\alpha)$ (with respect to the Lebesgue measure) on the expansion parameter. Here $H(d\alpha)$ is the Haar measure.

Remark 9. This result holds for all nonnegative and nondegenerate functions $p(\alpha|\theta)$ that, when combined with $H(d\alpha)$, give rise to a proper prior with respect to the Lebesgue measure. But the theorem does not directly apply if $p(\alpha|\theta)H(d\alpha)$ is improper. In the situation when $p(\alpha|\theta)H(d\alpha)$ can be expressed as the limit of a sequence of proper priors, the approach of Section 4.3 can be taken to show that the Haar measure is still optimal. It remains unclear, however, what a "cleaner" PX-DA algorithm would be in this case. The following result is a special case of the theorem, but we provide a different proof.

Corollary 2. Scheme 2, or, equivalently, Scheme 2.1, converges no slower than Scheme 1 or, equivalently, Scheme 1.1.

Proof. For any integrable function $h(\theta)$, we let $s(\mathbf{z}') = E_f\{h(\theta)|\mathbf{z}'\}$. Then, according to the proof of Theorem 2, we need only compare $E_f[\text{var}_{\text{PX-DA}}\{s(\mathbf{z}')|\mathbf{z}\}]$ under the two schemes.

Let $K_1(\mathbf{z}'|\mathbf{z})$ be the transition induced by step 2 of Scheme 1.1, and let $K_2(\mathbf{z}'|\mathbf{z})$ be that for Scheme 2.1, which is determined by (10). From Theorem 1, K_1 is invariant with respect to π. Because K_1 samples along the orbit of $\mathbf{z}$, it must also leave the conditional distribution invariant; that is, $\int K_1(\mathbf{z}''|\mathbf{z}')K_2(\mathbf{z}'|\mathbf{z})\,d\mathbf{z}' = K_2(\mathbf{z}''|\mathbf{z})$. Consequently,

$$\begin{aligned} &E_f[\text{var}_{K_2}\{s(\mathbf{z}')\}] \\ &\quad = E_f[\text{var}_{K_2}\{s(\mathbf{z}'')\}] \\ &\quad = E_f[E_{K_2}[\text{var}_{K_1}\{s(\mathbf{z}'')|\mathbf{z}'\}] + \text{var}_{K_2}[E_{K_1}\{s(\mathbf{z}'')|\mathbf{z}'\}]] \\ &\quad \geq E_f[\text{var}_{K_1}\{s(\mathbf{z}'')|\mathbf{z}'\}] = E_f[\text{var}_{K_1}\{s(\mathbf{z}')|\mathbf{z}\}]. \end{aligned}$$

Because reparameterization often corresponds to using a degenerate prior $p(\alpha) = \delta_{A(\theta)}$ in the PX-DA, we have the following result.

Corollary 3. Under Conditions (a) and (c), Scheme 2 is as good as or better than any reparameterization scheme that can be formulated as a PX-DA algorithm with degenerate prior.

Although using the Haar invariant prior for α in the PX-DA is optimal under Conditions (a) and (c), the resulting scheme may be difficult to implement. Moreover, $\mathcal{A}$ sometimes may not induce a data-transformation mechanism or have a group structure. The use of a proper prior, $p(\alpha|\theta)$, and the limiting argument in Section 4.3 is still valuable in these occasions.

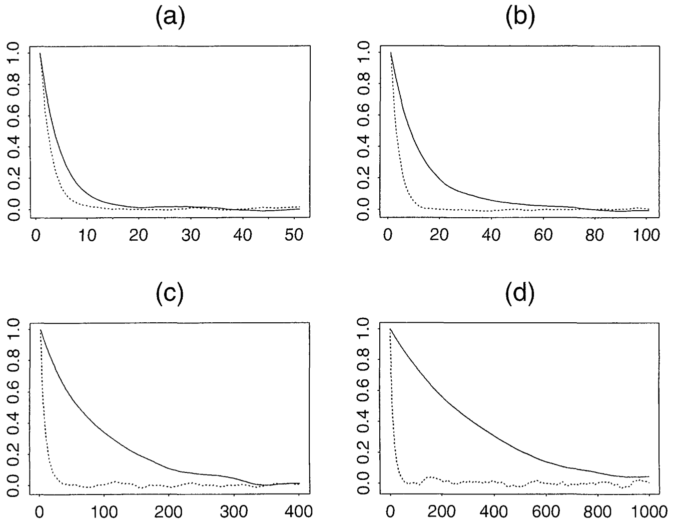

Figure 2. Autocorrelation Functions for DA (Solid Lines) and PX-DA (Dashed Lines) With Various Values of β_1: (a) β_1 = 1; (b) β_1 = 2; (c) β_1 = 4; (d) β_1 = 8.

6. A NUMERICAL EXAMPLE: PROBIT REGRESSION

Let $\mathbf{y} = (y_1, \ldots, y_n)$ be a set of iid binary observations from the probit model

$$Y_i|\theta \sim \text{Bernoulli}\{\Phi(X_i'\theta)\},$$

where $X_i(p \times 1)$ are the covariates, θ is the unknown regression coefficient, and Φ is the standard Gaussian cumulative distribution function. Of interest is the posterior distribution of θ under, say, a flat prior. A popular way to ease computation is to introduce a complete-data model in which a set of latent variables, $z_1, \ldots, z_n$, is augmented so that

$$[z_i|\theta] = \text{N}(X_i'\theta, 1)$$

and

$$y_i = \text{sgn}(z_i),$$

where $\text{sgn}(z) = 1$ if $z > 0$ and $\text{sgn}(z) = 0$ otherwise. The standard DA algorithm iterates the following steps:

1. Draw from $[z_i|y_i,\theta]$. That is, $z_i \sim \text{N}(X_i'\theta, 1)$ subject to $z_i \geq 0$ if $y_i = 1$, and $z_i \sim \text{N}(X_i'\theta, 1)$ subject to $z_i < 0$ for $y_i = 0$.
2. Draw $[\theta|z_i] = \text{N}(\hat\theta, V)$, where $\hat\theta = (\sum_i X_iX_i')^{-1}\sum_i X_iz_i$ and $V = (\sum_i X_iX_i')^{-1}$. Both $\hat\theta$ and V can be computed using the SWEEP operator (see, e.g., Little and Rubin 1987).

The complete-data model can be expanded by introducing an expansion parameter α for residual variance, which is originally fixed at 1:

$$[w_i|\theta] = \text{N}(X_i'\theta\alpha, \alpha^2)$$

and

$$y_i = \text{sgn}(w_i).$$

It is clear that this model can be derived by using the group of scale transformation $\mathbf{z} = t_\alpha(\mathbf{w}) \equiv (w_1/\alpha, \ldots, w_n/\alpha)$. The corresponding Haar measure is $H(d\alpha) = \alpha^{-1}d\alpha$. Scheme 2 has the same first step as in the standard DA, but with slightly different later steps:

2. Draw $\hat\alpha^2 \sim \text{RSS}/\chi_n^2$, where $\text{RSS} = \sum_i(z_i - X_i'\hat\theta)^2$, which is a by-product of the SWEEP operator, provided that we have computed $\sum_i z_i^2$.
3. Draw $\theta \sim \text{N}(\hat\theta/\hat\alpha, V)$.

If we put an inverse gamma prior distribution on α^2—that is, $\alpha^2 \sim b_0/\text{gamma}(a_0)$, where a_0 and b_0 are two positive constants—then Scheme 1 is identical to Scheme 2 except that the second step is changed to

- Draw $\alpha_0^2 \sim b_0/\text{gamma}(a_0)$. Then draw $\alpha_1^2 \sim (b_0 + \alpha_0^2\text{RSS}/2)/\text{gamma}(a_0 + n/2)$. Compute $\hat\alpha^2 = \alpha_1^2/\alpha_0^2 = (\text{gamma}(a_0) + \text{RSS}/2)/\text{gamma}(a_0 + n/2)$.

As $a_0 \to 0$, Scheme 1 converges to Scheme 2.

We took $n = 100$ and $X_i = (1, x_i)'$, with x_i generated from N(0, 1). The y_i were generated from Bernoulli($\Phi(\beta_0 + \beta_1 x_i)$) with the true values $\beta_0 = 0$ and $\beta_1 = 1, 2, 4, 8$. We implemented both the DA algorithm and Scheme 2.

Figure 2 shows the autocorrelation functions of the draws of β_1 for both algorithms under different true values for β_1. It is clear that as the real value for β_1 increases, the improvement of the PX-DA algorithm over the DA algorithm becomes more significant. Or, to put it another way, the PX-DA algorithm is not significantly slowed by the increased value of β_1, whereas the DA algorithm is. This phenomenon can be understood as follows. For the DA algorithm, $\operatorname{var}(\theta|\mathbf{y},\mathbf{z}) = (\sum_i X_i X_i')^{-1}$, whereas for the PX-DA algorithm,

$$\operatorname{var}(\theta|\mathbf{y},\mathbf{z}) = \left(\sum_i X_i X_i'\right)^{-1} + E[\hat{\theta}\hat{\theta}' \operatorname{var}(\chi_n)/\mathrm{RSS}] \approx \left(\sum_i X_i X_i'\right)^{-1} + \theta\theta'/2n,$$

which increases with θ. What we observed in Figure 2 can be understood from the fact that the sample autocorrelation is determined by $1 - E\{\operatorname{var}(\theta|\mathbf{y},\mathbf{z})|\mathbf{y}\}/\operatorname{var}(\theta|\mathbf{y})$ (Liu et al. 1994). The comparison in the rates of convergence should reflect the comparison in real computing time, because implementation of the PX-DA algorithm needs only negligible computing overhead in comparison to the DA algorithm.

7. DISCUSSION

Similarities and differences between this article and the work of Meng and van Dyk (1999a) should be noted. On one hand, both articles target overparameterization and reparameterization methods for speeding up the Gibbs sampler, with ideas originating from recent development in the EM algorithm. Both intend to identify useful rules in guiding the use of such methods and to study their theoretical properties. On the other hand, we concentrate more on the PX-DA algorithm under general settings and on the identification of general conditions necessary for the PX-DA algorithm to be optimal among a class of similar ones, whereas Meng and van Dyk focus more on interesting cases and statistical insights derived from them. Moreover, we present a different viewpoint of some issues concerning the relationship between reparameterization and overparameterization, the use of the PX-DA versus the conditional PX-DA (called Scheme 2 by Meng and van Dyk), and the justification for using an improper prior on the expansion parameter. Liu (1998) recently presented a covariance adjustment method where "adjustment mapping" plays the role of the expanded parameter in our setting.

The expansion parameter in the PX-DA algorithm plays a different role than the working parameter in Meng and van Dyk's (1997) efficient data augmentation, because the expansion parameter is inferred along with the PX-DA iteration instead of being fixed at an optimal value (see also Green 1997). It is conceivable, however, that a working parameter could be introduced to index a transformation group to help find a good parameter expansion scheme (van Dyk 1998).

A subtle point relates the PX-DA algorithm to auxiliary variable techniques. In our case, of interest is the posterior distribution of θ [i.e., $f(\theta|\mathbf{y})$], whereas the posterior distribution of the augmented missing data can be arbitrarily distorted. Introduction of α actually makes the joint (posterior) distribution of $(\theta, \mathbf{w})$ differ from that of $(\theta, \mathbf{z})$. This special feature distinguishes the PX-DA algorithm from the usual auxiliary variable techniques in which the introduction of a new parameter does not change the previous joint distribution.

The implication of Theorems 1 and 4 goes beyond the PX-DA algorithm described in this article. In particular, both theorems can be viewed as a generalization of the Gibbs sampler—because every Gibbs sampling move can be viewed as a transformation acting on the current state. Some conditional sampling rules for choosing one among a set of possible transformations is given by the theorems, which are especially useful when one envisions certain favorable directions to move along in the space. Liu and Sabatti (1998, 1999) have presented general ways of using the theorems, together with the Metropolis–Hastings steps, to accelerate a MCMC algorithm.

Although we have shown only one real example, we would like to assure the reader that the PX-DA algorithm can be implemented for all the models of Liu et al. (1998) in parallel with their PX-EM implementations. In particular, the optimality result for using the Haar prior applies to the multivariate t model (Meng and van Dyk 1999a), probit regression models, and random-effects models. An interesting situation arises from a random-effects model example of Meng and van Dyk (1999b) where they want to draw from an affine transformation group, say $\mathcal{A} = \{(\alpha_1, \alpha_2): (\alpha_1, \alpha_2)\mathbf{z} = \alpha_1 + \alpha_2\mathbf{z}\}$. In this case the group element (α_1, α_2) can be sampled by imposing a left-Haar measure prior on it, $|\alpha_2|^{-2} d\alpha_1 d\alpha_2$ (Liu and Sabatti 1999). This choice is equivalent to the right-Haar prior of Meng and Van Dyk (1999b), who used a slightly different definition for the affine transformation. When the implementation of step 2 of the PX-DA algorithm is difficult, a Metropolis–Hastings step can be added. (Some caveats are discussed in Liu and Sabatti 1999.) Alternatively, one can reduce computational complexity by restricting α on a finite subgroup or subset of $\mathcal{A}$ in which an exhaustive search can be conducted.

In summary, we have formally defined the PX-DA algorithm when a proper prior for the expansion parameter α is used and its generalized versions under limiting improper priors or Haar invariant priors. We have further identified some conditions under which the PX-DA algorithm can be guaranteed to outperform the ordinary DA algorithm. We have shown that using a Haar invariant prior for the expansion parameter is optimal among a class of such algorithms. In particular, we find Condition (a) for the correspondence between the expansion parameter α and a data-transformation mechanism constructive. Liu et al. (1998) and Meng and van Dyk (1999) provide further practical advice on the search for a good parameter expansion scheme.

[Received September 1997. Revised April 1999.]

REFERENCES

Bondar, J. V. (1972), "Structural Distributions Without Exact Transitivity," *Annals of Mathematical Statistics*, 43, 326–333.

——— (1976), "Borel Cross-Sections and Maximal Invariants," *The Annals of Statistics*, 4, 866–877.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Fraser, D. A. S. (1961), "The Fiducial Method and Invariance," *Biometrika*, 48, 261–280.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995), "Efficient Parameterizations for Normal Linear Mixed Models," *Biometrika*, 82, 479–488.

———, and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo" (with discussion), *Statistical Science*, 7, 473–483.

Green, P. J. (1997), Discussion of "The EM Algorithm—An Old Folk Song Sung to the Fast Tune," by X. L. Meng and D. van Dyk, *Journal of the Royal Statistical Society*, Ser. B, 59, 554–555.

Higdon, D. M. (1998), "Auxiliary Variable Methods for Markov Chain Monte Carlo With Applications," *Journal of the American Statistical Association*, 93, 585–595.

Little, R. J. A., and Rubin, D. B. (1987). Statistical Analysis with Missing Data.

Liu, C. H. (1998), "Covariance Adjustment for Markov Chain Monte Carlo—A General Framework," technical report, Bell Laboratories, Lucent Technologies.

———, Rubin, D. B., and Wu, Y. N. (1998), "Parameter Expansion to Accelerate EM—The PX-EM Algorithm," *Biometrika*, 85, 755–770.

Liu, J. S. (1994), "The Collapsed Gibbs Sampler With Applications to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966.

———, and Sabatti, C. (1998), "Simulated Sintering: Markov Chain Monte Carlo With Spaces of Varying Dimensions" (with discussion), in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, pp. 389–413.

——— (2000), "Generalized Gibbs Sampler and Multigrid Monte Carlo for Bayesian Computation," *Biometrika*, 87, in press.

———, Wong, W. H., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler With Applications to Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40.

——— (1995), "Covariance Structure and the Convergence Rate of the Gibbs Sampler With Various Scans," *Journal of the Royal Statistical Society*, Ser. B, 57, 157–169.

Meng, X. L., and van Dyk, D. (1997), "The EM Algorithm—An Old Folk Song Sung to the Fast Tune" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 59, 511–567.

——— (1999a), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation," *Biometrika*, 86, 301–320.

——— (1999b), "The Art of Data Augmentation," technical report, University of Chicago, Dept. of Statistics.

Palais, R. S. (1961), "On the Existence of Slices for Actions of Noncompact Lie Groups," *Annals of Mathematics*, 73, 295–323.

Rao, M. M. (1987), *Measure Theory and Integration*, New York: Wiley.

Rubin, D. B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, 6, 34–58.

——— (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: Wiley.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (with discussion), *Journal of the American Statistical Association*, 82, 805–811.

van Dyk, D. (1998), Discussion of "Simulated Sintering: Markov Chain Monte Carlo With Spaces of Varying Dimension," by J. S. Liu and C. Sabatti, in *Bayesian Statistics, 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, 408–409.

Wijsman, R. A. (1966), "Existence of Local Cross-Sections in Linear Cartan G-Spaces Under the Action of Noncompact Groups," *Proceedings of the American Mathematical Society*, 17, 295–301.