NORMAL BELIEF FUNCTIONS AND THE KALMAN FILTER

ARTHUR P. DEMPSTER[1]
Harvard University

Abstract

The class of multivariate normal probability distributions is extended to the class of *normal belief functions*, including as special types nonprobabilistic linear restrictions and normal distributions across the members of a partition of the frame into parallel hyperplanes. Matrix representations are defined for the resulting models, and the corresponding matrix operations for marginalization, minimal extension, and combination are explained. Normal belief functions may be used to illustrate construction and computation with graphically structured belief function models, notably the linear models and associated computations of Kalman filtering.

Keywords: Graphical models; Kalman filter; linear model computations; multivariate normal distributions; normal belief functions; propagation and fusion; sweep and reverse sweep.

1 Introduction

Two important elements of the *belief function* generalization of Bayesian inference (Dempster, 1968; Shafer, 1976, 1982) are the representation of ignorance by *vacuous belief functions* and the resolution of complex representations of uncertainty into

[1]*Address for correspondence:* Department of Statistics, Harvard University, 1 Oxford Street, Cambridge MA 02138, USA.

components by *graphical models* (Kong, 1986). Together with the class of *normal belief functions* defined in Section 2, these elements show standard Gaussian linear model techniques in a new light, as explained in Section 3 and illustrated by the Kalman filter in Section 4.

A historical forerunner is provided by Fisher's fiducial argument (Dempster, 1990a). Fisher consistently rejected conventional uniform prior distributions on grounds similar to those of 19th century critics (Zabell, 1989) when such priors lacked adequate theoretical or empirical bases. A defence of improper priors is that the resulting posterior inferences may closely approximate those from proper priors with large scale parameters, such as normal priors with large variances. Such priors remain conceptually troubling, however, because one rarely finds them justified from formal or informal experience, and because they often imply vanishingly small prior probability for regions of practical interest. Fisher debated fiducial vs Bayes with Harold Jeffreys in the 1930s (Lane, 1980), and late in life continued to ridicule artificial priors using terms like "completely bogus" (Fisher, 1959) and "dream from fairyland" (Fisher, 1962). What was new in Fisher was the serious proposal of an alternative method. The mathematically rigorous foundation of belief function theory extends, unifies, and clarifies Fisher's strategy of broadening the base of methodology for reasoning to posterior distributions (Dempster, 1966). In particular, the absence of a prior distribution in Fisher's formulation may be replaced by the presence of a vacuous belief function prior, thus filling a void in the logical structure.

The belief function concept of *graphical model* (Kong, 1986; Shafer, Shenoy, and Meloulli, 1987; Dempster and Kong, 1988; Thoma, 1989; Almond, 1989) extends a parallel concept for probabilistic models (Lauritzen and Spiegelhalter, 1988). As illustrated below with Kalman filtering, the belief function extension treats both logical and probabilistic components of a constructed model, not as separate concepts, but as manifestations of a single concept. The belief function formulation, moreover, eliminates an awkward feature of the Bayesian method. Constructing a Bayesian graphical model requires conceiving and imposing directional structure on the graph whereby earlier components of uncertain knowledge are judged asymmetrically prior to later components that are described as conditional likelihoods. The specification of graphical belief function models is based on symmetric independence assumptions that are simpler and more directly susceptible to empirical checking than are typical Bayesian conditional probability assumptions that are assumed to hold given arbitrary choices of often long lists of unknown parameter values specifying families of both distributions and conditions.

In sum, the graphical belief function framework permits uncertain reasoning with reduced baggage in the form of conventional prior and conditional probability assumptions. Instead, it encourages concentration of modelling efforts on recognizing and incorporating independent components of real information that can be combined into a graphical model. The examples in Section 4 illustrate how familiar statistical models are naturally viewed in this way.

2 Normal Belief Functions

2.1 Basic Definitions

The term *frame of discernment* (Shafer, 1976), or simply *frame*, has the mathematical role in belief function theory that *sample space* has in probability theory. The alternative terminology is germane to circumstances where, as is typically the case with belief functions, logicist interpretation of formal probabilities is understood, including interpretation of *beliefs* (or *lower probabilities*) and *plausibilities* (or *upper probabilities*). In applications, the term frame of discernment means the same as the term state space as used by scientists and engineers. For example, in linear model applications the frame is a linear space spanned by a list of *variables* whose names, and whose classification into types such as observables, fixed effects, random effects, and errors, are governed by understanding of scientific context.

The notation N will be used to denote an n-dimensional frame for a general normal belief function. The mathematical essentials are best conveyed by first considering N to be a coordinate-free linear space. The structure of a normal belief function requires two further spaces, namely, an m-dimensional hyperplane M in N, and a p-dimensional partition Π of M consisting of a family of parallel $(m-p)$-dimensional hyperplanes in M. The mathematical specification of a normal belief function over N is completed by specifying an ordinary full rank p-dimensional normal probability distribution over Π, which can be done in a coordinate-free way (Dempster, 1969) by specifying a mean point μ in Π and a full rank covariance inner product Σ over Π.

A *belief function* (or BEL) over the frame N is defined in general (Shafer, 1976) by an ordinary probability measure called the *basic probability assignment*, or *bpa*, over a sample space consisting of a class of subsets of N called the *focal elements* of BEL. Thus the mathematical system of N, M, Π, and $N(\mu, \Sigma)$ defines the normal belief function whose set of focal elements is Π and whose distribution of focal elements is $N(\mu, \Sigma)$. The mapping from Π, considered as an abstract space, to its representation as a set of hyperplanes in M may be called a *multi-valued mapping* (Dempster, 1967). This suggests calling M the *range* of the normal belief function and Π its *domain*.

Normal belief functions are *special* in the sense that the focal elements are mutually exclusive, whereas in the general theory nonempty intersections are allowed. The special case is not as special as it first appears, however, because any belief function can be made special by embedding its frame in a larger frame defined by the product space of the original frame and the space of focal elements, with appropriate redefinitions of the domain and range in accord with the original multivalued mapping. Such special representations are often of little interest, because practical use typically focuses on projections to margins that do not preserve the property of being special. However, since normality is preserved under linear operators, the property of being special is likewise preserved under this fundamental class of mappings, whence the theory of normal belief functions deals mainly with special BELs.

A normal BEL allows three types of information. In general, there are $n-m$ dimensions known with certainty due to the linear restriction from N to M. There are p dimensions whose uncertainty is described probabilistically by the normal distribution over Π. And there are $m - p$ dimensions within focal elements that are completely unknown, i.e., represent variation that is implicitly assigned a vacuous belief function. It is the last set of dimensions that forces an extension from ordinary probability distributions to belief functions.

Six nontrivial varieties of normal BEL are obtained according as one or two of the dimension numbers $n-m$, p, or $m-p$ are set to 0. Such cases provide building blocks for more complex normal BELs. Thus, if $n = m$ and $p = 0$, then the normal BEL is vacuous. If $n = m = p$, then the normal BEL is an ordinary normal distribution. Or, if $m = p = 0$, then the true point in N is known with certainty, as might occur by direct observation. Similarly, if $p = 0$ while $n > m > 0$ then the normal BEL is equivalent to specifying $n - m$ linear equations, whereas if $n > m = p$ then BEL is a normal probability distribution limited to an m-dimensional hyperplane in N. Finally, $n = m > p$ implies a proper belief function whose domain is the full frame N. The last type occurs, for example, with $n = 2$ and $p = 1$, when the unknowns (θ, e) are a true value θ and a measurement error e, while nothing is known about θ and e has a $N(0, \sigma^2)$ distribution, where e is the "pivotal quantity" in Fisher's fiducial terminology.

2.2 Density Representation

If $p > 0$, then the $N(\mu, \Sigma)$ component of a general normal BEL can be specified by its probability density function over Π. Moreover, since each point in the range space M is associated with exactly one point in the domain space Π, the density extends to a "density" over M that takes a constant value on all the points of any member of the partition Π, namely the value of the ordinary density associated with the member of Π. The mathematical fact that the "density" is uniform over each $(m - p)$-dimensional hyperplane in Π can be regarded as explaining the close connections between inferential theory based on normal BELs and Bayesian inference with improper uniform priors. Note, however, that the extended "density" over M is not interpretable as an ordinary probability density function over M. In fact, the normal BEL is vacuous over each member of Π, not uniformly distributed as the constant "density" might seem to imply.

2.3 Marginalization and Minimal Extension

It is a familiar textbook property of normal distributions that if a set of random variables is normally distributed then any subset of the variables is also normally distributed. The same *marginalization* result holds in the extended class of normal BELs. Moreover, with normal BELs a converse result holds, namely, the *minimal extension* (Shafer, 1976,1982) of a normal BEL from a margin to the full space is also a normal BEL. Minimal extension followed by marginalization back to the original margin always reproduces the original normal BEL, but marginalization followed by minimal extension reproduces an original normal BEL if and only if $m = n$ for the larger space and p is the same for both spaces, i.e., if and only if no information is lost by marginalizing. More general versions of these results for linear mappings between frames can be easily stated and proved. Details are omitted.

2.4 Combination

The combination of two or more belief functions over a common frame is a fundamental tool of both model construction and inferential computation. The theory of belief functions has depended from the start (Dempster, 1966,1967) on a central tool called by Shafer, *Dempster's rule*, or the direct sum operator "⊕" (Shafer, 1976), or more recently the *product-intersection* rule (Dempster, 1990a,1990b).

The rule has been criticized (Walley, 1987; Zadeh, 1986), but without due regard for the essential qualification that the rule requires *independence* of the individual belief functions being combined. Independence is both a conceptual and a mathematical idea. The concept refers to a judgment made each time independence is invoked that basically extends the familiar notion of probabilistic independence to include mutual independence among both logical and probabilistic components of a constructed model of uncertain knowledge. The essential idea is that the introduction of a new independent component, from whatever source, is assumed not to duplicate or interact with other component BELs so as to alter the validity of previously adopted and combined components.

The mathematical representation of independence, namely the combination rule itself, marries the Boolean device of set intersection for combining logical assertions with the probabilistic device of multiplying probabilities of independent events to obtain the probability of their conjunction. Briefly stated the rule asserts that focal elements from the separate BELs are intersected to obtain the focal elements of the combination, while the associated basic probabilities are multiplied to obtain the basic probability assignment of the combination. Details are given in the basic references.

It follows immediately from the definition that the direct sum of any number of normal BELs is again a normal BEL. The combined range is simply the intersection of the component ranges, while the combined domain is the intersection of the restrictions to the combined range of the component partitions. Similarly, the log "density" of the combined normal BEL is the sum of the component log "densities" restricted to the combined range. Since the component log "densities" are nonnegative quadratic forms, so is their sum, whence normality is preserved. The preservation of normality in the belief function sense is a strong result because it applies flexibly to normal BELs that are ordinary normal probability measures, or to sets of linear restrictions that specify deterministic relations, or to mixtures of both types.

3 Coordinate-dependent Representation of Normal BELs

3.1 Graphical Models

Belief function technology took a large step forward when the principles of constructing and computing with graphical models were recognized (Shenoy and Shafer, 1986; Pearl, 1986,1988; Kong, 1986; Lauritzen and Spiegelhalter, 1988). Brief expositions of these topics are given in Sections 3.1 and 3.2. Details for normal BELs are sketched in Sections 3.3 and 3.4.

A frame N is usually constructed in practice as the product-space of possible outcomes associated with a set of variables $\{X_1, X_2, \ldots, X_n\}$. A graphical model over N is specified by (1) selecting a list $\Im$ of subsets of $\{X_1, X_2, \ldots, X_n\}$, (2) specifying an *initial* BEL over the frame associated with each member of $\Im$, (3) minimally extending each such initial BEL specified over a margin of N to the full frame N, and (4) combining the resulting belief functions over N into a single belief function. Note the implication that the belief functions created in step (3) are assumed in step (4) to be independent.

The representation of a graphical model is far from unique. Examples: the list $\Im$ can be extended trivially by adding subsets with vacuous BELs, members of $\Im$ can be absorbed into larger members by minimally extending their initial BELs and combining with other BELs assembled at the larger margin, and any normal BEL with $p > 0$ can be indefinitely subdivided into other normal BELs. Less trivial "factorizations" (Thoma, 1989) may be sought, either for computational reasons or to facilitate reductive analysis of the modelled phenomenon.

Several graph-theoretic interpretations are associated with the term *graphical*. The most familiar takes the single elements of $\{X_1, X_2, \ldots, X_n\}$ to be the n nodes of a hypergraph whose hyperedges are the elements of $\Im$. A more fruitful interpretation turns on *join-graphs* whose nodes are the elements of $\Im$, and whose edges connect pairs of nodes having a nonempty intersection. By convention, all such edges need not be included in a join-graph representation, provided there exists a connected path among nodes joining all appearances of any X_i. If a join-graph has disconnected parts, such parts may contain only disjoint subsets of $\{X_1, X_2, \ldots, X_n\}$ that have no initial dependencies across parts and hence determine their own marginal BELs.

If a connected portion of a join-graph has no loops, and so defines a *join-tree*, then the computation of nodal margins is simplified by the propagation and fusion schemes discussed below. It is always possible to combine nodes into fewer but larger nodes that make a join-graph into a join-tree, but controlling the cost of subsequent marginalization computations requires that combination be done sparingly so that node sizes are kept small. On the other hand, it may be useful in practice to complicate a join-tree by adding small nodes that represent subsets of $\{X_1, X_2, \ldots, X_n\}$ having vacuous initial components, because it is desired to compute final marginal belief functions for such nodes using a propagation and fusion algorithm. With such extended node sets, the placement of initial component belief functions may be a matter of choice.

3.2 Propagation and Fusion in Graphical Models

Suppose that a graphical BEL is specified in join-tree form, with an initial component BEL specified at each node (allowing the possibility of nodes with vacuous initial components). Two alternative propagation and fusion strategies are to compute the marginal BEL at one preselected node of the join-tree, or alternatively to compute simultaneously the marginal BELs at all of the nodes. The first requires a single pass through the tree, while the second requires an additional reverse pass. The order of steps within the procedures can be varied to some degree. In applications where component BELs enter the formulation in real time, as in some uses of the Kalman filter, the first pass through the tree may provide successive updatings of information about the current state of a system.

Suppose that $A \in \Im$ and $B \in \Im$ are two nodes of a join-tree that have a nonempty intersection $C \in \Im$ and are joined by an edge of the tree. *Propagation* from A to B along their joining edge is defined to mean marginalizing a BEL defined on the frame associated with A to the margin C, followed by minimal extension to B, yielding a BEL on the frame associated with B. Note that propagation is a generalization of both marginalization and minimal extension, reducing to marginalization when $B = C$, or to minimal extension when $A = C$. *Fusion* on the other hand is an operation that takes place at a particular node of the join-tree, and consists of combining an initial BEL at the node with BELs propagated from neighboring nodes.

To compute the marginal BEL at a selected node A of the join-tree, place arrows on all the edges of the tree in the direction that points to A. The nodes of a join-tree may now be classified into three types: (i) leaf nodes that have exactly one outgoing arrow, (ii) intermediate nodes with one or more incoming arrows and one outgoing arrow, and (iii) the selected node with one or more incoming arrows and no outgoing arrows. The algorithm specifies one propagation step along each edge in the direction of its arrow, and one fusion step at each of the nodes of types (ii) and (iii). The order of these operations is subject to the following conditions: propagation begins at the nodes (i) in any order or simultaneously; fusion may start at a node (ii) after one or more incoming BELs are recorded but is complete only after all the incoming BELs have been fused with the initial BEL; propagation from a node (ii) takes place after the complete fusion step at that node; and, finally, fusion at the selected node combines the initial BEL at the node with all incoming BELs.

The extension required to compute marginal belief functions at all the nodes of a join-tree follows immediately from the single node algorithm. Instead of a single arrow, imagine two arrows on each edge, one in each direction. Again, one propagation step is needed for each arrow, performed under the same conditions as in the single node algorithm. If there are r edges connected to a node, then $(r + 1)$ fusion products are required at the node, namely, r that fuse subsets of $r - 1$ incoming BELs with the initial BEL, preparatory to propagation along the remaining edge, plus one final combination of the initial BEL with all the incoming BELs, to obtain the desired marginal at that node. There is flexibility in choosing the exact order of steps. One scheme is to propagate inward to a preselected node, followed by outward propagation to the leaves. Alternatively, scheduling can be done under local control, since the conditions for fusing at a node or propagating from a node require only record-keeping about existing and incoming BELs at that node.

The formulation of graphical model used here was first described by Kong (1986) who also derived the basic marginalization algorithm just defined. Examples are given in Kong and Dempster (1988). Similar theory using different mathematical terms is presented in Shafer, Shenoy, and Meloulli (1987).

3.3 Partially Swept Moment Matrices of Normal BELs

Suppose that the vector of variables $\mathbf{X} = [X_1, X_2, \ldots, X_n]$ defines the frame of a general normal belief function. As shown below, matrix representations tied to the basis $\mathbf{X}$ are many and varied, but not all are available for a given normal BEL.

When $m = p$, $\mathbf{X}$ is a vector of random variables with a multivariate normal distribution. Such a distribution is commonly characterized by its mean vector $\boldsymbol{\mu}$ and rank m covariance matrix $\boldsymbol{\Sigma}$. The shorthand notation

$$(3.1) \qquad \mathbf{M}_0 = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\Sigma} \end{bmatrix}$$

will be used to represent such matrix pairs. When $m > p$, the X_i are no longer random variables, so the moments are typically undefined and (3.1) cannot be used.

An alternative representation

$$(3.2) \qquad \mathbf{M}_1 = \begin{bmatrix} \boldsymbol{\mu}\boldsymbol{\Sigma}^{-1} \\ -\boldsymbol{\Sigma}^{-1} \end{bmatrix}$$

may be called the *potential* form (Lauritzen and Spiegelhalter, 1988). Combining normal BELs is closely related to summing potentials, because the log of a normal density is linear in potentials. Hence form (3.2) is important in the calculus of normal belief functions. Form (3.2) can be used directly only if $\boldsymbol{\Sigma}$ exists and can be inverted, i.e., if $m = p$ and $n = m$. As shown below, form (3.2) has a natural extension to the case $n = m > p$, replacing $\boldsymbol{\Sigma}^{-1}$ by a nonnegative definite matrix of rank p that specializes to $\boldsymbol{\Sigma}^{-1}$ when $m = p$. When $p = 0$ and $n = m$, i.e., when the normal BEL is vacuous, the extension further specializes also to the basic building block

$$(3.2') \qquad \mathbf{M}_1 = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix},$$

where $\mathbf{0}$ denotes a vector or matrix of zeros. However, if $n > m > p$, then neither (3.1) nor an extension of (3.2) is available.

To obtain a class of representations valid in general, it is necessary to define and then extend representations intermediate between (3.1) and (3.2). Based on the partition $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, form (3.1) may be rewritten

$$\mathbf{M}_{00} = \begin{bmatrix} \boldsymbol{\mu}_1 & \boldsymbol{\mu}_2 \\ \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} . \tag{3.1'}$$

Familiar computations modify (1′) into

$$\mathbf{M}_{01} = \begin{bmatrix} \boldsymbol{\mu}_{1.2} & \boldsymbol{\mu}_2\boldsymbol{\Sigma}_{22}^{-1} \\ \boldsymbol{\Sigma}_{11.2} & \mathbf{B}_{12} \\ \mathbf{B}'_{12} & -\boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix} \tag{3.3}$$

where

$$\boldsymbol{\mu}_{1.2} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \tag{3.4}$$

$$\mathbf{B}_{12} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \tag{3.5}$$

and

$$\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} . \tag{3.6}$$

Form (3.3) contains potentials for the marginal distribution of $\mathbf{X}_2$, and provides the necessary elements of the moment form of the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$, namely, the adjusted mean $\boldsymbol{\mu}_{1.2}$, the regression coefficients $\mathbf{B}_{12}$ of $\mathbf{X}_1$ on $\mathbf{X}_2$, and the residual covariance matrix $\boldsymbol{\Sigma}_{11.2}$ of $\mathbf{X}_1$ given fixed $\mathbf{X}_2$. Form (3.3) may be regarded as a generalization that includes (3.1) when $\mathbf{X}_2$ is empty or (3.2) when $\mathbf{X}_1$ is empty.

Form (3.3) may be called a *partially swept* representation, in contrast with (3.1) and (3.2) which are *unswept* and *completely swept*, respectively. In the case of an ordinary normal distribution, the representation (3.3) is obtained from the moments (3.1) by *sweeping* on the indices corresponding to $\mathbf{X}_2$, and (3.1) is recovered from (3.3) by *reverse sweeping*, these being familiar operations of linear model computations. The particular versions of these algorithms used here are taken from Dempster (1969). The forward sweep, or SWP, carries $\mathbf{M}_{00}$ to $\mathbf{M}_{01}$. It is easily checked that the reverse sweep, or RSW, is algorithmically identical except for a sign change in the (2,2) and (3,1) positions. Note that SWP implies moving towards a more conditional representation, and RSW means moving to a more marginal representation. Hence, despite the formal similarity of the operations, their distinct statistical interpretations justifies maintaining separate terms. The key property of SWP (or RSW) is that successive sweeps (or reverse sweeps) on disjoint subsets of the indices yield the same result as a single SWP (or RSW) on the combined set of indices. Thus (3.2) may be found directly from (3.1) as indicated in the definition (3.2), or successive sweeps may be used to pass from (3.1) to (3.3), and then from (3.3) to (3.2). Similarly, passage from (3.2) to (3.1) may be conceived as a single RSW on all the indices, or as successive RSW operations passing from (3.2) to (3.3) to (3.1). In numerical implementations, it is simplest to proceed using repeated SWP or RSW operations on single indices.

As written, form (3.3) is defined only for normal distributions such that $\mathbf{X}_2$ has a full rank covariance matrix. But the definition has a useful extension to limited circumstances that allow both $m-p>0$ and $n-m>0$. An interpretation of $m-p$ is that it is the dimension of the largest subset of $\mathbf{X}$ that has a vacuous marginal BEL. Many or even all subsets of dimension $m-p$ may have this property, and at least one must have the property. Defining $\mathbf{X}_2$ to be such a subset implies that fixing $\mathbf{X}_2=\mathbf{x}_2$ for any choice of $\mathbf{x}_2$ effectively reduces M, N, and P to M', N', and P' of dimensions $n'=n-m+p, m'=p$, and $p'=p$, respectively, where $m'-p'=0$. In other words, combining the original normal BEL with the nonprobabilistic normal BEL specified by the linear relations $\mathbf{X}_2=\mathbf{x}_2$ yields a normal BEL whose $\mathbf{X}_1$ margin is a normal probability distribution that can be represented in form (3.1). Adopting (3.2′) as the extension of (3.2) appropriate for a vacuous normal BEL, leads to

$$\text{(3.3}'\text{)} \qquad \mathbf{M}_{01}=\begin{bmatrix} \mu_{1.2} & \mathbf{0} \\ \mathbf{\Sigma}_{11.2} & \mathbf{B}_{12} \\ \mathbf{B}'_{12} & \mathbf{0} \end{bmatrix}$$

as the corresponding extension of (3.3), where $\mathbf{B}_{12}$ is the linear mapping from $\mathbf{X}_2=\mathbf{x}_2$ to the values of $\mathbf{X}_1$ on $\mathbf{M}$, and $\mu_{1.2}$ and $\mathbf{\Sigma}_{11.2}$ specify the moments of the normal distribution of $\mathbf{X}_1$ conditioned on $\mathbf{X}_2=\mathbf{x}_2$.

The terminology *partially swept on indices* $\mathbf{X}_2$ will continue to be used for (3.3′) even when no form (3.1) exists from which (3.3′) can be produced using SWP. Correspondingly, RSW cannot be applied to the indices $\mathbf{X}_2$ of (3.3′) to produce an unswept representation, because the pivotal elements in the (3,2) position of (3.3′) are zeros. Form (3.3′) may be called a *maximally marginal representation*, or *mmr* for short, because there remain no more negative diagonal pivots to which RSW can be applied. Note that it is only a matter of convenience in displays that the notation used above to define SWP and RSW implies that the swept indices are subsets chopped from the right end of $\{1,2,\ldots,n\}$. The definitions are not dependent on a selected order. Thus there exists an mmr for every choice of $\mathbf{X}_2$ consisting of $m-p$ variables whose marginal BEL is vacuous. The mmr form is important for practical computation because it is a natural vehicle for expressing the operations of marginalization and minimal extension.

In the mmr (3.3′) there are $n-m+p$ variables in the unswept list $\mathbf{X}_1$, but the conditional covariance matrix $\mathbf{\Sigma}_{11.2}$ of these variables has rank p. There must exist a subset of p elements of $\mathbf{X}_1$ whose conditional covariance has full rank p, and typically there exist many such subsets. If such a subset is chosen, and SWP is applied to these indices, then the covariance matrix associated with the remaining indices of $\mathbf{X}_1$ is reduced to $\mathbf{0}$, and if the partition of (3.3′) is modified to include the selected indices with $\mathbf{X}_2$ instead of with $\mathbf{X}_1$, then (3.3′) acquires the form

$$\text{(3.3}''\text{)} \qquad \mathbf{M}_{01}=\begin{bmatrix} \nu_{1.2} & \nu_{2}\mathbf{\Lambda}_{22} \\ \mathbf{0} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}'_{12} & -\mathbf{\Gamma}_{22} \end{bmatrix}$$

where Λ_{22} is a nonnegative definite $m \times m$ matrix of rank p. The form (3.3″) may be called a *maximally conditioned representation*, or *mcr*, because there are no further positive pivotal elements available for sweeping. The *mcr* form is important in practical computations for combination or fusion steps.

The general form of representation that includes *mmr* and *mcr* as the special cases $r = 0$ and $r = p$, respectively, is obtained by choosing any r on $0 \leq r \leq p$ and using the same definition as with (3.3″) except that only r indices are chosen from $\mathbf{X}_1$ for SWP. The result is a representation (3.3*) based on a partition where $\mathbf{X}_1$ has dimension $n - m + p - r$ and $\mathbf{X}_2$ has dimension $m - p + r$, and the inner product matrices in the (2,1) and (3,2) positions now have ranks $p - r$ and r, respectively. Form (3.3*) is not shown explicitly since notation for its components is not needed in the sequel.

Passage among different versions of (3.3*), including choices of types (3.3′) and (3.3″), means moving various indices from swept to unswept subsets, and can in most cases be accomplished numerically by repeated SWP or RSW operations on single indices. For example, in order to interchange the roles of two indices, one in the swept group and one in the unswept group, then a SWP applied to the one in the unswept group, followed by an RSW applied to the index in the swept group, would accomplish the task. In fact, these operations might be combined into a single INT operation. Zero pivotal elements can interfere with one or both of the successive SWP and RSW operations used by INT, but, in cases where the interchange is well-defined despite the zero pivotals, the definition of INT can be extended by the device of placing $\epsilon > 0$ in place of the zero pivotal, carrying out the SWP and RSW as functions of ϵ, then letting $\epsilon \to 0$. For example, if $p = 0$, the normal BEL is defined by $m - n$ linear restrictions, and there are no nonzero pivotals, yet INT can be defined and used for switching variables among left and right sides of the linear equations.

3.4 Algorithms for Propagation and Fusion of Normal BELs

Combination of two normal BELs over a common frame N is simplest when N itself is the domain of both BELs, since the intersection part of the product-intersection rule is then trivial (N intersected with itself to yield N), and only the product of "densities" need be computed. Moreover, when N is the domain of a BEL, the unswept pivotals of any general representation (3.3*) are always positive, so the representation can be completely unswept to yield a unique mcr form for the representation. Since the log "density" is linear in the elements of the mcr, combination is accomplished simply by adding the mcr representations of the two BELs to obtain the mcr of their direct sum.

The next simplest case arises when one BEL has domain properly contained in N, while the other has domain N. To derive an algorithm for this case, one may put a small $\epsilon > 0$ in place of the zero unswept pivotal elements of an mcr of the first BEL, then apply RSW to these indices to obtain the completely unswept mcr of the modified BEL. Thereafter combination is simply addition. After the addition, apply SWP to the indices that now are approximately ϵ^{-1} on the diagonals, and where σ_i, ρ and κ_{ij} are defined in (2.1) and (2.2). These results extend and simplify Jensen's (1986) results.

3 Inference for the Inverse Gaussian Distribution

We now use the results of Section 2 to examine inference for the inverse Gaussian model (Barndorff-Nielsen and Cox, 1994). Let Y_i $(i = 1, \cdots, n)$ be independent observations from the inverse Gaussian distribution $IG(\psi, \lambda)$ with probability density function

$$p(y; \psi, \lambda) = \frac{\sqrt{\psi}}{\sqrt{2\pi}} e^{\sqrt{\psi\lambda}} y^{-3/2} exp\{-\frac{1}{2}(\psi y^{-1} + \lambda y)\} .$$

We take ψ as the parameter of interest, with λ as the nuisance parameter. The log likelihood function is

$$\ell(\psi, \lambda) = \frac{n}{2} log(\psi) + n(\psi\lambda)^{\frac{1}{2}} - \frac{1}{2}\psi \sum y_i^{-1} - \frac{1}{2}\lambda \sum y_i .$$

The maximum likelihood estimate of ψ and λ are $\hat{\psi} = (\sum y_i^{-1}/n - \bar{y}^{-1})^{-1}$ and $\hat{\lambda} = \hat{\psi}/\bar{y}^2$ respectively. The maximum likelihood estimate of λ for given ψ is $\hat{\lambda}_\psi = \hat{\lambda}\psi/\hat{\psi}$, while conditional log likelihood is

$$\ell_c(\psi) = \frac{n-1}{2} log(\psi) - \frac{n}{2}\psi\hat{\psi}^{-1} .$$

The point $\tilde{\psi}$ which maximizes the $\ell_c(\psi)$ is $\tilde{\psi} = (n-1)\hat{\psi}/n$ and the SCLR statistic is

$$\tilde{r} = sgn(\tilde{\psi} - \psi)\{(n-1)(log(\tilde{\psi}/\psi) + \psi/\tilde{\psi} - 1)\}^{\frac{1}{2}} .$$

We can assess and compare the standard normal approximation to the distribution of the signed log likelihood ratio statistic r, a saddlepoint type approximation formula of $r^* = r + r^{-1} log(u/r)$ (Barndorff-Nielsen and Cox, 1994, p278) and SCLR statistic $\tilde{r}$ by expansion (2.17) using the exact distribution result $n\psi\hat{\psi}^{-1} \sim \chi^2_{n-1}$. The exact and approximate p-values are given in following table 1 for a sample point $y = (1, 9)$ and sample size $n = 2$.

The p-values generated by (2.17) provide more accurate approximation to the actual p-values than other methods including the saddlepoint type approximation of the r^* formula. Note that the p-values generated by $\Phi(r)$ are very inaccurate for this small sample size case. The r^* formula is more convenient to use than the expansion (2.17), although it fails to record a value at the observed maximum likelihood value; this probability as given by (2.17) is $1/2 + \gamma_{30}/(6\sqrt{2\pi n})$.

Acknowledgements

I am grateful to the referee for thorough and helpful comments.

References

[1] Barndorff-Nielsen, O.E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343-365.

[2] Barndorff-Nielsen, O.E. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, 73, 307-322.

[3] Cox, D.R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion) *J. Roy. Statist. Soc. Ser. B*, **49** 1-39

[4] Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics.* London, Chapman and Hall.

[5] DiCiccio, T. J. (1984). On parameter transformations and interval estimation. *Biometrika*, **71**, 477-485.

[6] Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Biometrika*, **77**, 65-76.

[7] Fraser, D.A.S. and Reid, N. (1995). Ancillaries and third order significance, *Utilitas Mathematica*, **7**, 33-53.

[8] Jensen, J. L. (1986). Similar tests and the standardized log likelihood ratio statistic *Biometrika*, **73**, 567-572.

[9] Jensen, J. L. (1992). The modified signed log likelihood statistic and saddlepoint approximation. *Biometrika*, **79**, 693-704.

[10] McCullagh, P. (1982). Discussion of paper by A. C. Atkinson. *J. Roy. Statist. Soc. Ser. B*, **44**, 29-30

[11] McCullagh, P. (1984). Local sufficiency. *Biometrika*, **71**, 233-244.

ψ	0.001	0.05	0.25	1.0	3.0	5.0	8.0	10.0
$\Phi(r)$	0.9999	0.9932	0.9588	0.8113	0.4740	0.2623	0.1031	0.0543
$\Phi(r^*)$	0.9895	0.8996	0.7484	0.4763	0.1864	0.0802	0.0241	0.0110
(2.17)	0.9879	0.8976	0.7447	0.4607	0.159	0.0584	0.0125	0.0041
Exact	0.9787	0.8504	0.6733	0.3990	0.1441	0.0593	0.0171	0.0077

Table 1: Inverse Gaussian model. Exact and approximate p-values: $n = 2, y = (1, 9)$